1 2	A needle in a haystack: metagenomic DNA sequencing to quantify <i>Mycobacterium tuberculosis</i> DNA and diagnose tuberculosis
3	Advisore Observ <sup>1</sup> Osserv Marus <sup>1</sup> Lis Audres Diservers Konsets <sup>1</sup> Jacob Leve <sup>1</sup> Dhilis
4 5	Adrienne Chang', Omary Mzava', Liz-Audrey Djomnang Kounatse', Joan Lenz', Philip Burnham <sup>1</sup> , Peter Kaplinsky <sup>1</sup> , Alfred Andama <sup>2</sup> , John Connelly <sup>3</sup> , Christine M. Bachman <sup>3</sup> , Adithya
6	Cattamanchi <sup>4</sup> , Amy Steadman <sup>3</sup> , Iwijn De Vlaminck <sup>1*</sup>
7	
8	Affiliations:
9	<sup>1</sup> Nancy E. and Peter C. Meinig School of Biomedical Engineering, Cornell University, Ithaca, New York,
10	USA
11	<sup>2</sup> Department of Internal Medicine, Makerere University College of Health Sciences, Kampala, Uganda
12 12	<sup>4</sup> Department of Medicine, University of California San Francisco, San Francisco, California, USA
13 14	Department of Medicine, Oniversity of California San Francisco, San Francisco, California, USA
15	*Corresponding author at: vlaminck@cornell.edu
16	
17	ABSTRACT
18	Background
19	Tuberculosis (TB) remains a significant cause of mortality worldwide. Metagenomic next-
20	generation sequencing has the potential to reveal biomarkers of active disease, identify
21	coinfection, and improve detection for sputum-scarce or culture-negative cases.
22	Methods
23	We conducted a large-scale comparative study of 427 plasma, urine, and oral swab samples
24	from 334 individuals from TB endemic and non-endemic regions to evaluate the utility of a
25	shotgun metagenomic DNA sequencing assay for tuberculosis diagnosis.
26	Findings
27	We found that the choice of a negative, non-TB control cohort had a strong impact on the
28	measured performance of the diagnostic test: the use of a control patient cohort from a
29	nonendemic region led to a test with nearly 100% specificity and sensitivity, whereas controls
30	from TB endemic regions exhibited a high background of nontuberculous mycobacterial DNA,
31	limiting the diagnostic performance of the test. Using mathematical modeling and quantitative
32	comparisons to matched qPCR data, we found that the burden of Mycobacterium tuberculosis
33	DNA constitutes a very small fraction (0.04 or less) of the total abundance of DNA originating
34	from mycobacteria in samples from TB endemic regions.
35	

#### 36 Interpretation

37 Our findings suggest that the utility of a metagenomic sequencing assay for tuberculosis 38 diagnostics is limited by the low burden of *M. tuberculosis* in extrapulmonary sites and an 39 overwhelming biological background of nontuberculous mycobacterial DNA.

#### 40 **Funding**

This work was supported by the National Institutes of Health, the Rainin Foundation, the
National Science Foundation, and the Bill and Melinda Gates Foundation.

- 43
- 44

## 45 Introduction

46 Despite the introduction of the World Health Organization's (WHO) End TB Strategy in 2015, 47 tuberculosis (TB) remains one of the top global causes of death due to a single infectious 48 pathogen. In 2021, the WHO reported the first rise in deaths due to tuberculosis in over a decade.<sup>1</sup> Recent advances in nucleic acid amplification tests, including the WHO-recommended 49 Xpert® MTB/RIF Ultra, have produced rapid tests with high positive predictive value<sup>2</sup>. However, 50 51 these assays are unable to replace the use of culture for bacteriological confirmation due to low 52 negative predictive value or implementation barriers present in low- and middle-income 53 countries that account for 98% of reported TB cases<sup>1</sup>. Though sputum culture has long been a aold standard for TB diagnostics, the time to positive detection can be affected by a number of 54 55 variables, including sputum volume, population characteristics (e.g. age, HIV coinfection), and method of specimen handling<sup>3,4</sup>. Variations due to these factors can lead to discrepancies 56 57 between quantitative and semiquantitative diagnostic assays. Thus, there is an unmet need for 58 robust, point-of-care tuberculosis diagnostics.

59

60 Metagenomics DNA sequencing has the capacity to detect all potential infectious pathogens in 61 a clinical sample using an unbiased approach. Numerous studies have demonstrated that DNA from Mycobacterium tuberculosis, the causative agent of tuberculosis, can be detected in 62 plasma, urine, and oral fluids<sup>5-7</sup>. However, its diagnostic performance in distinguishing positive 63 sputum culture from negative sputum culture samples has fallen short of the performance 64 standards set by the WHO (98% specificity, 80% sensitivity)<sup>8</sup>. We explored the utility of a 65 metagenomic sequencing assay for tuberculosis across three biofluids and four geographic 66 regions and found that diagnostic performance is highly influenced by the geographic origin of 67 68 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 69

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.



76

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

79

### 80 Results

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

82 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**)<sup>9,10</sup>. Compared to these biofluids, relatively 83 few studies have examined the properties and diagnostic potential of microbial DNA from oral 84 85 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, 86 and shipping blood samples, and provide a high yield of DNA<sup>11</sup>. We found that in contrast to the 87 88 microbial fragment length profiles of urinary and plasma cfDNA, DNA obtained from oral swabs 89 is longer and likely genomic in origin: the fragmentation pattern closely mirrors that of human 90 chromosomal DNA which arises from the tagmentation step in the library preparation protocol 91 (**Fig 2B**).

92

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**). We found that the abundance of *M. tuberculosis* DNA increases from

97 plasma, to urine, to oral swabs (Fig 2C). Within the tuberculosis cohort, there was no difference 98 in *M. tuberculosis* DNA abundance between sputum positive and sputum negative tuberculosis 99 samples. Setting a limit of detection of 0.1 molecule per million reads (MPM), we find that within 100 the tuberculosis cohort we detect M. tuberculosis DNA in 100% of the oral swab samples 101 (42/42), 95% of the urine samples (55/58), and 93% of the plasma samples (57/61). The median 102 abundance of *M. tuberculosis* DNA in the tuberculosis cohort was 0.512, 2.29, and 25.5 MPM 103 for plasma, urine, and oral swabs, respectively. Further, we found an average of 27.9-fold more 104 (adjusted p-value =  $3.3 \times 10^{-16}$ , Wilcoxon test) *M. tuberculosis* molecules in the oral swab 105 samples than in plasma samples.

106

107 As a point of comparison, we quantified the abundance of *M. tuberculosis* in plasma and urine 108 samples from two additional cohorts: 1) 141 urine samples and 40 plasma samples from patients living in the United States who received kidney or lung transplants, respectively<sup>12-</sup> 109 110 <sup>14</sup>(non-endemic cohort); and, 2) 62 plasma samples and 23 urine samples from pediatric 111 patients with suspected environmental enteropathy in Peru, a TB endemic region (endemic 112 cohort). We found no *M. tuberculosis* DNA signatures in 38/40 plasma non-endemic samples 113 and 89/141 urine non-endemic samples. In contrast, M. tuberculosis DNA was detected in 57/62 114 plasma samples and 16/23 urine samples from the endemic cohort. The median abundance of 115 M. tuberculosis DNA in the endemic cohort was 1.26 and 1.85 MPM in plasma and urine 116 samples, respectively. The higher abundance of *M. tuberculosis* DNA in the endemic cohort 117 relative to the non-endemic cohort was expected given that geography, ethnicity, and other 118 population-based factors have previously been shown to influence the microbiome composition 119 among healthy individuals<sup>15</sup>, and that the presence of an infectious organism does not 120 necessarily equate to a disease state<sup>16</sup>. Given that the abundance of *M. tuberculosis* DNA 121 across all biofluids in the endemic cohort was over 3.7-fold more than the non-endemic cohort (adjusted p-value =  $3.4 \times 10^{-14}$ , Wilcoxon test), but still significantly less than the tuberculosis 122 123 cohort (12.15-fold less across all biofluids; adjusted p-value =  $7 \times 10^{-4}$ , Wilcox test).

124

medRxiv preprint doi: https://doi.org/10.1101/2022.04.15.22273912; this version posted April 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.



126 Figure 2. A The fragment length distributions of *M. tuberculosis* DNA in plasma (red) and urine (blue). B 127 The fragment length distribution of *M. tuberculosis* DNA (purple) and host chromosomal DNA (green) in 128 oral swabs are similar, suggesting that the microbial DNA is genomic and the fragmentation profile is not 129 an intrinsic property but rather the consequence of sample preparation steps. C The abundance of M. 130 tuberculosis DNA across all cohorts and biofluids. The majority of non-endemic samples have little to no 131 detectable M. tuberculosis DNA, while the abundance of M. tuberculosis DNA in endemic and 132 tuberculosis samples increases from plasma to oral swab. Dashed line indicates a limit of detection cutoff 133 of 0.1 MPM.

134

125

#### 135 Diagnostic potential of M. tuberculosis DNA is influenced by choice of controls

136 We evaluated the diagnostic performance of *M. tuberculosis* DNA in oral swabs, plasma, and 137 urine. We found poor separation between sputum positive and sputum negative individuals 138 across all biofluid types (area under the curve [AUC] = 0.6, 0.6, and 0.5 for plasma, urine, and 139 oral swabs, respectively; Fig 3A). We found a modest increase in performance when comparing 140 the tuberculosis and endemic cohorts (AUC = 0.62 and 0.64 for urine and plasma, respectively; 141 Fig 3B). In contrast, we found almost perfect separation for the tuberculosis and non-endemic 142 cohorts (AUC = 0.93 and 0.97 for urine and plasma, respectively; Fig 3C). Moreover, we found 143 that the diagnostic performance was not influenced by the choice of metagenomic classifier, 144 reference database, or removal of confounding reads (see Table S1-S2, and Supplemental 145 Information).

146

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 153 non-endemic samples (Fig 3D). However, we saw no correlation when we performed the same 154 analysis using a mixture of sputum negative tuberculosis and endemic controls: the area under 155 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 156 157 but often overlooked criterium for metagenomic diagnostic test development: the choice of 158 control. Differences in diagnostic performance are highly influenced by the geographic origin of 159 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 160 161 samples are nearly indistinguishable unless the control cohort include samples from individuals 162 living in TB non-endemic regions (Table S3).





164

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.

170

171 Poor diagnostic performance can be attributed to a background of biological nontuberculous172 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*<sup>17</sup>. 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 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<sup>18</sup>.

180

181 To determine whether the poor diagnostic performance could be attributed to a biological 182 nontuberculous mycobacteria background, we simulated datasets by digitally spiking in M. 183 tuberculosis reads into datasets generated for non-endemic, endemic, and tuberculosis urine 184 samples. Across a range of 0 to 1000 spiked-in *M. tuberculosis* reads, we obtained nearly 185 perfect classification of the synthetic reads in non-endemic samples (0.987 ± 0.0210; Fig 4A). 186 However, the endemic and tuberculosis samples exhibited logarithmic relationships between the 187 number of spiked-in reads and precision. When we evaluated precision as a function of the 188 relative coverage of spiked-in *M. tuberculosis* reads to all *Mycobacterium* reads in a sample, we 189 found a strong correlation between the signal-to-noise ratio and the classification precision (Fig 190 4B). Non-endemic samples had little to no background DNA from Mycobacterium species and 191 exhibited high precision, while endemic and tuberculosis samples had a higher background of 192 nontuberculous mycobacteria that reduced the classification precision. To further test our hypothesis that background nontuberculous mycobacterial DNA drives poor classification 193 194 precision, we removed all Mycobacterium reads prior to spiking in M. tuberculosis, M. bovis, or 195 *M. avium* reads and found perfect classification across all samples, regardless of the number of 196 reads simulated (Tables S4-S6).

197



198

**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.

203

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<sup>19</sup>, whereas the IS1081 is present in

all *M. tuberculosis* complex species at a stable number of 5-7 repeats per genome<sup>20</sup>. We 209 210 evaluated the presence of IS1081 detected by metagenomic sequencing because the range in 211 copy numbers across different *M. tuberculosis* complex species is narrower. IS1081 was 212 detected by metagenomic sequencing in 5 of 27 sputum positive oral swab samples and was 213 not detected in any of the 15 sputum negative oral swab samples, as expected. Because 214 IS1081 is unique to *M. tuberculosis*, we were able to obtain a lower bound of the relative 215 abundance of *M. tuberculosis* versus *Mycobacterium* by comparing the per-base sequence 216 coverage of the IS1081 gene segment relative to the *M. tuberculosis* genome. Using the 217 minimum copy number for the IS1081 gene (five copies per genome), we found that the 218 coverage of nontuberculous mycobacteria relative to IS1081 was 210.397 (range of relative 219 coverage: 23-394). Given that IS1081 was not detected in 22 of the 27 sputum positive oral 220 swab samples and was not detected in any of the 15 sputum negative oral swab samples, this 221 range represents a lower bound. Thus, the burden of *M*. tuberculosis DNA represents 4.4% or 222 less of the total abundance of Mycobacterium DNA, indicating a significant background of 223 nontuberculous mycobacteria. Using species-specific insertion sequences revealed that the 224 background of Mycobacterium originates from a number of species, all of which are both 225 ubiquitous environmental mycobacteria and implicated in nontuberculous mycobacterial lung 226 infections (M. branderi, M. smegmatis, M. avium, M. celatum, M. gordonae, M. xenopi, M. fortuitum, *M. ulcerans*; **Table S7**)<sup>17,21</sup>. Further validation is required to determine if these species 227 228 are co-infectious and influence disease outcome.

229

#### 230 Discussion

231 We show that the utility of a metagenomic sequencing assay for tuberculosis diagnostics is 232 dependent on the geographic origin of control samples and limited by the low abundance of M. 233 tuberculosis in extrapulmonary sites. Such an assay is sensitive to the detection of 234 nontuberculous mycobacteria that arises from a lifelong exposure to species from the 235 Mycobacterium genus and contributes to the microbiome composition of samples originating 236 from TB endemic regions. Our findings demonstrate that the influence of geography on the 237 microbiome directly impacts the diagnostic performance: the inclusion of non-endemic samples 238 as controls invariably results in a near perfect test while poor diagnostic separation is obtained 239 when endemic samples are employed as controls. Mathematical modeling demonstrates that 240 the diagnostic potential is correlated with the abundance of *M. tuberculosis* DNA relative to the 241 background of nontuberculous mycobacterial DNA. Quantitative comparisons to matched 242 gPCR 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 sites, presents a major barrier for the implementation of an
 unbiased metagenomic DNA sequencing assay for tuberculosis diagnostics.

247

248 Detection of tuberculosis using a metagenomic sequencing assay for tuberculosis diagnostics is 249 thus akin to looking for a needle in a haystack: *M. tuberculosis* DNA constituted less than 4.4% 250 of the total abundance of *Mycobacterium* in samples from TB endemic regions included in this 251 study. Our work suggests that the median abundance of M. tuberculosis is lower than 0.06 and 252 0.42 copies/mL in blood and urine, respectively, and lower than 284 genome copies/ug of DNA 253 collected by oral swab, an estimate that is in agreement with previous reports quantifying the abundance of *M. tuberculosis* DNA through sequence-specific amplification<sup>7,22</sup>. Improvements to 254 255 a metagenomic sequencing assay for tuberculosis diagnostics could be made by increasing the volume of input biofluid<sup>23</sup>, choosing a sample preparation workflow with improved DNA 256 extraction and short read amplification<sup>10,22</sup>, or enriching for *M. tuberculosis*-specific sequences 257 using ultrashort PCR amplicons<sup>7,24</sup>. These approaches would minimize noise from 258 259 nontuberculous mycobacteria and increase the sequencing budget allocated to *M. tuberculosis*. 260 Additionally, further exploration of the nontuberculous mycobacteria fraction may reveal patterns 261 in disease outcome and provide new insights in the development of a robust geographic control. 262 Together, our results reveal challenges and opportunities for the development of a DNA-based 263 diagnostic tests for tuberculosis and provides a comprehensive characterization of M. 264 tuberculosis in extrapulmonary sites that can inform the development of molecular tests.

265

## 266 Methods

# 267 Study cohorts

268 A total of 427 datasets from plasma, urine, and oral swab samples were analyzed, 188 of which 269 were generated for this study (Table 2). Endemic plasma and urine samples were collected from 270 patients seeking treatment for environmental enteropathy in Peru. The study was approved by 271 the Johns Hopkins Bloomberg School of Public Health Institutional Review Board (protocol 272 00002185) and the Cornell University Institutional Review Board (protocol 1612006853). 273 Tuberculosis plasma samples were collected from patients presenting with respiratory 274 symptoms from tuberculosis clinics in Peru. The study was approved by the Foundation for 275 Innovative New Diagnostics' Clinical Trials Review Committee and the Cornell Institutional 276 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).

280

281 Additional datasets collected in the scope of previous studies were included as follows. Non-282 endemic urine samples were collected from kidney transplant patients who received care at 283 New York Presbyterian Hospital-Weill Cornell Medical Center. The study was approved by the 284 Weill Cornell Medicine Institutional Review Board (protocols 9402002786, 1207012730, 0710009490)<sup>12</sup>. Non-endemic plasma samples were collected from lung transplant patients who 285 received care at Stanford University Hospital. The study was approved by the Stanford 286 University Institutional Review Board (protocol 17666)<sup>13,14</sup>. Tuberculosis urine samples from 287 288 patients seeking treatment for tuberculosis in the Philippines were collected through a study 289 partly funded by the Department of Science and Technology - Philippine Council for Health 290 Research and Development (DOST-PCHRD). This study was approved by the University of the Philippines Manila Research Ethics Board (protocol UPMREB 2018-252-01)<sup>10</sup>. All patients 291 292 provided written informed consent.

293

294 Table 1. Overview of datasets included in this study. All data was generated for this study unless 295 otherwise indicated.

Cohort	Origin	Disease	Biofluid	Patients	Samples
Endemic	Peru	Environmental	Plasma	62	62
		Cheropathy	Urine	23	23
Non-endemic	USA	Kidney transplant <sup>12</sup>	Urine	82	141
		Lung transplant <sup>13,14</sup>	Plasma	6	40
Tuberculosis	Philippines <sup>10</sup>	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

296 297

# 298 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 3,000 x g on the same day for 30 minutes 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 3,000 x g for 30 minutes at ambient temperature within 30 minutes of specimen collection. The supernatant was similarly stored in 1 mL aliquots at -80°C.

306

307 Peripheral blood samples were collected in EDTA. Plasma was separated by centrifugation at 308  $1,600 \times g$  for 10 minutes followed by centrifugation at 16,000  $\times g$  for 10 minutes. Plasma was 309 stored in 1 mL aliquots at -80°C.

310

Oral swab samples were collected by swabbing the tongue for 15 seconds 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 1X TE buffer, vortexed for 30 seconds, and stored at -80°C.

315

# 316 DNA isolation and library preparation

317 DNA was extracted from plasma and urine using the QIAamp Circulating Nucleic Acid Kit 318 according to the manufacturer's instructions (Qiagen, Cat #55114). Sequencing libraries were 319 prepared using a single-stranded library preparation as described in Burnham et al<sup>9</sup>. DNA was 320 extracted from oral swab samples using the QIAamp UCP Pathogen Kit according to the 321 manufacturer's instructions (Qiagen, Cat #50214) and libraries were prepared using the Nextera 322 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 500platform (paired-end, 2x75 bp).

325

# 326 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)<sup>25</sup>. Reads were aligned (BWAmem<sup>26</sup>) to the human reference (UCSC hg19). Reads that did not align to the human genome reference were extracted and aligned to the bacteriophage phiX174.

331

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.

338

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)<sup>27,28</sup>. Reads aligning to the 16S/23S ribosomal RNA region were removed. Microbial abundance was quantified as Molecules per Million reads (MPM):

343

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

344 345

346 Quantification of insertion sequence and genome coverage

Reads that did not align to the human and phiX references were extracted and aligned (BWmem<sup>26</sup>) 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 S7**). 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:

354 
$$coverage = \frac{\Sigma(N \cdot L)}{G \cdot c}$$

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

- 357
- 358 ROC analysis

Receiver operating characteristic analyses were performed using the *roc* function the R package pROC<sup>29</sup>. For each biofluid, the abundance of *M. tuberculosis* DNA (in MPM) was compared between sputum positive tuberculosis and sputum negative tuberculosis samples, between tuberculosis and endemic cohorts, and between tuberculosis and non-endemic cohorts.

364

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 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.

370

# 371 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.

376

#### 377 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 x 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.

385

# 386 Acknowledgments

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

389 from the Bill and Melinda Gates Foundation INV-003145 (to I.D.V.). A.C. was supported by the 390 National Institutes of Health under the Ruth L. Kirschstein National Research Service Award 391 (6T32GM008267) from the National Institute of General Medical Sciences. P.B. was supported 392 by the National Science Foundation under the Graduate Research Fellowships Program (DGE-393 1144153). A special thanks to the late Dr. Anna Lena Lopez for her research supervision with 394 the Institute of Child Health and Human Development at the University of the Philippines 395 Manila's National Institutes of Health for samples collected and characterized in Manila, 396 Philippines.

397

## 398 Author's 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.

403

# 404 **Conflicts of interest**

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.

411

# 412 Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

# 415 References

- 416 1. Chakaya, J. et al. Global Tuberculosis Report 2020 Reflections on the Global TB burden,
- 417 treatment and prevention efforts. *Int. J. Infect. Dis.* S1201971221001934 (2021)
- 418 doi:10.1016/j.ijid.2021.02.107.

419 2. Moore, D. F., Guzman, J. A. & Mikhail, L. T. Reduction in turnaround time for laboratory

- 420 diagnosis of pulmonary tuberculosis by routine use of a nucleic acid amplification test.
- 421 Diagn. Microbiol. Infect. Dis. **52**, 247–254 (2005).
- 422 3. Ismail, N. A. et al. Optimizing Mycobacterial Culture in Smear-Negative, Human
- 423 Immunodeficiency Virus-Infected Tuberculosis Cases. *PLOS ONE* **10**, e0141851 (2015).
- 424 4. Bowness, R. *et al.* The relationship between Mycobacterium tuberculosis MGIT time to
- 425 positivity and cfu in sputum samples demonstrates changing bacterial phenotypes
- 426 potentially reflecting the impact of chemotherapy on critical sub-populations. *J. Antimicrob.*
- 427 *Chemother.* **70**, 448–455 (2015).
- 428 5. Fernández-Carballo, B. L., Broger, T., Wyss, R., Banaei, N. & Denkinger, C. M. Toward the
- 429 Development of a Circulating Free DNA-Based In Vitro Diagnostic Test for Infectious
- 430 Diseases: a Review of Evidence for Tuberculosis. *J. Clin. Microbiol.* **57**, e01234-18.
- 431 6. Oreskovic, A. *et al.* Diagnosing Pulmonary Tuberculosis by Using Sequence-Specific
  432 Purification of Urine Cell-Free DNA. *J. Clin. Microbiol.* **59**, e00074-21.
- 433 7. Oreskovic, A. *et al.* Characterizing the molecular composition and diagnostic potential of
- 434 Mycobacterium tuberculosis urinary cell-free DNA using next-generation sequencing. *Int. J.*
- 435 Infect. Dis. IJID Off. Publ. Int. Soc. Infect. Dis. **112**, 330–337 (2021).
- 436 8. Denkinger, C. M. *et al.* Guidance for the Evaluation of Tuberculosis Diagnostics That Meet
- 437 the World Health Organization (WHO) Target Product Profiles: An Introduction to WHO
- 438 Process and Study Design Principles. J. Infect. Dis. 220, S91–S98 (2019).
- 439 9. Burnham, P. *et al.* Single-stranded DNA library preparation uncovers the origin and diversity
- 440 of ultrashort cell-free DNA in plasma. *Sci. Rep.* **6**, 27859 (2016).

- 441 10. Chang, A. et al. Measurement Biases Distort Cell-Free DNA Fragmentation Profiles and
- 442 Define the Sensitivity of Metagenomic Cell-Free DNA Sequencing Assays. *Clin. Chem.*

443 (2021) doi:10.1093/clinchem/hvab142.

- 11. Daksis, J. I. & Erikson, G. H. Heteropolymeric Triplex-Based Genomic Assay® to Detect
- 445 Pathogens or Single-Nucleotide Polymorphisms in Human Genomic Samples. *PLOS ONE*

446 **2**, e305 (2007).

- 447 12. Burnham, P. *et al.* Urinary cell-free DNA is a versatile analyte for monitoring infections of the
  448 urinary tract. *Nat. Commun.* 9, 2412 (2018).
- 13. De Vlaminck, I. et al. Noninvasive monitoring of infection and rejection after lung
- 450 transplantation. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 13336–13341 (2015).
- 451 14. De Vlaminck, I. *et al.* Temporal response of the human virome to immunosuppression and
  452 antiviral therapy. *Cell* **155**, 1178–1187 (2013).
- 453 15. Gupta, V. K., Paul, S. & Dutta, C. Geography, Ethnicity or Subsistence-Specific Variations in
  454 Human Microbiome Composition and Diversity. *Front. Microbiol.* 8, 1162 (2017).
- 455 16. van Seventer, J. M. & Hochberg, N. S. Principles of Infectious Diseases: Transmission,
- 456 Diagnosis, Prevention, and Control. *Int. Encycl. Public Health* 22–39 (2017)
- 457 doi:10.1016/B978-0-12-803678-5.00516-6.
- 458 17. Adikaram, C. P. Overview of Non Tuberculosis Mycobacterial Lung Diseases. in
- 459 *Mycobacterium: Research and Development* (IntechOpen, 2018).
- 460 18. Garnier, T. et al. The complete genome sequence of Mycobacterium bovis. Proc. Natl.
- 461 *Acad. Sci.* **100**, 7877–7882 (2003).
- 462 19. Tanaka, M. M., Rosenberg, N. A. & Small, P. M. The control of copy number of IS6110 in
  463 Mycobacterium tuberculosis. *Mol. Biol. Evol.* 21, 2195–2201 (2004).
- 464 20. van Soolingen, D., Hermans, P. W., de Haas, P. E. & van Embden, J. D. Insertion element
- 465 IS1081-associated restriction fragment length polymorphisms in Mycobacterium

- 466 tuberculosis complex species: a reliable tool for recognizing Mycobacterium bovis BCG. J.
- 467 *Clin. Microbiol.* **30**, 1772–1777 (1992).
- 468 21. Primm, T. P., Lucero, C. A. & Falkinham, J. O. Health Impacts of Environmental
- 469 Mycobacteria. *Clin. Microbiol. Rev.* **17**, 98–106 (2004).
- 470 22. Oreskovic, A. & Lutz, B. R. Ultrasensitive hybridization capture: Reliable detection of <1
- 471 copy/mL short cell-free DNA from large-volume urine samples. *PLOS ONE* **16**, e0247851
- 472 (2021).
- 473 23. Labugger, I. et al. Detection of transrenal DNA for the diagnosis of pulmonary tuberculosis
- 474 and treatment monitoring. *Infection* **45**, 269–276 (2017).
- 475 24. Melkonyan, H. S. *et al.* Transrenal Nucleic Acids: From Proof of Principle to Clinical Tests.
- 476 Ann. N. Y. Acad. Sci. **1137**, 73–81 (2008).
- 477 25. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence
  478 data. *Bioinformatics* **30**, 2114–2120 (2014).
- 479 26. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform.
- 480 Bioinforma. Oxf. Engl. 25, 1754–1760 (2009).
- 481 27. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq
  482 quantification. *Nat. Biotechnol.* 34, 525–527 (2016).
- 28. Schaeffer, L., Pimentel, H., Bray, N., Melsted, P. & Pachter, L. Pseudoalignment for
  metagenomic read assignment. *Bioinformatics* 33, 2082–2088 (2017).
- 485 29. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC
- 486 curves. *BMC Bioinformatics* **12**, 77 (2011).

487

#### USA Non-endemic contro Plasma n=40 Urine n=141

Peru Endemic control Plasma n=62 Urine n=23 Tuberculosis Plasma n=61 Uganda Tuberculosis Oral swab n=42 Tuberculosis Urine n=58

















