Genomics 113 (2021) 1802–1815

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Original Article

A multi-phenotype genome-wide association study of clades causing tuberculosis in a Ghanaian- and South African cohort

Stephanie J. Müller [a,b,*,1], Haiko Schurz [a,b], Gerard Tromp [a,b], Gian D. van der Spuy [a,b], Eileen G. Hoal [a], Paul D. van Helden [a], Ellis Owusu-Dabo [c], Christian G. Meyer [d,e], Birgit Muntau [f], Thorsten Thye [g], Stefan Niemann [h], Robin M. Warren [a], Elizabeth Streicher [a], Marlo Möller [a], Craig Kinnear [a]

[a] DSI-NRF Centre of Excellence for Biomedical Tuberculosis Research, South African Medical Research Council Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa
[b] South African Tuberculosis Bioinformatics Initiative (SATBBI), Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa
[c] School of Public Health, College of Health Sciences, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
[d] Institute of Tropical Medicine, Eberhard-Karls University, Tübingen, Germany
[e] Faculty of Medicine, Duy Tan University, Da Nang, Vietnam
[f] National Reference Centre for Tropical Pathogens, Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany
[g] Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany
[h] German Centre for Infection Research (DZIF), Partner site Hamburg-Lübeck-Borstel, Borstel, Germany

ARTICLE INFO

Keywords:
Multi-phenotype GWAS
Mycobacterium tuberculosis complex
Imputation
South Africa
Ghana

ABSTRACT

Despite decades of research and advancements in diagnostics and treatment, tuberculosis remains a major public health concern. New computational methods are needed to interrogate the intersection of host- and bacterial genomes. Paired host genotype datum and infecting bacterial isolate information were analysed for associations using a multinomial logistic regression framework implemented in SNPTest. A cohort of 853 admixed South African participants and a Ghanaian cohort of 1359 participants were included. Two directly genotyped variants, namely rs529920 and rs41472447, were identified in the Ghanaian cohort as being statistically significantly associated with risk for infection with strains of different members of the MTBC. Thus, a multinomial logistic regression using paired host-pathogen data may prove valuable for investigating the complex relationships driving infectious disease.

1. Introduction

Tuberculosis (TB), a disease primarily affecting the lungs, is caused by pathogenic members of the *Mycobacterium tuberculosis* complex (MTBC) such as *M. africanum* (*M. africanum*) and *M. tuberculosis* (*M. tb*). Infection alone, however, is not sufficient to cause disease [3,28,39]. Each branch of the MTBC comprises several clades of specific strains with variable virulence and disease-causing mechanisms [5,19,23]. While *M. africanum* is the main cause of TB in West-African countries including Ghana, *M. tb* is responsible for TB cases in most other parts of the world, including South Africa [20]. In addition to socio-economic- and environmental factors [46,55], predisposing diseases [9,31], and the genetic make-up of the human host [10,28,40,45,51,66,67] have been shown to play pivotal roles in determining susceptibility to the disease.

Several associations between genomic loci and susceptibility to infectious diseases such as TB [3,21,67], malaria [49] and HIV [43] have been reported using candidate-gene association studies, linkage studies, and genome-wide association studies (GWAS). Using these approaches, a number of genes encoding proteins of the host immune system have been associated with susceptibility to TB, including human leukocyte antigens (*HLAs*), *NRAMP1*, mannose binding lectin (*MBL*), IFN-gamma (*IFN-γ*), and Vitamin D Receptor (*VDR*) [3,51,58,71]. While several studies have investigated the genetic association with TB

* Corresponding author at: DSI-NRF Centre of Excellence for Biomedical Tuberculosis Research, South African Medical Research Council Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa.
E-mail address: stephanie.muller@ibm.com (S.J. Müller).
[1] IBM Research Africa, 45 Juta street, Braamfontein, Johannesburg, South Africa

https://doi.org/10.1016/j.ygeno.2021.04.024
Received 20 August 2020; Received in revised form 26 March 2021; Accepted 11 April 2021
Available online 20 April 2021

[10,23,24,28,53], few have investigated the association between gene variants and susceptibility to particular strains of the MTBC [37,41].

Candidate gene studies compare allelic and genotyping frequencies of a specific genetic marker between a group of unrelated cases and controls. In a large cohort of 1916 sputum-positive Ghanaian TB patients genotyped for the *ALOX5* g.760G > A variant, individuals who were heterozygous for the polymorphism were found to be at increased risk for developing TB [21]. Furthermore, patients harboring the exonic variant (g.760A) had a greater association (OR = 1.70; [95% CI: 1.2–2.6]) with infection caused by *M. africanum* West Africa-2 [21]. Modelling a recessive mode of inheritance, a protective association (OR = 0.60; [95% CI: 0.4–0.9]) was identified between the occurrence of TB and the *MBL2* G57E variant in another cohort of Ghanaian patients [65]. TB patients belonging to the Ewe population were significantly more likely to be infected with *M. africanum* (OR = 3.02; [95% CI: 1.67–5.47]) and further stratification by lineage revealed that the association was strongly driven by infection with members of *M. africanum* West Africa-1 [2].

HLA types are also known to be important in the immune response to pathogens. In a South African candidate-gene association study of *HLA* alleles and the *M. tb* strain responsible for active TB, the *HLA-B27* allele was found to decrease risk for an additional disease episode due to a Beijing strain, following multiple episodes of disease caused by a Beijing strain [51]. In addition, specific *HLA* types were found to be associated with disease caused by the different strains investigated [51]. Finally, Caws and colleagues investigated the susceptibility of the human host to different *M. tb* strains. Using a candidate-gene approach, a cohort of 237 adult Vietnamese TB patients were analysed. The authors concluded that for this cohort, individuals carrying the C allele of the toll-like receptor-2 (*TLR2*) T597C polymorphism were significantly more likely to develop TB caused by mycobacteria belonging to the East-Asian/Beijing strain family (OR = 1.57 [95% CI 1.15–2.15]) [7].

A limitation of the candidate-gene study design however, is that it requires an *a priori* hypothesis regarding which genes to target in the association analysis. To address this limitation, GWAS have become a popular alternative for identifying genetic associations with disease. Through genotyping of many common genetic variants, GWA studies enable a global interrogation of a host's genome for associations to disease, without the limitation of predefined candidate genes [22]. While genome-wide associations between the human host and TB have been extensively studied in several populations, the susceptibility of the host to different members of the MTBC has only in recent years gained some attention.

The first GWAS to investigate genetic susceptibility to strains of different MTBC lineages aimed to identify genome-wide associations with TB onset, stratifying a Thai cohort by infecting MTBC lineage, and the age at onset [41]. The study initially attempted to identify age-related associations between five MTBC lineages and two age-stratified groups of TB participants, namely 219 young cases (under the age of 45), and 467 old cases (over the age of 45). To reduce the complexity of the association tests, the MTBC lineages were tested as one lineage versus a collective of all other lineages. After applying Bonferroni corrections, none of the genotyped single nucleotide polymorphisms (SNPs) reached genome-wide significance for association with either of the age-groups for any of the five MTBC lineages. However, when reducing the five lineages to two groups consisting only of 'Beijing' and 'non-Beijing' cases, and testing for age-related association to TB, the authors identified a single SNP on chromosome 1p13, rs1418425, reported to have a significant association to non-Beijing infected cases classified in the "old" age category ($P = 1.58 \times 10^{-7}$; OR = 1.62 [95% C.I.: 1.35–1.93]). The authors were able to replicate the SNP in two independent cohorts, further demonstrating the importance of performing GWAS with a specific focus on pathogen lineage [41].

Another study investigated the coevolution of *M. tb* and it's human host, with the hypothesis that the longstanding coexistence between the human genome and *M. tb* lineage may reduce the risk of progressing to active TB or minimize the severity of disease. The authors investigated TB severity (as measured by the TBScore) in two cohorts from Uganda with paired *M. tb*-human DNA available to determine if interactions between *M. tb* lineage and human genetic variants exist. Although no association was found between lineage and disease severity, an interaction between a SNP in SLC11A1 and the L4-Ugandan lineage were identified in both cohorts. In addition several IL12B polymorphisms were found to be associated with disease severity [37].

In order to improve our understanding of the genetic susceptibility to the MTBC clades, this study leveraged genome-wide genotyping data from the host and pathogen data to perform a genome-wide screen for clade-specific genetic associations in cohorts originating from two distinct populations.

2. Results

2.1. Defining MTBC clades and superclades

2.1.1. South Africa

The MTBC strains obtained in the South African cohort contained strains of eight of the 12 lineages, namely Beijing, CAS (represented as CAS1 in the infection database), Haarlem, Haarlem-like, Low-copy Clade (LCC), T, Quebec, and "Other" (Fig. 2A). During the grouping strategy, Beijing and CAS were clustered to form the "BeijingCAS1" superclade (Fig. 2B). Similarly, Haarlem, Haarlem-like, and LCC cases were clustered to form the "HaarlemsLCC" superclade (Fig. 2B). A clade denoted as "Other" was also present in the South African cohort but does not appear on the phylogenetic tree (Fig. 1) and thus was kept as a distinct member during the grouping strategy. The T superclade was excluded from subsequent analysis due to low frequency in the cohort after clustering into superclades, leaving five superclades represented by this cohort (Fig. 2B). MTBC clade distributions were dominated by the LAM clade and closely followed by Beijing. Superclade distributions showed similar frequencies for LAM, BeijingCAS1, and HaarlemsLCC, while the Quebec and "Other" clades were the least abundant (Fig. 2B). After grouping the clades into superclades, the strains of the LAM, HaarlemsLCC, and BeijingCAS1 superclades occurred most frequently in the cohort with all three superclades having a frequency greater than 125 in the dataset (Fig. 2B), while the "Other" and Quebec superclades were in least abundance, with frequencies of 60, and 45, respectively (Fig. 2D).

2.1.2. Ghana

The strains obtained in the Ghanaian cohort contained 12 clade annotations obtained from spoligotyping. The afri-181 and afri-438 clades were represented by *M. africanum* on the phylogenetic tree and were subsequently grouped with EAI at the point of divergence, and named as the EAI_afri superclade (Fig. 1). Beijing and CAS were merged, as were Haarlem and X, into the "BeijingCAS', and "HaarlemX" superclades, respectively. T and U clades were also clustered (Fig. 1). The "Ghana-2″ clade was kept as a distinct superclade, while LAM and CAM were grouped based on the similarity in their spoligotyping patterns as illustrated in [60]. This cohort thus contained 12 clades and six superclades after clustering (Fig. 2C, and Fig. 2D, respectively).

After grouping the clades into superclades, the strains of the EAI_afri and LAM_CAM superclades occurred most frequently in the cohort with both superclades having a frequency greater than 400 in the dataset (Fig. 2D). The HaarlemX, and T_U superclades had a frequency of approximately 160, and 250 in the cohort, respectively, while the BeijingCAS and Ghana-2 superclades were in least abundance, with frequencies less than 50 (Fig. 2D).

2.2. Genotype data quality control, haplotype phasing, and genotype imputation

For the South African cohort, analysis using Genotype Harmonizer

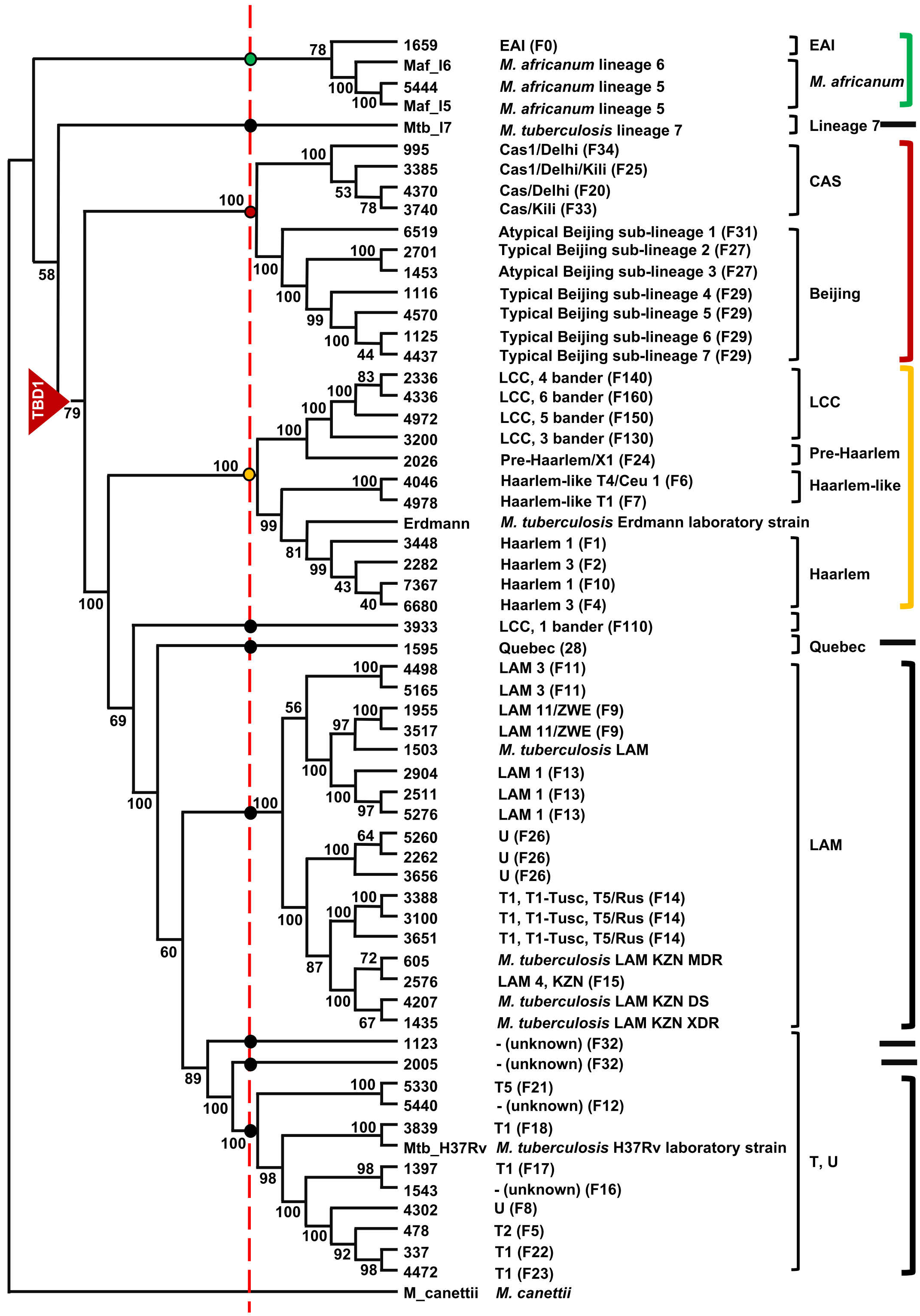

Fig. 1. Clustering of MTBC clades on the SNP–based phylogenetic tree (adapted from original tree sourced with permission from [17]). MTBC clades were grouped into superclades near a point of divergence on the phylogenetic tree. Clustering reduced 12 distinct clades into seven closely related superclades. TB cases identified to be due to infection with the East African Indian (EAI) and *M. africanum* clades (green bracket), were merged into the "EAI_afri" superclade. Similarly, CAS and Beijing clades were merged into the "BeijingCAS1" superclade (red bracket). The LCC, Pre-Haarlem, Haarlem-like, and Haarlem clades merged into one superclade designated as "HaarlemsLCC" (orange bracket), while the Quebec, Latin-American Mediterranean (LAM), T, and Lineage 7 clades remained unchanged due to the lack of a common progenitor on the red dotted line, were subsequently treated as individual superclades and are indicated by black brackets. The clades denoted as LCC 1 bander (F110) and 'unknown (F32)' were not clustered into superclades. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

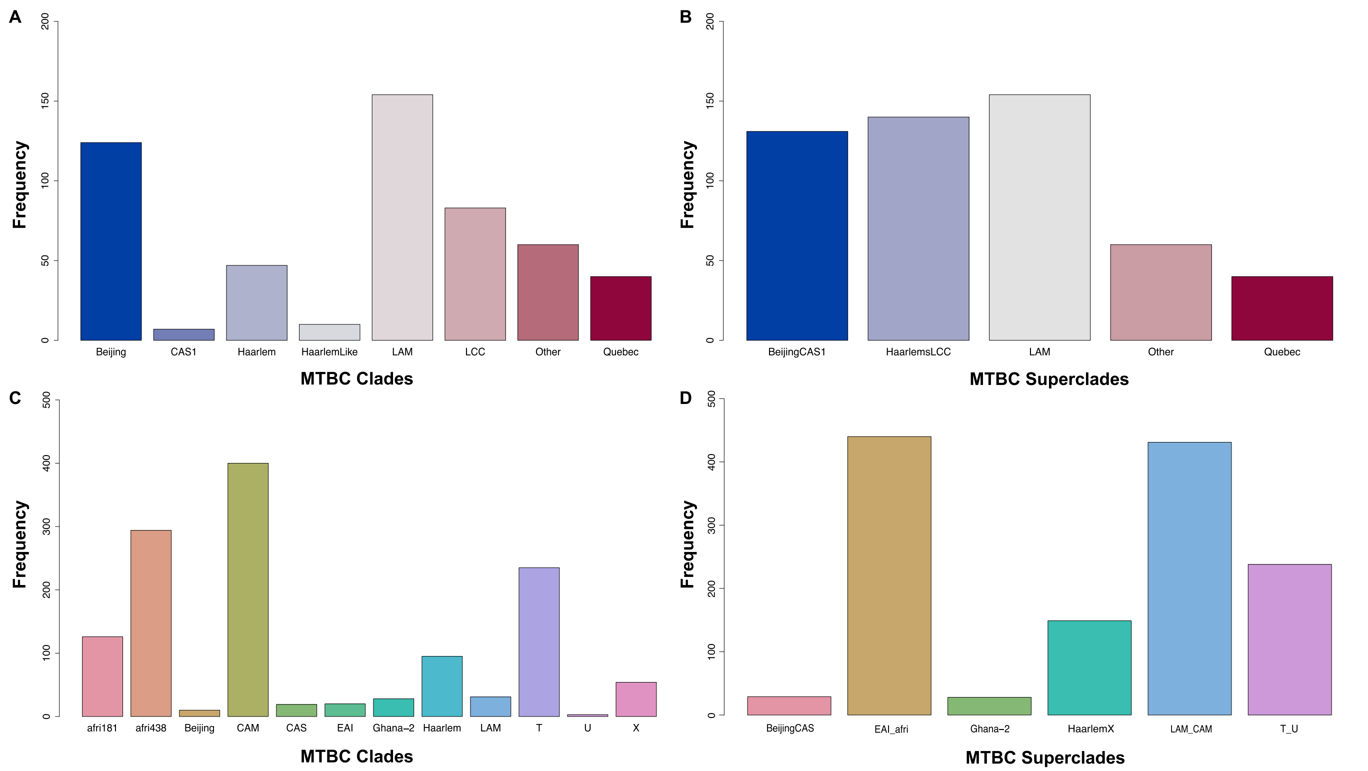

Fig. 2. Frequency distributions of MTBC Clades and Superclades. Frequency distributions of A: MTBC clades and B: superclades for the South African cohort and C: clades and D: superclades for the Ghanaian cohort. The South African cohort was dominated by the LAM superclade, while the Ghanaian cohort was dominated by the EAI_afri and LAM_CAM superclades.

yielded a dataset of 947 participants with 356,165 variants. After quality control filters were applied, 919 participants and 239,612 variants remained. For the Ghanaian cohort, similar analysis yielded a dataset of 3311 participants with 713,223 variants, and genotype QC filters resulted in a further reduction to 617,409 variants. No additional QC steps were used for the IH protocol, while the results of the additional data preparation steps for both cohorts imputed using the MIS tool are detailed in Table 2.

2.3. Selection of high-quality imputed genotype data

Imputation results are presented for Chromosomes 1, and X for the South African cohort, while for the Ghanaian cohort, Chromosome X data was not available, and thus the imputation of two autosomes are reported Table 3. For the South African cohort, the SIS workflow using the AGR resource imputed the highest proportion of SNPs, whereas for the Ghanaian cohort, the MIS workflow using the CAAPA resource imputed the greatest proportion of SNPs with a quality metric greater than 0.45 (Table 3).

For the South African cohort, genotype datum imputed with the AGR reference panel was selected as the dataset with the highest imputation quality across all reference panels assessed. After removal of monomorphic sites and filtering on the INFO or Rsq score, 28,566,283 SNPs for 919 participants remained. After removing the 136 related individuals identified prior to imputation, and filtering for SNP- and sample missingness, and MAF, a dataset of 7,145,406 variants for 783 participants remained. Of the 525 clade-matched samples, 445 were extracted from the dataset of samples passing QC and used in the association analysis.

For the Ghanaian cohort, despite the CAAPA resource imputing the greatest SNP density (Fig. 3D and Fig. 4D), the IH workflow using the 1000G reference panel imputed the highest quality of SNPs per MAF bin but was very closely followed by the other workflows and reference panels from the 20–30% MAF bin upwards (Fig. 5). Thus, the IH dataset, imputed with the 1000G reference panel was selected as the best dataset. After filtering monomorphic SNPs, INDELS, and variants not reaching the INFO score threshold, 25,968,622 SNPs remained for 3239 samples. After filtering for MAF, SNP- and sample missingness, and removal of 93 related individuals identified pre-imputation, the Ghanaian dataset comprised of 5,275,890 variants for 1273 clade-matched samples.

2.4. Covariable data

For the South African cohort, the covariables age, sex, and ancestry proportions were available for all samples. Ancestry proportions were for the European, African, San, South-Asian, and East-Asian ancestries. Of the 445 clade-matched samples that passed the post-imputation QC, ancestry proportions were available for 357 samples as generated previously using the Affymetrix genotype datum for this cohort and ADMIXTURE software [13]. For the remaining 88 samples, ancestry proportions were calculated from genotype datum generated by the MEGA array using ADMIXTURE. The East-Asian ancestry, being the smallest contributing ancestry proportion, was not included as a covariable in the analysis. Variances were calculated for each of the four remaining ancestry proportions and determined to be 0.027 (San), 0.035 (African), 0.014 (European), and 0.009 (South-Asian). As the variances were greater than the minimum cut-off of 0.001, they were included as covariables in the analysis.

For the Ghanaian cohort, age, sex, and ethnicity were included as covariables. Further examination of admixture for the Ghanaian dataset revealed that the cohort was not highly admixed, and thus ethnicity in the form of principal components (PC) was evaluated. One sample passing QC filters did not have one of the covariables and was excluded from the dataset leaving 1272 samples for the association analysis. The

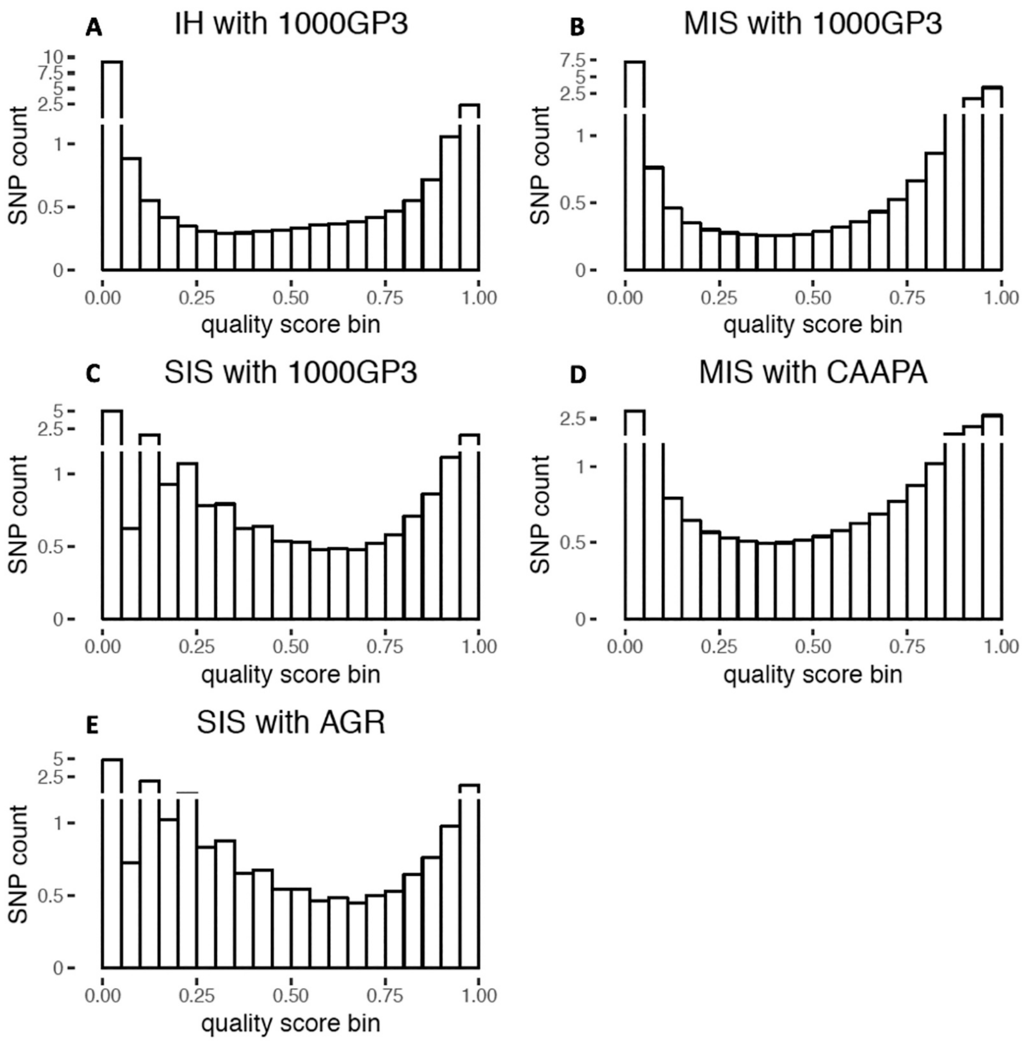

Fig. 3. SNP density plots for Chromosome 1 of the Ghanaian cohort post-imputation using the five workflows. A: IH with 1000G, B: MIS with 1000G, C: SIS with 1000G, D: MIS with CAAPA, E: SIS with AGR. The MIS workflow using the CAAPA resource imputed the greatest proportion of SNPs with a quality metric greater than 0.45.

variance in the PCs provided for the cohort was calculated to be 0.0002 (PC1), 0.0003 (PC2), and 0.0004 (PC3) and therefore determined to be insufficient for inclusion in the association analysis as covariables.

2.5. Multi-phenotype GWAS

2.5.1. South Africa

An MLR was conducted for the 445 superclade-matched South African samples using SNPTEST under an additive model. All results were reported using the LAM superclade as the baseline. Although none of the SNPs passed the GWAS cut-off for significance, 4631 SNPs had an LRT *p*-value less than 0.0005 and eleven SNPs had an LRT p-value less than 1×10^{-6} (Table 4). Odds ratios are reported (Table 4) and standard errors of the odds ratios are shown in Fig. 6A. A single SNP, rs9389610, located on chromosome 6, had a p-value of 1.60×10^{-7}. Individuals with the A allele of this SNP were twice as likely to be have TB due to infection with a member of the BeijingCAS1 superclade (OR: 2.19) than due to either the HaarlemsLCC (OR: 1.07) or LAM superclades. Individuals with the A allele were only slightly more at risk of being infected with a member of the 'Other' superclade (OR: 2.78), when compared to the BeijingCAS1 superclade (OR: 2.19), and were very unlikely to be infected with a member of the Quebec superclade (OR: 0.25).

For the four SNPs located on chromosome 5, individuals with the risk allele doubled the chances of being infected with the Quebec superclade (OR: ~2) while halving the risk of being infected with a member of the "Other" superclade (OR: ~0.5) (Table 4). Lastly, for the six SNPs located on chromosome 17, the risk allele was shown to double the risk of being infected with a member of the LAM superclade than with the HaarlemsLCC (OR: ~0.50) superclade. Individuals with the risk allele of these six SNPs were also equally at risk of being infected with a member of the BeijingCAS1 or LAM superclade and were twice as likely to be infected with the member of the "Other" superclade (OR: ~ 2) when compared to the BeijingCAS1 superclade (OR: ~ 1) (Table 4).

While these SNPs are unlikely to have a direct effect on the gene expression itself, the SNP may be in linkage disequilibrium with other nearby SNPs which do have a direct effect on the gene. For the South African cohort, the most significantly associated SNP was rs9389610 (g.139039029G > A), located on chromosome 6. This SNP is an imputed SNP and its two closest directly genotyped SNPs were rs4896385 (g.139011266G > T), and rs7742202 (g.139074280A > G). The rs4896385 SNP is located in *NHSL-1*, while the rs7742202 SNP is located in *GVQW2*. Neither of these genes have been previously shown to be involved in the pathogenesis of TB. The four SNPs located on chromosome 5 (Table 4) were annotated to the StAR Related Lipid Transfer Domain Containing 4 gene *(STARD4)* using the VEP tool, while the six SNPs located on chromosome 17 were annotated to the *TANC2* gene.

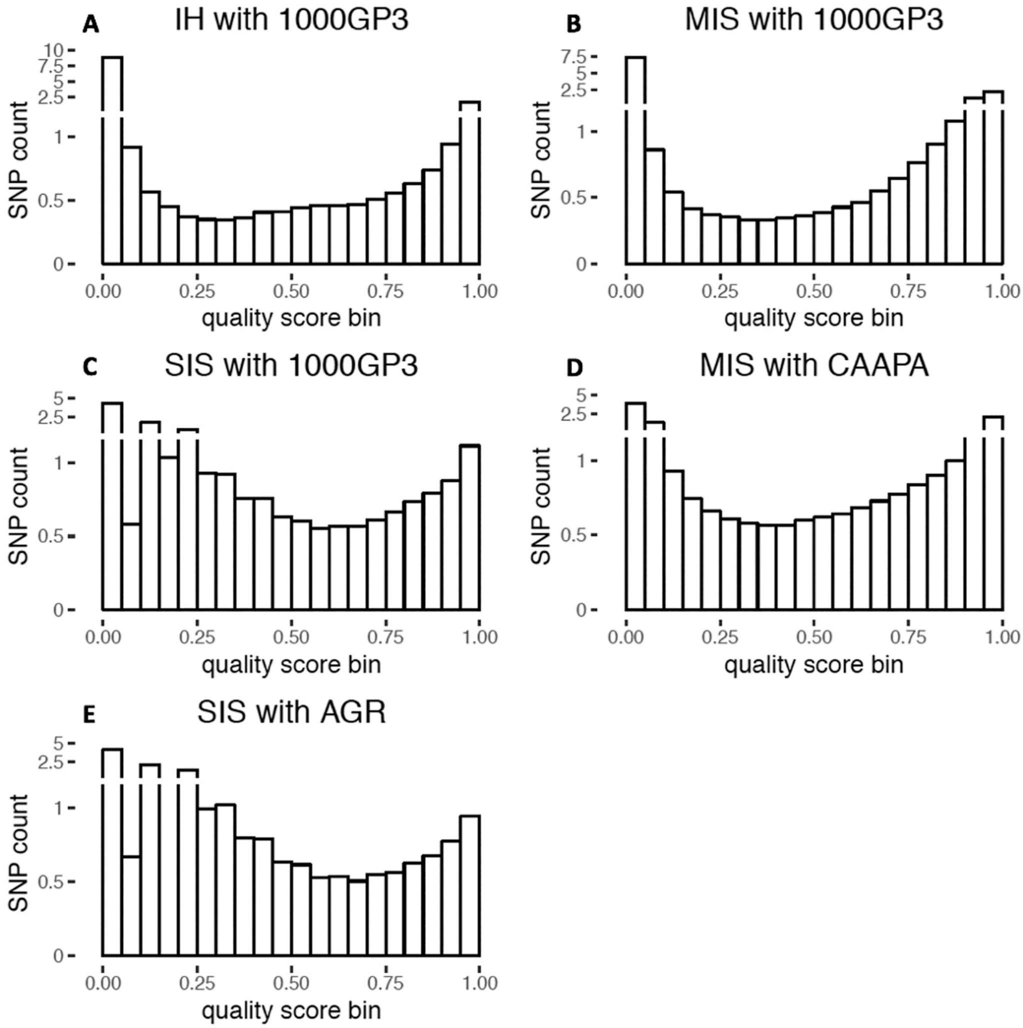

Fig. 4. SNP density plots for Chromosome 22 of the Ghanaian cohort post-imputation using the five workflows. A: IH with 1000G, B: MIS with 1000G, C: SIS with 1000G, D: MIS with CAAPA, E: SIS with AGR.

2.5.2. Ghana

Using SNPTEST under an additive model, an MLR was also conducted for 1272 superclade-matched Ghanaian samples. For this cohort, all association results were reported using the LAM_CAM superclade as the baseline. In summary, a total of 32 SNPs had an LRT *p*-value less than 1×10^{-6} (Table 5) and were significantly associated with the MTBC superclades.

Several imputed SNPs were shown to dramatically increase the risk of being infected by a particular superclade. For example, the risk allele of the SNP rs577800201 (g.20476046C > T) was shown to increase the risk of being infected with the EAI_afri superclade by 93 times, compared to the baseline LAM_CAM superclade, and was annotated by the VEP to map to *ACSM2A*. The risk allele of the SNP rs374315920 (g.38496435C > T) located on chromosome 17, was also found to increase an individual's risk of being infected with the EAI_afri superclade by more than 500 times, as compared to the LAM_CAM reference superclade, and the MLR specific for this SNP was highly significant with an LRT p-value of 1.68×10^{-255}. Using the VEP tool, this SNP was annotated to lie within the retinoic acid receptor alpha (RARA) gene.

Further analysis revealed that of the 32 SNPs identified here, two, namely the intergenic variant rs529920 on chromosome 6, and the intron variant rs41472447 on chromosome 12, were directly genotyped. The risk allele of the SNP rs529920 (g.153835125A > G) was found to halve the risk of being infected with a member of the BeijingCAS and EAI_afri superclades when compared to the LAM_CAM baseline superclade. Individuals with the risk allele for this SNP were also equally at risk of being infected with a member of the Ghana2, HaarlemX, or T_U superclades (OR: 1.04–1.24) (Table 5). Additionally, this SNP had an MAF of 0.4721 in the study cohort, which is similar to the MAF observed in African populations in the 1000G (MAF: 0.571) and in gnomAD genome (MAF:0.582). However, this SNP has no known gene consequence to date.

Similarly, the risk allele of the intron variant rs41472447 (g. 41708761A > G) was found to nearly triple the risk of infection with a member of the BeijingCAS superclade (OR: 2.56, C.I.: (1.48–4.41)) or the Ghana2 superclade (OR:2.94, C.I.: (1.71–5.08)). In contrast, the risk allele halved the risk of infection with the T_U superclade (OR:0.59, C.I.: 0.44–0.79) and had no effect on the risk for infection with members of the EAI_afri or HaarlemX superclades (Table 5). This variant had an MAF of 0.2461 in the study dataset which is similar to the MAF observed in African populations in the 1000G (MAF: 0.200) and in the gnomAD genome (MAF:0.165). Furthermore, rs41472447 maps to the *PDZRN4* gene in dbSNP, but has to date not been linked to susceptibility to tuberculosis.

2.6. Validation of imputed variants and significant associations

Three SNPs with significant associations (rs73497904, rs77139740, rs77641928) from the Ghanaian cohort were successfully genotyped. The genotypes of these three SNPs were used to assess the accuracy of

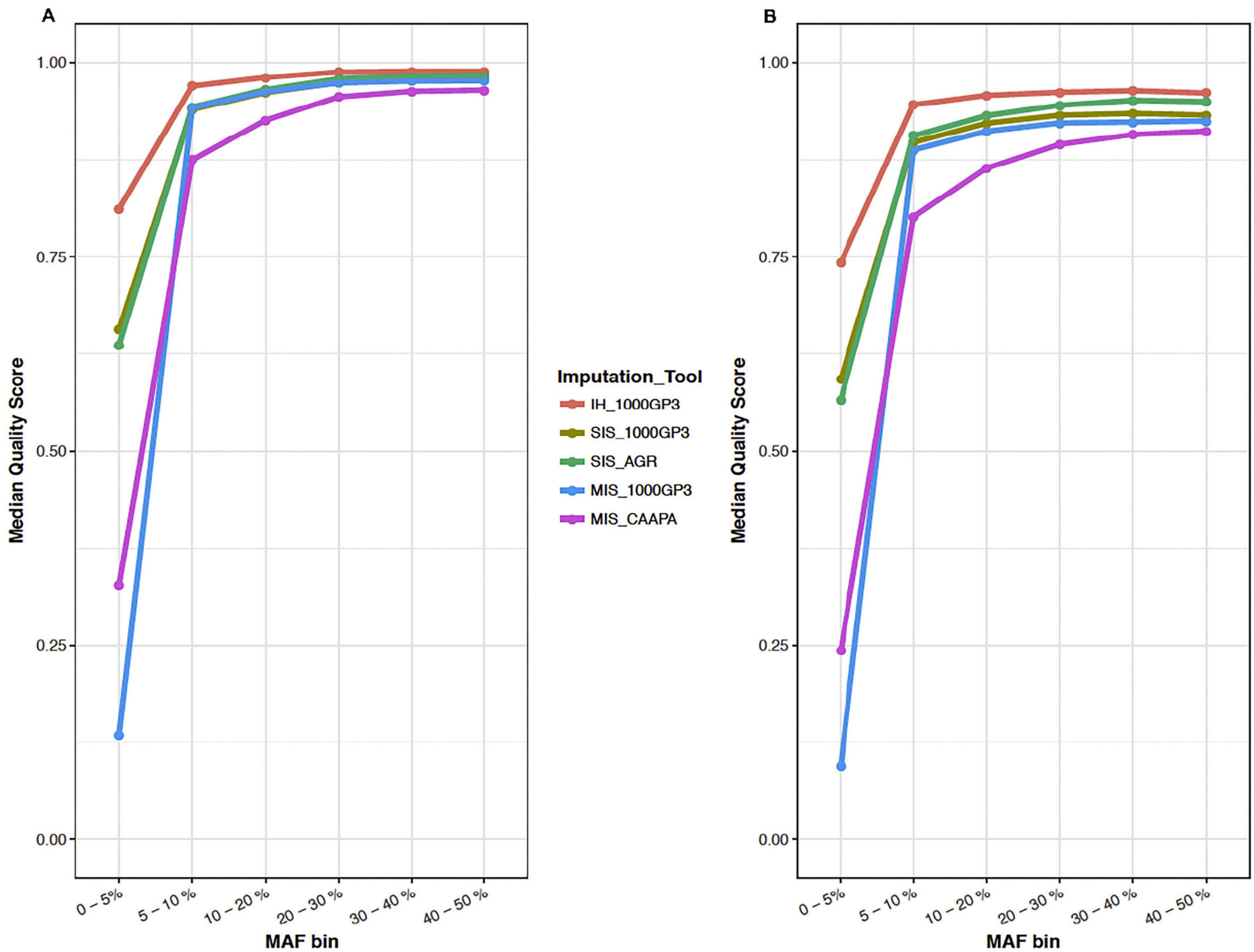

Fig. 5. Median quality scores across MAF bins for the Ghanaian cohort, using the five protocols. A: Chromosome 1 B: Chromosome 22. For both these representative chromosomes, the IH imputation protocol using the 1000G reference panel outperformed the other four workflows.

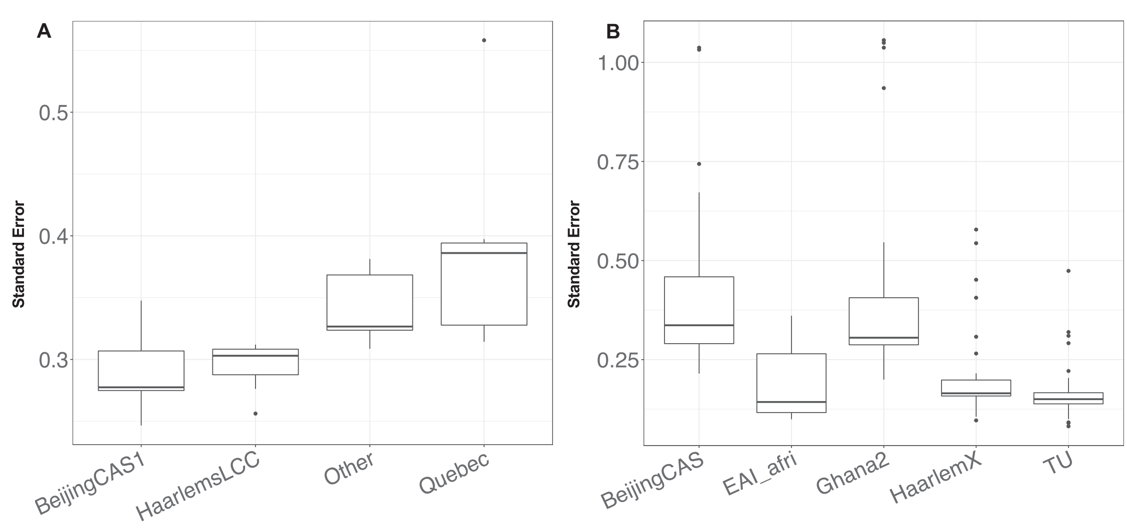

Fig. 6. Standard errors of odds ratios calculated for each superclade against the reference LAM superclade. A: South African cohort B: Ghanaian cohort. For both cohorts, the smallest superclades had the largest variation in their standard errors.

imputation as well as validate the significant associations. When comparing imputed to genotyped alleles for the three variants on an individual level, an overall imputation error rate of 2.25% was achieved (2320 samples with 1323 cases and 997 controls). While the imputation error rate was acceptable, the associations for each variant did not replicate and two of the variants, namely rs77139740 and rs77641928, had a MAF below 5% (Table 6). A possible reason for the associations not replicating is the loss of power as the Ghanaian cohort's sample size was significantly reduced in the original analysis.

3. Discussion

TB is a highly infectious disease affecting millions of people each year. The genetic susceptibility of the host to disease progression has been extensively studied using linkage analysis, candidate gene studies, and GWAS. Furthermore, a number of selected genes have been investigated for their contribution to genetic susceptibility of the host to strains of different lineages of the MTBC. However, to date, no genome-wide analysis of genetic markers affecting susceptibility to strains of different MTBC lineages has been performed.

3.1. Imputation

While several reference panels exist to facilitate genotype imputation, most of these panels focus on representing populations of European ancestry, and little representation has been made for African populations. Therefore, the present study focused on evaluating the quality of imputation attainable for the five-way admixed South African population and the Ghanaian cohort using the 1000G, AGR, and CAAPA reference panels.The more admixed a population is, the greater the heterogeneity in its haplotype block structure. This genetic complexity requires large reference panels with suitable ancestry to facilitate accurate genotype imputation [32]. The five-way admixed South African population contains genetic contributions from Bantu-speaking Africans, Europeans, KhoeSan, and South- and East-Asians [13,14]. Imputation was previously performed for this population; however, it was done using the 1000G Phase 1 [63] and the HapMap3 release 2 (The International HapMap 3 Consortium 2010) reference panels. The 1000G panel has been expanded substantially since whereas the HapMap3 reference panel, representing individuals mostly of European ancestry [10], has since been deprecated. However, previous work by Schurz et al. [53] showed that a genotype imputation accuracy of at least 88% could be achieved when evaluating the five-way admixed South African population, providing an alternative resource for obtaining a more comprehensive genomic dataset.

When evaluating the Ghanaian cohort, although imputation of the 1000G reference panel with the IH method performed the best, there was very little difference in the median quality scores for the different workflows seen for SNPs with an MAF of 10–50% (Fig. 5). For rare variants (MAF 0–5%) however, the IH method outperformed all others with a median quality score above 0.75, whereas both analyses with the MIS produced a median score below the cut-off of 0.45. Thus, for the Ghanaian cohort, all reference panels and methods tested could be considered viable options for imputing common variants with an MAF of 10–50% but should be considered carefully for variants with an MAF below 10%.

In contrast to the AGR which contains no individuals recruited from West-African countries, the CAAPA resource contains 88 individuals recruited from the West-African country of Nigeria [35]. Thus, it was unsurprising that the CAAPA resource performed well when imputing SNPs with a MAF above 10% (Fig. 5). From an MAF of 20–50%, the CAAPA resource performed similarly to the other three reference panels and may thus be considered suitable for imputing cohorts of West-African ancestry, such as the Ghanaian cohort used in this study.

3.2. Multi-phenotype GWAS

In this study, the MLR functionality within SNPTEST enabled the genome-wide investigation of genetic markers for association to a number of MTBC superclades. Although none of the SNPs passed the GWAS *p*-value cut-off, the most significant associations for the South African cohort imputed with the AGR reference panel were reported. This was likely due to the small sample size resulting in a subsequent reduction in statistical power. In contrast, 32 SNPs passed the GWAS cut-off for the Ghanaian cohort imputed with the 1000G reference panel and may be considered as potential targets for further investigation of host-directed therapies suitable for individuals of West-African descent. None of the SNPs with an LRT p-value less than 0.0005 in either cohort were found in the other, demonstrating the population-specific association of SNPs with the strains of different MTBC superclades, which has been previously shown in the investigation of TLRs and their association with cases of TB in populations of different ethnicities [52].

3.3. Potential drug targets

For the Ghanaian cohort, 32 SNPs with significant LRT *p*-values were identified as being associated with the MTBC superclades investigated (Table 5). Nine of the SNPs located on chromosome 12 mapped to the *PDZRN4* gene. For these nine SNPs, the risk allele increased the chances of individuals being infected with the BeijingCAS superclades 2.5 times, and in the region of three times for the Ghana2 superclade, while the risk allele halved the chances of being infected with the T_U superclade. Due to the low frequencies of the BeijingCAS and Ghana2 superclades observed for this cohort (Fig. 2D), it is possible that these odds ratios were inflated because of small sample sizes. Notably, the two SNPs which were directly genotyped, and not imputed, were found to be significantly associated with infection with particular MTBC superclades. This provided crucial evidence that despite the vast amount of imputed genotype data included in the MLR, the association analysis was able to detect two directly genotyped SNPs with potentially significant associations with the MTBC superclades.

Few studies have described the direct influence of *M. tb* infection on the expression of the *STARD4* and *RARA* genes, and no studies have investigated the outcomes of infection with strains belonging to different MTBC clades on these genes. The *STARD4* encodes the StAR-related lipid transfer protein which plays a crucial role in the transmembrane trafficking of lipids (such as cholesterol) - an important source of energy for *M. tb* [59]. Infection of macrophages with pathogens such as *M. tb* stimulates the process of lipid droplet formation [11]. It has been hypothesised that *M. tb* initiates this process in order to secure a reliable source of carbon to fuel bacterial growth [6]. Additionally, the accumulation of cholesterol in the bacterial cell wall drastically reduces the permeability of the cell wall, subsequently reducing the penetrating capability of the anti-TB drug Rifampicin [6]. However, a recent study has contradicted the notion that lipid droplet formation is a bacteria-driven process. Instead, it was proposed that the formation of lipids is an immune system-activated process, and does not occur as a result of direct stimulation by *M. tb*, but rather via the IFN- γ, H1F-alpha-dependent pathway of the host immune system [29].

All-trans retinoic acid (*RARA)*, the active form of Vitamin A, plays an essential role in the normal functioning of the adaptive and innate immune systems. The oral administration of retinoic acid to rats resulted in inhibition of the *M. tb* growth, following in vitro infection [70], thus making this gene a potential target for anti-TB therapies. The results of this study have highlighted several SNPs which possibly significantly increased the risk of individuals with Ghanaian ethnicity to being infected with the endemic TB strain of *M. africanum*. Given the burden of disease, and the dominance of *M. africanum* strains in Ghana, it may be worthwhile exploring the functional effect of these SNPs on the biological processes described.

3.4. Recommended improvements for future multi-phenotype GWAS

Several limitations may have affected this study. Obtaining a suitable sample size is a problem inherent in GWA studies making use of logistic regression modelling. Furthermore, the many phenotypes being analysed in this study, demanded a sufficient number of cases for each class. The frequency of each MTBC clade however, is dependent not only on the host, but is also affected by the virulence of the bacterium. Thus, with all these considered, the sample sizes included in the MLR should be sufficient for inclusion in the analysis but will likely also reflect the distribution in the population. Another limitation of the study is that at the time of analysis, the AGR reference panel was not publicly available for download to a local machine. Thus, its use in this study could only be facilitated via the SIS, a freely accessible online imputation server. Through this, we were able to obtain high-quality imputed data for the South African dataset, but it was necessary to be mindful when drawing comparisons as the other workflows made use of different imputation software.

With the current trajectory of the TB epidemic, novel methods are needed to augment current therapies for TB and combat the disease. This study provides the groundwork for future GWAS wishing to investigate the relationship between the host and the many members of the MTBC causing disease. Furthermore, the SNPs identified in this study may be evaluated in functional studies to assess their viability as targets for host-directed therapies. This study would not have been possible were it not for the collection of paired samples of blood and sputum from study participants. Thus, future studies of this kind will require that both samples be collected from participants in order to perform this association analysis. Although it was necessary to exclude low frequency MTBC cases at the superclade level to prevent the reduction in statistical power of the association test, this may have brought in a weakness in interpreting the odds ratios derived from the model. Thus, the odds ratios derived may only be interpreted at the superclade level and does not provide further granularity to association with specific clades. Therefore, future studies employing this method should consider excluding low-frequency clades before clustering as well as Bayesian analyses that allow inclusion of prior probability distribution for strain prevalence. Additionally, incorporation of more diverse reference panels, such as the AGR, and new algorithms for imputation could improve association results.

4. Material and methods

4.1. Study design

To perform genome-wide association analyses between human host genotypes and the infecting member of the MTBC, host genotypes with paired MTBC isolate information were sourced for two geographically distinct cohorts, namely a South African cohort, and a Ghanaian cohort. As these two cohorts are geographically distinct, and possess vastly different admixture profiles, we did not aim to replicate our findings within these two cohorts.

The South African cohort consisted of study participants recruited in the Western Cape Province of South Africa during the period of January 1993 through December 2004. Participants were recruited from suburbs where the TB incidence was high (28.9% in 2005) and the prevalence of HIV was a low 2% [30,57]. All study participants in this cohort self-identified as belonging to a five-way admixed South African population, were HIV-negative, and provided written informed consent. Blood samples were collected for SNP genotyping of the host and sputum samples were collected for bacterial culture on Loewenstein-Jensen (L-J) media.

For the Ghanaian cohort, participants were enrolled between September 2001 and July 2004 at the Korle Bu Teaching Hospital in Accra, Komfo Anokye Teaching Hospital in Kumasi, and at 15 additional hospitals and polyclinics in Accra and Kumasi, as well as regional district hospitals. All cases were HIV-negative and confirmed to have pulmonary TB by sputum microscopy, performing solid mycobacterial cultures using L-J media and also by two independent radiologists [66,67]. Of the Ghanaian samples included in this study, 1359 were TB cases and 69% of the participants were male (Table 1). Principal Component Analysis of the Ghanaian cohort showed contributing ethnicities from the Akan, Ga-Adangbe, Exe, and several other ethnic groups from northern Ghana [66]. All subsequent research was conducted in accordance with the principles expressed in the Declaration of Helsinki [69].

4.2. Host SNP genotyping

4.2.1. South Africa

SNP genotyping was performed using DNA extracted from blood samples. All participants in the South African cohort were genotyped using the GeneChip Human Mapping 500 K SNP array which contains 500,000 SNP markers (Affymetrix, California, United States), while a subset of this cohort was also genotyped using the Infinium Multi-Ethnic Genotyping Array (MEGA), which is comprised of 1.7 million SNP markers (Illumina, California, United States). Genotype-calling was performed using the Affymetrix Power Tools pipeline (V1.10.0) as previously described [14,53]. Genotype datum was made available in PLINK format (Purcell and Chang, n.d.; [8]). Following standard genotyping quality control (QC), ancestry proportions for the South African cohort on both the Affymetrix and MEGA arrays were estimated [13,53] using the unsupervised algorithm implemented in ADMIXTURE [1].

4.2.2. Ghana

DNA extracted from blood samples was genotyped using the Affymetrix SNP 6.0 array at the Affymetrix Services Laboratory in California, and at ATLAS Biolabs GmbH in Berlin. Participants were successfully genotyped for 783,338 variants (Table 1). Genotype datum was made available in PLINK format, genotypes were called using the Birdseed version 2 algorithm and ancestry proportions in the form of principal components were derived using the Eigenstrat software [66].

4.3. Bacterial SNP genotyping

For the South African cohort, MTBC isolates were genotyped using spoligotyping and IS*6110* Restriction Fragment Length Polymorphism (RFLP) methods as previously described [68]. For the Ghanaian cohort, MTBC isolates extracted from sputum samples were cultured on L-J media at the Kumasi Centre for Collaborative Research [42] and strains were identified using IS*6110* RFLP and spoligotyping [62]. All MTBC isolate information were captured on manually-curated infection databases for archiving and made available for this study.

4.4. Defining MTBC clades and superclades

The term "superclades" was used to describe the grouping of clades using a SNP-based phylogenetic tree of the MTBC. Where a common progenitor was shared, MTBC clades were grouped into superclades near a point of divergence (Fig. 1) [17]. This was performed to reduce the number of clades of low frequency, as low frequency groups are known to induce an unfavourable collinearity effect on logistic regression

Table 1
Summary of patient recruitment for the South African and Ghanaian cohorts.

Cohort	South Africa	Ghana
Cases	853	1359
Male	516	2087
Female	431	1224
Cases + Male	469 (55%)	933 (69%)
Cases + Female	384 (45%)	426 (31%)
Total number of participants	947	3311
Total number of variants	397,337	7838

Table 2
Summary of data pre–processing on the Michigan Imputation Server.

Cohort	South Africa		Ghana	
Reference panel	1000G	CAAPA	1000G	CAAPA
Chromosomes	1–23	1–22	1–22	1–22
Samples	919	919	3239	3239
Sex undefined	3	0	0	0
SNPs	239,612	233,309	617,409	617,409
Alternative allele frequency > 0.5	166	0	0	0
Reference Overlap	99.49	97.94	99.55	97.23
Match	164,643	153,863	420,805	412,495
Allele Switch	74,116	68,912	176,073	172,432
Strand flip	1	0	0	0
Strand flip and allele switch	0	0	1	0
A/T, C/G genotypes	6090	5723	15,683	15,330
Filtered sites				
Filter flag set	0	0	0	0
Invalid alleles	0	0	0	0
Duplicated sites	0	0	0	0
NonSNP sites	0	0	0	0
Monomorphic sites	0	0	0	0
Allele mismatch	54	8	2051	25
SNPs call rate < 90%	0	0	0	0
Excluded sites in total	55	8	2052	25
Sites remaining (before imputation)	239,477	228,498	612,561	600,257
Samples remaining	916	916	3239	3239

Table 3
Percentage proportion of SNPs with a quality metric greater than 0.45.

Cohort	South Africa		Ghana	
Chromosome	Chr 1	Chr X	Chr 1	Chr 22
IH–1000G[1]	39	29	38	39
MIS–1000G[2]	32	18	49	45
MIS–CAAPA[3]	22	–	56	50
SIS–1000G[4]	36	32	41	40
SIS–AGR[5]	43	40	38	36

[1] IH–1000G: In–House workflow using 1000G reference panel.
[2] MIS–1000G: Michigan Imputation Server workflow using 1000G reference panel.
[3] MIS–CAAPA: Michigan Imputation Server workflow using CAAPA reference panel.
[4] SIS–1000G: Sanger Imputation Server workflow using 1000G reference panel.
[5] SIS–AGR: Sanger Imputation Server workflow using AGR reference panel.

models [4]. Additionally, clustering of clades into superclades mitigated the class imbalance which would negatively affect the statistical models. Clades not amalgamated with other clades were also referred to as "superclades" after clustering and those not represented on the phylogenetic tree were kept as distinct superclades and not clustered with any of the members on the existing tree, except when a suitable reference phylogeny was found. After clustering, superclades with a frequency less than ten in the study dataset were excluded from subsequent analyses.

4.5. Genotype data quality control, haplotype phasing, and genotype imputation

Quality control, haplotype phasing and genotype imputation were performed using the methods described in Schurz et al. [54]. Briefly, genotype data were iteratively filtered for 2% SNP genotype missingness, 10% sample missingness and 5% SNP minor allele frequency (MAF), until no additional samples or variants were removed. Additionally, variants lacking chromosome or base pair information were updated using the 1000 Genomes Phase 3 (1000G) reference panel and the dbSNP [56] database. Imputation processes are strengthened by sample size. Thus, all available samples, regardless of relatedness or whether the sample had matching MTBC information, were included in the imputation. Related individuals were noted, but not removed in order to maximise the number of haplotypes available for the imputation process. The genotype QC procedure concluded with a sex concordance check as well as strand-alignment to the 1000G reference panel [human genome build 37 [61]] using the Genotype Harmonizer (version 1.4.15) tool [15].

Following the initial QC, haplotype phasing was performed using the ShapeITv2 [16] software set at default parameters. Although some studies have reported that pre-phasing reduces imputation accuracy [50], it is known to significantly speed up the computationally intensive process of genotype imputation [25,27]. To maximise the number of variants tested in the association analysis, the cleaned genotype data were imputed using five protocols - and three different reference panels - to determine which panel best served the given dataset in imputing missing variants, as detailed in Schurz et al. [54]. In addition, to maximise the amount of informative genotype datum available, all genotyped samples, namely the 947 participants in the South African cohort and the 3311 participants in the Ghanaian cohort, were submitted to the imputation process.

The In-House (IH) protocol made use of the 1000G reference panel with the IMPUTE2 imputation software [26]. The Sanger Imputation Server (SIS) [36] makes use of the Positional Burrows-Wheeler Transformation (PBWT) algorithm [18], with two options for the reference panel: the African Genome Resource (AGR) and the 1000G. Lastly, the Michigan Imputation Server (MIS) [12] makes use of the Minimac3 algorithm [26] and provides access to imputation using the 1000G and the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA) [35] reference panels. Of note, both the CAAPA and AGR reference panels were not publicly accessible for download at the time of this study and thus could only be accessed using these two online imputation server platforms. The key differences between the three

Table 4
Top 11 SNPs identified by MLR to be associated with strains of different MTBC superclades in the South African cohort.

		Allele		Odd Ratio (95% C.I.)				
Chr	SNP ID	Ref[a]	Risk	BeijingCAS1	HaarlemsLCC	Quebec	Other	LRT p–val[b]
5	rs17458866	C	T	0.34 (0.19–0.61)	0.44 (0.26–0.76)	1.99 (1.07–3.68)	0.46 (0.23–0.95)	10e–07
5	rs13355101	G	A	0.31 (0.17–0.57)	0.46 (0.27–0.79)	2.00 (1.08–3.70)	0.47 (0.23–0.96)	6.43e–07
5	rs12518239	C	A	0.29 (0.15–0.56)	0.39 (0.22–0.70)	1.91 (1.01–3.62)	0.51 (0.25–1.05)	9.41e–07
5	rs28769614	C	T	0.27 (0.13–0.53)	0.37 (0.20–0.68)	1.92 (1.00–3.66)	0.48 (0.23–1.01)	3.03e–07
6	rs9389610	G	A	2.19 (1.35–3.55)	1.07 (0.64–1.76)	0.25 (0.08–0.73)	2.78 (1.52–5.08)	1.60e–07
17	rs78022196	G	A	1.04 (0.60–1.80)	0.59 (0.32–1.09)	5.31 (2.44–11.57)	1.96 (1.03–3.73)	5.13e–07
17	rs72843143	C	T	0.98 (0.57–1.69)	0.57 (0.31–1.04)	4.81 (2.26–10.25)	1.89 (1.00–3.58)	8.18e–07
17	rs8071332	A	G	0.94 (0.55–1.60)	0.59 (0.33–1.06)	4.77 (2.22–10.28)	2.03 (1.08–3.81)	6.54e–07
17	rs10438776	T	C	0.99 (0.58–1.71)	0.55 (0.30–1.01)	4.99 (2.30–10.85)	1.94 (1.02–3.69)	3.93e–07
17	rs17682747	G	A	0.98 (0.57–1.68)	0.56 (0.31–1.03)	4.85 (2.27–10.33)	1.90 (1.01–3.58)	5.93e–07
17	rs7208461	T	C	0.94 (0.55–1.61)	0.54 (0.30–0.98)	4.68 (2.17–10.10)	1.77 (0.94–3.34)	8.64e–07

[a] Reference allele.
[b] LAM was used as the reference superclade.

Table 5
Top 32 SNPs identified by MLR to be significantly associated with strains of different MTBC superclades in the Ghanaian cohort.

		Allele		Odd Ratio (95% C.I.)					
Chr	SNP ID	Ref[a]	Risk	BeijingCAS	EAI_afri	Ghana2	HaarlemX	T_U	LRT p–val[b]
6	rs529920[*]	A	G	0.40 (0.22–0.70)	0.69 (0.57–0.84)	1.04 (0.60–1.80)	1.10 (0.84–1.44)	1.24 (0.98–1.56)	1.86e–07
12	rs73418916	A	G	0.30 (0.11–0.81)	34.15 (20.00–58.33)	1.38 (0.64–2.96)	0.76 (0.51–1.12)	0.88 (0.63–1.22)	2.31e–97
12	rs138396290	T	C	0.32 (0.13–0.77)	4.59 (3.63–5.82)	1.31 (0.73–2.37)	0.87 (0.63–1.19)	0.91 (0.70–1.19)	3.32e–62
12	rs75717431	T	C	2.46 (1.40–4.35)	0.98 (0.78–1.24)	3.30 (1.88–5.82)	0.95 (0.69–1.31)	0.59 (0.44–0.80)	8.55e–09
12	rs77428482	G	A	2.56 (1.45–4.52)	1.01 (0.81–1.28)	3.20 (1.81–5.64)	0.95 (0.69–1.32)	0.60 (0.45–0.81)	1.93e–08
12	rs77562721	G	A	2.50 (1.42–4.41)	1.04 (0.82–1.30)	3.12 (1.77–5.50)	0.93 (0.68–1.29)	0.60 (0.45–0.81)	2.53e–08
12	rs41524146	C	G	2.38 (1.36–4.18)	1.00 (0.80–1.26)	3.19 (1.82–5.58)	0.97 (0.71–1.34)	0.61 (0.45–0.81)	2.62e–08
12	rs7299395	G	A	2.16 (1.22–3.82)	0.99 (0.80–1.24)	3.12 (1.77–5.52)	0.93 (0.69–1.27)	0.64 (0.48–0.84)	2.50e–07
12	rs74550821	G	A	2.59 (1.47–4.57)	1.05 (0.84–1.32)	3.20 (1.82–5.63)	1.01 (0.74–1.39)	0.63 (0.47–0.84)	2.89e–08
12	rs144335343	C	T	2.33 (1.32–4.11)	1.08 (0.87–1.35)	3.63 (2.05–6.41)	1.00 (0.73–1.36)	0.69 (0.52–0.92)	9.15e–08
12	rs6582329	A	T	2.58 (1.46–4.57)	1.07 (0.85–1.34)	3.42 (1.94–6.02)	1.04 (0.76–1.43)	0.67 (0.50–0.90)	6.00e–08
12	rs12296167	T	G	2.42 (1.25–4.70)	2.03 (1.60–2.59)	3.33 (1.71–6.48)	1.03 (0.74–1.42)	1.41 (1.07–1.85)	2.02e–09
12	rs544003050	A	G	0.5 (0.24–1.05)	4.72 (3.47–6.43)	1.42 (0.68–2.98)	0.75 (0.52–1.07)	0.74 (0.54–1.00)	4.17e–40
12	rs58262822	C	G	0.58 (0.34–0.99)	8.50 (6.29–11.48)	0.75 (0.47–1.20)	0.81 (0.65–1.01)	0.93 (0.78–1.11)	1.96e–124
12	rs11108508	T	C	1.09 (0.56–2.09)	6.61 (5.26–8.32)	1.33 (0.73–2.44)	0.71 (0.48–1.05)	1.11 (0.85–1.46)	1.04e–112
12	rs41472447[*]	A	G	2.56 (1.48–4.41)	0.97 (0.78–1.21)	2.94 (1.71–5.08)	0.92 (0.68–1.26)	0.59 (0.44–0.79)	5.70e–09
13	rs549053537	A	T	0.63 (0.41–0.96)	5.51 (4.12–7.39)	1.23 (0.83–1.82)	0.86 (0.71–1.04)	0.91 (0.78–1.07)	8.12e–79
13	rs73497904[†]	C	G	1.34 (0.69–2.59)	65.39 (40.67–105.13)	1.20 (0.60–2.39)	0.96 (0.67–1.39)	1.23 (0.92–1.65)	2.23e–244
13	rs9524738	G	C	1.88 (0.86–4.14)	0.31 (0.26–0.38)	1.32 (0.68–2.55)	1.34 (0.97–1.84)	1.16 (0.90–1.50)	2.67e–52
15	rs551641937	G	A	0.52 (0.07–3.97)	275.89 (152.64–498.69)	0.52 (0.07–3.98)	0.39 (0.14–1.14)	1.11 (0.59–2.08)	3.61e–236
15	rs35799802	C	T	0.76 (0.36–1.62)	3.08 (2.29–4.13)	0.89 (0.42–1.88)	1.09 (0.75–1.59)	1.19 (0.87–1.64)	3.94e–15
15	rs55747528	C	T	0.53 (0.24–1.15)	28.58 (15.28–53.45)	1.47 (0.66–3.25)	0.72 (0.49–1.05)	0.79 (0.57–1.08)	1.49e–69
16	rs577800201	C	T	0.31 (0.07–1.33)	93.05 (54.17–159.85)	0.87 (0.32–2.36)	0.67 (0.4–1.13)	1.01 (0.68–1.50)	1.20e–167
16	rs187181146	C	T	0.64 (0.35–1.17)	8.65 (6.61–11.31)	0.94 (0.59–1.49)	0.83 (0.66–1.06)	0.92 (0.76–1.12)	9.77e–152
16	rs35868343	G	A	0.72 (0.46–1.11)	10.65 (7.14–15.87)	1.01 (0.66–1.55)	0.91 (0.74–1.12)	0.92 (0.77–1.10)	1.90e–93
17	rs143309838	G	A	0.57 (0.28–1.17)	9.38 (7.20–12.23)	0.93 (0.57–1.53)	0.84 (0.65–1.08)	0.99 (0.81–1.21)	2.08e–165
17	rs144224512	C	G	0.36 (0.05–2.69)	220.49 (118.29–411.01)	0.50 (0.08–3.13)	0.42 (0.17–1.01)	0.74 (0.40–1.36)	1.21e–215
17	rs374315920	C	T	3.52 (0.94–13.13)	548.96 (270.65–1113.45)	1.11 (0.14–8.79)	0.85 (0.27–2.63)	0.93 (0.37–2.36)	1.68e–255
17	rs77139740[†]	A	G	0.34 (0.12–1.01)	44.94 (26.85–75.20)	0.97 (0.43–2.22)	0.73 (0.48–1.12)	0.89 (0.63–1.26)	3.74e–120
17	rs77641928[†]	G	A	1.24 (0.37–4.17)	57.29 (33.69–97.42)	0.46 (0.06–3.62)	0.67 (0.30–1.49)	0.94 (0.53–1.67)	4.98e–282
22	rs553728019	A	C	0.51 (0.23–1.09)	23.66 (12.10–46.25)	1.05 (0.47–2.33)	0.76 (0.52–1.11)	0.79 (0.57–1.09)	2.83e–55
22	rs60153275	C	T	0.93 (0.32–2.71)	61.01 (36.66–101.54)	0.93 (0.32–2.71)	0.70 (0.38–1.28)	0.98 (0.63–1.51)	7.19e–281

[a] Reference allele.
[b] LAM_CAM was used as the reference superclade.
[*] Genotyped SNPs.
[†] SNPs successfully genotyped for validation.

Table 6
Association testing results for validated variants.

SNP	Ref[a]	Risk	MAF	HWE p-value (controls)	OR	P-value
rs73497904	C	G	0.15	0.35	0.74	0.18
rs77139740	A	G	0.02	0.59	0.99	0.99
rs77641928	G	A	0.03	0.27	1.06	0.72

[a] Reference allele.

protocols using the 1000G reference panel was the imputation software used, as well as additional strict QC filters imposed on the study dataset by the MIS and SIS platforms.

The three human genome reference panels used in this study offered access to a wide variety of genotype datum. Spanning 26 populations across the world, the 1000G offers one of the most diverse reference panels to have been compiled to date by including samples sourced from African, American, European, South- and East-Asian countries. Continental African populations contributing to the reference panel include individuals from the Esan ethnic group in Nigeria, Luhya in Kenya, the Yoruban people, as well as participants from The Gambia. Approximately half of the AGR resource is comprised of samples from the 1000G, while around 2000 samples were sourced from regions in the East-African country of Uganda. Around 100 samples were sourced from several regions in Ethiopia, as well as from Egypt, the Zulu people in South Africa, and the Nama/Khoesan people in Namibia. Lastly, the CAAPA resource [35] is comprised of approximately one third of the samples on the 1000G and just over a fifth of the number of samples on the AGR reference panel. The resource includes individuals self-reporting as having African ancestry and were recruited from nine cities in the United States, four populations in the Caribbean, four in Central- and South America, and two populations representing West Africa.

4.6. Selection of high-quality imputed genotype data

The quality control procedure for the imputed data was implemented as described in Schurz et al. [54]. Briefly, following five imputation protocols, imputed data were filtered using a genotype calling threshold of 0.7, and the internal quality metric produced by the imputation process [54]. SNPs with an INFO or R-squared (Rsq) value greater than 0.45 were prioritised for the association analysis and filtered iteratively for a maximum of 2% SNP genotype missingness, 10% sample missingness, and 5% SNP MAF using PLINK. Related individuals identified prior to imputation were removed followed by a second round of iterative filters for SNP- and sample missingness and MAF [54]. For both chromosomes 1 and X, imputation using either the 1000G or the CAAPA resource with the MIS performed the worst for the South African cohort [54] with the maximum median quality score only reaching 0.82 at a MAF of 50%. In comparison, the SIS-AGR workflow outperformed all other workflows, and the result correlated with the AGR imputing the highest SNP density for chromosome 1 and chromosome X [54]. Post-imputation QC concluded with extracting MTBC clade-matched samples from the remaining samples which had passed all QC filters.

4.7. Covariable data

All available covariables were obtained including sex, and age at time of active TB and subsequent recruitment into the study. To correct for differences in ethnicity among participants, either ancestry proportions or principal components were calculated and included as covariables. SNPTEST is unable to include covariables when the variance in the values provided is "too small" as indicated [34]. At the time of this study, SNPTEST was still under development, and it had not yet been established, or recorded in the software manual, to what degree of variance covariable data would not be accepted for inclusion in the logistic regression. For developing the method, it was established through trial and error that if the variance was below 0.001, these covariables could not be included.

4.8. Multi-phenotype GWAS

For the association analysis, a multinomial logistic regression (MLR) analysis using an additive genetic model was performed using SNPTEST v2.5.2 [33]. Two discrete variables, namely sex and superclade, as well as continuous variables, namely age at TB onset and ancestry proportions or principal components were included in the analysis. The phenotype tested was specified as the MTBC superclade. Thus, the MLR model specified was the occurrence of the MTBC superclade as a function of the baseline covariables given, as well as the host genotypes supplied.

The standard genome-wide significance cut-off of alpha = 5×10^{-8} was used when reporting significance of SNPs [44,64]. Odds ratios (OR) for the multiple phenotypes tested were calculated against a baseline phenotype by setting the odds of that phenotype occurring, given the genotype, to 1. For this study, the baseline phenotype was specified as the dominant superclade in the cohort, or a common superclade of intermediate frequency if more than one cohort was being studied. Thus, the LAM- and LAM_CAM superclades were used as the baseline phenotype for the association analyses of the South African, and Ghanaian cohorts, respectively.

SNPs with a Likelihood Ratio Threshold (LRT) *p*-value of less than 5×10^{-4} were selected and analysed in R (R [48]) and OR's were calculated from the beta values generated by SNPTEST. SNPs with a standard error greater than 1.5 for their OR's were excluded and SNPs with an LRT p-value less than 1×10^{-6} were prioritised for further investigation. These thresholds were chosen pragmatically to facilitate the completion of method development. Finally, the Variant Effect Predictor (VEP) Tool [38] was used to retrieve gene annotations for the SNPs of interest.

4.9. Validation of imputed variants and significant associations

Selected SNPs with significant associations for a particular cohort were genotyped on assays designed using the the ProbeDesign software by Roche. The genotypes were used to assess the accuracy of imputation as well as to validate the significant associations by comparing imputed to genotyped alleles for the variants on an individual level. Logistic regression was performed using PLINK [8,47], with the inclusion of age and sex as covariables.

Ethics approval and consent to participate

For the South African cohort, blood and sputum samples were collected from study participants as approved by the Health Research Ethics Committee of Stellenbosch University (Project numbers: S17/01/013 and 95/072). For the Ghanaian cohort, ethics for the study protocol was granted by the Committee on Human Research, Publications and Ethics, School of Medical Sciences, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana, and the Ethics Committee of the Ghana Health Service, Accra, Ghana [66]. Venous blood samples were taken only after a detailed explanation of the aims of the study, and consent was obtained of individuals enrolled or their parents/guardians by signature or by thumbprint in case of illiteracy.

Availability of data and materials

Summary statistics for the South African cohort's data can be made available to researchers who meet the criteria for access to confidential data after application to the Health Research Ethics Committee of Stellenbosch University. Requests may be sent to: Prof Craig Kinnear, *E*-mail: gkin@sun.ac.za.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Acknowledgements

The authors would like to acknowledge and thank the study participants for their contribution and participation. This research was partially funded by the South African government through the South African Medical Research Council. This work was also supported by the National Research Foundation of South Africa (grant number 93460) to E.H. and by a Strategic Health Innovation Partnership grant from the South African Medical Research Council and Department of Science and Innovation/South African Tuberculosis Bioinformatics Initiative to G.T. The content is solely the responsibility of the authors and does not necessarily represent the official views of the South African Medical Research Council.

References

[1] D.H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals, Genome Res. (2009) 1655–1664, https://doi.org/10.1101/gr.094052.109.vidual.
[2] A. Asante-Poku, D. Yeboah-Manu, I.D. Otchere, S.Y. Aboagye, D. Stucki, J. Hattendorf, S. Borrell, J. Feldmann, E. Danso, S. Gagneux, Mycobacterium africanum is associated with patient ethnicity in Ghana, PLoS Negl. Trop. Dis. 9 (2015), e3370, https://doi.org/10.1371/journal.pntd.0003370.
[3] R. Bellamy, Genetic susceptibility to tuberculosis in human populations, Thorax 53 (1998) 588–593.
[4] J.S. Bergtold, E.A. Yeager, A. Featherstone, Sample Size and Robustness of Inferences from Logistic Regression in the Presence of Nonlinearity and Multicollinearity, in: Presented at the Agricultural & Applied Economics Association's 2011 AAEA & NAREA Joint Annual Meeting, Pittsburgh, Pennsylvania, 2011.
[5] D. Brites, S. Gagneux, Co-evolution of Mycobacterium tuberculosis and Homo sapiens, Immunol. Rev. 264 (2015) 6–24, https://doi.org/10.1111/imr.12264.
[6] A. Brzostek, J. Pawelczyk, A. Rumijowska-Galewicz, B. Dziadek, J. Dziadek, Mycobacterium tuberculosis Is Able To Accumulate and Utilize Cholesterol, J. Bacteriol. 191 (2009) 6584–6591, https://doi.org/10.1128/JB.00488-09.
[7] M. Caws, G. Thwaites, S. Dunstan, T.R. Hawn, N. Thi Ngoc Lan, N.T.T. Thuong, K. Stepniewska, M.N.T. Huyen, N.D. Bang, T. Huu Loc, S. Gagneux, D. van Soolingen, K. Kremer, M. van der Sande, P. Small, P. Thi Hoang Anh, N.T. Chinh, H. Thi Quy, N. Thi Hong Duyen, D. Quang Tho, N.T. Hieu, E. Torok, T.T. Hien, N. H. Dung, N. Thi Quynh Nhu, P.M. Duy, N. van Vinh Chau, J. Farrar, The influence of host and bacterial genotype on the development of disseminated disease with *Mycobacterium tuberculosis*, PLoS Pathog. (2008) 4, https://doi.org/10.1371/journal.ppat.1000034.
[8] C.C. Chang, C.C. Chow, L.C. Tellier, S. Vattikuti, S.M. Purcell, J.J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets, GigaScience 4 (2015), https://doi.org/10.1186/s13742-015-0047-8.
[9] M.P. Cheng, C.N. Abou Chakra, C.P. Yansouni, S. Cnossen, I. Shrier, D. Menzies, C. Greenaway, Risk of active tuberculosis in patients with Cancer: a systematic review and meta-analysis, Clin. Infect. Dis. (2016) ciw838, https://doi.org/10.1093/cid/ciw838.
[10] E.R. Chimusa, N. Zaitlen, M. Daya, M. Möller, P.D. van Helden, J.M. Nicola, A. L. Price, E.G. Hoal, Genome-wide association study of ancestry-specific TB risk in the south African coloured population, Hum. Mol. Genet. 23 (2014) 796–809, https://doi.org/10.1093/hmg/ddt462.
[11] J. Daniel, H. Maamar, C. Deb, T.D. Sirakova, P.E. Kolattukudy, Mycobacterium tuberculosis uses host triacylglycerol to accumulate lipid droplets and acquires a dormancy-like phenotype in lipid-loaded macrophages, PLoS Pathog. 7 (2011), e1002093, https://doi.org/10.1371/journal.ppat.1002093.
[12] S. Das, L. Forer, S. Schönherr, C. Sidore, A.E. Locke, A. Kwong, S.I. Vrieze, E. Y. Chew, S. Levy, M. McGue, D. Schlessinger, D. Stambolian, P.-R. Loh, W. G. Iacono, A. Swaroop, L.J. Scott, F. Cucca, F. Kronenberg, M. Boehnke, G. R. Abecasis, C. Fuchsberger, Next-generation genotype imputation service and methods, Nat. Genet. 48 (2016) 1284–1287, https://doi.org/10.1038/ng.3656.
[13] M. Daya, L. Van Der Merwe, U. Galal, M. Möller, M. Salie, E.R. Chimusa, J. M. Galanter, P.D. Van Helden, B.M. Henn, C.R. Gignoux, E. Hoal, A panel of

ancestry informative markers for the complex five-way admixed South African Coloured population, PLoS One 8 (2013) 12, https://doi.org/10.1371/journal.pone.0082224.
[14] E. De Wit, W. Delport, C.E. Rugamika, A. Meintjes, M. Moller, P.D. Van Helden, C. Seoighe, E.G. Hoal, Genome-wide analysis of the structure of the south African Coloured population in the Western cape, Hum. Genet. 128 (2010) 145–153, https://doi.org/10.1007/s00439-010-0836-1.
[15] P. Deelen, M.J. Bonder, K.J. van der Velde, H.-J. Westra, E. Winder, D. Hendriksen, L. Franke, M.A. Swertz, Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration, BMC Res. Notes 7 (2014) 901.
[16] O. Delaneau, C. Coulonges, J.-F. Zagury, Shape-IT: new rapid and accurate algorithm for haplotype inference, BMC Bioinformatics 9 (2008) 540, https://doi.org/10.1186/1471-2105-9-540.
[17] A. Dippenaar, A phylogenomic-and proteomic investigation into the evolution and biological characteristics of the members of the group 2 Latin-American Mediterranean (LAM) genotype of *Mycobacterium tuberculosis* (PhD Thesis), Stellenbosch University, tellenbosch, 2014.
[18] R. Durbin, Efficient haplotype matching and storage using the positional burrows–wheeler transform (PBWT), Bioinformatics 30 (2014) 1266–1272, https://doi.org/10.1093/bioinformatics/btu014.
[19] S. Gagneux, Host-pathogen coevolution in human tuberculosis, Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. 367 (2012) 850–859, https://doi.org/10.1098/rstb.2011.0316.
[20] S. Gagneux, P.M. Small, Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development, Lancet Infect. Dis. 7 (2007) 328–337, https://doi.org/10.1016/S1473-3099(07)70108-1.
[21] F. Herb, T. Thye, S. Niemann, E.N.L. Browne, M.A. Chinbuah, J. Gyapong, I. Osei, E. Owusu-Dabo, O. Werz, S. Rusch-Gerdes, R.D. Horstmann, C.G. Meyer, ALOX5 variants associated with susceptibility to human pulmonary tuberculosis, Hum. Mol. Genet. 17 (2007) 1052–1060, https://doi.org/10.1093/hmg/ddm378.
[22] J.N. Hirschhorn, M.J. Daly, Genome-wide association studies for common diseases and complex traits, Nat. Rev. Genet. 6 (2005) 95–108, https://doi.org/10.1038/nrg1521.
[23] E.G. Hoal, A. Dippenaar, C. Kinnear, P.D. van Helden, M. Möller, The arms race between man and Mycobacterium tuberculosis : time to regroup, Infect. Genet. Evol. (2017), https://doi.org/10.1016/j.meegid.2017.08.021.
[24] E.P. Hong, M.J. Go, H.-L. Kim, J.W. Park, Risk prediction of pulmonary tuberculosis using genetic and conventional risk factors in adult Korean population, PLoS One 12 (2017), e0174642, https://doi.org/10.1371/journal.pone.0174642.
[25] B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, G.R. Abecasis, Fast and accurate genotype imputation in genome-wide association studies through pre-phasing, Nat. Genet. 44 (2012) 955–959, https://doi.org/10.1038/ng.2354.
[26] B.N. Howie, P. Donnelly, J. Marchini, A flexible and accurate genotype imputation method for the next generation of genome-wide association studies, PLoS Genet. 5 (2009), e1000529, https://doi.org/10.1371/journal.pgen.1000529.
[27] A. Kanterakis, P. Deelen, F. van Dijk, H. Byelas, M. Dijkstra, M.A. Swertz, Molgenis-impute: imputation pipeline in a box, BMC Res. Notes 8 (2015), https://doi.org/10.1186/s13104-015-1309-3.
[28] C. Kinnear, E.G. Hoal, H. Schurz, P.D. Van Helden, M. Moller, The role of human host genetics in tuberculosis resistance, Expert Rev. Respir. Med. 11 (2017) 721–737, https://doi.org/10.1080/17476348.2017.1354700.
[29] M. Knight, J. Braverman, K. Asfaha, K. Gronert, S. Stanley, Lipid droplet formation in Mycobacterium tuberculosis infected macrophages requires IFN-γ/HIF-1α signaling and supports host defense, PLoS Pathog. 14 (2018), e1006874, https://doi.org/10.1371/journal.ppat.1006874.
[30] F.E. Kritzinger, S. den Boon, S. Verver, D.A. Enarson, C.J. Lombard, M. W. Borgdorff, R.P. Gie, N. Beyers, No decrease in annual risk of tuberculosis infection in endemic area in Cape Town, South Africa, Trop. Med. Int. Health 14 (2009) 136–142, https://doi.org/10.1111/j.1365-3156.2008.02213.x.
[31] C.H. Lim, H.-H. Chen, Y.-H. Chen, D.-Y. Chen, W.-N. Huang, J.-J. Tsai, T.-Y. Hsieh, C.-W. Hsieh, W.-T. Hung, C.-T. Lin, K.-L. Lai, K.-T. Tang, C.-W. Tseng, Y.-M. Chen, The risk of tuberculosis disease in rheumatoid arthritis patients on biologics and targeted therapy: a 15-year real world experience in Taiwan, PLoS One 12 (2017), e0178035, https://doi.org/10.1371/journal.pone.0178035.
[32] M. Lin, C. Caberto, P. Wan, Y. Li, A. Lum-Jones, M. Tiirikainen, L. Pooler, B. Nakamura, X. Sheng, J. Porcel, U. Lim, V.W. Setiawan, L. Le Marchand, L. R. Wilkens, C.A. Haiman, I. Cheng, C.W.K. Chiang, Population-specific reference panels are crucial for genetic analyses: an example of the CREBRF locus in native Hawaiians, Hum. Mol. Genet. 29 (2020) 2275–2284, https://doi.org/10.1093/hmg/ddaa083.
[33] J. Marchini, SNPTEST v2 Technical Details 10, 2010.
[34] J. Marchini, SNPTest, 2021. https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=oxstatgen;f8f2270b.1207.
[35] R.A. Mathias, M.A. Taub, C.R. Gignoux, W. Fu, S. Musharoff, T.D. O'Connor, C. Vergara, D.G. Torgerson, M. Pino-Yanes, S.S. Shringarpure, L. Huang, N. Rafaels, M.P. Boorgula, H.R. Johnston, V.E. Ortega, A.M. Levin, W. Song, R. Torres, B. Padhukasahasram, C. Eng, D.-A. Mejia-Mejia, T. Ferguson, Z.S. Qin, A.F. Scott, M. Yazdanbakhsh, J.G. Wilson, J. Marrugo, L.A. Lange, R. Kumar, P.C. Avila, L. K. Williams, H. Watson, L.B. Ware, C. Olopade, O. Olopade, R. Oliveira, C. Ober, D. L. Nicolae, D. Meyers, A. Mayorga, J. Knight-Madden, T. Hartert, N.N. Hansel, M. G. Foreman, J.G. Ford, M.U. Faruque, G.M. Dunston, L. Caraballo, E.G. Burchard, E. Bleecker, M.I. Araujo, E.F. Herrera-Paz, K. Gietzen, W.E. Grus, M. Bamshad, C. D. Bustamante, E.E. Kenny, R.D. Hernandez, T.H. Beaty, I. Ruczinski, J. Akey, K. C. Barnes, A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome, Nat. Commun. 7 (2016), https://doi.org/10.1038/ncomms12522.
[36] S. McCarthy, S. Das, W. Kretzschmar, O. Delaneau, A.R. Wood, A. Teumer, H. M. Kang, C. Fuchsberger, P. Danecek, K. Sharp, Y. Luo, C. Sidore, A. Kwong, N. Timpson, S. Koskinen, S. Vrieze, L.J. Scott, H. Zhang, A. Mahajan, J. Veldink, U. Peters, C. Pato, C.M. van Duijn, C.E. Gillies, I. Gandin, M. Mezzavilla, A. Gilly, M. Cocca, M. Traglia, A. Angius, J.C. Barrett, D. Boomsma, K. Branham, G. Breen, C.M. Brummett, F. Busonero, H. Campbell, A. Chan, S. Chen, E. Chew, F.S. Collins, L.J. Corbin, G.D. Smith, G. Dedoussis, M. Dorr, A.-E. Farmaki, L. Ferrucci, L. Forer, R.M. Fraser, S. Gabriel, S. Levy, L. Groop, T. Harrison, A. Hattersley, O.L. Holmen, K. Hveem, M. Kretzler, J.C. Lee, M. McGue, T. Meitinger, D. Melzer, J.L. Min, K. L. Mohlke, J.B. Vincent, M. Nauck, D. Nickerson, A. Palotie, M. Pato, N. Pirastu, M. McInnis, J.B. Richards, C. Sala, V. Salomaa, D. Schlessinger, S. Schoenherr, P. E. Slagboom, K. Small, T. Spector, D. Stambolian, M. Tuke, J. Tuomilehto, L.H. Van den Berg, W. Van Rheenen, U. Volker, C. Wijmenga, D. Toniolo, E. Zeggini, P. Gasparini, M.G. Sampson, J.F. Wilson, T. Frayling, P.I.W. de Bakker, M. A. Swertz, S. McCarroll, C. Kooperberg, A. Dekker, D. Altshuler, C. Willer, W. Iacono, S. Ripatti, N. Soranzo, K. Walter, A. Swaroop, F. Cucca, C.A. Anderson, R.M. Myers, M. Boehnke, M.I. McCarthy, R. Durbin, G. Abecasis, J. Marchini, the Haplotype Reference Consortium, A reference panel of 64,976 haplotypes for genotype imputation, Nat. Genet. 48 (2016) 1279–1283, https://doi.org/10.1038/ng.3643.
[37] M.L. McHenry, J. Bartlett, R.P. Igo, E. Wampande, P. Benchek, H. Mayanja-Kizza, K. Fluegge, N.B. Hall, S. Gagneux, S.A. Tishkoff, C. Wejse, G. Sirugo, W.H. Boom, M. Joloba, S.M. Williams, C.M. Stein, Interaction between host genes and M. tuberculosis lineage can affect tuberculosis severity: evidence for coevolution, bioRxiv (2019) 769448, https://doi.org/10.1101/769448.
[38] W. McLaren, L. Gil, S.E. Hunt, H.S. Riat, G.R.S. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The Ensembl variant effect predictor, Genome Biol. 17 (2016), https://doi.org/10.1186/s13059-016-0974-4.
[39] M. Möller, E.G. Hoal, Current findings, challenges and novel approaches in human genetic susceptibility to tuberculosis, Tuberculosis 90 (2010) 71–83, https://doi.org/10.1016/j.tube.2010.02.002.
[40] N.O. Oki, A.A. Motsinger-Reif, P.R. Antas, S. Levy, S.M. Holland, T.R. Sterling, Novel human genetic variants associated with extrapulmonary tuberculosis: a pilot genome wide association study, BMC Res. Notes 4 (2011) 28, https://doi.org/10.1186/1756-0500-4-28.
[41] Y. Omae, L. Toyo-oka, H. Yanai, S. Nedsuwan, S. Wattanapokayakit, N. Satproedprai, N. Smittipat, P. Palittapongarnpim, P. Sawanpanyalert, W. Inunchot, E. Pasomsub, N. Wichukchinda, T. Mushiroda, M. Kubo, K. Tokunaga, S. Mahasirimongkol, Pathogen lineage-based genome-wide association study identified CD53 as susceptible locus in tuberculosis, J. Hum. Genet. (2017), https://doi.org/10.1038/jhg.2017.82.
[42] E. Owusu-Dabo, O. Adjei, C.G. Meyer, R.D. Horstmann, A. Enimil, T.F. Kruppa, F. Bonsu, E.N.L. Browne, M.A. Chinbuah, I. Osei, J. Gyapong, C. Berberich, T. Kubica, S. Niemann, S. Ruesch-Gerdes, Mycobacterium tuberculosis drug resistance, Ghana. Emerg. Infect. Dis. 12 (2006) 1170–1172, https://doi.org/10.3201/eid1207.051028.
[43] T. Pastinen, K. Liitsola, P. Niini, M. Salminen, A.-C. SyväNen, Contribution of the CCR5 and MBL genes to susceptibility to HIV type 1 infection in the Finnish population, AIDS Res. Hum. Retrovir. 14 (1998) 695–698, https://doi.org/10.1089/aid.1998.14.695.
[44] I. Pe'er, R. Yelensky, D. Altshuler, M.J. Daly, Estimation of the multiple testing burden for genomewide association studies of nearly all common variants, Genet. Epidemiol. 32 (2008) 381–385, https://doi.org/10.1002/gepi.20303.
[45] E. Png, B. Alisjahbana, E. Sahiratmadja, S. Marzuki, R. Nelwan, Y. Balabanova, V. Nikolayevskyy, F. Drobniewski, S. Nejentsev, I. Adnan, E. van de Vosse, M. L. Hibberd, R. van Crevel, T.H. Ottenhoff, M. Seielstad, A genome wide association study of pulmonary tuberculosis susceptibility in Indonesians, BMC Med. Genet. (2012) 13, https://doi.org/10.1186/1471-2350-13-5.
[46] R.D. Pratiwi, Socio-economic and environmenys risk factors of tuberculosis in Wonosobo, Central Java, Indonesia, in: Graduate Studies in Public Health, Graduate Program, Sebelas Maret University Jl. Ir Sutami 36A, Surakarta 57126. Telp/Fax: (0271) 632 450 ext.208 First, 2016, https://doi.org/10.26911/theicph.2016.027.
[47] S. Purcell, C. Chang, PLINK 1.9, 2021. www.cog-genomics.org/plink/1.9/.
[48] R. Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2017.
[49] K.A. Rockett, G.M. Clarke, K. Fitzpatrick, C. Hubbart, A.E. Jeffreys, K. Rowlands, R. Craik, M. Jallow, D.J. Conway, K.A. Bojang, M. Pinder, S. Usen, F. Sisay-Joof, G. Sirugo, O. Toure, M.A. Thera, S. Konate, S. Sissoko, A. Niangaly, B. Poudiougou, V.D. Mangano, E.C. Bougouma, S.B. Sirima, D. Modiano, L.N. Amenga-Etego, A. Ghansah, K.A. Koram, M.D. Wilson, A. Enimil, J. Evans, O. Amodu, S. Olaniyan, T. Apinjoh, R. Mugri, A. Ndi, C.M. Ndila, S. Uyoga, A. Macharia, N. Peshu, T. N. Williams, A. Manjurano, E. Riley, C. Drakeley, H. Reyburn, V. Nyirongo, D. Kachala, M. Molyneux, S.J. Dunstan, N.H. Phu, N.T. Ngoc Quyen, C.Q. Thai, T. T. Hien, L. Manning, M. Laman, P. Siba, H. Karunajeewa, S. Allen, A. Allen, T.M. E. Davis, P. Michon, I. Mueller, A. Green, S. Molloy, K.J. Johnson, A. Kerasidou, V. Cornelius, L. Hart, A. Vanderwal, M. SanJoaquin, G. Band, S.Q. Le, M. Pirinen, N. Sepúlveda, C.C.A. Spencer, T.G. Clark, T. Agbenyega, E. Achidi, O. Doumbo, J. Farrar, K. Marsh, T. Taylor, D.P. Kwiatkowski, Reappraisal of known malaria resistance loci in a large multi-centre study, Nat. Genet. 46 (2014) 1197–1204, https://doi.org/10.1038/ng.3107.
[50] N.R. Roshyara, K. Horn, H. Kirsten, P. Ahnert, M. Scholz, Comparing performance of modern genotype imputation methods in different ethnicities, Sci. Rep. 6 (2016), https://doi.org/10.1038/srep34386.

[51] M. Salie, L. Van Der Merwe, M. Moller, M. Daya, G.D. Van Der Spuy, P.D. Van Helden, M.P. Martin, X.J. Gao, R.M. Warren, M. Carrington, E.G. Hoal, Associations between human leukocyte antigen class i variants and the mycobacterium tuberculosis subtypes causing disease, J. Infect. Dis. 209 (2014) 216–223, https://doi.org/10.1093/infdis/jit443.

[52] H. Schurz, M. Daya, M. Möller, E.G. Hoal, M. Salie, TLR1, 2, 4, 6 and 9 variants associated with tuberculosis susceptibility: a systematic review and meta-analysis, PLoS One 10 (2015), e0139711, https://doi.org/10.1371/journal.pone.0139711.

[53] H. Schurz, C.J. Kinnear, C. Gignoux, G. Wojcik, P.D. van Helden, G. Tromp, B. Henn, E.G. Hoal, M. Möller, A sex-stratified genome-wide association study of tuberculosis using a multi-ethnic genotyping Array, Front. Genet. 9 (2019), https://doi.org/10.3389/fgene.2018.00678.

[54] H. Schurz, S.J. Müller, P.D. van Helden, G. Tromp, E.G. Hoal, C.J. Kinnear, M. Möller, Evaluating the accuracy of imputation methods in a five-way admixed population, Front. Genet. 10 (2019), https://doi.org/10.3389/fgene.2019.00034.

[55] J.A. Seddon, A.C. Hesseling, P. Godfrey-Faussett, K. Fielding, H.S. Schaaf, Risk factors for infection and disease in child contacts of multidrug-resistant tuberculosis: a cross-sectional study, BMC Infect. Dis. 13 (2013), https://doi.org/10.1186/1471-2334-13-392.

[56] S.T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, K. Sirotkin, dbSNP: the NCBI database of genetic variation, Nucleic Acids Res. 29 (2001) 308–311.

[57] O. Shisana, T. Rhele, L.C. Simbayi, K. Zuma, S. Jooste, N. Zungu, D. Labadarios, D. Onoya, South African national HIV prevalence, Incidence and Behaviour Survey 2012 (2012).

[58] C. Søborg, H.O. Madsen, Å.B. Andersen, T. Lillebaek, A. Kok-Jensen, P. Garred, Mannose-binding lectin polymorphisms in clinical tuberculosis, J. Infect. Dis. 188 (2003) 777–782, https://doi.org/10.1086/377183.

[59] R.E. Soccio, R.M. Adams, M.J. Romanowski, E. Sehayek, S.K. Burley, J.L. Breslow, The cholesterol-regulated StarD4 gene encodes a StAR-related lipid transfer protein with two closely related homologues, StarD5 and StarD6, Proc. Natl. Acad. Sci. 99 (2002) 6943–6948, https://doi.org/10.1073/pnas.052143799.

[60] D. Stucki, D. Brites, L. Jeljeli, M. Coscolla, Q. Liu, A. Trauner, L. Fenner, L. Rutaihwa, S. Borrell, T. Luo, Q. Gao, M. Kato-Maeda, M. Ballif, M. Egger, R. Macedo, H. Mardassi, M. Moreno, G.T. Vilanova, J. Fyfe, M. Globan, J. Thomas, F. Jamieson, J.L. Guthrie, A. Asante-Poku, D. Yeboah-Manu, E. Wampande, W. Ssengooba, M. Joloba, W.H. Boom, I. Basu, J. Bower, M. Saraiva, S.E. G. Vasconcellos, P. Suffys, A. Koch, R. Wilkinson, L. Gail-Bekker, B. Malla, S.D. Ley, H.-P. Beck, B.C. de Jong, K. Toit, E. Sanchez-Padilla, M. Bonnet, A. Gil-Brusola, M. Frank, V.N. Penlap Beng, K. Eisenach, I. Alani, P.W. Ndung'u, G. Revathi, F. Gehre, S. Akter, F. Ntoumi, L. Stewart-Isherwood, N.E. Ntinginya, A. Rachow, M. Hoelscher, D.M. Cirillo, G. Skenders, S. Hoffner, D. Bakonyte, P. Stakenas, R. Diel, V. Crudu, O. Moldovan, S. Al-Hajoj, L. Otero, F. Barletta, E.J. Carter, L. Diero, P. Supply, I. Comas, S. Niemann, S. Gagneux, Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages, Nat. Genet. 48 (2016) 1535–1543, https://doi.org/10.1038/ng.3704.

[61] P.H. Sudmant, E.J. Gardner, R.E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. Hsi-Yang Fritz, M.K. Konkel, A. Malhotra, A.M. Stütz, X. Shi, F. Paolo Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M.J. P. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H.Y.K. Lam, X. Jasmine Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J.M. Kidd, Y. Kong, E.-W. Lameijer, S. McCarthy, P. Flicek, R.A. Gibbs, G. Marth, C.E. Mason, A. Menelaou, D.M. Muzny, B.J. Nelson, A. Noor, N.F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E.E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalin, A. Untergasser, J.A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M.A. Batzer, S.A. McCarroll, R. E. Mills, M.B. Gerstein, A. Bashir, O. Stegle, S.E. Devine, C. Lee, E.E. Eichler, J. O. Korbel, An integrated map of structural variation in 2,504 human genomes, Nature 526 (2015) 75–81, https://doi.org/10.1038/nature15394.

[62] P. Supply, C. Allix, S. Lesjean, M. Cardoso-Oelemann, S. Rusch-Gerdes, E. Willery, E. Savine, P. de Haas, H. van Deutekom, S. Roring, P. Bifani, N. Kurepina, B. Kreiswirth, C. Sola, N. Rastogi, V. Vatin, M.C. Gutierrez, M. Fauville, S. Niemann, R. Skuce, K. Kremer, C. Locht, D. van Soolingen, Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of Mycobacterium tuberculosis, J. Clin. Microbiol. 44 (2006) 4498–4510, https://doi.org/10.1128/JCM.01392-06.

[63] The 1000 Genomes Project Consortium, An integrated map of genetic variation from 1,092 human genomes, Nature 491 (2012) 56–65, https://doi.org/10.1038/nature11632.

[64] The International HapMap Consortium, A haplotype map of the human genome, Nature 437 (2005) 1299–1320, https://doi.org/10.1038/nature04226.

[65] T. Thye, S. Niemann, K. Walter, S. Homolka, C.D. Intemann, M.A. Chinbuah, A. Enimil, J. Gyapong, I. Osei, E. Owusu-Dabo, S. Rüsch-Gerdes, R.D. Horstmann, S. Ehlers, C.G. Meyer, Variant G57E of mannose binding lectin associated with protection against tuberculosis caused by Mycobacterium africanum but not by M. tuberculosis, PLoS ONE 6 (2011) e20908, https://doi.org/10.1371/journal.pone.0020908.

[66] T. Thye, E. Owusu-Dabo, F.O. Vannberg, R. Van Crevel, J. Curtis, E. Sahiratmadja, Y. Balabanova, C. Ehmen, B. Muntau, G. Ruge, J. Sievertsen, J. Gyapong, V. Nikolayevskyy, P.C. Hill, G. Sirugo, F. Drobniewski, E. Van De Vosse, M. Newport, B. Alisjahbana, S. Nejentsev, T.H.M. Ottenhoff, A.V.S. Hill, R. D. Horstmann, C.G. Meyer, Common variants at 11p13 are associated with susceptibility to tuberculosis, Nat. Genet. 44 (2012) 257–259, https://doi.org/10.1038/ng.1080.

[67] T. Thye, F.O. Vannberg, S.H. Wong, E. Owusu-Dabo, I. Osei, J. Gyapong, G. Sirugo, F. Sisay-Joof, A. Enimil, M.A. Chinbuah, S. Floyd, D.K. Warndorff, L. Sichali, S. Malema, A.C. Crampin, B. Ngwira, Y.Y. Teo, K. Small, K. Rockett, D. Kwiatkowski, P.E. Fine, P.C. Hill, M. Newport, C. Lienhardt, R.A. Adegbola, T. Corrah, A. Ziegler, A.P. Morris, C.G. Meyer, R.D. Horstmann, A.V.S. Hill, Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2, Nat. Genet. 42 (2010) 739–741, https://doi.org/10.1038/ng.639.

[68] G.D. van der Spuy, K. Kremer, S.L. Ndabambi, N. Beyers, R. Dunbar, B.J. Marais, P. D. van Helden, R.M. Warren, Changing Mycobacterium tuberculosis population highlights clade-specific pathogenic characteristics, Tuberculosis 89 (2009) 120–125, https://doi.org/10.1016/j.tube.2008.09.003.

[69] WHO, World Medical Association Declaration of Helsinki, 2001.

[70] H. Yamada, S. Mizuno, A.C. Ross, I. Sugawara, Retinoic acid therapy attenuates the severity of tuberculosis while altering lymphocyte and macrophage numbers and cytokine expression in rats infected with Mycobacterium tuberculosis, J. Nutr. 137 (2007) 2696–2700, https://doi.org/10.1093/jn/137.12.2696.

[71] J.J. Yim, P. Selvaraj, Genetic susceptibility in tuberculosis, Respirology 15 (2010) 241–256, https://doi.org/10.1111/j.1440-1843.2009.01690.x.