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

Dayna Croock[1], Yolandi Swart[1], Haiko Schurz[1], Desiree C. Petersen[1], Marlo Möller[1,2], Caitlin Uren[1,2*]

[1]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

[2]Centre 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 00 individuals 15 years and older (Cudahy et al., 2020)], most controls have likely been exposed to *M.tb* 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.

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RSA(Worcester)	H3Africa array	SAC	61 cases	Unpublished
RSA(Xhosa)	Whole genome sequencing	IsiXhosa	44/120	Unpublished

A list of sites genotyped on the Infinium™ H3Africa array (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 ($F_{estimate} < 0.2$ for females and $F_{estimate} > 0.8$ 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

ADMIXTURE was used to determine the correct number of contributing ancestral proportions in our multi-way admixed population cohort (Alexander & Lange, 2011). ADMIXTURE 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.

Population	n	Source
European (British – GBR)	40	1000 Genomes (1000G) phase 3 (1000 Genomes Project Consortium et al., 2015)
East Asian (Chinese – CHB)	40	1000G phase 3

Bantu-speaking African (Yoruba – YRI)	40	1000G phase 3
Southeast Asian (Malaysian)	38	Singapore Sequencing Malay Project (SSMP) (Wong et al., 2013)
KhoeSan (Nama)	33	African Genome Variation Project (AGVP/ADRP) (Gurdasani et al., 2015)

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.

(4) Local ancestry adjusted allelic (LAAA) model:

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 x 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.

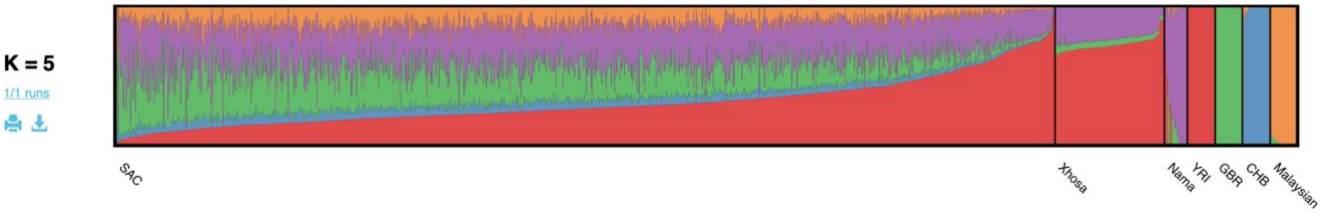


Figure 1. Genome-wide ancestral proportions of all individuals in the merged dataset. Ancestral proportions for each individual are plotted vertically with different colours representing different contributing ancestries.

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.

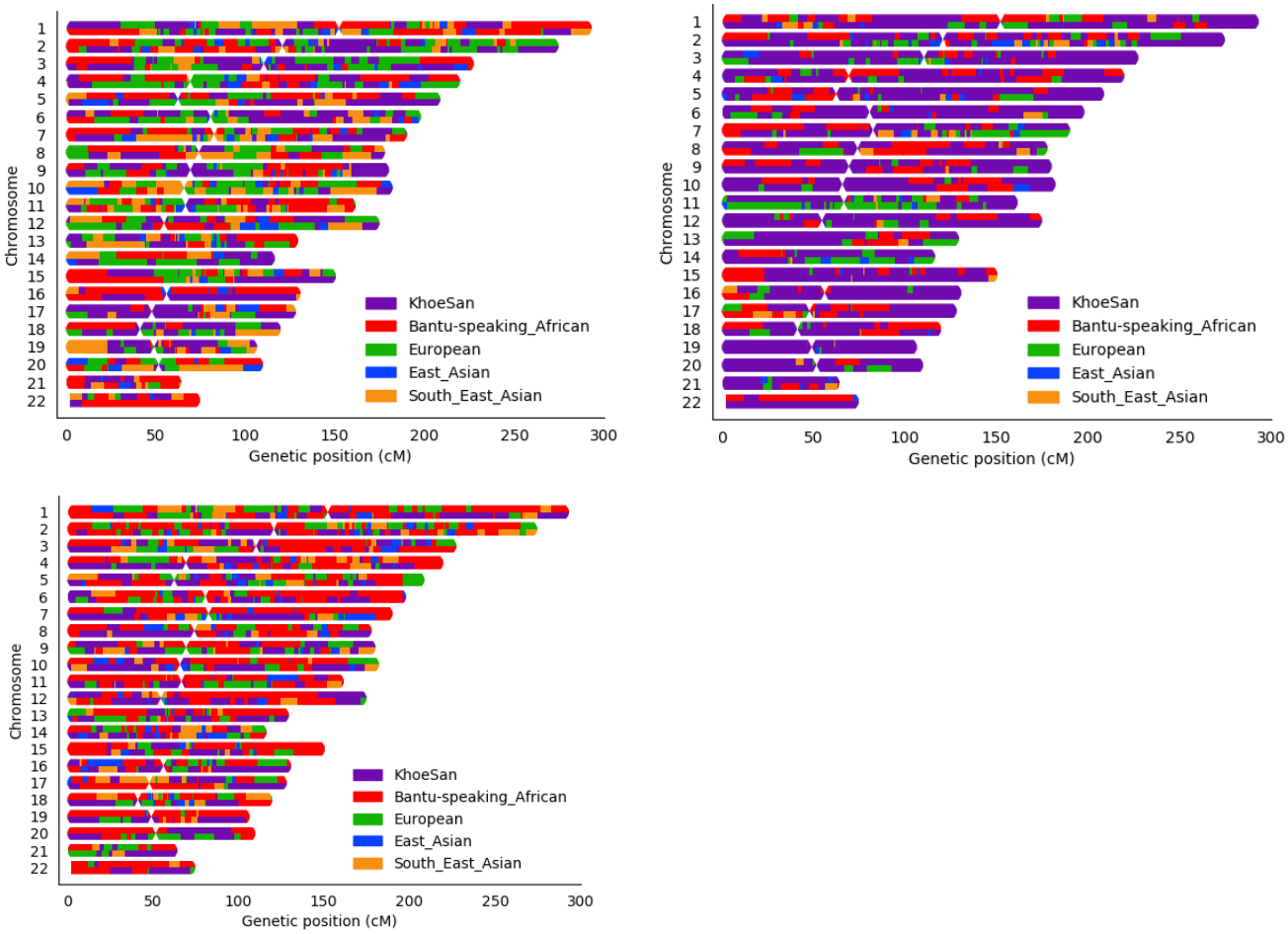


Figure 2. Local ancestry karyograms of three admixed individuals from the SAC population. Each admixed individual has unique local ancestry patterns generated by admixture among geographically distinct ancestral population groups.

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.

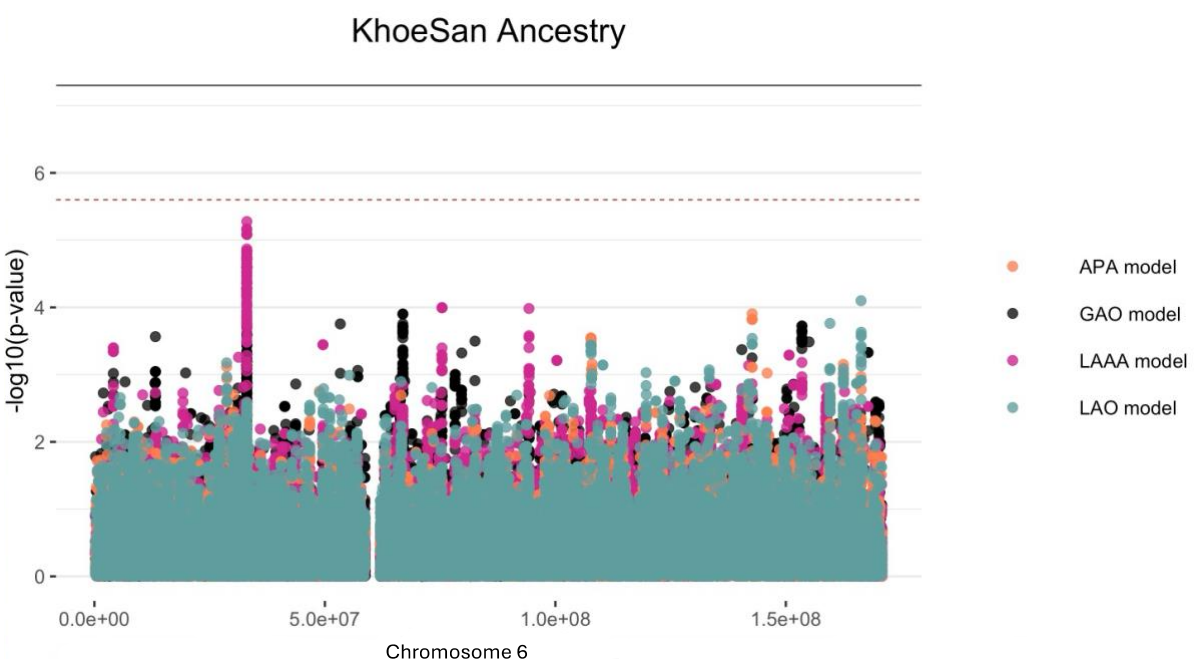


Figure 3. Log transformation of association signals obtained for KhoeSan ancestry whilst using the LAAA model on chromosome 6. The dashed red line represents the significant threshold for admixture mapping calculated with the software STEAM (p-value = 2.5 x 10^{-6}) and the black solid line represents the genome wide significant threshold (p-

value = 5×10^{-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).

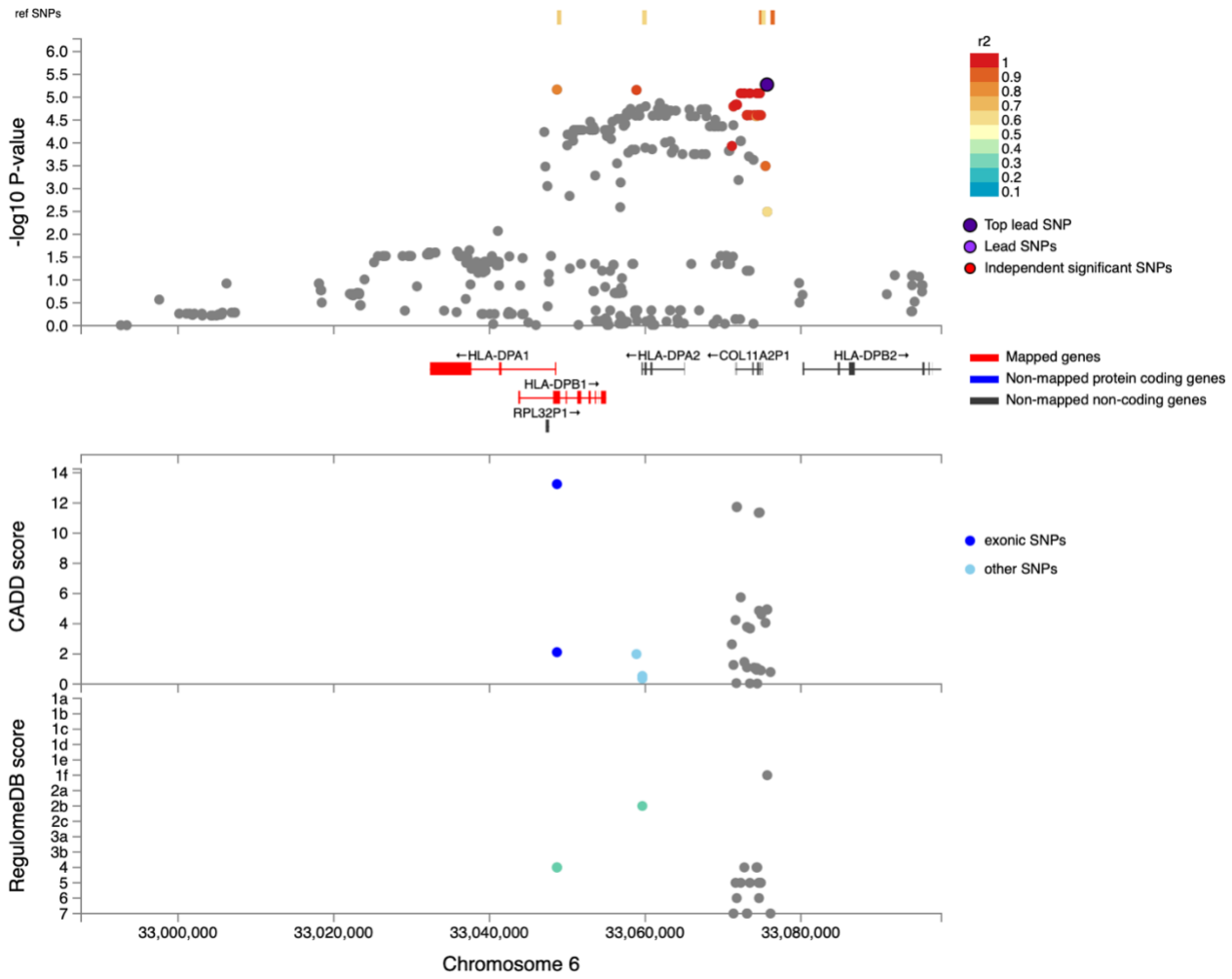


Figure 4. Regional plot indicating the nearest genes in the region of the lead variant (*rs3117230*) observed on chromosome 6. SNPs in linkage disequilibrium (LD) with the lead variant are coloured red/orange. The lead variant is indicated in purple. Functional protein-coding genes are coded in red and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

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

Position	Marker name	Ref	Alt	AltFreq	OR	SE	*p*-value	Gene	Location
33075635	*rs3117230*	A	G	0.370	0.437	0.182	$5.292e^{-06}$	*HLA-DPB1*	Intergenic
33048661	*rs1042151*	A	G	0.325	0.437	0.184	$6.806e^{-06}$	*HLA-DPB1*	Exonic
33058874	*rs2179920*	C	T	0.369	0.445	0.180	$6.960e^{-06}$	*HLA-DPB1*	Intergenic
33072266	*rs2064478*	C	T	0.371	0.447	0.181	8.222^{e-06}	*HLA-DPB1*	Intergenic
33072729	*rs3130210*	G	T	0.371	0.447	0.181	8.222^{e-06}	*HLA-DPB1*	Intergenic
33073440	*rs2064475*	G	A	0.371	0.447	0.181	$8.222e^{-06}$	*HLA-DPB1*	Intergenic
33074348	*rs3117233*	T	C	0.371	0.447	0.181	$8.222e^{-06}$	*HLA-DPB1*	Intergenic
33074707	*rs3130213*	G	A	0.371	0.447	0.181	$8.222e^{-06}$	*HLA-DPB1*	Intergenic

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-DP, -DR* 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.

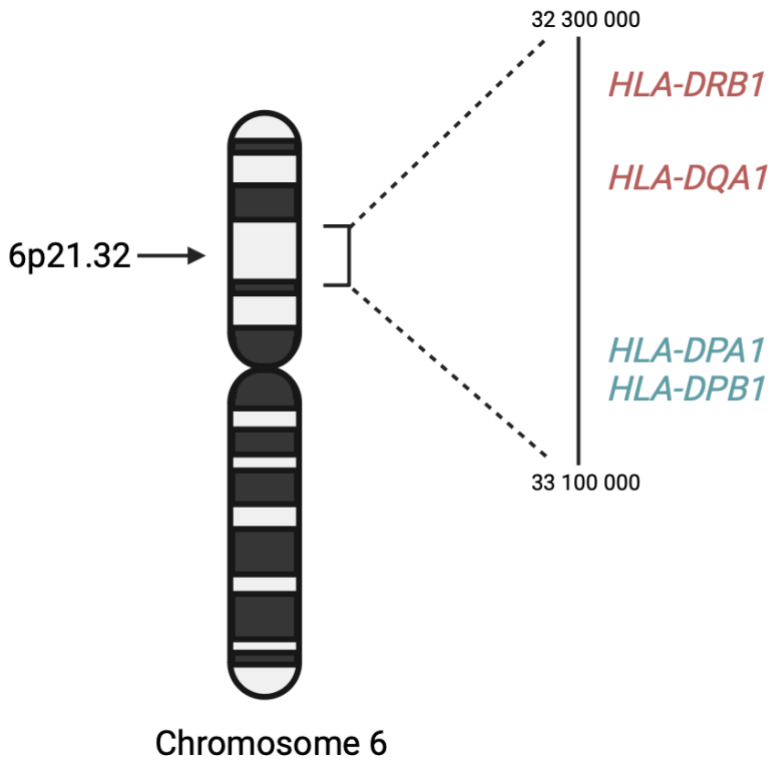


Figure 5. A schematic diagram the location of HLA-II genes associated with TB susceptibility. Genes in red were identified by the ITHGC. Genes in blue were identified by this study.

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

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, *12*, 246. https://doi.org/10.1186/1471-2105-12-246

Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P., & Ramachandran, S. (2016). pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics*, *32*(18), 2817–2823. https://doi.org/10.1093/bioinformatics/btw327

Cai, L., Li, Z., Guan, X., Cai, K., Wang, L., Liu, J., & Tong, Y. (2019). The research progress of host genes and tuberculosis susceptibility. *Oxidative Medicine and Cellular Longevity*, *2019*, 9273056. https://doi.org/10.1155/2019/9273056

Chang, S. T., Linderman, J. J., & Kirschner, D. E. (2005). Multiple mechanisms allow Mycobacterium tuberculosis to continuously inhibit MHC class II-mediated antigen presentation by macrophages. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(12), 4530–4535. https://doi.org/10.1073/pnas.0500362102

Chen, D., Tashman, K., Palmer, D. S., Neale, B., Roeder, K., Bloemendal, A., Churchhouse, C., & Ke, Z. T. (2022). A data harmonization pipeline to leverage external controls and boost power in GWAS. *Human Molecular Genetics*, *31*(3), 481–489. https://doi.org/10.1093/hmg/ddab261

Chihab, L. Y., Kuan, R., Phillips, E. J., Mallal, S. A., Rozot, V., Davis, M. M., Scriba, T. J., Sette, A., Peters, B., Lindestam Arlehamn, C. S., & SATVI Study Group. (2023). Expression of specific HLA class II alleles is associated with an increased risk for active tuberculosis and a distinct gene expression profile. *HLA : Immune Response Genetics*, *101*(2), 124–137. https://doi.org/10.1111/tan.14880

Chimusa, E. R., Daya, M., Möller, M., Ramesar, R., Henn, B. M., van Helden, P. D., Mulder, N. J., & Hoal, E. G. (2013). Determining ancestry proportions in complex admixture scenarios in South Africa using a novel proxy ancestry selection method. *Plos One*, *8*(9), e73971. https://doi.org/10.1371/journal.pone.0073971

Chimusa, E. R., Zaitlen, N., Daya, M., Möller, M., van Helden, P. D., Mulder, N. J., Price, A. L., & Hoal, E. G. (2014). Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. *Human Molecular Genetics*, *23*(3), 796–809. https://doi.org/10.1093/hmg/ddt462

Choudhury, A., Sengupta, D., Ramsay, M., & Schlebusch, C. (2021). Bantu-speaker migration and admixture in southern Africa. *Human Molecular Genetics*, *30*(R1), R56–R63. https://doi.org/10.1093/hmg/ddaa274

Cudahy, P. G. T., Wilson, D., & Cohen, T. (2020). Risk factors for recurrent tuberculosis after successful treatment in a high burden setting: a cohort study. *BMC Infectious Diseases*, *20*(1), 789. https://doi.org/10.1186/s12879-020-05515-4

Dawkins, B. A., Garman, L., Cejda, N., Pezant, N., Rasmussen, A., Rybicki, B. A., Levin, A. M., Benchek, P., Seshadri, C., Mayanja-Kizza, H., Iannuzzi, M. C., Stein, C. M., & Montgomery, C. G. (2022). Novel HLA associations with outcomes of Mycobacterium tuberculosis exposure and sarcoidosis in individuals of African ancestry using nearest-neighbor feature selection. *Genetic Epidemiology*, *46*(7), 463–474. https://doi.org/10.1002/gepi.22490

Daya, M., van der Merwe, L., Galal, U., Möller, M., Salie, M., Chimusa, E. R., Galanter, J. M., van Helden, P. D., Henn, B. M., Gignoux, C. R., & Hoal, E. (2013). A panel of ancestry informative markers for the complex five-way admixed South African coloured population. *Plos One*, *8*(12), e82224. https://doi.org/10.1371/journal.pone.0082224

de Sá, N. B. R., Ribeiro-Alves, M., da Silva, T. P., Pilotto, J. H., Rolla, V. C., Giacoia-Gripp, C. B. W., Scott-Algara, D., Morgado, M. G., & Teixeira, S. L. M. (2020). Clinical and genetic markers associated with tuberculosis, HIV-1 infection, and TB/HIV-immune reconstitution inflammatory syndrome outcomes. *BMC Infectious Diseases*, *20*(1), 59. https://doi.org/10.1186/s12879-020-4786-5

Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F., & Marchini, J. (2013). Haplotype estimation using sequencing reads. *American Journal of Human Genetics*, *93*(4), 687–696. https://doi.org/10.1016/j.ajhg.2013.09.002

Duan, Q., Xu, Z., Raffield, L. M., Chang, S., Wu, D., Lange, E. M., Reiner, A. P., & Li, Y. (2018). A robust and powerful two-step testing procedure for local ancestry adjusted allelic association analysis in admixed populations. *Genetic Epidemiology*, *42*(3), 288–302. https://doi.org/10.1002/gepi.22104

Durbin, R. (2014). Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics*, *30*(9), 1266–1272. https://doi.org/10.1093/bioinformatics/btu014

Escombe, A. R., Ticona, E., Chávez-Pérez, V., Espinoza, M., & Moore, D. A. J. (2019). Improving natural ventilation in hospital waiting and consulting rooms to reduce nosocomial tuberculosis transmission risk in a low resource setting. *BMC Infectious Diseases*, *19*(1), 88. https://doi.org/10.1186/s12879-019-3717-9

Gallant, C. J., Cobat, A., Simkin, L., Black, G. F., Stanley, K., Hughes, J., Doherty, T. M., Hanekom, W. A., Eley, B., Beyers, N., Jaïs, J. P., van Helden, P., Abel, L., Alcaïs, A., Hoal, E. G., & Schurr, E. (2010). Impact of age and sex on mycobacterial immunity in an area of high tuberculosis incidence. *The International Journal of Tuberculosis and Lung Disease*, *14*(8), 952–959.

Glaziou, P., Floyd, K., & Raviglione, M. C. (2018). Global epidemiology of tuberculosis. *Seminars in Respiratory and Critical Care Medicine*, *39*(3), 271–285. https://doi.org/10.1055/s-0038-1651492

Grinde, K. E., Brown, L. A., Reiner, A. P., Thornton, T. A., & Browning, S. R. (2019). Genome-wide Significance Thresholds for Admixture Mapping Studies. *American Journal of Human Genetics*, *104*(3), 454–465. https://doi.org/10.1016/j.ajhg.2019.01.008

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M. O., Choudhury, A., Ritchie, G. R. S., Xue, Y., Asimit, J., Nsubuga, R. N., Young, E. H., Pomilla, C., Kivinen, K., Rockett, K., Kamali, A., ... Sandhu, M. S. (2015). The African Genome Variation Project shapes medical genetics in Africa. *Nature*, *517*(7534), 327–332. https://doi.org/10.1038/nature13997

Harishankar, M., Selvaraj, P., & Bethunaickan, R. (2018). Influence of genetic polymorphism towards pulmonary tuberculosis susceptibility. *Frontiers in Medicine*, *5*, 213. https://doi.org/10.3389/fmed.2018.00213

Houben, R. M. G. J., & Dodd, P. J. (2016). The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. *PLoS Medicine*, *13*(10), e1002152. https://doi.org/10.1371/journal.pmed.1002152

Kroon, E. E., Kinnear, C. J., Orlova, M., Fischinger, S., Shin, S., Boolay, S., Walzl, G., Jacobs, A., Wilkinson, R. J., Alter, G., Schurr, E., Hoal, E. G., & Möller, M. (2020). An observational study identifying highly tuberculosis-exposed, HIV-1-positive but persistently TB, tuberculin and IGRA negative persons with M. tuberculosis specific antibodies in Cape Town, South Africa. *EBioMedicine*, *61*, 103053. https://doi.org/10.1016/j.ebiom.2020.103053

Kuhn, R. M., Haussler, D., & Kent, W. J. (2013). The UCSC genome browser and associated tools. *Briefings in Bioinformatics*, *14*(2), 144–161. https://doi.org/10.1093/bib/bbs038

Laghari, M., Sulaiman, S. A. S., Khan, A. H., Talpur, B. A., Bhatti, Z., & Memon, N. (2019). Contact screening and risk factors for TB among the household contact of children with active TB: a way to find source case and new TB cases. *BMC Public Health*, *19*(1), 1274. https://doi.org/10.1186/s12889-019-7597-0

Lehohla, P. (2012). *South African Census 2011 Meta-data* (Report No. 03-01-47; p. 130). South African Census.

Li, M., Hu, Y., Zhao, B., Chen, L., Huang, H., Huai, C., Zhang, X., Zhang, J., Zhou, W., Shen, L., Zhen, Q., Li, B., Wang, W., He, L., & Qin, S. (2021). A next generation sequencing combined genome-wide association study identifies novel tuberculosis susceptibility loci in Chinese population. *Genomics*, *113*(4), 2377–2384. https://doi.org/10.1016/j.ygeno.2021.05.035

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867–2873. https://doi.org/10.1093/bioinformatics/btq559

Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *American Journal of Human Genetics*, *93*(2), 278–288. https://doi.org/10.1016/j.ajhg.2013.06.020

Matose, M., Poluta, M., & Douglas, T. S. (2019). Natural ventilation as a means of airborne tuberculosis infection control in minibus taxis. *South African Journal of Science*, *115*(9/10). https://doi.org/10.17159/sajs.2019/5737

Möller, M., Kinnear, C. J., Orlova, M., Kroon, E. E., van Helden, P. D., Schurr, E., & Hoal, E. G. (2018). Genetic Resistance to Mycobacterium tuberculosis Infection and Disease. *Frontiers in Immunology*, *9*, 2219. https://doi.org/10.3389/fimmu.2018.02219

Möller, M., & Kinnear, C. J. (2020). Human global and population-specific genetic susceptibility to Mycobacterium tuberculosis infection and disease. *Current Opinion*

in Pulmonary Medicine, *26*(3), 302–310. https://doi.org/10.1097/MCP.0000000000000672

Nyamundanda, G., Poudel, P., Patil, Y., & Sadanandam, A. (2017). A novel statistical method to diagnose, quantify and correct batch effects in genomic studies. *Scientific Reports*, *7*(1), 10849. https://doi.org/10.1038/s41598-017-11110-6

Oliveira-Cortez, A., Melo, A. C., Chaves, V. E., Condino-Neto, A., & Camargos, P. (2016). Do HLA class II genes protect against pulmonary tuberculosis? A systematic review and meta-analysis. *European Journal of Clinical Microbiology & Infectious Diseases*, *35*(10), 1567–1580. https://doi.org/10.1007/s10096-016-2713-x

Oyageshio, O. P., Myrick, J. W., Saayman, J., van der Westhuizen, L., Al-Hindi, D., Reynolds, A. W., Zaitlen, N., Uren, C., Möller, M., & Henn, B. M. (2023). Strong effect of demographic changes on tuberculosis susceptibility in south africa. *MedRxiv*. https://doi.org/10.1101/2023.11.02.23297990

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Ravikumar, M., Dheenadhayalan, V., Rajaram, K., Lakshmi, S. S., Kumaran, P. P., Paramasivan, C. N., Balakrishnan, K., & Pitchappan, R. M. (1999). Associations of HLA-DRB1, DQB1 and DPB1 alleles with pulmonary tuberculosis in south India. *Tubercle and Lung Disease : The Official Journal of the International Union against*

Tuberculosis and Lung Disease, *79*(5), 309–317. https://doi.org/10.1054/tuld.1999.0213

Robinson, J., Barker, D. J., Georgiou, X., Cooper, M. A., Flicek, P., & Marsh, S. G. E. (2020). IPD-IMGT/HLA Database. *Nucleic Acids Research*, *48*(D1), D948–D955. https://doi.org/10.1093/nar/gkz950

Schurz, H., Kinnear, C. J., Gignoux, C., Wojcik, G., van Helden, P. D., Tromp, G., Henn, B., Hoal, E. G., & Möller, M. (2018). A Sex-Stratified Genome-Wide Association Study of Tuberculosis Using a Multi-Ethnic Genotyping Array. *Frontiers in Genetics*, *9*, 678. https://doi.org/10.3389/fgene.2018.00678

Schurz, H., Müller, S. J., van Helden, P. D., Tromp, G., Hoal, E. G., Kinnear, C. J., & Möller, M. (2019). Evaluating the Accuracy of Imputation Methods in a Five-Way Admixed Population. *Frontiers in Genetics*, *10*, 34. https://doi.org/10.3389/fgene.2019.00034

Schurz, H., Naranbhai, V., Yates, T. A., Gilchrist, J. J., Parks, T., Dodd, P. J., Möller, M., Hoal, E. G., Morris, A. P., Hill, A. V. S., & International Tuberculosis Host Genetics Consortium. (2024). Multi-ancestry meta-analysis of host genetic susceptibility to tuberculosis identifies shared genetic architecture. *ELife*, *13*. https://doi.org/10.7554/eLife.84394

Selvaraj, P., Raghavan, S., Swaminathan, S., Alagarasu, K., Narendran, G., & Narayanan, P. R. (2008). HLA-DQB1 and -DPB1 allele profile in HIV infected patients

with and without pulmonary tuberculosis of south India. *Infection, Genetics and Evolution*, *8*(5), 664–671. https://doi.org/10.1016/j.meegid.2008.06.005

Smith, M. H., Myrick, J. W., Oyageshio, O., Uren, C., Saayman, J., Boolay, S., van der Westhuizen, L., Werely, C., Möller, M., Henn, B. M., & Reynolds, A. W. (2023). Epidemiological correlates of overweight and obesity in the Northern Cape Province, South Africa. *PeerJ*, *11*, e14723. https://doi.org/10.7717/peerj.14723

Sveinbjornsson, G., Gudbjartsson, D. F., Halldorsson, B. V., Kristinsson, K. G., Gottfredsson, M., Barrett, J. C., Gudmundsson, L. J., Blondal, K., Gylfason, A., Gudjonsson, S. A., Helgadottir, H. T., Jonasdottir, A., Jonasdottir, A., Karason, A., Kardum, L. B., Knežević, J., Kristjansson, H., Kristjansson, M., Love, A., ... Stefansson, K. (2016). HLA class II sequence variants influence tuberculosis risk in populations of European ancestry. *Nature Genetics*, *48*(3), 318–322. https://doi.org/10.1038/ng.3498

Swart, Y., Uren, C., Eckold, C., Cliff, J. M., Malherbe, S. T., Ronacher, K., Kumar, V., Wijmenga, C., Dockrell, H. M., van Crevel, R., Walzl, G., Kleynhans, L., & Möller, M. (2022). *cis* -eQTL mapping of TB-T2D comorbidity elucidates the involvement of African ancestry in TB susceptibility. *BioRxiv*. https://doi.org/10.1101/2022.10.19.512814

Swart, Y., Uren, C., van Helden, P. D., Hoal, E. G., & Möller, M. (2021). Local ancestry adjusted allelic association analysis robustly captures tuberculosis susceptibility loci. *Frontiers in Genetics*, *12*, 716558. https://doi.org/10.3389/fgene.2021.716558

Swart, Y., van Eeden, G., Sparks, A., Uren, C., & Möller, M. (2020). Prospective avenues for human population genomics and disease mapping in southern Africa. *Molecular Genetics and Genomics*, *295*(5), 1079–1089. https://doi.org/10.1007/s00438-020-01684-8

Swart, Y., van Eeden, G., Uren, C., van der Spuy, G., Tromp, G., & Moller, M. (2022). GWAS in the southern African context. *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/2022.02.16.480704

Ugarte-Gil, C., Alisjahbana, B., Ronacher, K., Riza, A. L., Koesoemadinata, R. C., Malherbe, S. T., Cioboata, R., Llontop, J. C., Kleynhans, L., Lopez, S., Santoso, P., Marius, C., Villaizan, K., Ruslami, R., Walzl, G., Panduru, N. M., Dockrell, H. M., Hill, P. C., Mc Allister, S., ... van Crevel, R. (2020). Diabetes Mellitus Among Pulmonary Tuberculosis Patients From 4 Tuberculosis-endemic Countries: The TANDEM Study. *Clinical Infectious Diseases*, *70*(5), 780–788. https://doi.org/10.1093/cid/ciz284

Uren, C, Hoal, E. G., & Möller, M. (2020). Putting RFMix and ADMIXTURE to the test in a complex admixed population. *BMC Genetics*, *21*(1), 40. https://doi.org/10.1186/s12863-020-00845-3

Uren, Caitlin, Henn, B. M., Franke, A., Wittig, M., van Helden, P. D., Hoal, E. G., & Möller, M. (2017). A post-GWAS analysis of predicted regulatory variants and tuberculosis susceptibility. *Plos One*, *12*(4), e0174738. https://doi.org/10.1371/journal.pone.0174738

Uren, Caitlin, Hoal, E. G., & Möller, M. (2021). Mycobacterium tuberculosis complex and human coadaptation: a two-way street complicating host susceptibility to TB. *Human Molecular Genetics*, *30*(R1), R146–R153. https://doi.org/10.1093/hmg/ddaa254

Verhein, K. C., Vellers, H. L., & Kleeberger, S. R. (2018). Inter-individual variation in health and disease associated with pulmonary infectious agents. *Mammalian Genome*, *29*(1–2), 38–47. https://doi.org/10.1007/s00335-018-9733-z

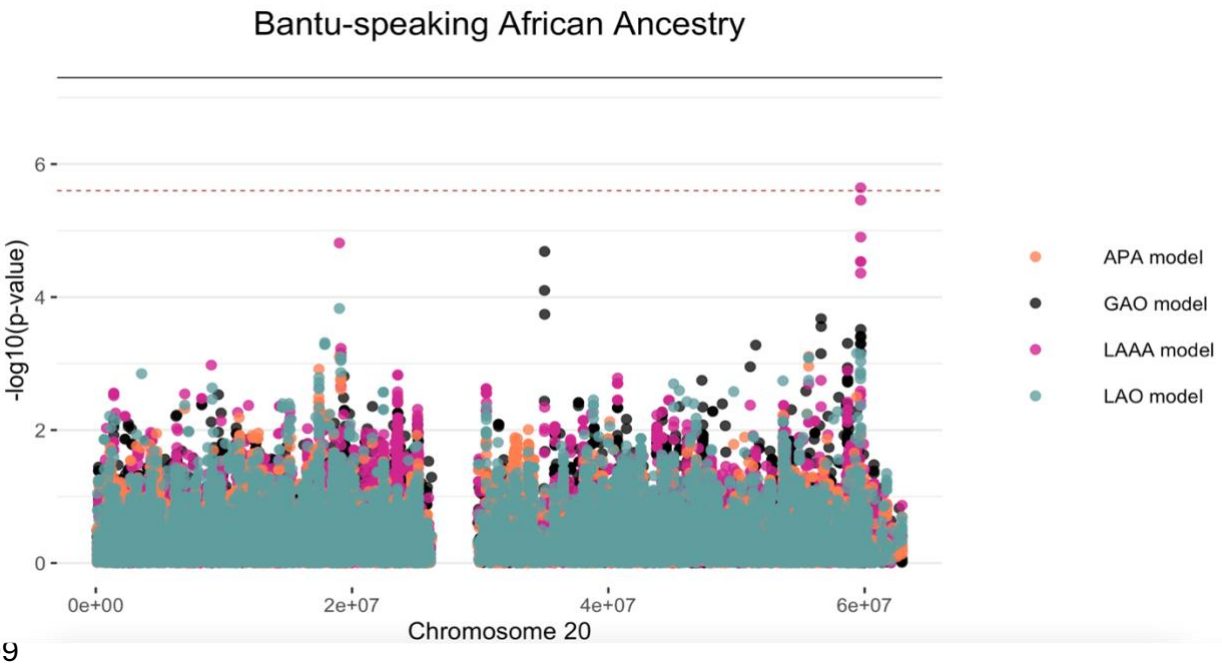
Witek, J., & Mohiuddin, S. S. (2024). Biochemistry, Pseudogenes. In *StatPearls*. StatPearls Publishing.

Wong, L.-P., Ong, R. T.-H., Poh, W.-T., Liu, X., Chen, P., Li, R., Lam, K. K.-Y., Pillai, N. E., Sim, K.-S., Xu, H., Sim, N.-L., Teo, S.-M., Foo, J.-N., Tan, L. W.-L., Lim, Y., Koo, S.-H., Gan, L. S.-H., Cheng, C.-Y., Wee, S., … Teo, Y.-Y. (2013). Deep whole-genome sequencing of 100 southeast Asian Malays. *American Journal of Human Genetics*, *92*(1), 52–66. https://doi.org/10.1016/j.ajhg.2012.12.005

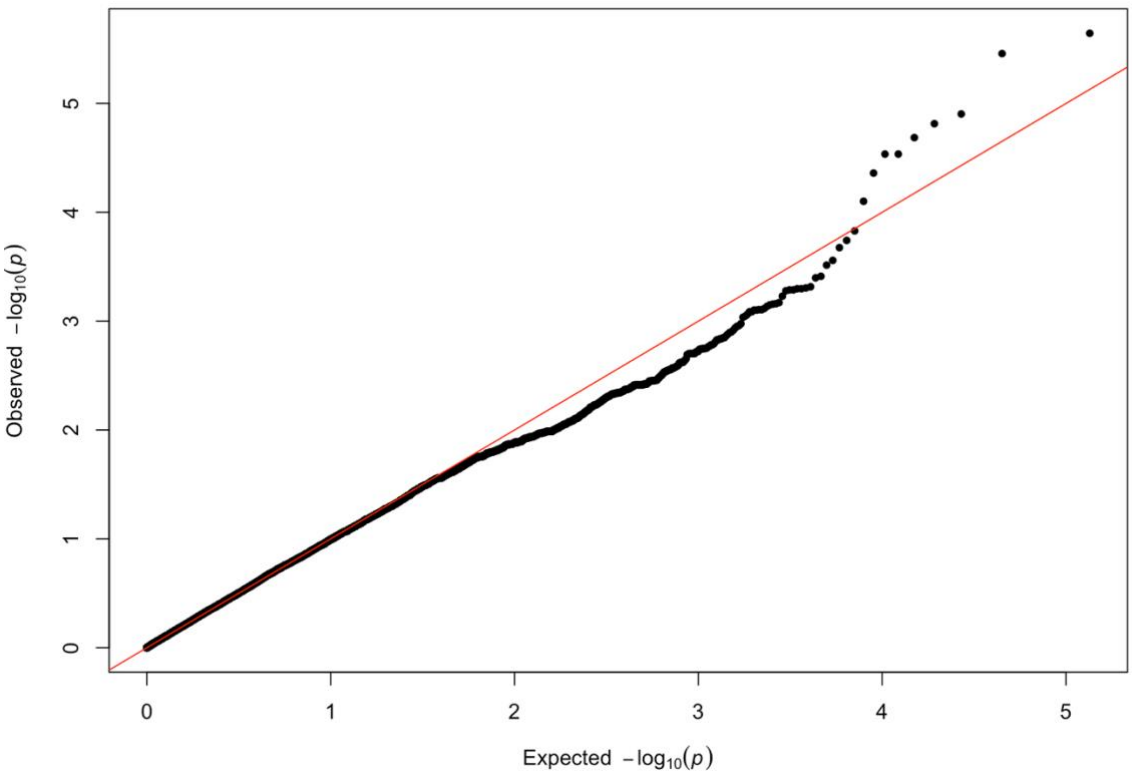
World Health Organization. (2023). *Global Tuberculosis Report 2023* (World Health Organization, Ed.; p. 75). World Health Organization.

Zheng, R., Li, Z., He, F., Liu, H., Chen, J., Chen, J., Xie, X., Zhou, J., Chen, H., Wu, X., Wu, J., Chen, B., Liu, Y., Cui, H., Fan, L., Sha, W., Liu, Y., Wang, J., Huang, X., … Ge, B. (2018). Genome-wide association study identifies two risk loci for tuberculosis in Han Chinese. *Nature Communications*, *9*(1), 4072. https://doi.org/10.1038/s41467-018-06539-w

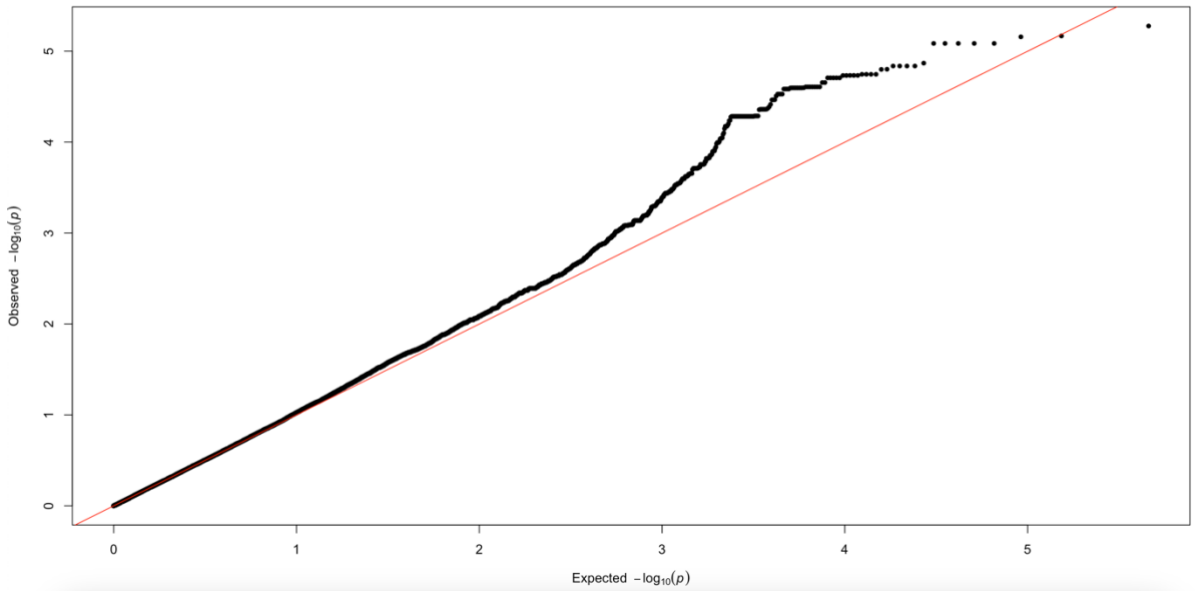
Supplementary Material



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 × 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.