scientific reports

OPEN

Metagenomic DNA sequencing to quantify *Mycobacterium tuberculosis* DNA and diagnose tuberculosis

Adrienne Chang[1], Omary Mzava[1], Liz-Audrey Kounatse Djomnang[1], Joan Sesing Lenz[1], Philip Burnham[1], Peter Kaplinsky[1], Alfred Andama[2], John Connelly[3], Christine M. Bachman[3], Adithya Cattamanchi[4], Amy Steadman[3] & Iwijn De Vlaminck[1]✉

Tuberculosis (TB) remains a significant cause of mortality worldwide. Metagenomic next-generation sequencing has the potential to reveal biomarkers of active disease, identify coinfection, and improve detection for sputum-scarce or culture-negative cases. We conducted a large-scale comparative study of 428 plasma, urine, and oral swab samples from 334 individuals from TB endemic and non-endemic regions to evaluate the utility of a shotgun metagenomic DNA sequencing assay for tuberculosis diagnosis. We found that the composition of the control population had a strong impact on the measured performance of the diagnostic test: the use of a control population composed of individuals from a TB non-endemic region led to a test with nearly 100% specificity and sensitivity, whereas a control group composed of individuals from TB endemic regions exhibited a high background of nontuberculous mycobacterial DNA, limiting the diagnostic performance of the test. Using mathematical modeling and quantitative comparisons to matched qPCR data, we found that the burden of *Mycobacterium tuberculosis* DNA constitutes a very small fraction (0.04 or less) of the total abundance of DNA originating from mycobacteria in samples from TB endemic regions. Our findings suggest that the utility of a minimally invasive metagenomic sequencing assay for pulmonary tuberculosis diagnostics is limited by the low burden of *M. tuberculosis* and an overwhelming biological background of nontuberculous mycobacterial DNA.

Abbreviations

TB	Tuberculosis
cfDNA	Cell-free DNA
AUC	Area under the curve

Despite the introduction of the World Health Organization's (WHO) End TB Strategy in 2015, tuberculosis (TB) remains one of the top global causes of death due to a single infectious pathogen. In 2021, the WHO reported the first rise in deaths due to tuberculosis in over a decade[1]. Recent advances in nucleic acid amplification tests, including the WHO-recommended Xpert® MTB/RIF Ultra, have produced rapid tests with high positive predictive value[2]. However, these assays are unable to replace the use of culture for bacteriological confirmation due to low negative predictive value or implementation barriers present in low- and middle-income countries that account for 98% of reported TB cases[1]. Though sputum culture has long been a gold standard for TB diagnostics, the time to positive detection can be affected by a number of variables, including sputum volume, population characteristics (e.g., age, HIV coinfection), and method of specimen handling[3,4]. Variations due to these factors can lead to discrepancies between quantitative and semiquantitative diagnostic assays. Thus, there is an unmet need for robust, point-of-care tuberculosis diagnostics.

Metagenomic DNA sequencing has the capacity to detect all potential infectious pathogens in a clinical sample using an unbiased approach. Crucially, previous studies have shown that noninvasively or minimally sampled

[1]Nancy E. and Peter C. Meinig School of Biomedical Engineering, Cornell University, Ithaca, NY, USA. [2]Department of Internal Medicine, Makerere University College of Health Sciences, Kampala, Uganda. [3]Global Health Labs, Bellevue, WA, USA. [4]Center for Tuberculosis and Division of Pulmonary and Critical Care Medicine, University of California San Francisco, San Francisco, CA, USA. ✉email: vlaminck@cornell.edu

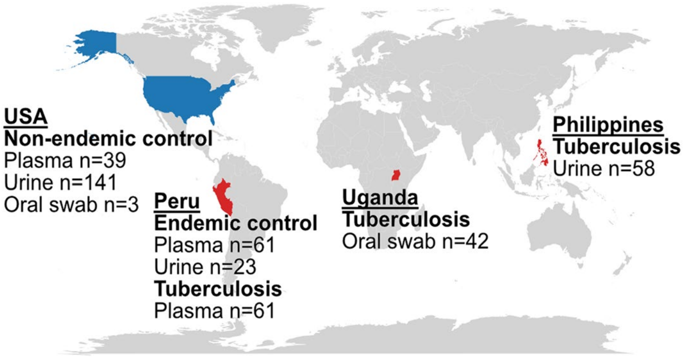

Figure 1. Geographic distribution of samples included in this study. High tuberculosis burden countries are shaded in red.

Cohort	Origin	Disease	Biofluid	Patients	Samples
Endemic	Peru	Environmental enteropathy	Plasma	61	61
			Urine	23	23
Non-endemic	USA	Kidney transplant[10]	Urine	82	141
		Lung transplant[5,11]	Plasma	6	39
		Healthy	Oral swab	2	3
Tuberculosis	Philippines[12]	Sputum positive	Urine	30	30
		Sputum negative	Urine	28	28
	Peru	Sputum positive	Plasma	17	17
		Sputum negative	Plasma	44	44
	Uganda	Sputum positive	Oral swab	27	27
		Sputum negative	Oral swab	15	15

Table 1. Overview of datasets included in this study. All data was generated for this study unless otherwise indicated.

biofluids (e.g., urine and plasma) can be used to monitor infection even if the biofluid is not downstream of or sampled from the site of injury. For example, circulating plasma DNA has been used to monitor infection and rejection after lung transplantation[5]. Numerous studies have demonstrated that DNA from *Mycobacterium tuberculosis*, the causative agent of tuberculosis, can be detected in plasma, urine, and oral fluids[6–8]. However, its diagnostic performance in distinguishing positive sputum culture from negative sputum culture samples has fallen short of the performance standards set by the WHO (98% specificity, 80% sensitivity)[9].

We explored the utility of a metagenomic sequencing assay for tuberculosis across three biofluids and four geographic regions and found that diagnostic performance is highly influenced by the geographic origin of the control cohort (Fig. 1). Using simulation and modeling, we found that the diagnostic performance is correlated with the abundance of *M. tuberculosis* DNA relative to the background of DNA from nontuberculous mycobacteria. The background of DNA originating from nontuberculous mycobacteria is low in samples from TB non-endemic regions but overwhelms and obscures the *M. tuberculosis* signal in samples from TB endemic regions. Our study provides insight into the burden and properties of *M. tuberculosis* in different biofluids and can inform the development of molecular tests that achieve the requisite standards for sensitivity and specificity.

Methods

Study cohorts.

A total of 428 datasets from plasma, urine, and oral swab samples were analyzed, 191 of which were generated for this study (Table 1). Endemic plasma and urine samples were collected from patients seeking treatment for environmental enteropathy in Peru. The study was approved by the Johns Hopkins Bloomberg School of Public Health Institutional Review Board (protocol 00002185) and the Cornell University Institutional Review Board (protocol 1612006853). Tuberculosis plasma samples were collected from patients presenting with respiratory symptoms from tuberculosis clinics in Peru. The study was approved by the Foundation for Innovative New Diagnostics' Clinical Trials Review Committee and the Cornell Institutional Review Board (protocol 1612006851). Oral swab samples were collected from individuals who presented with symptoms of respiratory illness at outpatient clinics in Uganda. The study was approved by the Makerere University School of Medicine Research and Ethics Committee (protocol 2017-020). Additional oral swab samples were collected from healthy volunteers at Cornell University. The study was approved by the Cornell University Institutional Review Board (protocol 1910009101).

Additional datasets collected in the scope of previous studies were included as follows. Non-endemic urine samples were collected from kidney transplant patients who received care at New York Presbyterian

Hospital-Weill Cornell Medical Center. The study was approved by the Weill Cornell Medicine Institutional Review Board (protocols 9402002786, 1207012730, 0710009490)[10]. Non-endemic plasma samples were collected from lung transplant patients who received care at Stanford University Hospital. The study was approved by the Stanford University Institutional Review Board (protocol 17666)[5,11]. Tuberculosis urine samples from patients seeking treatment for tuberculosis in the Philippines were collected through a study partly funded by the Department of Science and Technology-Philippine Council for Health Research and Development (DOST-PCHRD). This study was approved by the University of the Philippines Manila Research Ethics Board (protocol UPMREB 2018-252-01)[12]. All patients provided written informed consent and all experiments were performed in accordance with relevant guidelines and regulations.

The geographic distribution of the datasets included in the study were mapped using the *geom_map()* function from the R package ggplot2.

Sample collection. Urine samples were collected via the conventional clean-catch midstream method. For the endemic cohort, approximately 50 mL of urine was centrifuged at 3000×*g* on the same day for 30 min and the supernatant was stored in 1 mL aliquots at −80 °C. For the tuberculosis cohort, approximately 10 mL of urine was mixed with 2 mL Streck Cell-Free DNA Urine Preserve (Streck, Cat #230604) and centrifuged at 3000×*g* for 30 min at ambient temperature within 30 min of specimen collection. The supernatant was similarly stored in 1 mL aliquots at −80 °C.

Peripheral blood samples were collected in EDTA. Plasma was separated by centrifugation at 1600×*g* for 10 min followed by centrifugation at 16,000×*g* for 10 min. Plasma was stored in 1 mL aliquots at −80 °C.

Oral swab samples were collected by swabbing the tongue for 15 s with rotation using a Copan regular flocked swab with molded breakpoint at 30 mm (Copan, Cat #520CS01). The swab head was then broken off into a collection tube containing 1× TE buffer, vortexed for 30 s, and stored at −80 °C.

DNA isolation and library preparation. DNA was extracted from plasma and urine using the QIAamp Circulating Nucleic Acid Kit according to the manufacturer's instructions (Qiagen, Cat #55114). Sequencing libraries were prepared using a single-stranded library preparation as described in Burnham et al.[13]. DNA was extracted from oral swab samples using the QIAamp UCP Pathogen Kit according to the manufacturer's instructions (Qiagen, Cat #50214) and libraries were prepared using the Nextera XT DNA Library Prep Kit (Illumina, Cat #FC-131-1024). All libraries were characterized using the AATI Fragment Analyzer before pooling and sequencing on the Illumina NextSeq 500 platform (paired-end, 2×75 bp).

Fragment length distribution and quantification of M. tuberculosis. Low-quality bases and Illumina-specific sequences were trimmed (Trimmomatic-0.32, LEADING:25 TRAILING:25 SLIDINGWINDOW:4:30 MINLEN:15)[14]. Reads were aligned (bwa mem[15]) to the human reference (UCSC hg19). Reads that did not align to the human genome reference were extracted and aligned to the bacteriophage phiX174.

To obtain the fragment length distribution for *M. tuberculosis* DNA, unaligned reads were extracted again and aligned to the circularized *M. tuberculosis H37Rv* genome (edited from NC_000962.3). Reads aligning with a minimum mapping quality of 30 were extracted, and the lengths computed as the absolute difference between the start and end coordinates. The fragment length density profile was computed using the *hist* function from the R Graphics Package.

To quantify the abundance of *M. tuberculosis*, reads that did not align to the human and phiX references were extracted and aligned to a curated list of bacterial and viral reference genomes using kallisto (--pseudobam)[16,17]. Reads aligning to the 16S/23S ribosomal RNA region were removed. Microbial abundance was quantified as Reads per Million reads (RPM):

$$RPM = \frac{Microbial\ Reads \times 10^6}{Total\ Reads}.$$

Quantification of insertion sequence and genome coverage. Reads that did not align to the human and phiX references were extracted and aligned (bwa mem[15]) to the circularized *M. tuberculosis H37Rv* genome (edited from NC_000962.3), to the IS6110 and IS1081 sequences retrieved from the GenBank repository for *M. tuberculosis* H37Rv (NC_000962.3), and to additional nontuberculous mycobacteria-specific insertion sequences (Table S9). Reads aligning with a minimum mapping quality of 30 were extracted, and the lengths computed as the absolute difference between the start and end coordinates. The coverage was calculated using the following equation:

$$coverage = \frac{\Sigma(N \cdot L)}{G \cdot c},$$

where N is the number of reads of length L, G is the sequence length, and c is the sequence copy number.

ROC analysis. Receiver operating characteristic analyses were performed using the *roc* function from the R package pROC[18]. For each biofluid, the abundance of *M. tuberculosis* DNA (in RPM) was compared between sputum positive tuberculosis and sputum negative tuberculosis samples, between tuberculosis and endemic cohorts, and between tuberculosis and non-endemic cohorts.

To evaluate the effects of using non-endemic and endemic samples on the diagnostic performance for tuberculosis, reads that aligned to *M. tuberculosis* were extracted from each sample. Sputum positive tuberculosis

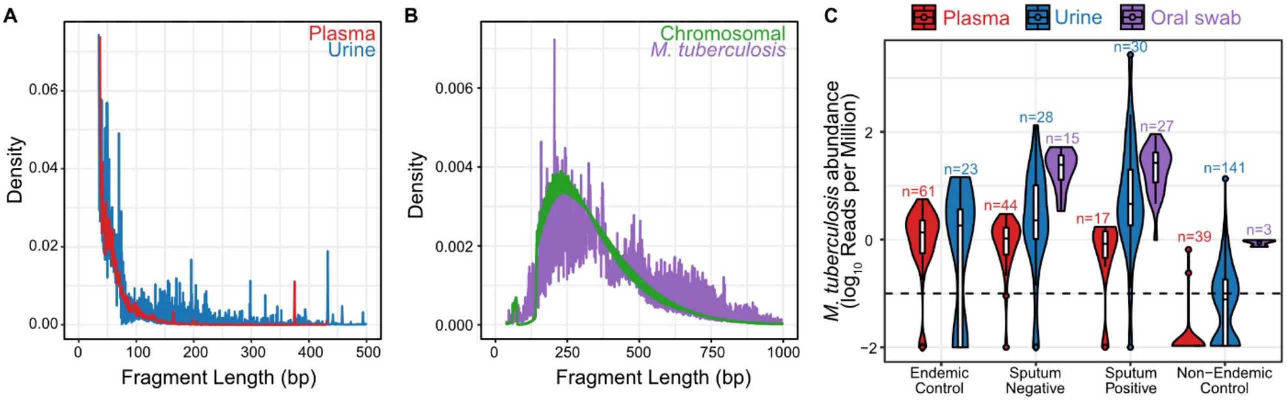

Figure 2. (**A**) The fragment length distributions of *M. tuberculosis* DNA in plasma (red) and urine (blue). (**B**) The fragment length distribution of *M. tuberculosis* DNA (purple) and host chromosomal DNA (green) in oral swabs are similar, suggesting that the microbial DNA is genomic and the fragmentation profile is not an intrinsic property but rather the consequence of sample preparation steps. (**C**) The abundance of *M. tuberculosis* DNA across all cohorts and biofluids. The majority of non-endemic samples have little to no detectable *M. tuberculosis* DNA, while the abundance of *M. tuberculosis* DNA in endemic and tuberculosis samples increases from plasma to oral swab. Dashed line indicates a limit of detection cutoff of 0.1 RPM.

samples were compared to a known mixture of sputum negative tuberculosis samples and either non-endemic or endemic samples. For each of 50 sampling rounds, 15 samples were randomly chosen for ROC analysis.

Simulating microbial DNA reads. Paired-end reads were simulated according to microbial cfDNA fragment length distribution from the *M. tuberculosis H37Rv* (NC_000962.3), *M. bovis BCG str. Pasteur 1173P2* genome (NC_008769.1), and *M. avium subsp. Hominissuis strain OCU464* (NZ_CP009360.4) genomes using a custom python script.

Statistical analysis and data availability. All statistical analyses were performed in R 3.5.0. Boxes in the boxplots indicate the 25th and 75th percentiles, the band in the box represents the median, and whiskers extend to 1.5 × interquartile range of the hinge. The sequence data for the non-endemic urine cohort was deposited in the database of Genotypes and Phenotypes (dbGaP, accession number phs001564v3.p1). The sequence data for the non-endemic plasma cohort was deposited in the Sequence Read Archive (accession number PRJNA263522). The sequence data generated in the scope of this study will be deposited in the Sequence Read Archive.

Ethics approval and consent to participate. The study was approved by the Johns Hopkins Bloomberg School of Public Health Institutional Review Board (protocol 00002185), the Cornell University Institutional Review Board (protocol 1612006853, 1910009101), the Foundation for Innovative New Diagnostics' Clinical Trials Review Committee and the Cornell Institutional Review Board (protocol 1612006851), the Makerere University School of Medicine Research and Ethics Committee (protocol 2017-020), the Weill Cornell Medicine Institutional Review Board (protocols 9402002786, 1207012730, 0710009490), the Stanford University Institutional Review Board (protocol 17666), and by the University of the Philippines Manila Research Ethics Board (protocol UPMREB 2018-252-01). All patients provided written informed consent.

Results

Biophysical properties of M. tuberculosis *DNA in urine, plasma, and oral swabs.*

Microbial cell-free DNA (cfDNA) in urine and plasma has been shown to be ultrashort, with an average fragment length of less than 100 bp (Fig. 2A)[12,13]. Compared to these biofluids, relatively few studies have examined the properties and diagnostic potential of microbial DNA from oral swabs. Oral swabs samples are an attractive new avenue for metagenomic DNA assays because they can be obtained noninvasively, are more cost effective than obtaining, storing, and shipping blood samples, and provide a high yield of DNA[19]. We found that in contrast to the microbial fragment length profiles of urinary and plasma cfDNA, DNA obtained from oral swabs is longer and likely genomic in origin: the fragmentation pattern closely mirrors that of human chromosomal DNA which arises from the tagmentation step in the library preparation protocol (Fig. 2B).

We set out to quantify the abundance of *M. tuberculosis* DNA detected in 161 samples obtained from patients presenting with symptoms of respiratory illness at tuberculosis clinics in the Philippines and Uganda (tuberculosis cohort), approximately half of which were sputum positive for tuberculosis (Table 1). Across all samples, we obtained an average of 31,831,284 reads (range: 1,908,898 to 155,693,161 reads; see Supplemental Information). We found that the abundance of *M. tuberculosis* DNA increases from plasma, to urine, to oral swabs (Fig. 2C). Within the tuberculosis cohort, there was no difference in *M. tuberculosis* DNA abundance between sputum positive and sputum negative tuberculosis samples. Setting a limit of detection of 0.1 reads per million reads (RPM), we find that within the tuberculosis cohort we detect *M. tuberculosis* DNA in 100% of the oral swab samples (42/42), 95% of the urine samples (55/58), and 93% of the plasma samples (57/61; Table S1). The median

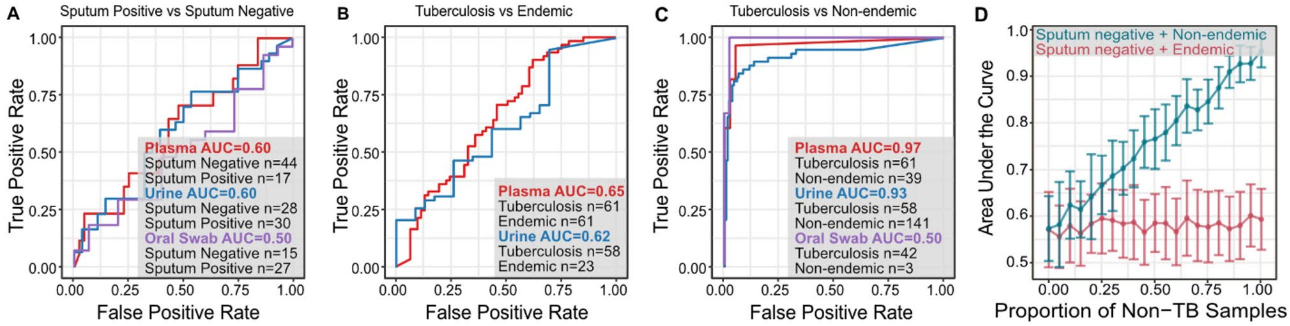

Figure 3. The performance of the metagenomic assay in discriminating (**A**) sputum positive versus sputum negative samples, (**B**) tuberculosis versus endemic cohorts, and (**C**) tuberculosis versus non-endemic cohorts (AUC = area under the curve). (**D**) Evaluating the effects of incorporating non-endemic samples (blue) and endemic samples (red) on the diagnostic performance via simulation show that the inclusion of non-endemic samples skews the assay's sensitivity.

abundance of *M. tuberculosis* DNA in the tuberculosis cohort was 0.512, 2.29, and 25.5 RPM for plasma, urine, and oral swabs, respectively. Further, we found an average of 27.9-fold more (adjusted p-value = 3.3×10^{-16}, Wilcoxon test) *M. tuberculosis* molecules in the oral swab samples than in plasma samples.

As a point of comparison, we quantified the abundance of *M. tuberculosis* in plasma and urine samples from two additional cohorts: 1) 141 urine samples and 39 plasma samples from patients living in the United States who received kidney or lung transplants, respectively[5,10,11] (non-endemic cohort); and, 2) 61 plasma samples and 23 urine samples from pediatric patients with suspected environmental enteropathy in Peru, a TB endemic region (endemic cohort). We found no *M. tuberculosis* DNA signatures in 37/39 plasma non-endemic samples and 89/141 urine non-endemic samples. In contrast, *M. tuberculosis* DNA was detected in 57/61 plasma samples and 16/23 urine samples from the endemic cohort. The median abundance of *M. tuberculosis* DNA in the endemic cohort was 1.26 and 1.85 RPM in plasma and urine samples, respectively. The higher abundance of *M. tuberculosis* DNA in the endemic cohort relative to the non-endemic cohort was expected given that geography, ethnicity, and other population-based factors have previously been shown to influence the microbiome composition among healthy individuals[20], and that the presence of an infectious organism does not necessarily equate to a disease state[21]. Given that the abundance of *M. tuberculosis* DNA across all biofluids in the endemic cohort was over 9.5-fold more than the non-endemic cohort (adjusted p-value = 2.1×10^{-23}, Wilcoxon test), but still significantly less than the tuberculosis cohort (14.15-fold less across all biofluids; adjusted p-value = 1.1×10^{-3}, Wilcox test).

Diagnostic potential of *M. tuberculosis* DNA is influenced by choice of controls. We evaluated the diagnostic performance of *M. tuberculosis* DNA in oral swabs, plasma, and urine. We found poor separation between sputum positive and sputum negative individuals across all biofluid types (area under the curve [AUC] = 0.6, 0.6, and 0.5 for plasma, urine, and oral swabs, respectively; Fig. 3A and Table S2). We found a modest increase in performance when comparing the tuberculosis and endemic cohorts (AUC = 0.62 and 0.65 for urine and plasma, respectively; Fig. 3B). In contrast, we found almost perfect separation for the tuberculosis and non-endemic cohorts (AUC = 0.93, 0.97, and 0.99 for urine, plasma, and oral swab, respectively; Fig. 3C). Moreover, we found that the diagnostic performance was not influenced by the choice of metagenomic classifier, reference database, or removal of confounding reads (see Table S3–S4).

To explore the effects of a biological background on the diagnostic performance, we randomly selected combinations of 15 urine samples composed of sputum positive and a known mixture of sputum negative and non-endemic controls. We performed 50 sampling rounds and found a positive correlation between the proportion of non-endemic samples and the diagnostic performance, with a mean AUC of 0.57 when the negative control consisted of only sputum negative samples and a mean AUC of 0.95 when the negative control was composed entirely of non-endemic samples (Fig. 3D). However, we saw no correlation when we performed the same analysis using a mixture of sputum negative tuberculosis and endemic controls: the area under the curve remained relatively constant, fluctuating between 0.56 and 0.60 across all negative control mixtures of sputum negative and endemic samples. This observation highlights a crucial but often overlooked criterium for metagenomic diagnostic test development: the choice of control population. Differences in diagnostic performance are highly influenced by the geographic origin of the samples. This is supported by the performance of published studies assaying nucleic acids in blood or urine for tuberculosis diagnosis: sputum culture positive and sputum culture negative samples are nearly indistinguishable unless the control cohort include samples from individuals living in TB non-endemic regions (Table S5).

Poor diagnostic performance can be attributed to a background of biological nontuberculous mycobacteria. We hypothesized that the poor diagnostic performance could be attributed to geographic factors: individuals living in TB endemic regions are exposed to nontuberculous mycobacteria, such as *M. avium*, *M. abscessus*, and *M. kansasii*[22]. This is further supported by observations that the abundance of *M. tuberculosis* DNA is positively correlated with age (Fig. S1). This exposure results in a nontuberculous mycobacteria

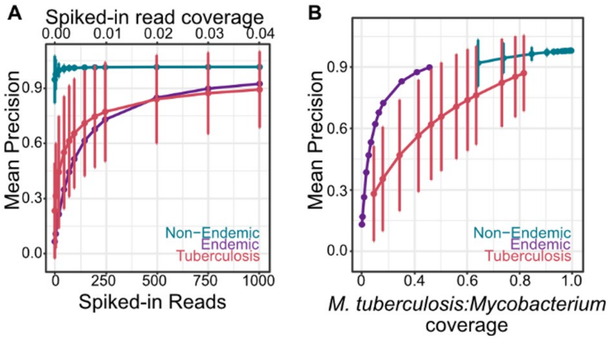

Figure 4. (**A**) Classification precision of synthetic *M. tuberculosis* reads spiked into non-endemic urine samples (n = 29, 5 replicates per sample), endemic urine samples (n = 23, 5 replicates per sample), and tuberculosis urine samples (n = 66, 5 replicates per sample). (**B**) The classification precision is correlated with the relative coverage of *M. tuberculosis* to total *Mycobacterium* in the sample.

background that is indistinguishable from a true disease signal due to the low abundance of *M. tuberculosis* and the high sequence similarity between *Mycobacterium* genomes, which can exceed 99% nucleotide similarity[23].

To determine whether the poor diagnostic performance could be attributed to a biological nontuberculous mycobacteria background, we simulated datasets by digitally spiking in *M. tuberculosis* reads into datasets generated for non-endemic, endemic, and tuberculosis urine samples. Across a range of 0 to 1000 spiked-in *M. tuberculosis* reads, we obtained nearly perfect classification of the synthetic reads in non-endemic samples (0.987 ± 0.0210; Fig. 4A). However, the endemic and tuberculosis samples exhibited logarithmic relationships between the number of spiked-in reads and precision. When we evaluated precision as a function of the relative coverage of spiked-in *M. tuberculosis* reads to all *Mycobacterium* reads in a sample, we found a strong correlation between the signal-to-noise ratio and the classification precision (Fig. 4B). Non-endemic samples had little to no background DNA from *Mycobacterium* species and exhibited high precision, while endemic and tuberculosis samples had a higher background of nontuberculous mycobacteria that reduced the classification precision. To further test our hypothesis that background nontuberculous mycobacterial DNA drives poor classification precision, we removed all *Mycobacterium* reads prior to spiking in *M. tuberculosis*, *M. bovis*, or *M. avium* reads and found perfect classification across all samples, regardless of the number of reads simulated (Tables S6–S8).

The availability of sputum PCR results for the oral swab samples provided further opportunity to evaluate the *M. tuberculosis* signal relative to the nontuberculous mycobacterial background. The Xpert® MTB/RIF Ultra assay (Cepheid, Sunnyvale, USA) targets three segments of the *M. tuberculosis* complex, two of which are the IS6110 and IS1081 insertion sequences. *M. tuberculosis* isolates contain between 0–25 copies of IS6110[24], whereas the IS1081 is present in all *M. tuberculosis* complex species at a stable number of 5–7 repeats per genome[25]. We evaluated the presence of IS1081 detected by metagenomic sequencing because the range in copy numbers across different *M. tuberculosis* complex species is narrower. IS1081 was detected by metagenomic sequencing in 5 of 27 sputum positive oral swab samples and was not detected in any of the 15 sputum negative oral swab samples, as expected. Because IS1081 is unique to *M. tuberculosis*, we were able to obtain a lower bound of the relative abundance of *M. tuberculosis* versus *Mycobacterium* by comparing the per-base sequence coverage of the IS1081 gene segment relative to the *M. tuberculosis* genome. Using the minimum copy number for the IS1081 gene (five copies per genome), we found that the coverage of nontuberculous mycobacteria relative to IS1081 was 210.397 (range of relative coverage: 23–394). Given that IS1081 was not detected in 22 of the 27 sputum positive oral swab samples and was not detected in any of the 15 sputum negative oral swab samples or in the 3 non-endemic oral swab samples, this range represents a lower bound. Thus, the burden of *M.* tuberculosis DNA represents 4.4% or less of the total abundance of *Mycobacterium* DNA, indicating a significant background of nontuberculous mycobacteria. Mapping to species-specific insertion sequences revealed that the background of *Mycobacterium* originates from a number of species, all of which are both ubiquitous environmental mycobacteria and implicated in nontuberculous mycobacterial lung infections (*M. branderi*, *M. smegmatis*, *M. avium*, *M. celatum*, *M. gordonae*, *M. xenopi*, *M. fortuitum*, *M. ulcerans*; Table S9)[22,26]. Further validation is required to determine if these species are co-infectious and influence disease outcome.

Discussion

We show that the utility of a minimally invasive metagenomic sequencing assay for pulmonary tuberculosis diagnostics is dependent on the geographic origin of control samples and limited by the low abundance of *M. tuberculosis* in extrapulmonary sampling sites. Such an assay is sensitive to the detection of nontuberculous mycobacteria that arises from a lifelong exposure to species from the *Mycobacterium* genus and contributes to the microbiome composition of samples originating from TB endemic regions. Our findings demonstrate that the influence of geography on the microbiome directly impacts the diagnostic performance: the inclusion of non-endemic samples in the control cohort invariably results in a near perfect test while poor diagnostic separation is obtained for geographically-controlled TB positive and TB negative individuals. Mathematical modeling demonstrates that the diagnostic potential is correlated with the abundance of *M. tuberculosis* DNA relative to the background of nontuberculous mycobacterial DNA. Quantitative comparisons to matched qPCR reveals that nontuberculous mycobacterial DNA is 23-fold or more abundant than the abundance *M. tuberculosis* DNA in the samples investigated here. The overwhelming biological background of *Mycobacterium* in samples of interest, in combination with the low abundance of *M. tuberculosis* in extrapulmonary sampling sites, presents a major barrier for the implementation of an unbiased metagenomic DNA sequencing assay for pulmonary tuberculosis diagnostics.

Detection of tuberculosis using a metagenomic sequencing assay for tuberculosis diagnostics is thus akin to looking for a needle in a haystack: *M. tuberculosis* DNA constituted less than 4.4% of the total abundance of *Mycobacterium* in samples from TB endemic regions included in this study. Our work suggests that the median abundance of *M. tuberculosis* is lower than 0.06 and 0.42 copies/mL in blood and urine, respectively, and lower than 284 genome copies/μg of DNA collected by oral swab, an estimate that is in agreement with previous reports quantifying the abundance of *M. tuberculosis* DNA through sequence-specific amplification[8,27]. Improvements to a metagenomic sequencing assay for tuberculosis diagnostics could be made by increasing the volume of input biofluid[28], choosing a sample preparation workflow with improved DNA extraction and short read amplification[12,27], or enriching for *M. tuberculosis*-specific sequences using ultrashort PCR amplicons[8,29]. These approaches would minimize noise from nontuberculous mycobacteria and increase the sequencing budget allocated to *M. tuberculosis*. Additionally, further exploration of the nontuberculous mycobacteria fraction may reveal patterns in disease outcome and provide new insights in the development of a robust geographic control. Together, our results reveal challenges and opportunities for the development of a DNA-based diagnostic tests for pulmonary tuberculosis and provides a comprehensive characterization of *M. tuberculosis* in extrapulmonary sites that can inform the development of molecular tests.

Data availability

The sequence data for the non-endemic urine cohort was deposited in the database of Genotypes and Phenotypes (dbGaP, accession number phs001564v3.p1). The sequence data for the non-endemic plasma cohort was deposited in the Sequence Read Archive (accession number PRJNA263522). The sequence data generated in the scope of this study will be deposited in the Sequence Read Archive.

Received: 3 June 2022; Accepted: 26 September 2022
Published online: 10 October 2022

References

1. Chakaya, J. *et al.* Global Tuberculosis Report 2020—Reflections on the Global TB burden, treatment and prevention efforts. *Int. J. Infect. Dis.* https://doi.org/10.1016/j.ijid.2021.02.107 (2021).
2. Moore, D. F., Guzman, J. A. & Mikhail, L. T. Reduction in turnaround time for laboratory diagnosis of pulmonary tuberculosis by routine use of a nucleic acid amplification test. *Diagn. Microbiol. Infect. Dis.* **52**, 247–254 (2005).
3. Ismail, N. A. *et al.* Optimizing mycobacterial culture in smear-negative, human immunodeficiency virus-infected tuberculosis cases. *PLoS ONE* **10**, e0141851 (2015).
4. Bowness, R. *et al.* The relationship between *Mycobacterium tuberculosis* MGIT time to positivity and cfu in sputum samples demonstrates changing bacterial phenotypes potentially reflecting the impact of chemotherapy on critical sub-populations. *J. Antimicrob. Chemother.* **70**, 448–455 (2015).
5. De Vlaminck, I. *et al.* Noninvasive monitoring of infection and rejection after lung transplantation. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 13336–13341 (2015).
6. Fernández-Carballo, B. L., Broger, T., Wyss, R., Banaei, N. & Denkinger, C. M. Toward the development of a circulating free DNA-based in vitro diagnostic test for infectious diseases: A review of evidence for tuberculosis. *J. Clin. Microbiol.* **57**, e01234-18 (2019).
7. Oreskovic, A. *et al.* Diagnosing pulmonary tuberculosis by using sequence-specific purification of urine cell-Free DNA. *J. Clin. Microbiol.* **59**, e00074-e121 (2021).
8. Oreskovic, A. *et al.* Characterizing the molecular composition and diagnostic potential of *Mycobacterium tuberculosis* urinary cell-free DNA using next-generation sequencing. *Int. J. Infect. Dis.* **112**, 330–337 (2021).
9. Denkinger, C. M. *et al.* Guidance for the evaluation of tuberculosis diagnostics that meet the World Health Organization (WHO) target product profiles: An introduction to WHO process and study design principles. *J. Infect. Dis.* **220**, S91–S98 (2019).
10. Burnham, P. *et al.* Urinary cell-free DNA is a versatile analyte for monitoring infections of the urinary tract. *Nat. Commun.* **9**, 2412 (2018).
11. De Vlaminck, I. *et al.* Temporal response of the human virome to immunosuppression and antiviral therapy. *Cell* **155**, 1178–1187 (2013).
12. Chang, A. *et al.* Measurement biases distort cell-free DNA fragmentation profiles and define the sensitivity of metagenomic cell-free DNA sequencing assays. *Clin. Chem.* https://doi.org/10.1093/clinchem/hvab142 (2021).
13. Burnham, P. *et al.* Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. *Sci. Rep.* **6**, 27859 (2016).
14. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
15. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760 (2009).
16. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
17. Schaeffer, L., Pimentel, H., Bray, N., Melsted, P. & Pachter, L. Pseudoalignment for metagenomic read assignment. *Bioinformatics* **33**, 2082–2088 (2017).
18. Robin, X. *et al.* pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* **12**, 77 (2011).
19. Daksis, J. I. & Erikson, G. H. Heteropolymeric triplex-based genomic assay® to detect pathogens or single-nucleotide polymorphisms in human genomic samples. *PLoS ONE* **2**, e305 (2007).
20. Gupta, V. K., Paul, S. & Dutta, C. Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. *Front. Microbiol.* **8**, 1162 (2017).
21. van Seventer, J. M. & Hochberg, N. S. Principles of infectious diseases: Transmission, diagnosis, prevention, and control. In *International Encyclopedia of Public Health,* 22–39 (2017). https://doi.org/10.1016/B978-0-12-803678-5.00516-6.
22. Adikaram, C. P. Overview of non tuberculosis mycobacterial lung diseases. In *Mycobacterium: Research and Development* (IntechOpen, 2018).

23. Garnier, T. *et al.* The complete genome sequence of *Mycobacterium bovis*. *Proc. Natl. Acad. Sci.* **100**, 7877–7882 (2003).
24. Tanaka, M. M., Rosenberg, N. A. & Small, P. M. The control of copy number of IS6110 in *Mycobacterium tuberculosis*. *Mol. Biol. Evol.* **21**, 2195–2201 (2004).
25. van Soolingen, D., Hermans, P. W., de Haas, P. E. & van Embden, J. D. Insertion element IS1081-associated restriction fragment length polymorphisms in *Mycobacterium tuberculosis* complex species: A reliable tool for recognizing *Mycobacterium bovis* BCG. *J. Clin. Microbiol.* **30**, 1772–1777 (1992).
26. Primm, T. P., Lucero, C. A. & Falkinham, J. O. Health impacts of environmental mycobacteria. *Clin. Microbiol. Rev.* **17**, 98–106 (2004).
27. Oreskovic, A. & Lutz, B. R. Ultrasensitive hybridization capture: Reliable detection of <1 copy/mL short cell-free DNA from large-volume urine samples. *PLoS ONE* **16**, e0247851 (2021).
28. Labugger, I. *et al.* Detection of transrenal DNA for the diagnosis of pulmonary tuberculosis and treatment monitoring. *Infection* **45**, 269–276 (2017).
29. Melkonyan, H. S. *et al.* Transrenal nucleic acids: From proof of principle to clinical tests. *Ann. N. Y. Acad. Sci.* **1137**, 73–81 (2008).

Acknowledgements

A special thanks to the late Dr. Anna Lena Lopez for her research supervision with the Institute of Child Health and Human Development at the University of the Philippines Manila's National Institutes of Health for samples collected and characterized in Manila, Philippines.

Author contributions

A. Chang, A.S., and I.D.V. contributed to the study design. A. Chang, O.M., L.D.K., J.L., and P.B. performed the experiments. A Chang, P.K., and I.D.V. analyzed the data. A.A., A. Cattamanchi, C.M.B., and J.C. facilitated data collection. A. Chang and I.D.V. wrote the manuscript. All authors provided comments and edits.

Funding

This work was supported by NIH Awards 1DP2AI138242 (to I.D.V.), R01AI146165 (to I.D.V.), 1R01AI151059 (to I.D.V.), a Synergy award from the Rainin Foundation (to I.D.V.), and a grant from the Bill and Melinda Gates Foundation INV-003145 (to I.D.V.). A.C. was supported by the National Institutes of Health under the Ruth L. Kirschstein National Research Service Award (6T32GM008267) from the National Institute of General Medical Sciences. P.B. was supported by the National Science Foundation under the Graduate Research Fellowships Program (DGE-1144153). The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Competing interests

I.D.V. has received research grants from the Bill and Melinda Gates Foundation, the National Institutes of Health, and the Rainin Foundation. P.B. has received research grants from the National Science Foundation. I.D.V. and P.B. are inventors on the patent US-2020-0048713-A1 titled "Methods of Detecting Cell-Free DNA in Biological Samples". The rights to the patent were licensed by Eurofins through Cornell University. I.D.V. is a Member of the Advisory Board and has stock in Karius Inc. A.C., O.M., L.D.K., J.L., P.K., A.A., J.C., C.M.B., A.C., and A.S. declare no potential conflict of interest.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-21244-x.

Correspondence and requests for materials should be addressed to I.D.V.