Confirmation of HLA-II associations with TB susceptibility in admixed African samples

Dayna Croock¹, Yolandi Swart¹, Haiko Schurz¹, Desiree C. Petersen¹, Marlo Möller¹,², Caitlin Uren¹,²*

¹DSI-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
²Centre for Bioinformatics and Computational Biology, Stellenbosch University
*Corresponding author: caitlinu@sun.ac.za

Abstract

The International Tuberculosis Host Genetics Consortium (ITHGC) demonstrated the power of large-scale GWAS analysis across diverse ancestries in identifying tuberculosis (TB) susceptibility loci. Despite identifying a significant genetic correlate in the human leukocyte antigen (HLA)-II region, this association did not replicate in the African ancestry-specific analysis, due to small sample size and the inclusion of admixed samples. Our study aimed to build upon the findings from the ITHGC and identify TB susceptibility loci in an admixed South African cohort using the local ancestry allelic adjusted association (LAAA) model. We identified a near-genome-wide significant association (rs3117230, p-value = 5.292 x10⁻⁶, OR = 0.437, SE = 0.182) in the HLA-DPB1 gene originating from KhoeSan ancestry. These findings extend the work of the ITHGC, underscore the need for innovative strategies in studying complex admixed populations, and confirm the role of the HLA-II region in TB susceptibility in admixed South African samples. [148/150 words]

Keywords

Human leukocyte antigen (HLA)-II, tuberculosis (TB), local ancestry, admixture, KhoeSan ancestry

Introduction

Tuberculosis (TB) is a communicable disease caused by Mycobacterium tuberculosis (M.tb) (World Health Organization, 2023). M.tb infection has a wide range of clinical manifestations from asymptomatic, non-transmissible or so-called “latent” infections to active TB (Möller...
et al., 2018). Approximately 1/4 of the global population is infected with \( M. tb \), but only 5-15% of infected individuals will develop active TB (Houben & Dodd, 2016). Several factors increase the risk of progressing to active TB, including co-infection with human immunodeficiency virus (HIV) and comorbidities, such as diabetes mellitus, asthma and other airway and lung diseases (Glaziou et al., 2018). Socio-economic factors including smoking, malnutrition, alcohol abuse, intravenous drug use, prolonged residence in a high burdened community, overcrowding, informal housing and poor sanitation also influence \( M. tb \) transmission and infection (Cudahy et al., 2020; Escombe et al., 2019; Laghari et al., 2019; Matose et al., 2019). Additionally, individual variability in infection and disease progression has been attributed to variation in the host genome (Uren et al., 2021; Verhein et al., 2018).

Numerous genome-wide association studies (GWASs) investigating TB susceptibility have been conducted across different population groups. Compared to other well-studied diseases, GWASs investigating TB susceptibility are sparse and the results from these studies do not replicate well across populations (Möller & Kinnear, 2020; Möller et al., 2018; Uren et al., 2017). This lack of replication could be caused by small sample sizes, variation in phenotype definitions among studies, variation in linkage disequilibrium (LD) patterns across different population groups and the presence of population-specific effects (Möller & Kinnear, 2020). Additionally, complex LD patterns within population groups, produced by admixture, impede the detection of statistically significant loci when using traditional GWAS methods (Swart et al., 2020).

The International Tuberculosis Host Genetics Consortium (ITHGC) performed a meta-analysis of TB GWAS results including 14 153 TB cases and 19 536 controls of African, Asian and European ancestries (Schurz et al., 2024). The multi-ancestry meta-analysis identified one genome-wide significant variant (rs28383206) in the human leukocyte antigen (HLA)-II region (\( p = 5.2 \times 10^{-9} \), OR = 0.89, 95% CI = 0.84-0.95). The association peak at the HLA-II locus encompassed several genes encoding crucial antigen presentation proteins (including HLA-DR and HLA-DQ). While ancestry-specific association analyses in the European and Asian cohorts also produced suggestive peaks in the HLA-II region, the African ancestry-specific association test did not yield any associations or suggestive peaks. The authors described possible reasons for the lack of associations, including the smaller sample size compared to...
the other ancestry-specific meta-analyses, increased genetic diversity within African individuals and population stratification produced by two admixed cohorts from the South African Coloured (SAC) population. The SAC population (as termed in the South African census (Lehohla, 2012)) form part of a multi-way (up to five-way) admixed population with ancestral contributions from Bantu-speaking African (~30%), KhoeSan (~30%), European (~20%), and East (~10%) and Southeast Asian (~10%) populations (Chimusa et al., 2013). The diverse genetic background of admixed individuals can lead to population stratification, potentially introducing confounding variables. However, the power to detect statistically significant loci in admixed populations can be improved by leveraging admixture-induced local ancestry (Swart et al., 2021; Swart, van Eeden, et al., 2022). Since previous computational algorithms were not able to include local ancestry as a covariate for GWASs, the local ancestry allelic adjusted association model (LAAA) was developed to overcome this limitation (Duan et al., 2018). The LAAA model identifies ancestry-specific alleles associated with the phenotype by including the minor alleles and the corresponding ancestry of the minor alleles (obtained by local ancestry inference) as covariates. The LAAA model has been successfully applied in a cohort of multi-way admixed SAC individuals to identify novel variants associated with TB susceptibility (Swart et al., 2021; Swart, van Eeden, et al., 2022).

Our study builds upon the findings from the ITHGC (Schurz et al., 2024) and aim to resolve the challenges faced in African ancestry-specific association analysis. Here, we explore host genetic correlates of TB in a complex admixed SAC population using the LAAA model.

**Methods**

**Data**

This study included the two SAC admixed datasets from the ITHGC analysis [RSA(A) and RSA(M)] as well as four additional TB case-control datasets obtained from admixed South African population groups (Table 1). Like the SAC population, the Xhosa population are admixed with rain-forest forager and KhoeSan ancestral contributions (Choudhury et al., 2021). All datasets were collected over the past 30 years under different research projects (Daya et al., 2013; Kroon et al., 2020; Schurz et al., 2018; Smith et al., 2023; Ugarte-Gil et al., 2020) and individuals that were included in the analyses consented to the use of their
data in future research regarding TB host genetics. Across all datasets, TB cases were bacteriologically confirmed (culture positive) or diagnosed by GeneXpert. Controls were healthy individuals with no previous or current history of TB disease or treatment. However, given the high prevalence of TB in South Africa [852 cases (95% CI 679-1026) per 100 000 individuals 15 years and older (Cudahy et al., 2020)], most controls have likely been exposed to \textit{M.tuberculosis} at some point (Gallant et al., 2010). For all datasets, cases and controls were obtained from the same community and thus share similar socio-economic status and health care access.

Table 1. Summary of the datasets included in analysis.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Genotyping platform</th>
<th>Self-reported ethnicity</th>
<th>Cases/controls</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSA(A)</td>
<td>Affymetrix 500k</td>
<td>SAC</td>
<td>642/91</td>
<td>(Daya et al., 2013)</td>
</tr>
<tr>
<td>RSA(M)</td>
<td>MEGA array 1.1M</td>
<td>SAC</td>
<td>555/440</td>
<td>(Schurz et al., 2018; Swart et al., 2021)</td>
</tr>
<tr>
<td>RSA(TANDEM)</td>
<td>H3Africa array</td>
<td>SAC and Bantu-speaking African</td>
<td>161/133</td>
<td>(Swart, Uren, et al., 2022)</td>
</tr>
<tr>
<td>RSA(NCTB)</td>
<td>H3Africa array</td>
<td>SAC</td>
<td>49/111</td>
<td>(Oyageshio et al., 2023)</td>
</tr>
<tr>
<td>RSA(Worcester)</td>
<td>H3Africa array</td>
<td>SAC</td>
<td>61 cases</td>
<td>Unpublished</td>
</tr>
<tr>
<td>RSA(Xhosa)</td>
<td>Whole genome sequencing</td>
<td>IsiXhosa</td>
<td>44/120</td>
<td>Unpublished</td>
</tr>
</tbody>
</table>

A list of sites genotyped on the Infinium\textsuperscript{TM} H3Africa array ([https://chipinfo.h3abionet.org/browse](https://chipinfo.h3abionet.org/browse)) were extracted from the whole-genome sequenced [RSA(Xhosa)] dataset and treated as genotype data in subsequent analyses. Quality control (QC) of raw genotype data was performed using PLINK v1.9 (Purcell et al., 2007). In all datasets, individuals were screened for sex concordance and discordant sex information was corrected based on X chromosome homozygosity estimates ($F_{\text{estimate}} < 0.2$ for females and $F_{\text{estimate}} > 0.8$ for males). In the event that sex information could not be corrected based on homozygosity estimates, individuals with missing or discordant sex information were removed. Individuals with genotype call rates less than 90\% and SNPs with more than 5\% missingness were removed. Monomorphic sites were removed. Individuals were screened for deviations in HWE for each SNP and sites deviating from the HWE threshold of $10^{-5}$ were removed. Sex chromosomes were excluded from the analysis. The genome coordinates across all datasets were checked for consistency and, if necessary, converted to GRCh37 using the UCSC liftOver tool (Kuhn et al., 2013).
Genotype datasets were pre-phased using SHAPEIT v2 (Delaneau et al., 2013) and imputed using the Positional Burrows-Wheeler Transformation (PBWT) algorithm through the Sanger Imputation Server (SIS) (Durbin, 2014). The African Genome Resource (AGR) panel (n=4,956), accessed via the SIS, was used as the reference panel for imputation (Gurdasani et al., 2015) since it has been shown that the AGR is the best reference panel for imputation of missing genotypes for samples from the SAC population (Schurz et al., 2019). Imputed data were filtered to remove sites with imputation quality INFO scores less than 0.95. Individual datasets were screened for relatedness using KING software (Manichaikul et al., 2010) and individuals up to second degree relatedness were removed. A total of 7,544,769 markers overlapped across all six datasets. This list of intersecting markers was extracted from each dataset using PLINK --extract flag. The datasets were then merged using the PLINK v1.9. After merging, all individuals missing more than 10% genotypes were removed, markers with more than 5% missing data were excluded and a HWE filter was applied to controls (threshold <10$^{-5}$). The merged dataset was screened for relatedness using KING and individuals up to second degree relatedness were subsequently removed. The final merged dataset after QC and data filtering (including the removal of related individuals) consisted of 1,544 individuals (952 TB cases and 592 healthy controls). A total of 7,510,057 variants passed QC and filtering parameters.

Global ancestry inference

ADMIIXTURE was used to determine the correct number of contributing ancestral proportions in our multi-way admixed population cohort (Alexander & Lange, 2011). ADMIIXTURE estimates the number of contributing ancestral populations (denoted by K) and population allele frequencies through cross-validation (CV). All 1,544 individuals were grouped into running groups of equal size together with 191 reference populations (Table 2). Running groups were created to ensure approximately equal numbers of reference populations and admixed populations. Xhosa and SAC samples were divided into separate running groups.

Table 2. Ancestral populations included for global ancestry deconvolution.

<table>
<thead>
<tr>
<th>Population</th>
<th>n</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>European (British – GBR)</td>
<td>40</td>
<td>1000 Genomes (1000G) phase 3 (1000 Genomes Project Consortium et al., 2015)</td>
</tr>
<tr>
<td>East Asian (Chinese – CHB)</td>
<td>40</td>
<td>1000G phase 3</td>
</tr>
</tbody>
</table>
Redundant SNPs were removed by PLINK through LD pruning by removing each SNP with LD $r^2 > 0.1$ within a 50-SNP sliding window (advanced by 10 SNPs at a time). Ancestral proportions were inferred in an unsupervised manner for $K = 3$-6 (1 iteration). The best value of $K$ for the data was selected by choosing the $K$ value with the lowest CV error across all running groups. Ten iterations of $K = 3$ and $K = 5$ was run for the Xhosa and SAC individuals respectively. Since it has been shown that RFMix (Maples et al., 2013) outperforms ADMIXTURE in determining global ancestry proportions (C Uren et al., 2020), RFMix was also used to refine inferred global ancestry proportions. Global ancestral proportions were visualised using PONG (Behr et al., 2016).

Local ancestry inference

The merged dataset and the reference file (containing reference populations from Table 2) were phased separately using SHAPEIT2. The local ancestry for each position in the genome was inferred using RFMix (Maples et al., 2013). Default parameters were used, but the number of generations since admixture was set to 15 for the SAC individuals and 20 for the Xhosa individuals (as determined by previous studies) (Uren et al., 2016). RFMix was run with three expectation maximisation iterations and the --reanalyse-reference flag.

Batch effect screening and correction

Merging separate datasets generated at different timepoints and/or facilities, as we have done here, will undoubtedly introduce batch effects. Principal component analysis (PCA) is a common method used to visualise batch effects, where the first two principal components (PCs) are plotted with each sample coloured by batch, and a separation of colours is indicative of a batch effect (Nyamundanda et al., 2017). However, it is difficult to differentiate between separation caused by population structure and separation caused by batch effect using PCA alone. An alternative method to detect batch effects (Chen et al., 2022) involves coding case/control status by batch followed by running an association
analysis testing each batch against all other batches. If any single dataset has more positive
signals compared to the other datasets, then batch effects may be responsible for producing
spurious results. Batch effects can be resolved by removing those SNPs which pass the
genome-wide significance threshold from the merged dataset. We have adapted this batch
effect correction method for application in a highly admixed cohort with complex population
structure (Croock et al., 2024). Our modified method was used to remove SNPs affected by
batch effects from our merged dataset.

Local ancestry allelic adjusted association analysis

The LAAA association model was used to investigate if there are allelic, ancestry-specific or
ancestry-specific allelic associations with TB susceptibility in our merged dataset. Global
ancestral components inferred by RFMix, age and sex were included as covariates in the
association tests. Variants with minor allele frequency (MAF) < 1% were removed to improve
the stability of the association tests. Dosage files, which code the number of alleles of a
specific ancestry at each locus across the genome, were compiled. Separate regression
models for each ancestral contribution were fitted to investigate which ancestral
contribution is associated with TB susceptibility. Details regarding the models have been
described elsewhere (Swart, van Eeden, et al., 2022); but in summary, four regression
models were tested to detect the source of the association signals observed:

1) Null model or global ancestry (GA) model:
The null model only includes global ancestry, sex and age covariates. This test investigates
whether an additive allelic dose exerts an effect on the phenotype (without including local
ancestry of the allele).

2) Local ancestry (LA) model:
The LA model includes the number of alleles of a specific ancestry at a locus as covariates.
This model is used in admixture mapping to identify ancestry-specific variants associated
with a specific phenotype.

3) Ancestry plus allelic (APA) model:
The APA model simultaneously performs model (1) and (2). This model tests whether an
additive allelic dose exerts an effect of the phenotype whilst adjusting for local ancestry.
The LAAA model is an extension of the APA model, which models the combination of the minor allele and ancestry of the minor allele at a specific locus and the effect this interaction has on the phenotype.

The R package STEAM (Significance Threshold Estimation for Admixture Mapping) (Grinde et al., 2019) was used to determine the genome-wide significance threshold given the global ancestral proportions of each individual and the number of generations since admixture ($g = 15$). STEAM permuted these factors 10,000 times to derive a threshold for significance. Results were visualised in RStudio. A genome-wide significance threshold of $p$-value $< 2.5 	imes 10^{-6}$ was deemed significant by STEAM.

Results

Global and local ancestry inference

After close inspection of global ancestry proportions generated using ADMIXTURE, the K number of contributing ancestries (the lowest k-value determined through cross-validation) was K = 3 for the Xhosa individuals and K = 5 for the SAC individuals (Figure 1). This is consistent with previous global ancestry deconvolution results (Chimusa et al., 2014; Choudhury et al., 2021). It is evident that our cohort is a complex, highly admixed group with ancestral contributions from the indigenous KhoeSan (~22 - 30%), Bantu-speaking African (~30 - 72%), European (~5 - 24%), Southeast Asian (~11%) and East Asian (~5%) population groups.

Figure 1. Genome-wide ancestral proportions of all individuals in the merged dataset. Ancestral proportions for each individual are plotted vertically with different colours representing different contributing ancestries.

Local ancestry was estimated for all individuals. Admixture between geographically distinct populations creates complex ancestral and admixture-induced LD blocks, which can be
visualised using local ancestry karyograms. Figure 2 shows karyograms for three individuals from the merged dataset. It is evident that, despite individuals being from the same population group, each possesses unique patterns of local ancestry arising from differing numbers and lengths of ancestral segments.

Figure 2. Local ancestry karyograms of three admixed individuals from the SAC population. Each admixed individual has unique local ancestry patterns generated by admixture among geographically distinct ancestral population groups.

Local ancestry-allelic adjusted analysis

A total of 784,557 autosomal markers (with MAF > 1%) and 1,544 unrelated individuals (952 TB cases and 592 healthy controls) were included in logistic regression models to assess whether any loci and/or ancestries were significantly associated with TB status (whilst adjusting for sex, age, and global ancestry proportions). LAAA models were successfully applied for all five contributing ancestries (KhoeSan, Bantu-speaking African, European, East Asian and Southeast Asian). Only one variant (rs74828248) was significantly associated with
TB status (p-value < 2.5 x 10^{-6}) whilst utilising the LAAA model and whilst adjusting for Bantu-speaking African ancestry on chromosome 20 (p-value = 2.272 x 10^{-6}, OR = 0.316, SE = 0.244) (Supplementary Figure 1). No genomic inflation was detected in the QQ-plot for this region (Supplementary Figure 2). However, this variant is located in an intergenic region and the link to TB susceptibility is unclear.

Although no other variants passed the genome-wide significance threshold, multiple lead SNPs were identified. Notably, an appreciable peak was identified in the HLA-II region of chromosome 6 when using the LAAA model and adjusting for KhoeSan ancestry (Figure 3). The QQ-plot suggested minimal genomic inflation, which was verified by calculating the genomic inflation factor (λ = 1.05289) (Supplementary Figure 3). The lead variants identified using the LAAA model whilst adjusting for KhoeSan ancestry in this region on chromosome 6 are summarised in Table 3. The association peak encompasses the HLA-DPA1/B1 (major histocompatibility complex, class II, DP alpha 1/beta 1) genes (Figure 4). It is noteworthy that without the LAAA model, this association peak would not have been observed for this cohort. This highlights the importance of utilising the LAAA model in future association studies when investigating disease susceptibility loci in admixed individuals, such as the SAC population.
value = $5 \times 10^{-8}$). The four different models are represented in black (global ancestry only - GAO), blue (local ancestry effect - LAO), orange (ancestry plus allelic effect - APA) and pink (local ancestry adjusted allelic effect - LAAA).

![Figure 4. Regional plot indicating the nearest genes in the region of the lead variant (rs3117230) observed on chromosome 6. SNPs in linkage disequilibrium (LD) with the lead variant are coloured red/orange. The lead variant is indicated in purple. Functional protein-coding genes are coded in red and non-functional (pseudo-genes) are indicated in black.](image)

The lead variant lies within COL11A2P1 (collagen type X1 alpha 2 pseudogene 1). COL11A2P1 is an unprocessed pseudogene (ENSG00000228688). Unprocessed pseudogenes are seldomly transcribed and translated into functional proteins (Witek & Mohiuddin, 2024). HLA-DPB1 and HLA-DPA1 are the closest functional protein-coding genes to our lead variants.
Table 3. Suggestive associations (p-value < 1e-5) for the LAAA analysis adjusting for KhoeSan local ancestry on chromosome 6

<table>
<thead>
<tr>
<th>Position</th>
<th>Marker name</th>
<th>Ref</th>
<th>Alt</th>
<th>AltFreq</th>
<th>OR</th>
<th>SE</th>
<th>p-value</th>
<th>Gene</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>33075635</td>
<td>rs3117230</td>
<td>A</td>
<td>G</td>
<td>0.370</td>
<td>0.437</td>
<td>0.182</td>
<td>5.292e-06</td>
<td>HLA-DPB1</td>
<td>Intergenic</td>
</tr>
<tr>
<td>33048661</td>
<td>rs1042151</td>
<td>A</td>
<td>G</td>
<td>0.325</td>
<td>0.437</td>
<td>0.184</td>
<td>6.806e-06</td>
<td>HLA-DPB1</td>
<td>Exonic</td>
</tr>
<tr>
<td>33058874</td>
<td>rs2179920</td>
<td>C</td>
<td>T</td>
<td>0.369</td>
<td>0.445</td>
<td>0.180</td>
<td>6.960e-06</td>
<td>HLA-DPB1</td>
<td>Intergenic</td>
</tr>
<tr>
<td>33072266</td>
<td>rs2064478</td>
<td>C</td>
<td>T</td>
<td>0.371</td>
<td>0.447</td>
<td>0.181</td>
<td>8.222e-06</td>
<td>HLA-DPB1</td>
<td>Intergenic</td>
</tr>
<tr>
<td>33072729</td>
<td>rs3130210</td>
<td>G</td>
<td>T</td>
<td>0.371</td>
<td>0.447</td>
<td>0.181</td>
<td>8.222e-06</td>
<td>HLA-DPB1</td>
<td>Intergenic</td>
</tr>
<tr>
<td>33073440</td>
<td>rs2064475</td>
<td>G</td>
<td>A</td>
<td>0.371</td>
<td>0.447</td>
<td>0.181</td>
<td>8.222e-06</td>
<td>HLA-DPB1</td>
<td>Intergenic</td>
</tr>
<tr>
<td>33074348</td>
<td>rs3117233</td>
<td>T</td>
<td>C</td>
<td>0.371</td>
<td>0.447</td>
<td>0.181</td>
<td>8.222e-06</td>
<td>HLA-DPB1</td>
<td>Intergenic</td>
</tr>
<tr>
<td>33074707</td>
<td>rs3130213</td>
<td>G</td>
<td>A</td>
<td>0.371</td>
<td>0.447</td>
<td>0.181</td>
<td>8.222e-06</td>
<td>HLA-DPB1</td>
<td>Intergenic</td>
</tr>
</tbody>
</table>

Ref, reference allele; Alt, alternate allele; AltFreq, alternate allele frequency; OR, odds ratio; SE, standard error

Discussion

The LAAA analysis of host genetic susceptibility to TB, involving 942 TB cases and 592 controls, identified one suggestive association peak adjusting for KhoeSan local ancestry. The association peak identified in this study encompasses the HLA-DPB1 gene, a highly polymorphic locus, with over 2 000 documented allelic variants (Robinson et al., 2020). This association is noteworthy given that HLA-DPB1 alleles have been associated with TB resistance (Dawkins et al., 2022; Ravikumar et al., 1999; Selvaraj et al., 2008). The direction of effect the lead variants in our study (Table 3) similarly suggest a protective effect against developing active TB. However, variants in HLA-DPB1 were not identified in the ITHGC meta-analysis.

Population stratification arising from the highly heterogeneous admixed cohorts might have masked this association signal in the African ancestry-specific association analysis. The association peak in the HLA-II region was only captured using the LAAA model whilst adjusting for KhoeSan local ancestry. This underscores the importance of incorporating global and local ancestry in association studies investigating complex multi-way admixed
individuals, as the genetic heterogeneity present in admixed individuals (produced as a result of admixture-induced and ancestral LD patterns) may cause association signals to be missed when using traditional association models (Duan et al., 2018; Swart, van Eeden, et al., 2022).

We did not replicate the significant association signal in HLA-DRB1 identified by the ITHGC. However, the ITHGC also did not replicate this association in their own African ancestry-specific analysis. The significant association, rs28383206, identified by the ITHGC appears to be tagging the HLA-DQA1*02:1 allele, which is associated with TB in Icelandic and Asian populations (Li et al., 2021; Sveinbjornsson et al., 2016; Zheng et al., 2018). It is possible that this association signal is specific to non-African populations, but additional research is required to verify this hypothesis. Both our study and the ITHGC independently pinpointed variants associated with TB susceptibility in different genes within the HLA-II locus (Figure 5).

The HLA-II region spans ~0.8Mb on chromosome 6p21.32 and encompasses the HLA-DR, -DP and -DQ alpha and beta chain genes. The HLA-II complex is the human form of the major histocompatibility complex class II (MHC-II) proteins on the surface of antigen presenting cells, such as monocytes, dendritic cells and macrophages. The innate immune response against M.tb involves phagocytosis by alveolar macrophages. In the phagosome, mycobacterial antigens are processed for presentation on MHC-II on the surface of the antigen presenting cell. Previous studies have suggested that M.tb interferes with the MHC-II pathway to enhance intracellular persistence and delay activation of the adaptive immune response (Oliveira-Cortez et al., 2016). For example, M.tb can inhibit phagosome maturation and acidification, thereby limiting antigen processing and presentation on MHC-II molecules (Chang et al., 2005). Given that MHC-II plays an essential role in the adaptive immune response to TB and numerous studies have identified HLA-II variants associated with TB (Cai et al., 2019; Chihab et al., 2023; de Sá et al., 2020; Harishankar et al., 2018; Schurz et al., 2024; Selvaraj et al., 2008), additional research is required to elucidate the effects of HLA-II variation on TB risk status.
In conclusion, application of the LAAA to a highly admixed SAC cohort revealed a suggestive association signal in the HLA-II region associated with protection against TB. Our study builds on the results of the ITHGC by demonstrating an alternative method to identify association signals in cohorts with complex genetic ancestry. This analysis shows the value of including individual global and local ancestry in genetic association analyses. Furthermore, we confirm HLA-II loci associations with TB susceptibility in an admixed South African population and hope that this publication will encourage greater appreciation for the role of the adaptive immune system in TB susceptibility and resistance.

Acknowledgements

We acknowledge the support of the DSI-NRF Centre of Excellence for Biomedical Tuberculosis Research, South African Medical Research Council Centre for Tuberculosis Research (SAMRC CTR), Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa. We also
acknowledge the Centre for High Performance Computing (CHPC), South Africa, for providing computational resources. This research was partially funded by the South African government through the SAMRC and the Harry Crossley Research Foundation.

Author ORCIDs
Dayna Croock: 0000-0002-5107-8006
Yolandi Swart: 0000-0002-9840-3646
Haiko Schurz: 0000-0002-0009-3409
Desiree C. Petersen: 0000-0002-0817-2574
Marlo Möller: 0000-0002-0805-6741
Caitlin Uren: 0000-0003-2358-0135

Ethics
Ethics approval was granted by the Health Research Ethics Committee (HREC) of Stellenbosch University, South Africa (project number S22/02/031).

Competing interests
None declared.

References


566 sequencing combined genome-wide association study identifies novel tuberculosis
568 https://doi.org/10.1016/j.ygeno.2021.05.035
569
571 Robust relationship inference in genome-wide association studies. Bioinformatics,
572 26(22), 2867–2873. https://doi.org/10.1093/bioinformatics/btq559
573
discriminative modeling approach for rapid and robust local-ancestry inference.
575 American Journal of Human Genetics, 93(2), 278–288.
576 https://doi.org/10.1016/j.ajhg.2013.06.020
577
579 airborne tuberculosis infection control in minibus taxis. South African Journal of
581
582 Möller, M., Kinnear, C. J., Orlova, M., Kroon, E. E., van Helden, P. D., Schurr, E., & Hoal,
583 E. G. (2018). Genetic Resistance to Mycobacterium tuberculosis Infection and
584 Disease. Frontiers in Immunology, 9, 2219.
585 https://doi.org/10.3389/fimmu.2018.02219
586
587 Möller, M., & Kinnear, C. J. (2020). Human global and population-specific genetic
588 susceptibility to Mycobacterium tuberculosis infection and disease. Current Opinion
in Pulmonary Medicine, 26(3), 302–310.

https://doi.org/10.1097/MCP.0000000000000672


Supplementary Figure 1. Log transformation of association signals obtained for Bantu-speaking African ancestry whilst using the LAAA model on chromosome 20. The dashed red line represents the significant threshold for admixture mapping calculated with the software STEAM (p-value = 2.5 x 10^{-6}) and the black solid line represents the genome-wide significant threshold (p-value = 5 x 10^{-8}). The four different models are represented in black (global ancestry only - GAO), blue (local ancestry effect - LAO), orange (ancestry plus allelic effect - APA) and pink (local ancestry-adjusted allelic effect - LAAA).

Supplementary Figure 2. QQ-plot of expected p-values and observed p-values for the association signals obtained for Bantu-speaking African ancestry located on chromosome 20.
Supplementary Figure 3. QQ-plot of expected $p$-values and observed $p$-values for the association signals obtained for Khoisan ancestry located on chromosome 6.