BMJ Open

Protocol for a biomarker discovery study to identify correlates of risk for future tuberculosis disease progression in South African children (INTREPID)

Melissa Shea Hamilton ,[1,2] Iena P Derks ,[3] Myrsini Kaforou,[1,2] Rory Dunbar,[3] Ryan P McNamara,[4,5] Sarah M Fortune,[4,5] Robin Basu-Roy,[6] Anneen van Deventer,[3] Corne Bosch,[3] Claire Dunican ,[1,2] Marieke M van der Zalm ,[3] Michael Levin ,[1,2] H Simon Schaaf ,[3] John A Altin,[7] Anneke C Hesseling,[3] James A Seddon[1,2,3]

To cite: Hamilton MS, Derks IP, Kaforou M, *et al*. Protocol for a biomarker discovery study to identify correlates of risk for future tuberculosis disease progression in South African children (INTREPID). *BMJ Open* 2026;**16**:e099348. doi:10.1136/bmjopen-2025-099348

► Prepublication history for this paper is available online. To view these files, please visit the journal online (https://doi.org/10.1136/bmjopen-2025-099348).

Received 15 January 2025
Accepted 12 January 2026

For numbered affiliations see end of article.

Correspondence to
Dr Melissa Shea Hamilton; s.hamilton@imperial.ac.uk

ABSTRACT

Introduction Young children and children living with HIV are at high risk of progressing to tuberculosis (TB) disease following *Mycobacterium tuberculosis* (*Mtb*) exposure and infection, and also of developing severe forms of disease and TB-related mortality. Identifying children who have very early (sub-clinical) TB disease, prior to progression to clinically apparent TB, would mean that TB preventive treatment (TPT) could be more efficiently targeted to this group. Identifying biomarker changes on drug therapy in children with *Mtb* infection or very early disease could pave the way for the development of tests that can identify which children have viable bacilli and are therefore at increased risk of disease progression.

Methods and analysis The INTREPID study will use already collected samples taken from well-phenotyped paediatric cohorts in three clinical studies conducted in South Africa in children <5 years, including a drug-resistant TPT trial (TB-CHAMP), an observational household contact study (interferon-gamma release assay studies) and a prospective diagnostic study (Umoya), all conducted in a setting with a high burden of TB and HIV. We will employ transcriptomic, proteomic, metabolomic and serology approaches to analyse changes in host blood profiles at every stage along the TB continuum, from *Mtb* exposure to disease and from children treated for *Mtb* infection and early TB disease, as well as targeted *Mtb* antibody analysis. Data on viral co-infections and relevant clinical and epidemiological parameters will be integrated and evaluated to identify the optimal biosignatures that can predict future progression to clinically overt disease in children below 5 years of age, including those living with HIV.

Ethics and dissemination The study protocol received ethical approval from the Stellenbosch University Health Research Ethics Committee (N23/03/025). The study findings will be disseminated through peer-reviewed publications, scientific conferences and formal presentations to healthcare professionals and to local communities, in collaboration with the Desmond Tutu TB Centre Community Advisory Board.

STRENGTHS AND LIMITATIONS TO THE STUDY

⇒ A key strength of this study is the use of well-curated samples from rigorously conducted studies in children <5 years of age, including a drug-resistant tuberculosis (TB) preventive treatment trial (TB-CHAMP), an observational household contact study (interferon-gamma release assay studies) and a prospective hospital-based diagnostic study (Umoya), all conducted in a setting with high burden of TB and HIV.

⇒ We are uniquely placed to investigate children sampled at every stage along the TB continuum and have longitudinal samples from children who have progressed from *Mycobacterium tuberculosis* (*Mtb*) exposure to disease and from children treated for *Mtb* infection and early TB disease.

⇒ Complementary biomarker signatures that predict transitions between relevant clinical TB states will be identified using transcriptomic (RNA-sequencing), proteomic (SomaScan), metabolomic (targeted and untargeted) and antibody approaches (Systems Serology).

⇒ Data from multi-omic approaches will be integrated and evaluated in the context of clinical and epidemiological co-variates, including co-infection with other pathogens (PepSeq).

⇒ Children were enrolled using both community-based and hospital-based recruitment strategies, and recruitment was limited to one country, which are limitations of this study.

INTRODUCTION

Globally, 70 million children are estimated to have immunological sensitisation to *Mycobacterium tuberculosis* (*Mtb*), classified as *Mtb* infection.[1] Of these, 1.25 million children are estimated to progress to tuberculosis (TB) disease each year, with nearly a quarter of a million dying, almost all <5 years of age and undiagnosed.[2] Young children are at

the highest risk of progression from *Mtb* infection to TB disease and of developing severe forms of TB, including TB meningitis.[2] They therefore represent a priority group for TB preventive treatment (TPT).

Although TPT is effective in preventing TB disease in children,[3] many healthy children need TPT to prevent one TB case.[4] Providing TPT to all children <5 years following *Mtb* exposure is globally recommended,[2] yet this is a blunt and inefficient approach, as most TB-exposed children do not progress to TB disease, leading to substantial overtreatment with associated costs and risks. Risks include both those to the individual (toxicity, pill burden), as well as to society (antimicrobial resistance).[5] A more targeted approach, where TPT is provided to children at the highest risk of disease progression, would reduce unnecessary TPT courses and decrease the burden on health services and families.

The traditional binary conceptualisation of TB clinical states as *Mtb* infection or TB disease is increasingly challenged with wide recognition of a continuum including *Mtb* exposure, infection, asymptomatic disease and clinically apparent or overt disease.[6 7] Most research in adults identifying biomarkers predicting future disease progression has identified metabolic and immunological activity of early TB disease.[8 9] There are no data identifying these subtle signatures in young children or in children living with HIV, who are at different stages of immunological development and are thus likely to have different signatures.[10]

Viral co-infections are likely to have a profound influence on the interaction between the host immune system and *Mtb*.[11] Therefore, the identification of *Mtb* biomarkers in children should also consider the relevant clinical and epidemiological context, including age and exposure to other pathogens. Viral infections are common in young children and several viruses, including CMV and HIV, influence the risk of progression to TB, and likely the performance of TB biomarkers, many of which reflect the activity of type I IFNs which are also produced in the context of viral infection.[12–15] In addition to impacting the risk of progression, co-infections may also change the clinical presentation of TB disease, including symptoms and disease severity, and biomarker performance. In the Correlate of Risk Targeted Intervention Study (CORTIS) trial, the RISK11 transcriptomic signature score, a TB blood biomarker, was significantly altered by the presence of viruses in the upper respiratory tract in South African adults.[16 17]

Currently, it is not possible to determine which child contacts have viable mycobacteria in their body, a necessary pre-condition for developing incident or future TB disease.[6] It is increasingly suggested that many individuals with immunological evidence of *Mtb* sensitisation have eradicated all organisms and have no viable bacilli.[18] If it were possible to identify the subset of children with viable bacilli, then an alternative approach for targeted TPT would be to provide therapy only to these children. As the death of *Mtb* can be detected through immunological assays,[19] it may be possible to identify signatures indicating *Mtb* death that would allow for more targeted TPT approaches. This might include baseline and early sampling on treatment to allow a stop/go decision-making process for continuation of TPT dependent on identification of *Mtb* death.

Study objectives

In this study, we aim to identify biomarkers and combinations of biomarkers that will predict which children are at high risk of progressing to future TB disease. We will leverage samples from children <5 years of age in three clinical studies conducted in South Africa. We will carefully select samples from our extensive repository to enable (1) cross-sectional comparison of different well-defined clinical states, (2) nested case-control analysis of children followed over time who transition from one clinical state to the next and (3) analysis of children treated for *Mtb* infection and early TB disease. We will use transcriptomic, proteomic, metabolomic and antibody approaches to fully interrogate these samples in their clinical and epidemiological context to identify the best biomarkers that accurately describe these changes (figure 1). Finally, we will undertake an integrated analysis of these approaches, together with analysis of concurrent viral exposure/infection and important clinical co-variates to arrive at the most complete and accurate signatures to fully understand the biology of *Mtb,* particularly as it relates to future disease progression in young children including those living with and those exposed to HIV.

METHODS AND ANALYSIS

Description of cohorts and setting

We will use samples from three complementary studies led and overseen by the Desmond Tutu TB Centre (DTTC), Stellenbosch University, and conducted in epidemiologically diverse settings across South Africa: TB-CHAMP, interferon-gamma release assay (IGRA) studies and Umoya (table 1).

TB-CHAMP

TB-CHAMP (Tuberculosis Child Multidrug-Resistant Preventative Therapy) was a randomised, placebo-controlled community-based trial of TPT for children <5 years old exposed to multidrug-resistant (MDR)-TB in their household. MDR-TB is defined as TB disease caused by *Mtb* resistant to at least rifampicin and isoniazid. The trial was implemented at five sites in South Africa in five different provinces, with recruitment ending on 31 July 2022 and follow-up completing by 31 January 2023. *Mtb* exposure was quantified using a standard measure.[20] Children were randomised to 6 months of either daily levofloxacin or matched placebo. By trial closure, 922 children (839 under 5 years) had been enrolled (2.1% HIV-positive and 35.8% HIV-exposed) and 21 TB endpoints had been reached. At trial entry, prevalent TB

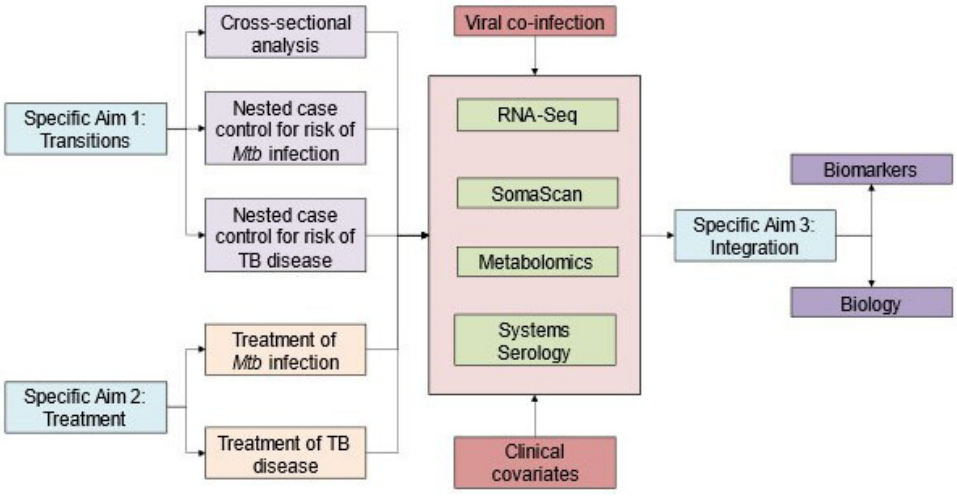

Figure 1 Overview of the INTREPID study. *Mtb*, *Mycobacterium tuberculosis*; RNA-Seq, RNA-sequencing; TB, tuberculosis.

disease was excluded through comprehensive history, examination, chest X-ray (CXR) and microbiology. IGRA testing (QFT-GIT) was conducted at baseline, but children were randomised regardless of IGRA status. The primary outcome was incident TB disease (confirmed or unconfirmed, as defined by the 2015 National Institutes of Health (NIH) consensus definitions)[21] and TB death, by 48 weeks post-randomisation. For the end point review, all CXRs were dual read by experts blinded to clinical details and trial arm. All end points were reviewed by a blinded independent end-point review committee. Blood samples (PAXgene and serum) were collected at baseline, 8, 16, 24 and 48 weeks, and at the time of incident TB disease and stored at −80°C to preserve RNA and protein integrity for long-term storage and downstream applications.[22 23]

Table 1 Overview of TB-CHAMP, IGRA studies and Umoya cohorts

	TB-CHAMP	IGRA studies	Umoya
Number of children <5 years	839	534	475
Number HIV-positive	12 (1.4%)	108 (20.2%)	36 (7.6%)
PAXgene available	Baseline, 8, 16, 24 and 48 weeks	None	Baseline
Serum available	Baseline, 8, 16, 24 and 48 weeks	None	Baseline, 2, 8, 16 and 24 weeks
Plasma available	None	Baseline, 12, 24 and 60 weeks	None

TB-CHAMP, Tuberculosis Child Multidrug-Resistant Preventative Therapy trial
IGRA, interferon-gamma release assay; TB, tuberculosis.

The IGRA studies

The IGRA studies recruited children between December 2007 and June 2012 into a prospective, community-based household contact diagnostic study evaluating the diagnostic utility of IGRAs in three communities with a high burden of TB in Cape Town.[20] The cohort included children <15 years with known exposure to an adult with pulmonary drug-susceptible TB, together with matched children from control households with no known TB exposure and enriched for HIV through recruitment from HIV clinics. *Mtb* exposure was quantified using a standard measure.[24] Children were evaluated for TB disease at baseline using history, examination, CXR and microbiology, and TB disease was classified based on a standard protocol case definition.[25] Prevalent TB was defined as disease diagnosed within 30 days of enrolment. Of 1088 children enrolled, 534 were <5 years. Of those <5 years, 20.2% were HIV-positive and 30% were HIV-exposed. For end point review, all CXRs were dual read by experts blinded to clinical details and all end points were reviewed by a blinded independent end-point review committee. As all relevant clinical, radiological and microbiological data were collected, reclassification of TB endpoints is feasible using the 2015 NIH case definition.[21] All children <5 years and all those living with HIV were referred for TPT as per local standard of care. Many did not attend local clinics and poor adherence was common. 34 children <5 years developed incident TB, mostly children who did not receive TPT. Longitudinal data collection included serial measures of *Mtb* infection (IGRA and tuberculin skin test (TST)). IGRA supernatant and plasma were collected and stored at −80°C.

Umoya

Umoya is an ongoing prospective cohort study that recruits children presenting with presumptive pulmonary TB to two hospitals (one secondary and another tertiary level) in Cape Town. Standard TB investigations are completed at enrolment, including clinical examination,

respiratory sampling for TB microbiology, CXR, TST and TB exposure status. TB status is defined as confirmed, unconfirmed and unlikely TB, using the 2015 NIH case definition.[21] For the end point review, all CXRs are dual read by experts blinded to clinical details. To measure immune responses in children with very early, limited disease, we have identified children in this cohort who have TB exposure, have acute symptoms (<2 weeks, to include cough, fever, respiratory distress, wheeze) and have either a normal CXR or a CXR with only hilar lymph nodes (early radiological disease), regardless of microbiological confirmation. We have also identified children with no symptoms and normal radiology but with respiratory samples positive for *Mtb*. Serum and PAXgene samples are collected at enrolment and stored at −80°C, with serum also collected at 2, 8, 16 and 24 weeks, and at unscheduled visits. By the end of December 2024, 610 children had been enrolled, with 475 <5 years. Recruitment and follow-up will continue through 2025, including long-term assessment of lung health; however, we included samples from patients enrolled before 2025 for this study. To date, ±40% of children have confirmed or unconfirmed TB,[21] 50% are symptomatic controls (initially presenting with presumptive TB but not started on TB treatment) and 10% are healthy sibling controls. Of those under 5 years, 7.6% are HIV-positive and 20% are HIV-exposed uninfected.

Laboratory analysis and data generation

Transcriptomics

From stored PAXgene tubes, total RNA will be extracted using PAXgene Blood RNA kits and RNA quantity and quality measured. Following library preparation, including globin and ribosomal RNA depletion, RNA-Seq will be done using an Illumina HiSeq-2500 platform at 100 bp paired-end configurations at the South African Medical Research Council Genomics Centre, generating a raw read count of 40 million reads per sample. After quality control, sequence reads will be mapped and quantified, including the identification of transcript isoforms. We will perform differential expression analysis using linear models to compare features between clinical groups in a pairwise manner, accounting for confounding factors using DeSeq2 in R.[26] We will also perform time-course and predictive analysis for outcomes of *Mtb* infection or TB disease,[27] as well as cell type adjustments using cell type deconvolution (DeconRNAseq).[28]

Proteomics

Serum and plasma proteomic profiles will be measured on the SomaScan platform, which uses the Slow Off-rate Modified Aptamer reagents consisting of short single-stranded DNA sequences that allow tight and specific binding of proteins to their target. This specificity has enabled the current SomaScan platform (V.4.1) to accurately resolve over 7000 proteins in complex clinical samples such as plasma and serum. We will send samples of serum and plasma (75 µL) to SomaLogic, Inc (Boulder, Colorado, USA). Samples will be analysed in a single run at three serum/plasma concentrations (2%, 0.2%, 0.005%) to ensure the precise measurement of low, medium and highly abundant proteins and the raw fluorescence data returned for analysis. Data validation and quality control steps will be performed using internal assay controls to correct for variations in aptamer hybridisation efficiency and inter- and intra-assay variability.

Systems Serology

We will use the Systems Serology platform to characterise biophysical and functional properties of antigen-specific serum/plasma antibodies at the Systems Serology Lab (Harvard School of Public Health, Boston, USA). For biophysical profiling, we will use multiplexed Luminex technology to interrogate serum/plasma responses to a panel of established *Mtb* antigens, including purified protein derivative, antigen 85, early secretory antigen target 6, culture filtrate protein 10, heat shock protein X, GroES and lipoarabinomannan. Influenza haemagglutinin and Ebola glycoprotein will be included as positive and negative reactogenicity controls, respectively. Fluorescently barcoded microspheres will be incubated with diluted serum/plasma, and fluorescent reagents will be used to detect (1) antigen-specific antibody isotype and subclass, using commercially available fluorescent secondary antibodies against human IgG1–IgG4, IgA1, IgA2 and IgM; (2) binding of Fc receptors, using Fc gamma receptor binding antibodies (FcγR): FcγR2A, FcγR2B, FcγR3A and FcγR3B; Fc alpha receptor I; and Fc mu receptor available through Duke Protein Production Facility and covalently coupled to phycoerythrin; and (3) Fc glycosylation, using commercially available fluorescently coupled *Sambucus nigra* agglutinin (recognises sialic acid) and *Ricinus communis* agglutinin I (recognises galactose). Median fluorescence intensity will be measured for all features of interest, and further protocol information can be found elsewhere.[29] All assays will be performed in duplicate. For functional profiling, we will evaluate antibody-mediated restriction of *Mtb* using sera/plasma from children who progressed to TB versus non-progressors. A luciferase reporter strain of H37Rv will be opsonised with diluted sera/plasma and incubated in donor blood for 5 days, a duration empirically determined to maximise detection of killing. Restriction will be measured by luminescence.

Metabolomics

We will use a combination of metabolomic technologies including liquid chromatography mass spectrometry and ^{1}H-nuclear mass resonance methods[30] to globally profile the metabolome with broad coverage.[31] This will enable the detection of small molecule metabolites and lipids and the opportunity for novel discovery in parallel to targeted data extraction of pre-defined metabolite panels. Hydrophilic interaction chromatography and two reversed-phase chromatography separations will enable detection of very polar and ionic metabolites, as well as

small hydrophilic and moderately hydrophobic species and neutral and complex lipids with both positive and negative mode ion detection. The National Phenome Centre (NPC), at Imperial College, London, UK, will analyse aliquots of serum and plasma with a prioritisation algorithm based on available samples volume and assay volume requirements. Where specific metabolic pathways or compound classes are implicated in differences between clinical states, between progressors or non-progressors or between children treated or untreated for *Mtb* infection or disease from broad untargeted analysis, targeted quantitative assays from the BioExact Assay Suite (https://phenomecentre.org/platform/) will be incorporated into the algorithm. Sample preparation, running the assays, quality control and assurance of the data and data pre-processing will be performed by the NPC using established standardised operating procedures (SOPs)[31] and an open-source pipeline.[32] The NPC will provide raw and pre-processed data along with a quality report. Metabolite identification of unknown chemicals and putative biomarkers identified during the discovery profiling phase will be applied to establish de novo structure elucidation of analytes of interest where not previously annotated.[33]

Viral Approaches

We will infer historical and contemporary exposures across the human virome using highly-multiplexed peptide-based serology on the PepSeq platform.[34] To that end, we will assay plasma IgG reactivity using the previously described 'HV2' library that consists of 15 000 DNA-barcoded 30mer peptides representing 80 commonly infecting virus species (empirically selected based on reactivity observed in a previously conducted virome-wide screen of 244 000 peptides, representing >200 species).[35] For longitudinal analyses, we will apply the previously described Peptide Set Enrichment Analysis method[36] to integrate high-dimensional peptide-level reactivity changes between consecutive timepoints into species-specific infection calls. We will also evaluate evidence of viral genomic material in the RNA-Seq data using MetaMix.[37] Using our pilot whole blood RNA-Seq data (40 million reads per sample), we have previously identified reads mapping to lytic viral genes (ie, EBV). Additionally, we will mine the data for host transcriptomic residual profiles that will indicate co-infections of interest (eg, CMV, SARS-CoV-2 and influenza). These will be identified by comparing the whole blood transcriptomic profiles generated in this study to those generated from children in previous studies with confirmed viral infections using multi-class, multi-label classification methods. We have shown that individual transcriptomic profiles can be assigned to specific causative pathogens, by interrogating a database of blood transcriptional profiles for multiple infectious diseases.[38]

Statistical analysis

The data generated from these approaches will undergo modality-specific pre-processing as outlined above. Once a feature abundance matrix is generated, abundance levels between different clinical groups of interest will be compared while accounting for covariates of interest (ie, age, sex, BCG vaccination status, TPT provision, site of recruitment). P values will be corrected for multiple testing via false discovery rate estimation. To investigate the underlying biology of disease, systems biology approaches including Ingenuity Pathway Analysis for the transcriptomic, proteomic and metabolomic data will be used to map the differentially expressed entities to pathways. Further analyses will be undertaken to determine important network characteristics, including hub molecules and molecular interactions. For biomarker identification, least absolute shrinkage and selection operator (LASSO) will be used to select features that best differentiate progressor and non-progressor groups.[39] To select the minimal list (signature) of discriminatory biomarkers with potential for translation for clinical use, we will employ our in-house variable selection (forward selection partial least squares), that results in minimal models—smaller compared with LASSO—while retaining the predictive accuracy.[40 41] Finally, we will assess the performance of the defined signatures by calculating the ROC AUC (Receiver Operating Characteristic- Area Under the Curve) based on the defined signatures, in each case using 80% of each cohort as a training set to develop the ROC curves, and the remaining 20% reserved as a test set.[42]

Specific aim 1: To identify biomarkers that predict progression to *Mtb* infection and/or TB disease in children in the context of relevant clinical and biological covariates

Cross-sectional analysis

We will carefully select children from each of five clinical categories (figure 2), matched for age and sex. Each group of 35 children will have 15 boys and 15 girls and there will be six children in each 1-year age band, matched for HIV exposure/infection status. Children will be selected from all three cohorts to populate each category, which assumes similar immune responses to drug-susceptible and drug-resistant *Mtb*. We will use the definition for limited, early pulmonary disease as outlined above, and

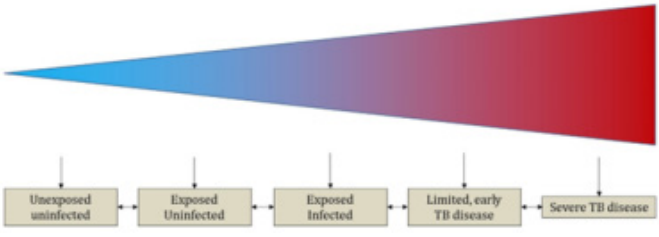

Figure 2 Cross-sectional analysis of children in different clinical states. Blue indicates no bacilli transitioning into large numbers of bacilli as indicated by the red colour. TB, tuberculosis.

severe/extensive pulmonary disease will include extensive parenchymal involvement or other features as described by Wiseman.[43] Any children with asymptomatic disease (no symptoms but an abnormal CXR or who are microbiologically positive) will also be included as an additional comparator group (although the numbers will be relatively small). A PAXgene tube from each child at baseline will undergo RNA-Seq and serum or plasma from each child taken at baseline will be sent for SomaScan, Systems Serology, metabolomic analysis and PepSeq. The aim of this analysis will be to identify differences in multi-omics data between the five clinical categories.

Risk of progression to *Mtb* infection indicated by IGRA conversion

167 children<5 years old from the IGRA studies were exposed in their household to an infectious TB case but were IGRA negative at baseline. Of these, 35 converted to IGRA-positive status. We will identify all samples from these 35 children (baseline and all follow-up samples until the timepoint when their IGRA test converted—at least three samples per child) and will match (on age, sex, exposure score, HIV status and weight-for-age z-score) these children (1:2) to 70 children who were IGRA-negative at baseline and remained IGRA-negative throughout the study (controls).

Samples from children progressing to *Mtb* infection and controls will undergo SomaScan, Systems Serology, metabolomic analysis and PepSeq (PAXgene tubes not available for the IGRA studies). We will realign timepoints for the cases, so that the time of IGRA positivity is time 0 and all samples taken prior are 0 minus x, y or z weeks. We will divide children randomly into a training and test group. In the training set, we will first perform differential abundance analysis between controls and cases at time 0 and we will then compare the controls to samples taken from the cases at several time-points: 0 minus x weeks, 0 minus y weeks, etc (figure 3). Once we have identified signatures in the training set (80%), we will evaluate them independently in the test set (20%). While we recognise that IGRA testing is an imperfect way of characterising *Mtb* infection, performance of IGRAs against a reference standard of confirmed TB is good in children <5 years, with a sensitivity of 0.85 (95% CI 0.79 to 0.89) and a specificity of 0.94 (95% CI 0.88 to 0.97),[44] suggesting that biomarkers that are associated with development of a positive IGRA will be associated with the development of *Mtb* infection.

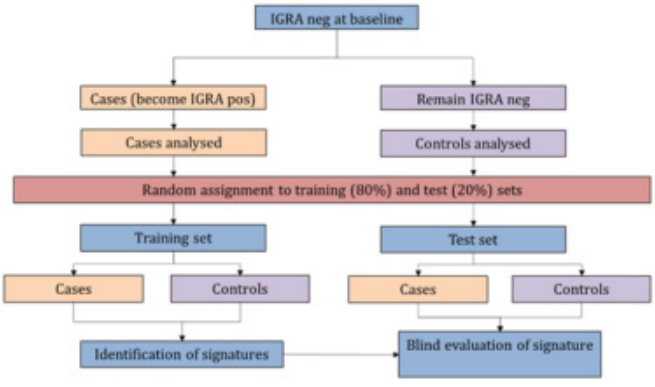

Figure 3 Nested case-control study to evaluate risk of progression to TB infection with training/test design. IGRA, interferon-gamma release assay; TB, tuberculosis.

Risk of progression to TB disease

For the gene expression analysis, we will use samples from TB-CHAMP: 21 children who were well at baseline that subsequently developed incident TB disease. We will compare the 21 children who progressed to disease with (1:2) 42 children, matched on age, sex, study site and trial arm, who remained disease free. We will evaluate baseline samples from those who remained well and samples from all time points in those who progressed to disease, including baseline and at the time of incident TB. We anticipate a median of 4 time points per child with incident TB; therefore, we will undertake RNA sequencing on 84 samples from children with incident TB and on 42 samples from controls. The nested case-control analysis uses a time course analysis to compare gene expression changes across time points (diagnosis vs 8 weeks prior) where the statistical power is derived from the total number of longitudinal samples allowing the use of 84 samples to power the analysis. As above, we will realign time points for the cases, so that the time of diagnosis is time 0 and compare gene expression between controls and cases at all time points. We will again divide children randomly into a training and test group (figure 4). For SomaScan, metabolomic analyses and Systems Serology, we will have an abundance of samples from the TB-CHAMP and IGRA studies. We will use a similar methodology to that described above. However, for these approaches, we will use the 21 incident TB cases from TB-CHAMP and the 34 incident TB disease cases from the IGRA studies, so we will compare 55 cases to 110 controls (using the same controls for TB-CHAMP as for the RNA-Seq experiment). We will evaluate samples taken from all time points for incident TB cases and compare them to controls at baseline. We estimate that there are at least three samples per incident TB case, so we will analyse 165 samples from children with incident TB. For each of these samples, analysis by PepSeq will also be completed.

Sample size and power calculation

Power calculation is provided here for the RNA-Seq modality, as it quantifies the largest number of features in comparison to the other modalities. We therefore performed a sample size calculation in R for the RNA-Seq experiment using RNASeqPower (V.1.12.0). *Cross-sectional analysis*:

To compare between each two of the five clinical stages (figure 2), based on the following assumptions: (i) RNA-sequencing depth of at least 30 million reads, (ii) a coefficient of variation of 0.3 within the group, (iii) a probability of 0.05 for a type I error, (iv) a probability of 0.1 for a type II error (ie, a power of 90%) and (v) an

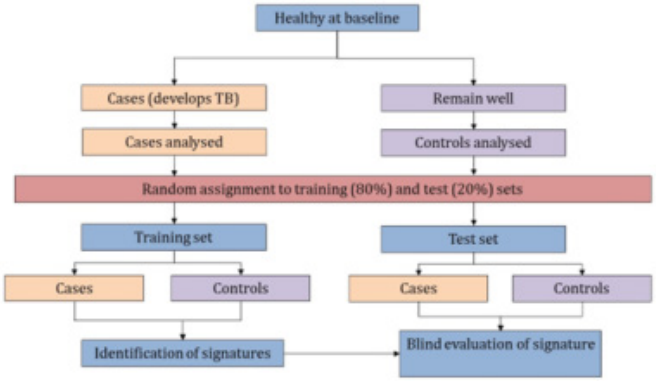

Figure 4 Nested case-control study to evaluate risk of progression to TB disease with training/test design. TB, tuberculosis.

effect size of 1.5, we would need to include at least 25 individuals in each group to demonstrate difference in RNA expression. *Nested case-control analysis*: To compare between cases (exposed uninfected vs exposed infected, and progressors vs non-progressors; figures 3,4) based on the previous assumptions, we need to include at least 25 cases and 25 controls.

Specific aim 2: To determine biomarkers of *Mtb* treatment in children with *Mtb* infection and early TB disease, in the context of co-infections and other relevant co-variates

Changes in host-omics profiles during treatment of *Mtb* infection

Levofloxacin is an antituberculosis drug and a broad-spectrum antibiotic expected to cause substantial and varied disruption in the host immune system of children enrolled in the TB-CHAMP trial, who were all exposed to MDR-TB in the household. At baseline, there are four groups of children, namely those on levofloxacin and those on placebo, and within these groups some with positive and some with negative IGRA tests. With the assumption that more children with a positive IGRA have viable *Mtb,* TB-CHAMP provides a unique opportunity to untangle the non-specific effects of chemotherapy from the specific effects on *Mtb* and study the change in biomarker signatures caused by mycobacterial death. We will identify 30 age- and sex-matched children in each of these four groups and undertake RNA-Seq, SomaScan, Systems Serology, metabolomics and PepSeq at baseline and 2 months (figure 5). We will carry out differential expression or abundance analysis to identify differences in genes, proteins, antibodies and metabolites between 2 months and baseline in each group. By accounting for the changes in immune profiles of IGRA-negative children given placebo from IGRA-negative children given levofloxacin, we will be able to define the non-specific effects of levofloxacin on immune responses in children without TB. By subtracting the change in immune profile of IGRA-positive children given placebo from IGRA-positive children treated with levofloxacin, we will be able to define the overall impact of levofloxacin on immune responses in children (including impact on *Mtb* and other effects). By subtracting the first difference from the second, it will be possible to define the specific effect of levofloxacin on *Mtb* death, potentially leading to a signature that can inform the use and duration of TPT. Understanding the impact of age, sex, HIV exposure and infection status, and other viral co-infections on these signatures will also inform how TPT regimens should be designed and implemented in the future.

Change in host-omics profiles when treating (early) TB disease

We will select 30 children from the Umoya cohort who have well-characterised early and limited TB disease (as previously defined) and carry out SomaScan, Systems Serology, metabolomics and PepSeq on samples at baseline and at 8 weeks (PAXgene not collected at 8 weeks). By defining the immunological change in treatment of children with early disease, we will be able to identify patterns of change during the treatment of early disease (asymptomatic microbiologically confirmed and asymptomatic clinically diagnosed), potentially leading to the development of a molecular biomarker that can identify sub-clinical TB and therefore children at high risk of progression to clinical disease. These analyses will also generate a better understanding of the biology of sub-clinical TB.

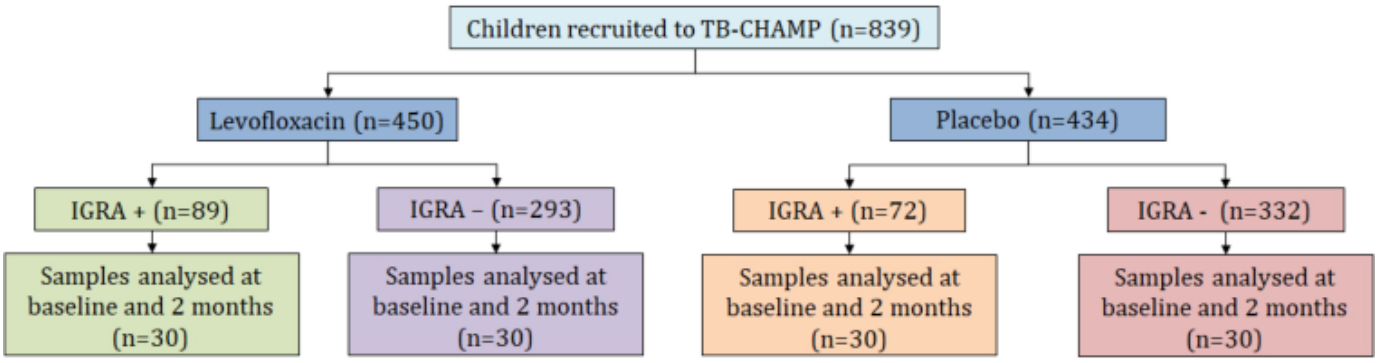

Figure 5 Evaluation of immunological response to treatment of individuals with *Mtb* infection, treated with TB preventive therapy. IGRA, interferon-gamma release assay; *Mtb*, *Mycobacterium tuberculosis*; TB, tuberculosis.

Specific aim 3: To integrate all biomarkers of disease progression and *Mtb* treatment to identify optimal biosignatures and to define the biological pathways implicated in these processes

Many methods of multi-omics' integration have been developed, including graph-based, clustering-based and factor analysis-based methods.[45–47] We will perform supervised and unsupervised integration analyses to identify the biological mechanisms that distinguish children who progress to *Mtb* infection from those who remain IGRA-negative, and critically, on those who progress to disease from those who remain TB disease-free. We will analyse each timepoint separately and then integrate all timepoints from the same child to identify biological trajectories that predict progression to *Mtb* infection or TB disease. First, the acquired RNA, protein, metabolite and antibody profiling data, along with the clinical and viral data, will be pre-processed and analysed on a single-omics level. Count data from the RNA-Seq dataset, abundance levels from the protein profiling, metabolite and antibody datasets, along with clinical and viral data, will be used as an input for the integration across all patients. Second, we aim to identify a multi-omics' set of features that distinguish progressors from non-progressors using DIABLO.[48] In addition, we will perform a data-driven hypothesis-free analysis to identify clusters of patients with similar-omics profiles. To this end, we will use multi-omics factor analysis (MOFA), which is a versatile generalisation of principle component analysis (PCA) to multi-omics data[45] as it infers a low-dimensional representation of the data. MOFA also allows for missing values to be imputed, and factors to be annotated through gene set enrichment analysis. The molecular features identified by MOFA from the inferred factor loadings will be used to complement the biomarkers identified by DIABLO. We will also use similarity network fusion (SNF) for unsupervised analysis and disease subtype identification.[49] SNF creates a matrix for each omic type/data modality (RNA, protein, metabolites, antibodies, viral exposure) to identify pairwise sample similarities and construct a network for each data type. The molecular features identified by MOFA from the inferred factor loadings could be used as a complement to the biomarkers identified by DIABLO. Finally, a single consensus network is produced which contains shared and complementary information from the data sources.[49] From this, clusters of patients will be identified with greater resolution than the single omics networks allow.

Risks and mitigation

Many reported TB biomarkers fail to replicate in subsequent studies. The main factors for such failure are inadequate sample size in the discovery process; failure to include a representative spectrum of patients during the biomarker discovery process and over-fitting of data due to failure to validate the performance of candidate biomarkers in independent cohorts. Additional risks for failure of biomarkers to replicate are variations between target populations in the frequency of co-factors such as malnutrition, viral or other co-infections, HIV and socio-economic conditions. Technical factors are inadequate sample handling and storage, lack of standardised laboratory procedures, failure to account for batch effects in processing and potential risk of bias due to unblinding to relevant clinical data. We have attempted to mitigate these risks by considering them in the design of our study and during our proposed programme of work. All three studies enrolled children in South Africa, using the same approaches and protocols, using standard case definitions for TB, and led by the same group. We have developed extremely well-characterised clinical phenotypes of disease based on clinical, radiological and microbiological data. We have included participants drawn from a wide spectrum of ethnic, geographical and epidemiological settings. TB-CHAMP recruited children from five geographically and ethnically diverse settings across five provinces. The other two studies recruited children from urban Cape Town, with extensive ethnic and genetic diversity in our study population.[50] We have addressed technical errors in biomarker validation, by ensuring that all steps in the discovery and validation process follow standardised operating procedures (SOPs). These include categorisation of patients into diagnostic groups: collection and recovery of plasma, serum and RNA and sample aliquoting and storage. Sample volumes are adequate for all planned analyses and sample integrity has been maintained. All laboratory analyses will be undertaken by experienced laboratory personnel and will follow established SOPs. All laboratory analyses will be blinded to clinical disease states and all other participant characteristics, reducing the potential for bias. We will consider batch effects, by randomising samples from different diagnostic groups and different cohorts between sample runs. Once biomarkers have been identified, we will carry out external validation. Discussions have already taken place with the lead investigators of the PREDICT (R01AI175315) and PROTECT (R01AI175312) projects to plan mutual cross-validation of resulting signatures.

Biomarkers discovered using one technological platform (eg, RNA-Seq) may fail to replicate when an alternative technology (eg, point of care chip-based PCR) is used for detection. To mitigate this risk, our pipeline ensures validation first by the same technology as used in the discovery process and then transfer to an alternative technology for large scale analysis. To mitigate a single-omic approach not identifying biological markers of TB progression, we have applied multi-omic approaches and modalities to address this clinical need.

Participant and public involvement

The DTTC Institutional Community Advisory Board (CAB) was involved in the development of the proposal and provided critical input into the study approach and dissemination plan.

Ethics and dissemination

The three parent studies were all approved by the Stellenbosch University Health Research Ethics

Committee (HREC): TB-CHAMP (M16/02/009), IGRA studies (N05/07/129 and N08/08/207) and Umoya (N17/08/083). The study protocol for INTREPID was approved by HREC on 19 April 2023 (N23/03/025). At recruitment into all three clinical studies, each participant was allocated a unique study number. Personal identifier details were recorded and stored securely, only available to selected delegated team members, and kept separately from any clinical and epidemiological data, laboratory samples or results from laboratory analysis. These data, along with all research data derived from the methods described in this protocol, will only be linked to the study number and stored on a secure database at DTTC with access restricted to the principal investigators only. Study findings will be disseminated through publication in peer-reviewed journals and at scientific conferences and disseminated to local TB and other healthcare services. We will work with the DTTC CAB to develop the most appropriate dissemination plan for communication with the communities from which participants were recruited.

Author affiliations
[1]Department of Infectious Disease, Imperial College London, London, England, UK
[2]Centre for Paediatrics and Child Health, Imperial College London, London, England, UK
[3]Desmond Tutu TB Centre, Department of Paediatrics and Child Health, Stellenbosch University Faculty of Medicine and Health Sciences, Cape Town, South Africa
[4]Ragon Institute of Massachusetts General Brigham, MIT and Harvard, Cambridge, Massachusetts, USA
[5]Department of Immunology and Infectious Diseases, Harvard TH Chan School of Public Health, Boston, Massachusetts, USA
[6]Centre for Genomics and Child Health, Queen Mary University of London, London, UK
[7]Translational Genomics Research Institute, Arizona State University, Tempe, Arizona, USA

Contributors The study concept and design were conceived by JAS and ACH. Method development from: JAS, MSH, RPM, SMF, RBR and JAA. Clinical recruitment of samples: JAS, MMZ and ACH. Clinical sample evaluation, screening and processing: JAS, MSH, IPD, RD, AVD, CB, MMZ and ACH. Data analysis plan: JAS, MK, MSH and CD. Preparation of the first draft of the manuscript: JAS, MSH, IPD and MK. All authors provided edits and critiqued the manuscript: JAS, MSH, IPD, MK, RD, RPM, SMF, RBR, AVD, CB, CD, MMZ, ML, HSS, JAA and ACH. Guarantor: JAS.

Funding This five-year study started in May 2023. The current protocol is version 1.0. INTREPID is funded by the US National Institutes of Health (R01AI175235). Infrastructure support for metabolomics is provided by the National Institute for Health Research (NIHR) Imperial Biomedical Research Centre (BRC).

Competing interests None declared.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Melissa Shea Hamilton https://orcid.org/0000-0001-6308-0253
Iena P Derks https://orcid.org/0000-0002-7070-5036
Claire Dunican https://orcid.org/0000-0001-5835-263X
Marieke M van der Zalm https://orcid.org/0000-0002-8787-9139
Michael Levin https://orcid.org/0000-0003-2767-6919
H Simon Schaaf https://orcid.org/0000-0001-5755-4133

REFERENCES

1 Dodd PJ, Sismanidis C, Seddon JA. Global burden of drug-resistant tuberculosis in children: a mathematical modelling study. *Lancet Infect Dis* 2016;16:1193–201.
2 Global tuberculosis report. 2023. Available: https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2023 [accessed 20 Oct 2024]
3 Smieja M, Marchetti C, Cook D, *et al*. Isoniazid for preventing tuberculosis in non-HIV infected persons. *Cochrane Database Syst Rev* 1999.
4 Ragonnet R, Trauer JM, McBryde ES, *et al*. Is IPT more effective in high-burden settings? Modelling the effect of tuberculosis incidence on IPT impact. *Int J Tuberc Lung Dis* 2017;21:60–6.
5 Chihota V, Gombe M, Gupta A, *et al*. Tuberculosis Preventive Treatment in High TB-Burden Settings: A State-of-the-Art Review. *Drugs (Abingdon Engl)* 2025;85:127–47.
6 Seddon JA, Whittaker E, Kampmann B, *et al*. The evolving research agenda for paediatric tuberculosis infection. *Lancet Infect Dis* 2019;19:e322–9.
7 Coussens AK, Zaidi SMA, Allwood BW, *et al*. Classification of early tuberculosis states to guide research for improved care and prevention: an international Delphi consensus exercise. *Lancet Respir Med* 2024;12:484–98.
8 Goletti D, Lee M-R, Wang J-Y, *et al*. Update on tuberculosis biomarkers: From correlates of risk, to correlates of active disease and of cure from disease. *Respirology* 2018;23:455–66.
9 Petruccioli E, Scriba TJ, Petrone L, *et al*. Correlates of tuberculosis risk: predictive biomarkers for progression to active tuberculosis. *Eur Respir J* 2016;48:1751–63.
10 Suliman S, Jaganath D, DiNardo A. Predicting Pediatric Tuberculosis: The Need for Age-Specific Host Biosignatures. *Clin Infect Dis* 2023;77:450–2.
11 Whittaker E, López-Varela E, Broderick C, *et al*. Examining the Complex Relationship Between Tuberculosis and Other Infectious Diseases in Children. *Front Pediatr* 2019;7:233.
12 Dodd PJ, Prendergast AJ, Beecroft C, *et al*. The impact of HIV and antiretroviral therapy on TB risk in children: a systematic review and meta-analysis. *Thorax* 2017;72:559–75.
13 Martinez L, Nicol MP, Wedderburn CJ, *et al*. Cytomegalovirus acquisition in infancy and the risk of tuberculosis disease in childhood: a longitudinal birth cohort study in Cape Town, South Africa. *Lancet Glob Health* 2021;9:e1740–9.
14 Stockdale L, Nash S, Farmer R, *et al*. Cytomegalovirus Antibody Responses Associated With Increased Risk of Tuberculosis Disease in Ugandan Adults. *J Infect Dis* 2020;221:1127–34.
15 Cobelens F, Nagelkerke N, Fletcher H. The convergent epidemiology of tuberculosis and human cytomegalovirus infection. *F1000Res* 2018;7:280.
16 Mendelsohn SC, Mbandi SK, Fiore-Gartland A, *et al*. Prospective multicentre head-to-head validation of host blood transcriptomic biomarkers for pulmonary tuberculosis by real-time PCR. *Commun Med* 2022;2:26.
17 Mulenga H, Musvosvi M, Mendelsohn SC, *et al*. Longitudinal Dynamics of a Blood Transcriptomic Signature of Tuberculosis. *Am J Respir Crit Care Med* 2021;204:1463–72.
18 Behr MA, Kaufmann E, Duffin J, *et al*. Latent Tuberculosis: Two Centuries of Confusion. *Am J Respir Crit Care Med* 2021;204:142–8.
19 Wilkinson KA, Kon OM, Newton SM, *et al*. Effect of treatment of latent tuberculosis infection on the T cell response to Mycobacterium tuberculosis antigens. *J Infect Dis* 2006;193:354–9.
20 Mandalakas AM, Kirchner HL, Walzl G, *et al*. Optimizing the Detection of Recent Tuberculosis Infection in Children in a High Tuberculosis–HIV Burden Setting. *Am J Respir Crit Care Med* 2015;191:820–30.
21 Graham SM, Cuevas LE, Jean-Philippe P, *et al*. Clinical Case Definitions for Classification of Intrathoracic Tuberculosis in Children: An Update. *Clin Infect Dis* 2015;61Suppl 3:S179–87.
22 Tang R, She Q, Lu Y, *et al*. Quality Control of RNA Extracted from PAXgene Blood RNA Tubes After Different Storage Periods. *Biopreserv Biobank* 2019;17:477–82.
23 Valo E, Colombo M, Sandholm N, *et al*. Effect of serum sample storage temperature on metabolomic and proteomic biomarkers. *Sci Rep* 2022;12:4571.

24 Mandalakas AM, Kirchner HL, Lombard C, *et al*. Well-quantified tuberculosis exposure is a reliable surrogate measure of tuberculosis infection. *Int J Tuberc Lung Dis* 2012;16:1033–9.
25 Wiseman CA, Mandalakas AM, Kirchner HL, *et al*. Novel application of NIH case definitions in a paediatric tuberculosis contact investigation study. *Int J Tuberc Lung Dis* 2015;19:446–53.
26 Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;15:550.
27 Berk M, Ebbels T, Montana G. A statistical framework for biomarker discovery in metabolomic time course data. *Bioinformatics* 2011;27:1979–85.
28 Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* 2013;29:1083–5.
29 Tong X, McNamara RP, Avendaño MJ, *et al*. Waning and boosting of antibody Fc-effector functions upon SARS-CoV-2 vaccination. *Nat Commun* 2023;14:4174.
30 Dona AC, Jiménez B, Schäfer H, *et al*. Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping. *Anal Chem* 2014;86:9887–94.
31 Lewis M, Chekmeneva E, Camuzeaux S, *et al*. An open platform for large scale lc-ms-based metabolomics. *ChemRxiv Cambridge Cambridge Open Engag* [Preprint].
32 Sands CJ, Wolfer AM, Correia GDS, *et al*. The nPYc-Toolbox, a Python module for the pre-processing, quality-control and analysis of metabolic profiling datasets. *Bioinformatics* 2019;35:5359–60.
33 Whiley L, Chekmeneva E, Berry DJ, *et al*. Systematic Isolation and Structure Elucidation of Urinary Metabolites Optimized for the Analytical-Scale Molecular Profiling Laboratory. *Anal Chem* 2019;91:8873–82.
34 Ladner JT, Henson SN, Boyle AS, *et al*. Epitope-resolved profiling of the SARS-CoV-2 antibody response identifies cross-reactivity with endemic human coronaviruses. *Cell Reports Medicine* 2021;2:100189.
35 Elko EA, Nelson GA, Mead HL, *et al*. COVID-19 vaccination elicits an evolving, cross-reactive antibody response to epitopes conserved with endemic coronavirus spike proteins. *Cell Rep* 2022;40:111022.
36 Kelley EJ, Henson SN, Rahee F, *et al*. Virome-wide detection of natural infection events and the associated antibody dynamics using longitudinal highly-multiplexed serology. *Nat Commun* 2023;14:1783.
37 Morfopoulou S, Plagnol V. Bayesian mixture analysis for metagenomic community profiling. *Bioinformatics* 2015;31:2930–8.
38 Habgood-Coote D, Wilson C, Shimizu C, *et al*. Diagnosis of childhood febrile illness using a multi-class blood RNA molecular signature. *Med* 2023;4:635–54.
39 Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med* 1997;16:385–95.
40 Herberg JA, Kaforou M, Wright VJ, *et al*. Diagnostic Test Accuracy of a 2-Transcript Host RNA Signature for Discriminating Bacterial vs Viral Infection in Febrile Children. *JAMA* 2016;316:835–45.
41 Coin L. Github code for FSPLS, Available: https://github.com/lachlancoin/fspls [Accessed 5 Aug 2024].
42 Jackson H, Menikou S, Hamilton S, *et al*. Kawasaki Disease Patient Stratification and Pathway Analysis Based on Host Transcriptomic and Proteomic Profiles. *Int J Mol Sci* 2021;22:5655.
43 Wiseman CA, Gie RP, Starke JR, *et al*. A Proposed Comprehensive Classification of Tuberculosis Disease Severity in Children. *Pediatr Infect Dis J* 2012;31:347–52.
44 Hirabayashi R, Nakayama H, Yahaba M, *et al*. Japanese TB Clinical Practice Guideline Systematic Review Task Force. Utility of interferon-gamma releasing assay for the diagnosis of active tuberculosis in children: A systematic review and meta-analysis. *J Infect Chemother* 2024;30:516–25.
45 Argelaguet R, Velten B, Arnol D, *et al*. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* 2018;14:e8124.
46 Abdi H, Williams LJ, Valentin D. Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *WIREs Computational Stats* 2013;5:149–79.
47 Singh A, Shannon CP, Gautier B, *et al*. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* 2019;35:3055–62.
48 Rohart F, Gautier B, Singh A, *et al*. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol* 2017;13:e1005752.
49 Wang B, Mezlini AM, Demir F, *et al*. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* 2014;11:333–7.
50 Choudhury A, Ramsay M, Hazelhurst S, *et al*. Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nat Commun* 2017;8:2062.