Online Supplementary Information

Ethnicity and mycobacterial lineage as determinants of tuberculosis disease phenotype

Manish Pareek PhD MRCP *Clinical Lecturer*[a,b,c]

Jason Evans MSc PhD *Clinical Scientist*[d]

John Innes FRCP *Honorary Consultant*[e]

E Grace Smith FRCP FRCPath *Consultant*[d]

Suzie Hingley-Wilson PhD *Research Associate*[a]

Kathryn E Lougheed PhD *Research Associate*[f]

Saranya Sridhar DPhil *Research Associate*[a]

Martin Dedicoat PhD MRCP *Consultant*[e]

Peter Hawkey PhD *Professor*[d,g]

Ajit Lalvani DM FRCP *Chair of Infectious Diseases and Director*[a]

[a]Tuberculosis Research Unit, Department of Respiratory Medicine, National Heart and Lung Institute, Imperial College London, UK
[b]Department of Infectious Disease Epidemiology, Imperial College London, UK
[c]Department of Infection, Immunity and Inflammation, University of Leicester, UK
[d]Health Protection Agency Midlands Laboratory, Heart of England NHS Foundation Trust, Bordesley Green East, Birmingham, UK
[e]Department of Infection and Tropical Medicine, Heart of England NHS Foundation Trust, Bordesley Green East, Birmingham, UK
[f]Department of Integrative Cell Biology, Division of Cell & Molecular Biology, Imperial College London, UK
[g]School of Immunity and Infection, The Medical School, University of Birmingham, Edgbaston, B15 2TT

Number of supplementary tables: 1 **Number of supplementary figures:** 3

Supplementary data analysis

Demographic characteristics of the cohort were described using median and interquartile range (IQR) (for continuous variables) and proportions/percentages (categorical variables); comparisons were made using the non-parametric Mann-Whitney U-test and Pearson's chi-square test (or Fishers exact test if appropriate) respectively.

We initially described the temporal dynamics and distribution of the global lineages in our cohort separately by the individuals' country of origin, region of origin (subdivided into Americas and Caribbean, Europe, Middle East/North Africa, South Africa, West and Central Africa, East Africa, Indian Subcontinent and Central Asia, Southeast and East Asia) and ethnic group. Comparisons of which lineages caused disease in UK- and foreign-born individuals was made using Fisher's exact test.

Factors associated with *Mtb* lineage were assessed by constructing separate regression models for each of the main lineages in our study - East Asian, Euro-America, East African Indian and Indo-Oceanic. We did not undertake regression analyses for two lineages (West African-1 and West African-2) due to the small number of cases caused by these lineages in our dataset. Univariate association of clinical and demographic factors individually associated with each lineage was assessed using logistic regression and reported as crude odds ratios (OR) and 95% confidence intervals (CI). To calculate adjusted odds ratios (and 95% CI) we mutually adjusted for year of notification, age, gender, ethnicity, country of birth, HIV status, previous history of TB, drug sensitivity and smear positivity.

For the next stage of the analysis we described the broad clinical phenotypes (pulmonary only, extrapulmonary only and concurrent pulmonary-extrapulmonary) associated with each of the lineages. We then compared, for each lineage, the proportion of cases that were pulmonary only as compared to extrapulmonary only, extrathoracic only and lymph node only respectively; comparisons were made using univariate logistic regression.

We then assessed the factors (especially *Mtb* lineage) associated with extrathoracic only, extrapulmonary only and lymph node only disease respectively versus exclusively pulmonary disease with three separate logistic regression models (for each clinical phenotype). Univariate, and multivariate, associations of factors individually associated with each of the three clinical phenotypes detailed above (compared against pulmonary cases only) was reported as crude, and adjusted, odds ratios (OR) respectively (and 95% confidence intervals). Adjusted odds ratios (and 95% CI) were obtained by mutually adjusting for the following factors in all three models: year of notification, age, gender, ethnicity, country of birth, time since arrival, HIV status, previous history of TB, drug sensitivity and lineage. In the descriptive analysis we categorised individuals as having fully sensitive disease, mono-resistant disease (in other words resistance to one/more chemotherapeutic agent but not multi-drug resistant) and multi-drug resistant disease (in other words resistance to rifampicin and isoniazid). For the logistic modelling due to small numbers of drug resistant TB we combined mono-resistant TB and multi-drug resistant TB into one category "drug-resistant". In the regression models, we concentrated on the four main lineages present in our cohort. We used the East Asian lineage as the reference category as this strain has been previously shown to be associated with extrapulmonary disease although,[14-16] as a sensitivity analysis, we reran the multivariate model using the Euro-American lineage (which was the commonest) as the reference lineage against which the other lineages were compared.

Analyses used STATA 9.2 (StataCorp, College Station, TX). All tests were two tailed; p-value of ≤ 0.05 was considered significant.

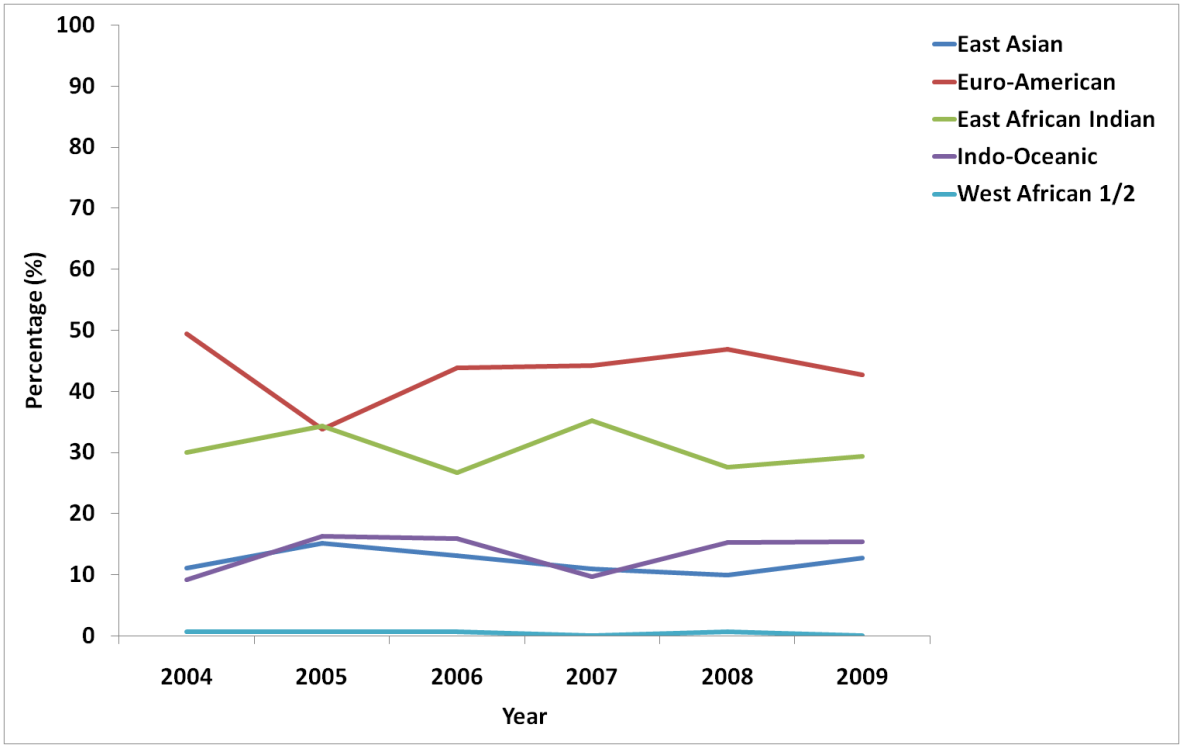
Supplementary tables

Supplementary table 1. Distribution of *M. tuberculosis* lineages by different demographic groups (stratified by ethnicity and location of birth) (percentages reflect the percentage of each lineage in each of the demographic groups – ie. Should be read across for each row)

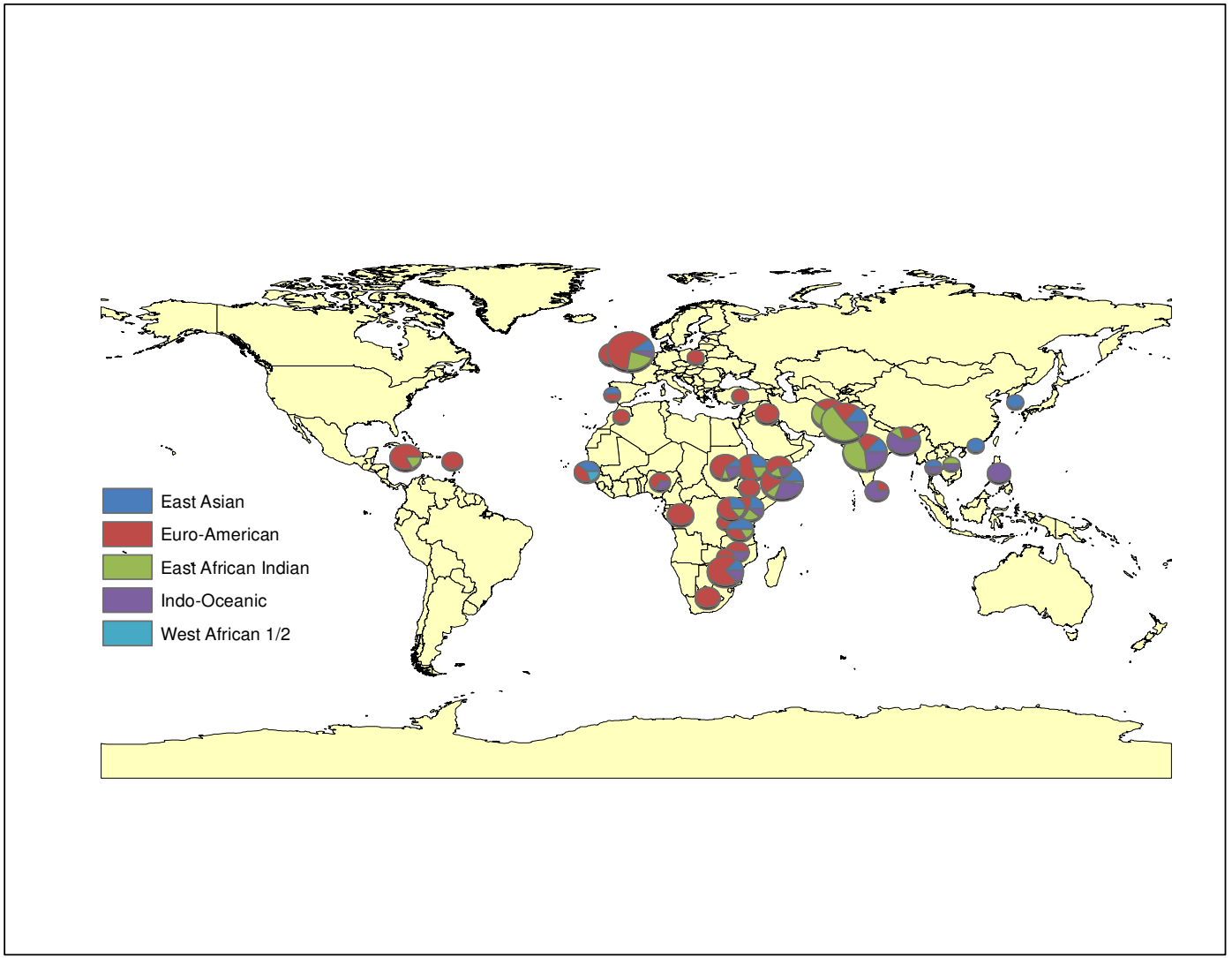
Location of birth/ethnicity	Lineage				
	East Asian (n=130)	Euro-American (n=468)	East African Indian (n=323)	Indo-Oceanic (n=145)	West African 1/2 (n=4)
Total White	**16 (10.8)**	**110 (74.3)**	**21 (14.2)**	**1 (0.7)**	**0 (0.0)**
UK born White	14 (10.2)	102 (74.5)	20 (14.6)	1 (0.7)	0 (0.0)
Foreign-born White	2 (18.2)	8 (72.7)	1 (9.1)	0 (0)	0 (0.0)
Total Indian Subcontinent	**69 (12.0)**	**151 (26.2)**	**262 (45.4)**	**94 (16.3)**	**1 (0.2)**
UK born Indian Subcontinent	8 (6.8)	50 (42.7)	45 (38.5)	13 (11.1)	1 (0.9)
Foreign-born Indian Subcontinent	61 (13.3)	101 (22.0)	217 (47.2)	81 (17.6)	0 (0.0)
Total Afro-Caribbean	**35 (12.6)**	**180 (64.8)**	**21 (7.6)**	**39 (14.0)**	**3 (1.1)**
UK born Afro-Caribbean	6 (11.1)	43 (79.6)	5 (9.3)	0 (0.0)	0 (0.0)
Foreign-born Afro-Caribbean	29 (13.0)	137 (61.2)	16 (7.1)	39 (17.4)	3 (1.4)
Total Oriental/Other Asia	**7 (30.4)**	**4 (17.4)**	**6 (26.1)**	**6 (26.1)**	**0 (0.0)**
UK born Oriental/Other Asia	1 (33.3)	1 (33.3)	0 (0.0)	1 (33.3)	0 (0.0)
Foreign-born Oriental/Other Asia	6 (30.0)	3 (15.0)	6 (30.0)	5 (25.0)	0 (0.0)
Total Other	**3 (6.8)**	**23 (52.3)**	**13 (29.6)**	**5 (11.4)**	**0 (0.0)**
UK born Other	1 (14.3)	4 (57.1)	2 (28.6)	0 (0.0)	0 (0.0)
Foreign-born Other	2 (5.4)	19 (51.4)	11 (29.7)	5 (13.5)	0 (0.0)

Supplementary figures

Supplementary figure 1. Temporal trends in lineages causing active tuberculosis between 2004 and 2009



Supplementary figure 2. Distribution of *M. tuberculosis* lineages by country of origin amongst the individuals in the cohort (size of pie charts is scaled relative to the number of cases from that country and the proportion of cases from each lineage) for all lineages


East Asian
Euro-American
East African Indian
Indo-Oceanic
West African 1/2

Supplementary figure 3. Distribution of mycobacterial lineages by time since arrival for cases amongst the foreign-born