Supplemental Information for:

Lineage specific histories of *Mycobacterium tuberculosis* dispersal in Africa and Eurasia

Mary B O'Neill[a,b,*], Abigail Shockey[b], Alex Zarley[c], William Aylward[d], Vegard Eldholm[e], Andrew Kitchen[f], Caitlin S Pepperell[b,g]

[a]Laboratory of Genetics, University of Wisconsin-Madison, Madison, WI 53706, USA
[b]Department of Medical Microbiology and Immunology, University of Wisconsin-Madison, Madison, WI 53706, USA
[c]Department of Geography, University of Wisconsin-Madison, WI 53706, USA
[d]Department of Classical and Ancient Near Eastern Studies, University of Wisconsin-Madison, Madison, WI 53706, USA
[e]Infection Control and Environmental Health, Norwegian Institute of Public Health, 0456 Oslo, Norway
[f]Department of Anthropology, University of Iowa, Iowa City, IA 52242, USA
[g]Department of Medicine, University of Wisconsin-Madison, Madison, WI 53706, USA
*Present address: Unit of Human Evolutionary Genetics, Institut Pasteur, 75015 Paris, France

Table of Contents

Appendix S1. Supplementary note about the relationship between genetic diversity and geographic distance from Addis Ababa.

Our finding of no significant decline in *M.tb* diversity as a function of distance from Addis Ababa conflicts with a previously published report (Comas et al. 2015). To ensure our differing results were not driven by a lack of samples from the Americas, we repeated the analysis including the same samples from the Americas used in Comas *et al*. (2015), and obtained results that were similar to the Old World collection alone (adjusted R-squared = 8.9 x 10^{-4}, p = 0.34 versus adjusted R-squared = -0.1, p = 0.88, respectively). We did not find a trend in diversity as a function of distance at the lineage specific level either (fig. S6, table S2), and speculate that the significance obtained by Comas *et al*. (2015) may have been driven by the lineage makeup of their defined regions (i.e., samples from the Americas consisted solely of L4 isolates while other regions harbored isolates from multiple lineages) and/or the larger number of isolates from Ethiopia relative to other regions. Our analysis using a larger sample size, finer geographic resolution, and the use of waypoints of human migration in calculating distances from Addis Ababa do not lend support to the hypothesis of serial bottlenecks related to out of Africa migrations having shaped diversity of non-African populations of *M.tb*.

Appendix S2. Supplementary note about the effects of selection on inference of migration.

The effects of population expansion, linkage, and purifying selection on *M.tb* genetic diversity have previously been demonstrated (Pepperell et al. 2013). Given these previous observations, we were curious about a potential impact of purifying selection on inference of migration. To address this question, we simulated data under demographic models with and without selection and migration, and then analyzed the resulting sequence alignments in BEAST. Results of our three population model suggested that purifying selection had a statistically negligible effect on migration rates, which can be observed from plots of the mean relative rates (fig. S13) or of the relative support of migration rates (fig. S14). We note that the discrete migration model implemented in BEAST was able to capture much of the asymmetry of our three population asymmetrical simulations as evidenced by the distribution of relative migration rates and Bayes factor (BF) support for said rates. BEAST also consistently produced similar BF support for rates estimated from data simulated under symmetrical migration models (i.e., those with global M = 0.5 or 0.0). Our simulations thus suggest that consistent purifying selection is unlikely to dramatically affect estimates of, or support for, migration rates between populations in these scenarios.

Appendix S3. Supplementary note about migration inference with ∂a∂i.

We attempted to use ∂a∂i to infer *M.tb* migration history as a complement to our analyses performed in BEAST. Briefly, EasySFS was used to convert the multi-sample VCF of L1 to a two-dimensional SFS (https://github.com/isaacovercast/easySFS). Populations were defined as India and the rest of the world (RoW) and projected from *n*=31,58 to *n*=25,25 (India and RoW, respectively). Migration inference with the synonymous SFS was performed using ∂a∂i. We modeled no split (standard neutral model), a split with no migration, a split with symmetric migration, a split with unidirectional migration (India to RoW), and a split with asymmetric migration and identified the best-fit model and maximal likelihood parameters of the migration model given our observed data. Parameters v1 and v2 were fixed according to their values estimated from a model of instantaneous expansion. Our parameter estimates m and τ, were optimized for each migration model. We used the Akaike information criterion (AIC) estimator for model selection and calculated the Poisson residuals between model and data for each best-fit model.

Results of these analyses indicated that all tested models fit poorly, as indicated by clustering of residuals from the 2D site frequency spectra, and the method was therefore unreliable for model selection (fig. S16, fig. S17, table S3).

We also used ∂a∂i to analyze simulated data for a fully linked genome under simple models of migration (see *Methods* for details of simulations). Briefly, EasySFS was used to convert the multi-sample VCF of simulated populations to a two-dimensional SFS (*n*=25,25). We modeled a split with no migration, a split with symmetric migration, a split with unidirectional migration (pop0 to pop1), and a split with asymmetric migration. We identified the best-fit model and maximal likelihood parameters of the migration models across replicates for each simulated condition. Our parameter estimates v1, v2, m and τ, were optimized for each migration model. We used the AIC estimator for model selection.

Results of these analyses also indicated that the methods performed poorly in this context, with the correct model only inferred in 26% of instances (table S4). These results are consistent with prior research indicating that SFS-based methods perform well for fully linked genomes when inference is done under very simple models but not more complex models (Pepperell et al. 2013).

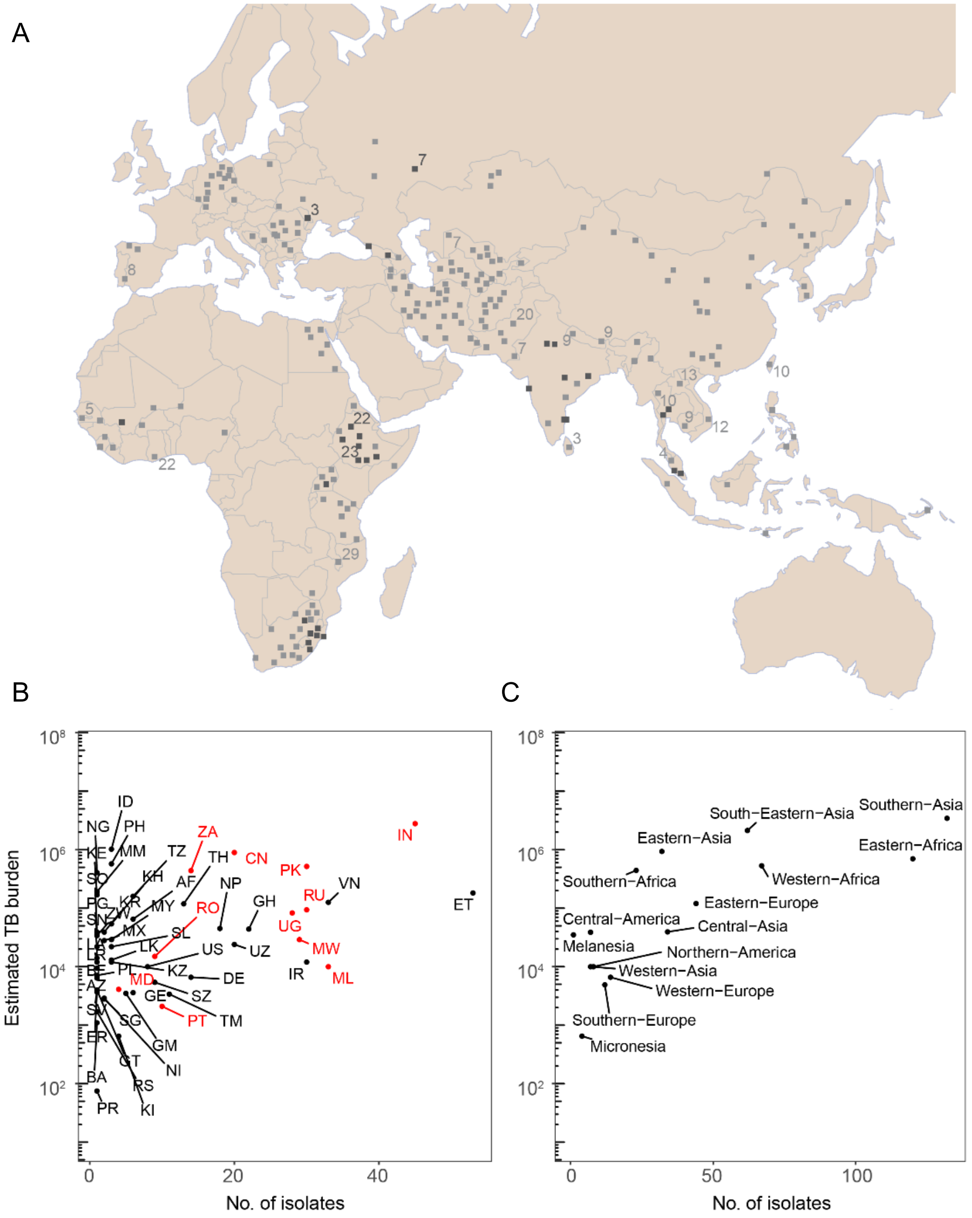

Fig. S1. Sampling strategy. (A) Map of the geographic locations for each of the 552 samples in the Old World collection. Black squares reflect the coordinates of the origin of an isolate when precise geographic information was available (e.g., city), whereas light grey squares reflect the randomized coordinates assigned to an isolate within the hospitable areas of the county from which it originated (see methods). Numbers next to squares designate the number of samples originating from a given location. (B) The number of isolates from each country in the Old World collection versus the estimated TB burden of that country. Estimated TB burden reflects the 2016 TB burden estimates by WHO. Iso2 codes for each country are labelled, and countries for which the total number of available genomes were downsampled (see *Methods*) are colored in red. (C) The number of isolates from each UN sub-region in the Old World collection versus the estimated TB burden for that sub-region. Sub-region TB burdens were determined by summing the estimated TB burden reported for all countries within the sub-region for which we had isolates in our Old World collection.

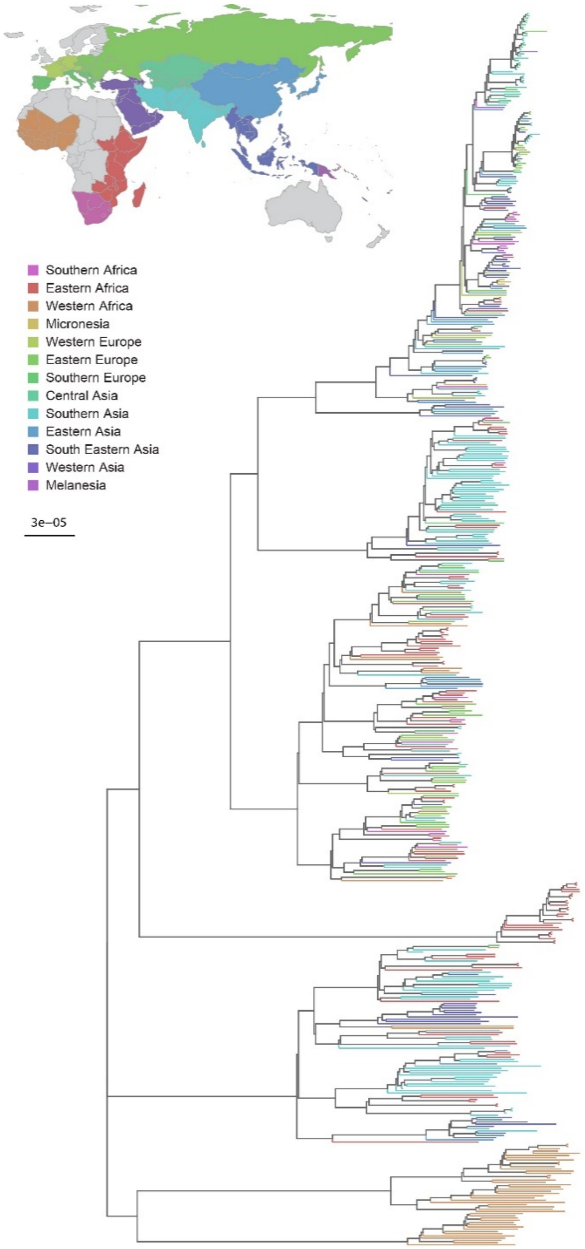

Fig. S2. Maximum likelihood phylogeny of the Old World collection of 552 Mycobacterium tuberculosis isolates. Phylogenetic analysis was performed with RAxML using the general time reversible model of nucleotide substitution under the Gamma model of rate heterogeneity on all sits where at least half of the isolates had confident genotypes (3,838,249bp). Rapid bootstrapping was performed with the -autoMR flag on the corresponding variant sites of the alignment (60,787), converging after 50 replicates. Tip labels are colored with respect to their geographic origin according to the UN geoscheme, as pictured in the legend.

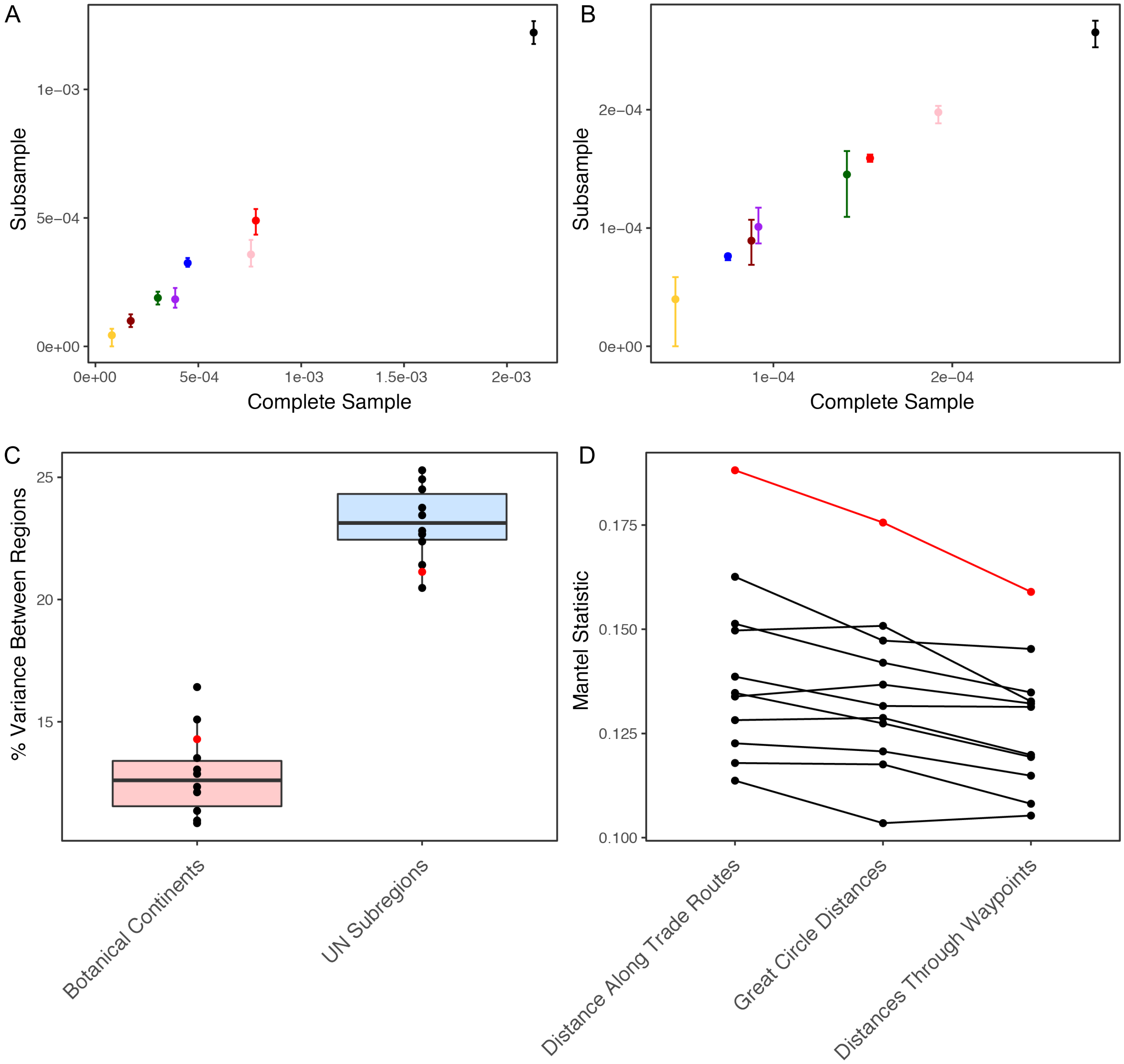

Fig. S3. Effect of subsampling on population genetic results. The Old World collection (n = 552) was down-sampled to 20 isolates per UN region without replacement a total of ten times, and all isolates for regions with less than 20 isolates were retained, resulting in alignments of 198 isolates. Population genetic analyses were repeated for each of these new datasets. (A) Relationship of (A) Watterson's ϴ values and (B) π values calculated for all M.tb isolates (black) and each lineage (L1, pink; L2, blue; L3, purple; L4, red; L5, dark-red; L6, dark-green; L7, yellow) when considering all 552 isolates (x-axis) versus the down-sampled datasets of 198 isolates (y-axis). Bars represent the minimum and maximum values obtained across the 10 random sub-samples. (C) Distribution of the variation attributable to between-region comparisons (botanical continents or UN subregions) in Analysis of Molecular Variance across 10 sub-sampled datasets (black) and the full Old World collection (red). (D) Mantel test statistic values calculated across 10 sub-samples (black) or the full Old World collection (red). Distance matrices were standardized.

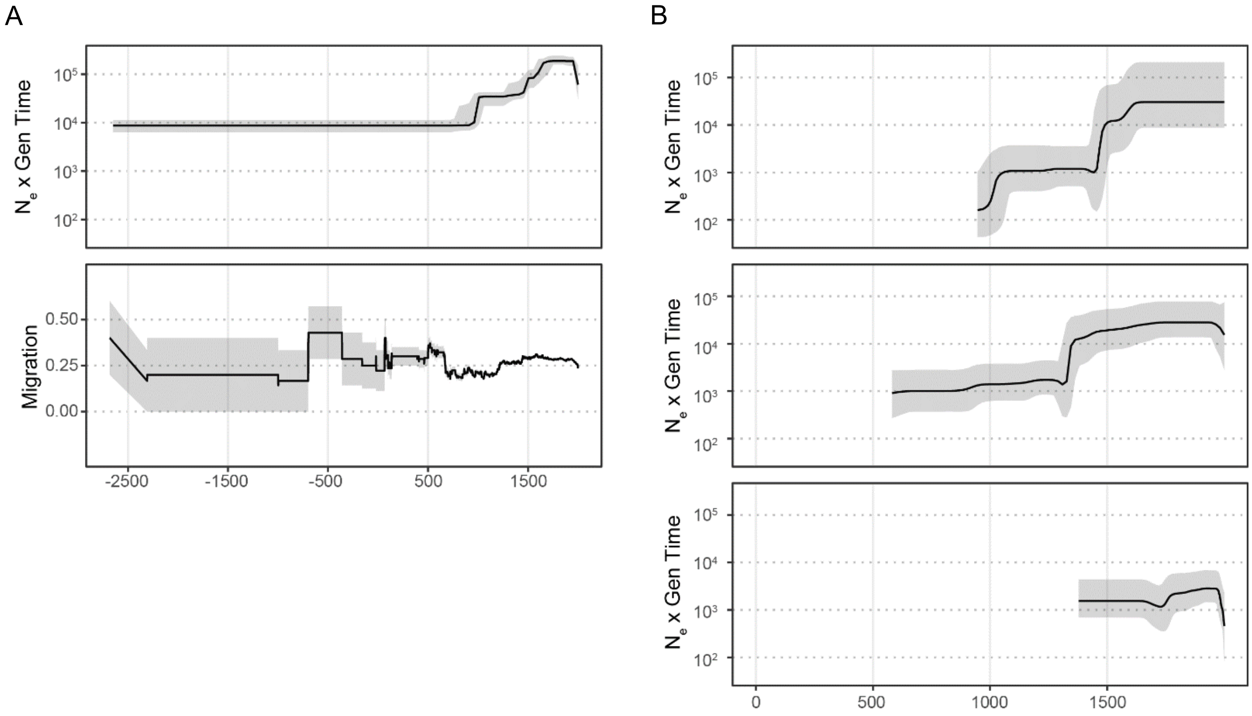

Fig. S4. Demographic histories and migration rates of *M.tb*. (A) Old World collection. Top panel – Bayesian skyline plot (BSP) shows the inferred change in effective population through time. Black lines denote median N_e and gray shading the 95% highest posterior density. Bottom panel - migration rate through time inferred from the phylogeographic analysis (see *Methods*). Grey shading depicts the rates inferred after the addition or subtraction of a single migration event, demonstrating the uncertainty of rate estimates from the early history of the phylogeny. (*B*) BSPs for Lineages 5-7 (top to bottom, respectively). Lineages 5-7 are only found in only one subregion each throughout their phylogenies resulting in a migration rate of zero through time. Dates are shown in calendar years and are based on scaling the phylogeny with a substitution rate of 5 x 10^{-8} substitutions/site/year.

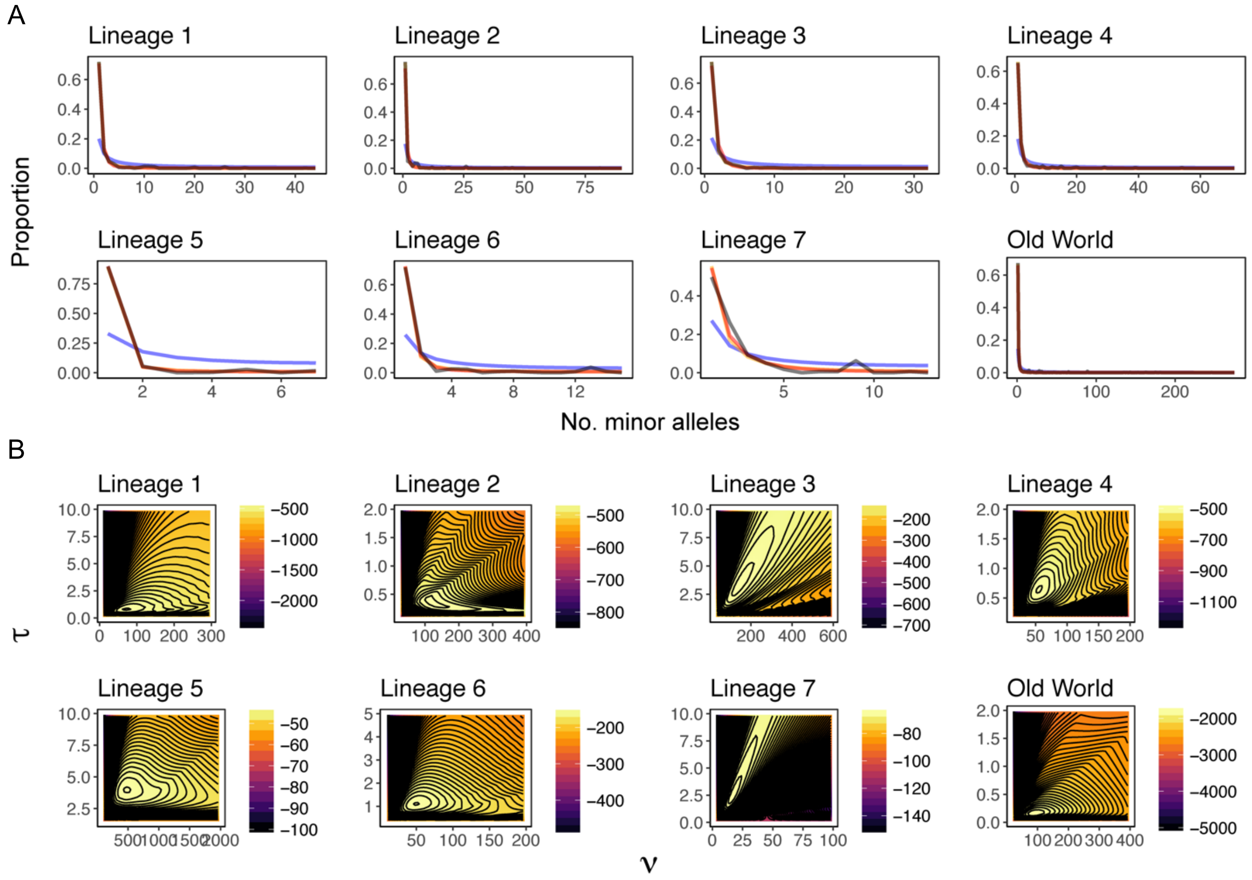

Fig. S5. Demographic analyses with ∂a∂i. (A) Folded site frequency spectrums. Demographic inference with the synonymous SFS for the entire Old World collection and each of the seven lineages was performed using methods implemented in ∂a∂i. The observed folded SFS is plotted in black, while the expected folded SFS under constant population size is plotted in blue, instantaneous expansion in red, and exponential growth in orange. (B) Likelihood surfaces. Heatmaps of log10 likelihood values (see scale bars) over a range of values for two demographic parameters in a model of instantaneous expansion are plotted: generations since expansion (τ) and N_e/N_{anc}(v).

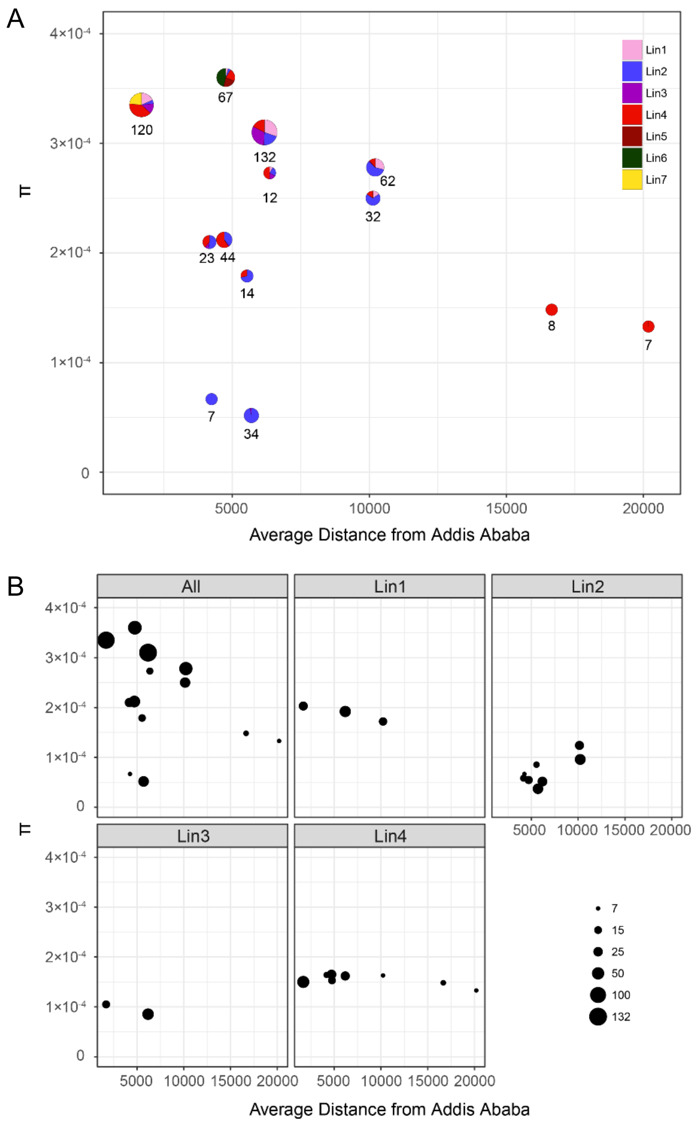

Fig. S6. Relationship between genetic diversity and geographic distance from Addis Ababa. (A) Diversity as a function of distance for all isolates among 13 UN subregions. Treating each UN subregion as a population, nucleotide diversity (π) was compared to the mean distance of isolates from the region to Addis Ababa (see *Methods*). Point size reflects the number of isolates per subregion and colors reflect the relative proportions of each lineage (see key). The number of isolates per subregion are denoted near points. (B) Diversity as a function of distance for all isolates and isolates belonging to particular lineages. Individual lineages in each UN subregion were treated as a population and nucleotide diversity (π) was compared to the mean distance of said isolates from the region to Addis Ababa (see *Methods*); population groupings resulting in less than seven isolates were not included. Point size reflects the number of isolates (see key). Data points are in Table S2.

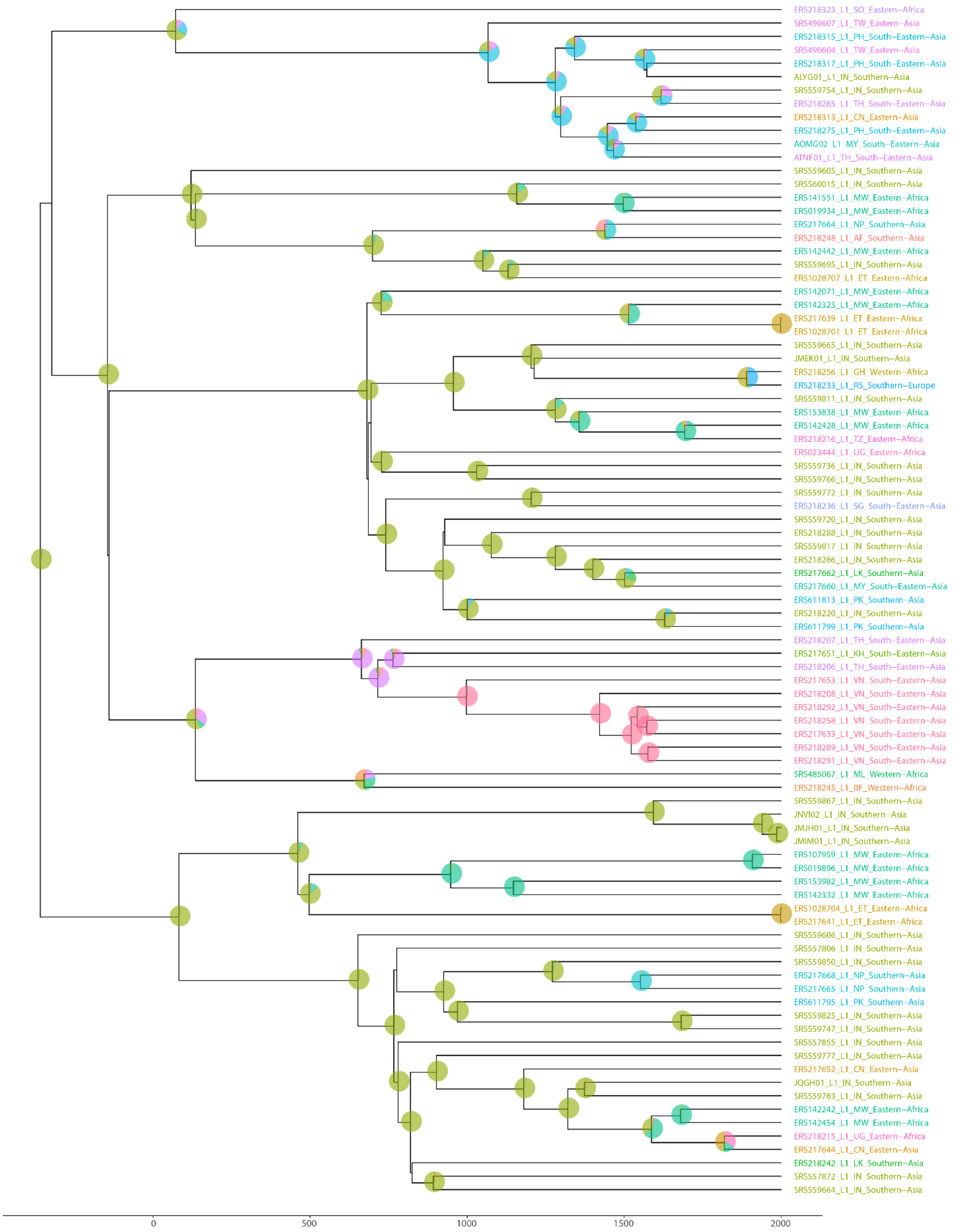
ERS218323_L1_SO_Eastern–Africa
SRS490607_L1_TW_Eastern–Asia
ERS218315_L1_PH_South–Eastern–Asia
SRS490604_L1_TW_Eastern–Asia
ERS218317_L1_PH_South–Eastern–Asia
ALYG01_L1_IN_Southern–Asia
SRS559754_L1_IN_Southern–Asia
ERS218265_L1_TH_South–Eastern–Asia
ERS218313_L1_CN_Eastern–Asia
ERS218275_L1_PH_South–Eastern–Asia
AOMG02_L1_MY_South–Eastern–Asia
ATNF01_L1_TH_South–Eastern–Asia
SRS559605_L1_IN_Southern–Asia
SRS560015_L1_IN_Southern–Asia
ERS141551_L1_MW_Eastern–Africa
ERS019934_L1_MW_Eastern–Africa
ERS217664_L1_NP_Southern–Asia
ERS218248_L1_AF_Southern–Asia
ERS142442_L1_MW_Eastern–Africa
SRS559695_L1_IN_Southern–Asia
ERS1028707_L1_ET_Eastern–Africa
ERS142071_L1_MW_Eastern–Africa
ERS142323_L1_MW_Eastern–Africa
ERS217639_L1_ET_Eastern–Africa
ERS1028701_L1_ET_Eastern–Africa
SRS559665_L1_IN_Southern–Asia
JMEK01_L1_IN_Southern–Asia
ERS218256_L1_GH_Western–Africa
ERS218233_L1_RS_Southern–Europe
SRS559811_L1_IN_Southern–Asia
ERS153838_L1_MW_Eastern–Africa
ERS142428_L1_MW_Eastern–Africa
ERS218216_L1_TZ_Eastern–Africa
ERS023444_L1_UG_Eastern–Africa
SRS559736_L1_IN_Southern–Asia
SRS559766_L1_IN_Southern–Asia
SRS559772_L1_IN_Southern–Asia
ERS218236_L1_SG_South–Eastern–Asia
SRS559720_L1_IN_Southern–Asia
ERS218288_L1_IN_Southern–Asia
SRS559817_L1_IN_Southern–Asia
ERS218286_L1_IN_Southern–Asia
ERS217662_L1_LK_Southern–Asia
ERS217660_L1_MY_South–Eastern–Asia
ERS611813_L1_PK_Southern–Asia
ERS218220_L1_IN_Southern–Asia
ERS611799_L1_PK_Southern–Asia
ERS218207_L1_TH_South–Eastern–Asia
ERS217651_L1_KH_South–Eastern–Asia
ERS218206_L1_TH_South–Eastern–Asia
ERS217653_L1_VN_South–Eastern–Asia
ERS218208_L1_VN_South–Eastern–Asia
ERS218292_L1_VN_South–Eastern–Asia
ERS218258_L1_VN_South–Eastern–Asia
ERS217633_L1_VN_South–Eastern–Asia
ERS218289_L1_VN_South–Eastern–Asia
ERS218291_L1_VN_South–Eastern–Asia
SRS485067_L1_ML_Western–Africa
ERS218245_L1_BF_Western–Africa
SRS559867_L1_IN_Southern–Asia
JNVI02_L1_IN_Southern–Asia
JMJH01_L1_IN_Southern–Asia
JMIM01_L1_IN_Southern–Asia
ERS107959_L1_MW_Eastern–Africa
ERS019896_L1_MW_Eastern–Africa
ERS153982_L1_MW_Eastern–Africa
ERS142332_L1_MW_Eastern–Africa
ERS1028704_L1_ET_Eastern–Africa
ERS217641_L1_ET_Eastern–Africa
SRS559606_L1_IN_Southern–Asia
SRS557806_L1_IN_Southern–Asia
SRS559850_L1_IN_Southern–Asia
ERS217668_L1_NP_Southern–Asia
ERS217665_L1_NP_Southern–Asia
ERS611795_L1_PK_Southern–Asia
SRS559825_L1_IN_Southern–Asia
SRS559747_L1_IN_Southern–Asia
SRS557855_L1_IN_Southern–Asia
SRS559777_L1_IN_Southern–Asia
ERS217652_L1_CN_Eastern–Asia
JQGH01_L1_IN_Southern–Asia
SRS559783_L1_IN_Southern–Asia
ERS142242_L1_MW_Eastern–Africa
ERS142454_L1_MW_Eastern–Africa
ERS218215_L1_UG_Eastern–Africa
ERS217644_L1_CN_Eastern–Asia
ERS218242_L1_LK_Southern–Asia
SRS557872_L1_IN_Southern–Asia
SRS559664_L1_IN_Southern–Asia
0
500
1000
1500
2000

Fig. S8. Maximum clade credibility tree of lineage 1 M.tb genomes. Pie charts at nodes are colored according to the location state probabilities (country). Tip labels are colored according to the country of origin where the isolate was obtained.

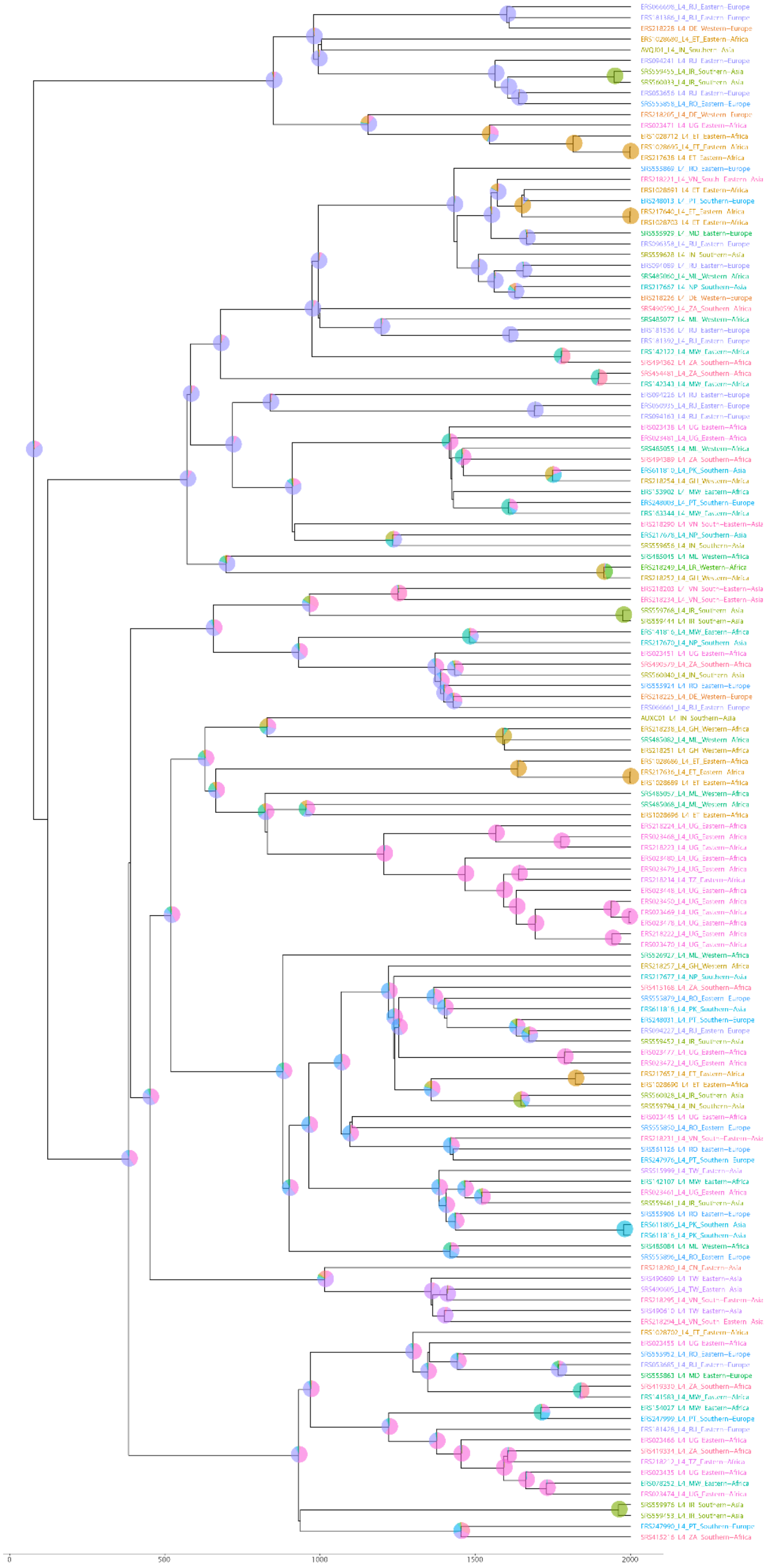
ERS066698_L4_RU_Eastern-Europe
ERS181386_L4_RU_Eastern-Europe
ERS218228_L4_DE_Western-Europe
ERS1028680_L4_ET_Eastern-Africa
AVQJ01_L4_IN_Southern-Asia
ERS094241_L4_RU_Eastern-Europe
SRS559455_L4_IR_Southern-Asia
SRS560033_L4_IR_Southern-Asia
ERS053656_L4_RU_Eastern-Europe
SRS555858_L4_RO_Eastern-Europe
ERS218205_L4_DE_Western-Europe
ERS023471_L4_UG_Eastern-Africa
ERS1028712_L4_ET_Eastern-Africa
ERS1028695_L4_ET_Eastern-Africa
ERS217638_L4_ET_Eastern-Africa
SRS555869_L4_RO_Eastern-Europe
ERS218221_L4_VN_South-Eastern-Asia
ERS1028691_L4_ET_Eastern-Africa
ERS248013_L4_PT_Southern-Europe
ERS217640_L4_ET_Eastern-Africa
ERS1028703_L4_ET_Eastern-Africa
SRS555929_L4_MD_Eastern-Europe
ERS096358_L4_RU_Eastern-Europe
SRS559628_L4_IN_Southern-Asia
ERS094089_L4_RU_Eastern-Europe
SRS485060_L4_ML_Western-Africa
ERS217667_L4_NP_Southern-Asia
ERS218226_L4_DE_Western-Europe
SRS490590_L4_ZA_Southern-Africa
SRS485077_L4_ML_Western-Africa
ERS181536_L4_RU_Eastern-Europe
ERS181392_L4_RU_Eastern-Europe
ERS142122_L4_MW_Eastern-Africa
SRS494362_L4_ZA_Southern-Africa
SRS454481_L4_ZA_Southern-Africa
ERS142343_L4_MW_Eastern-Africa
ERS094226_L4_RU_Eastern-Europe
ERS050935_L4_RU_Eastern-Europe
ERS094163_L4_RU_Eastern-Europe
ERS023438_L4_UG_Eastern-Africa
ERS023481_L4_UG_Eastern-Africa
SRS485055_L4_ML_Western-Africa
SRS494389_L4_ZA_Southern-Africa
ERS611810_L4_PK_Southern-Asia
ERS218254_L4_GH_Western-Africa
ERS153902_L4_MW_Eastern-Africa
ERS248003_L4_PT_Southern-Europe
ERS163344_L4_MW_Eastern-Africa
ERS218290_L4_VN_South-Eastern-Asia
ERS217678_L4_NP_Southern-Asia
SRS559656_L4_IN_Southern-Asia
SRS485045_L4_ML_Western-Africa
ERS218249_L4_LR_Western-Africa
ERS218252_L4_GH_Western-Africa
ERS218203_L4_VN_South-Eastern-Asia
ERS218234_L4_VN_South-Eastern-Asia
SRS559768_L4_IR_Southern-Asia
SRS559441_L4_IR_Southern-Asia
ERS141816_L4_MW_Eastern-Africa
ERS217670_L4_NP_Southern-Asia
ERS023451_L4_UG_Eastern-Africa
SRS490579_L4_ZA_Southern-Africa
SRS560040_L4_IN_Southern-Asia
SRS555924_L4_RO_Eastern-Europe
ERS218225_L4_DE_Western-Europe
ERS066661_L4_RU_Eastern-Europe
AUXC01_L4_IN_Southern-Asia
ERS218238_L4_GH_Western-Africa
SRS485082_L4_ML_Western-Africa
ERS218251_L4_GH_Western-Africa
ERS1028686_L4_ET_Eastern-Africa
ERS217636_L4_ET_Eastern-Africa
ERS1028689_L4_ET_Eastern-Africa
SRS485057_L4_ML_Western-Africa
SRS485068_L4_ML_Western-Africa
ERS1028696_L4_ET_Eastern-Africa
ERS218224_L4_UG_Eastern-Africa
ERS023468_L4_UG_Eastern-Africa
ERS218223_L4_UG_Eastern-Africa
ERS023480_L4_UG_Eastern-Africa
ERS023479_L4_UG_Eastern-Africa
ERS218214_L4_TZ_Eastern-Africa
ERS023448_L4_UG_Eastern-Africa
ERS023450_L4_UG_Eastern-Africa
ERS023469_L4_UG_Eastern-Africa
ERS023478_L4_UG_Eastern-Africa
ERS218222_L4_UG_Eastern-Africa
ERS023470_L4_UG_Eastern-Africa
SRS526927_L4_ML_Western-Africa
ERS218257_L4_GH_Western-Africa
ERS217677_L4_NP_Southern-Asia
SRS415168_L4_ZA_Southern-Africa
SRS555879_L4_RO_Eastern-Europe
ERS611818_L4_PK_Southern-Asia
ERS248031_L4_PT_Southern-Europe
ERS094227_L4_RU_Eastern-Europe
SRS559452_L4_IR_Southern-Asia
ERS023477_L4_UG_Eastern-Africa
ERS023472_L4_UG_Eastern-Africa
ERS217657_L4_ET_Eastern-Africa
ERS1028690_L4_ET_Eastern-Africa
SRS560028_L4_IR_Southern-Asia
SRS559794_L4_IN_Southern-Asia
ERS023445_L4_UG_Eastern-Africa
SRS555850_L4_RO_Eastern-Europe
ERS218231_L4_VN_South-Eastern-Asia
SRS561126_L4_RO_Eastern-Europe
ERS247976_L4_PT_Southern-Europe
SRS515999_L4_TW_Eastern-Asia
ERS142107_L4_MW_Eastern-Africa
ERS023461_L4_UG_Eastern-Africa
SRS559461_L4_IR_Southern-Asia
SRS555906_L4_RO_Eastern-Europe
ERS611805_L4_PK_Southern-Asia
ERS611816_L4_PK_Southern-Asia
SRS485084_L4_ML_Western-Africa
SRS555896_L4_RO_Eastern-Europe
ERS218280_L4_CN_Eastern-Asia
SRS490609_L4_TW_Eastern-Asia
SRS490605_L4_TW_Eastern-Asia
ERS218295_L4_VN_South-Eastern-Asia
SRS490610_L4_TW_Eastern-Asia
ERS218294_L4_VN_South-Eastern-Asia
ERS1028702_L4_ET_Eastern-Africa
ERS023455_L4_UG_Eastern-Africa
SRS555952_L4_RO_Eastern-Europe
ERS053685_L4_RU_Eastern-Europe
SRS555863_L4_MD_Eastern-Europe
SRS419330_L4_ZA_Southern-Africa
ERS141583_L4_MW_Eastern-Africa
ERS154027_L4_MW_Eastern-Africa
ERS247999_L4_PT_Southern-Europe
ERS181428_L4_RU_Eastern-Europe
ERS023466_L4_UG_Eastern-Africa
SRS419334_L4_ZA_Southern-Africa
ERS218212_L4_TZ_Eastern-Africa
ERS023435_L4_UG_Eastern-Africa
ERS078252_L4_MW_Eastern-Africa
ERS023474_L4_UG_Eastern-Africa
SRS559976_L4_IR_Southern-Asia
SRS559453_L4_IR_Southern-Asia
ERS247990_L4_PT_Southern-Europe
SRS415216_L4_ZA_Southern-Africa
0
500
1000
1500
2000

Fig. S9. Maximum clade credibility tree of lineage 4 M.tb genomes. Pie charts at nodes are colored according to the location state probabilities (country). Tip labels are colored according to the country of origin where the isolate was obtained.

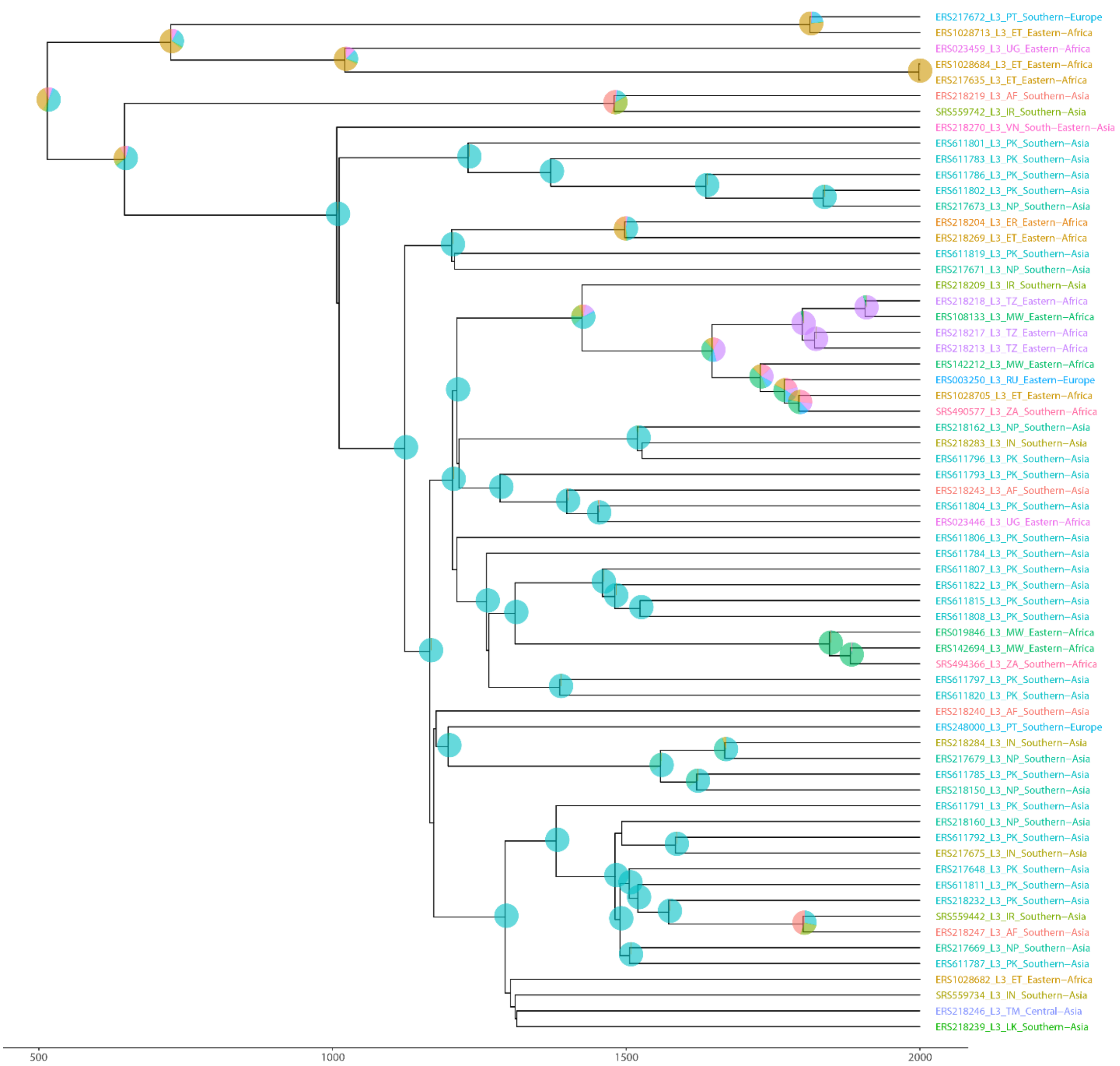
ERS217672_L3_PT_Southern–Europe
ERS1028713_L3_ET_Eastern–Africa
ERS023459_L3_UG_Eastern–Africa
ERS1028684_L3_ET_Eastern–Africa
ERS217635_L3_ET_Eastern–Africa
ERS218219_L3_AF_Southern–Asia
SRS559742_L3_IR_Southern–Asia
ERS218270_L3_VN_South–Eastern–Asia
ERS611801_L3_PK_Southern–Asia
ERS611783_L3_PK_Southern–Asia
ERS611786_L3_PK_Southern–Asia
ERS611802_L3_PK_Southern–Asia
ERS217673_L3_NP_Southern–Asia
ERS218204_L3_ER_Eastern–Africa
ERS218269_L3_ET_Eastern–Africa
ERS611819_L3_PK_Southern–Asia
ERS217671_L3_NP_Southern–Asia
ERS218209_L3_IR_Southern–Asia
ERS218218_L3_TZ_Eastern–Africa
ERS108133_L3_MW_Eastern–Africa
ERS218217_L3_TZ_Eastern–Africa
ERS218213_L3_TZ_Eastern–Africa
ERS142212_L3_MW_Eastern–Africa
ERS003250_L3_RU_Eastern–Europe
ERS1028705_L3_ET_Eastern–Africa
SRS490577_L3_ZA_Southern–Africa
ERS218162_L3_NP_Southern–Asia
ERS218283_L3_IN_Southern–Asia
ERS611796_L3_PK_Southern–Asia
ERS611793_L3_PK_Southern–Asia
ERS218243_L3_AF_Southern–Asia
ERS611804_L3_PK_Southern–Asia
ERS023446_L3_UG_Eastern–Africa
ERS611806_L3_PK_Southern–Asia
ERS611784_L3_PK_Southern–Asia
ERS611807_L3_PK_Southern–Asia
ERS611822_L3_PK_Southern–Asia
ERS611815_L3_PK_Southern–Asia
ERS611808_L3_PK_Southern–Asia
ERS019846_L3_MW_Eastern–Africa
ERS142694_L3_MW_Eastern–Africa
SRS494366_L3_ZA_Southern–Africa
ERS611797_L3_PK_Southern–Asia
ERS611820_L3_PK_Southern–Asia
ERS218240_L3_AF_Southern–Asia
ERS248000_L3_PT_Southern–Europe
ERS218284_L3_IN_Southern–Asia
ERS217679_L3_NP_Southern–Asia
ERS611785_L3_PK_Southern–Asia
ERS218150_L3_NP_Southern–Asia
ERS611791_L3_PK_Southern–Asia
ERS218160_L3_NP_Southern–Asia
ERS611792_L3_PK_Southern–Asia
ERS217675_L3_IN_Southern–Asia
ERS217648_L3_PK_Southern–Asia
ERS611811_L3_PK_Southern–Asia
ERS218232_L3_PK_Southern–Asia
SRS559442_L3_IR_Southern–Asia
ERS218247_L3_AF_Southern–Asia
ERS217669_L3_NP_Southern–Asia
ERS611787_L3_PK_Southern–Asia
ERS1028682_L3_ET_Eastern–Africa
SRS559734_L3_IN_Southern–Asia
ERS218246_L3_TM_Central–Asia
ERS218239_L3_LK_Southern–Asia
500
1000
1500
2000

Fig. S10. Maximum clade credibility tree of lineage 3 M.tb genomes. Pie charts at nodes are colored according to the location state probabilities (country). Tip labels are colored according to the country of origin where the isolate was obtained.

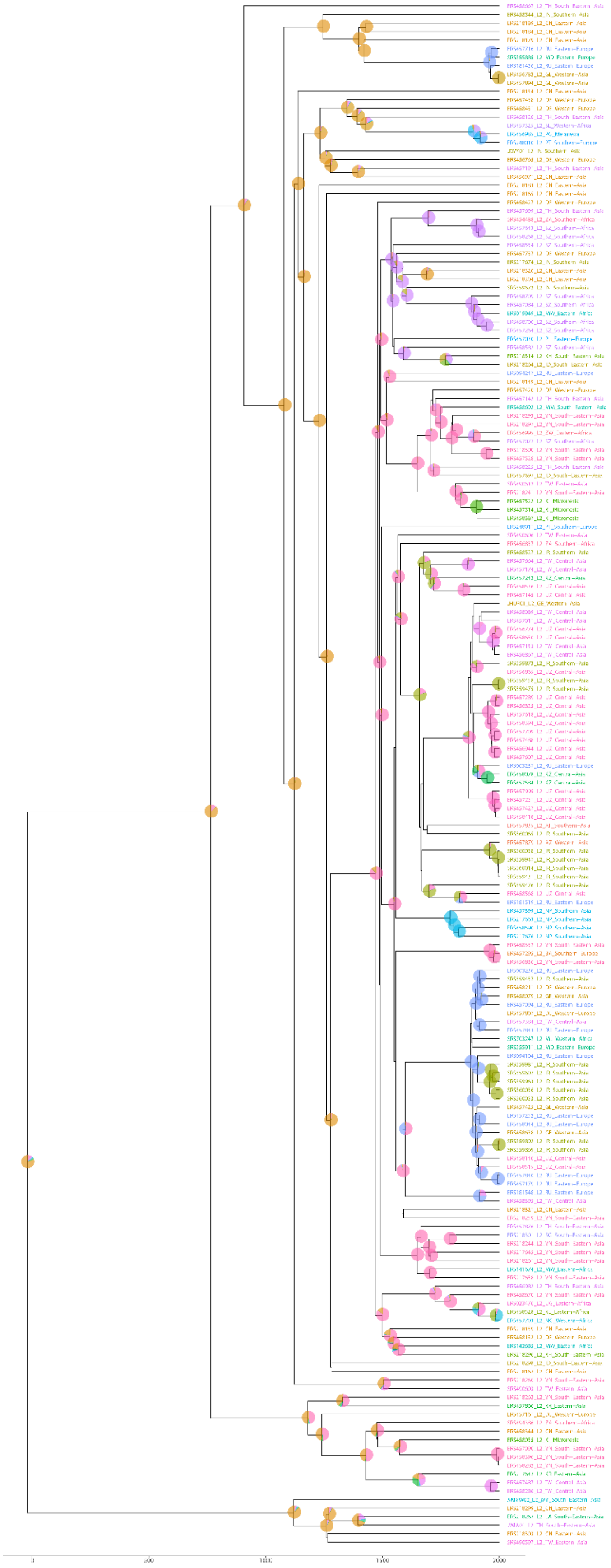

Fig. S11. Maximum clade credibility tree of lineage 2 M.tb genomes. Pie charts at nodes are colored according to the location state probabilities (country). Tip labels are colored according to the country of origin where the isolate was obtained.

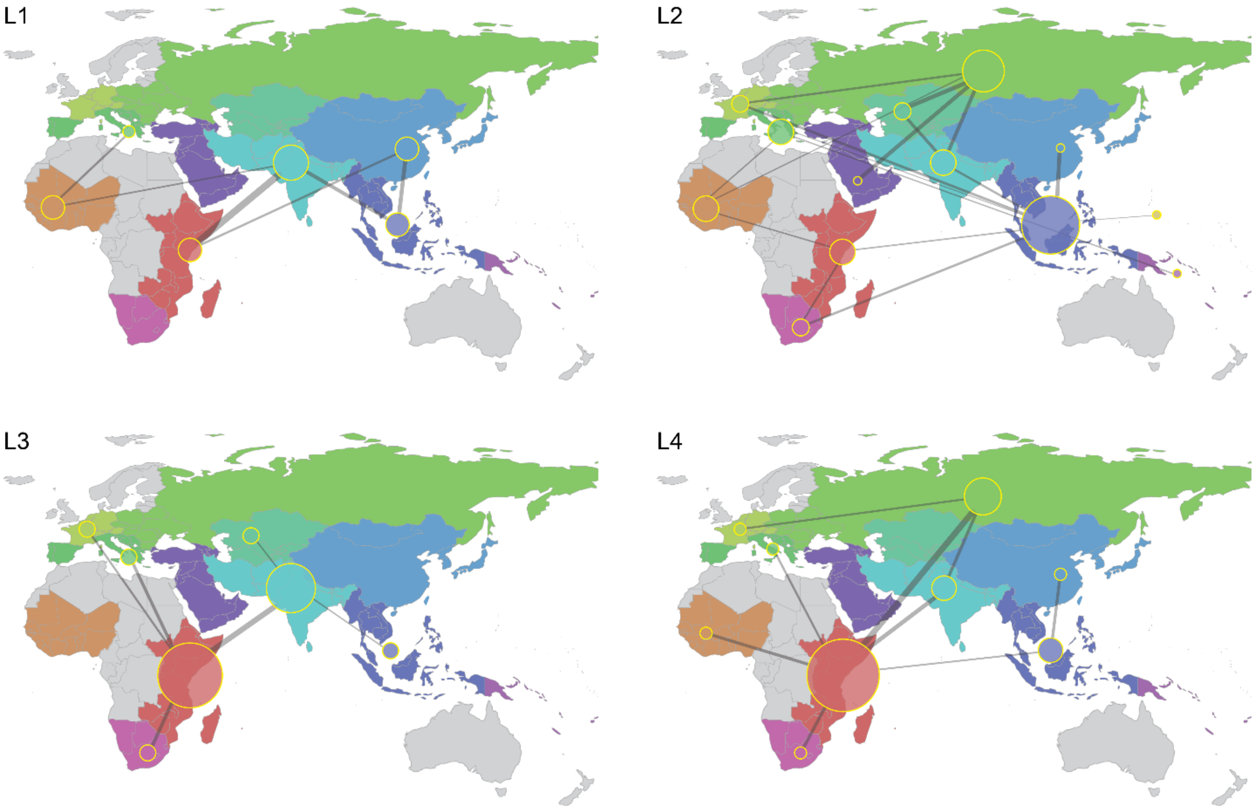

Fig. S12. Relative rates of migration between UN subregions. Relative median rates of migration between geographic regions sampled every 10,000 states in the Bayesian analysis for individual lineages 1-4 are displayed by the thickness of the line connecting the regions. Node size reflects the relative degree of connectivity within each analysis (how many connections the region shares).

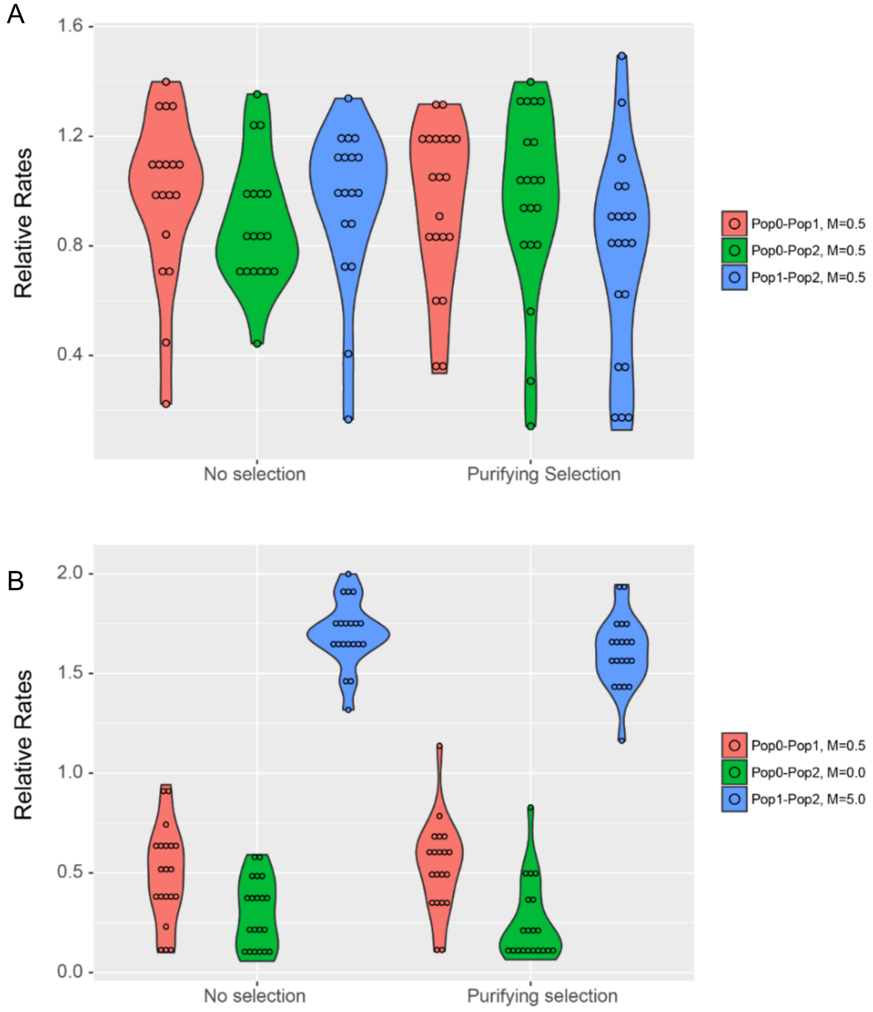

Fig. S13. Three Population Migration Simulation Rate Estimates. Sequence alignments were simulated using the forward-in-time simulation package SFS_CODE (Hernandez 2008) and analyzed in BEAST (Drummond & Rambaut 2007). Alignments were simulated under (A) symmetric and (B) asymmetric migration regimes, as well as with and without purifying selection. The three possible pairwise migration rates (between populations 0 and 1 = Pop0-Pop1, between 0 and 2 = Pop0-Pop2, and between 1 and 2 = Pop1-Pop2) were all set to 0.5 (i.e., *M* = 0.5) in the symmetric migration regime. In the asymmetric regime, the alignments were simulated with different pairwise population migration rates (Pop0-Pop1 = 0.5, Pop0-Pop2 = 0.0, and Pop1-Pop2 = 5.0). Twenty alignments were simulated with and without purifying selection under each migration regime (total simulated alignments with migration = 80); twenty alignments were also simulated under a scenario without pairwise migration (Pop0-Pop1 = Pop0-Pop2 = Pop1-Pop2 = 0.0) and without purifying selection. The distributions of the migration rate estimates from BEAST analysis are presented as violin plots for the symmetric migration simulations in the top plot and for the asymmetric migration simulations in the bottom plot. There were no statistically significant differences between migration rates estimated from sequences simulated with and without purifying selection.

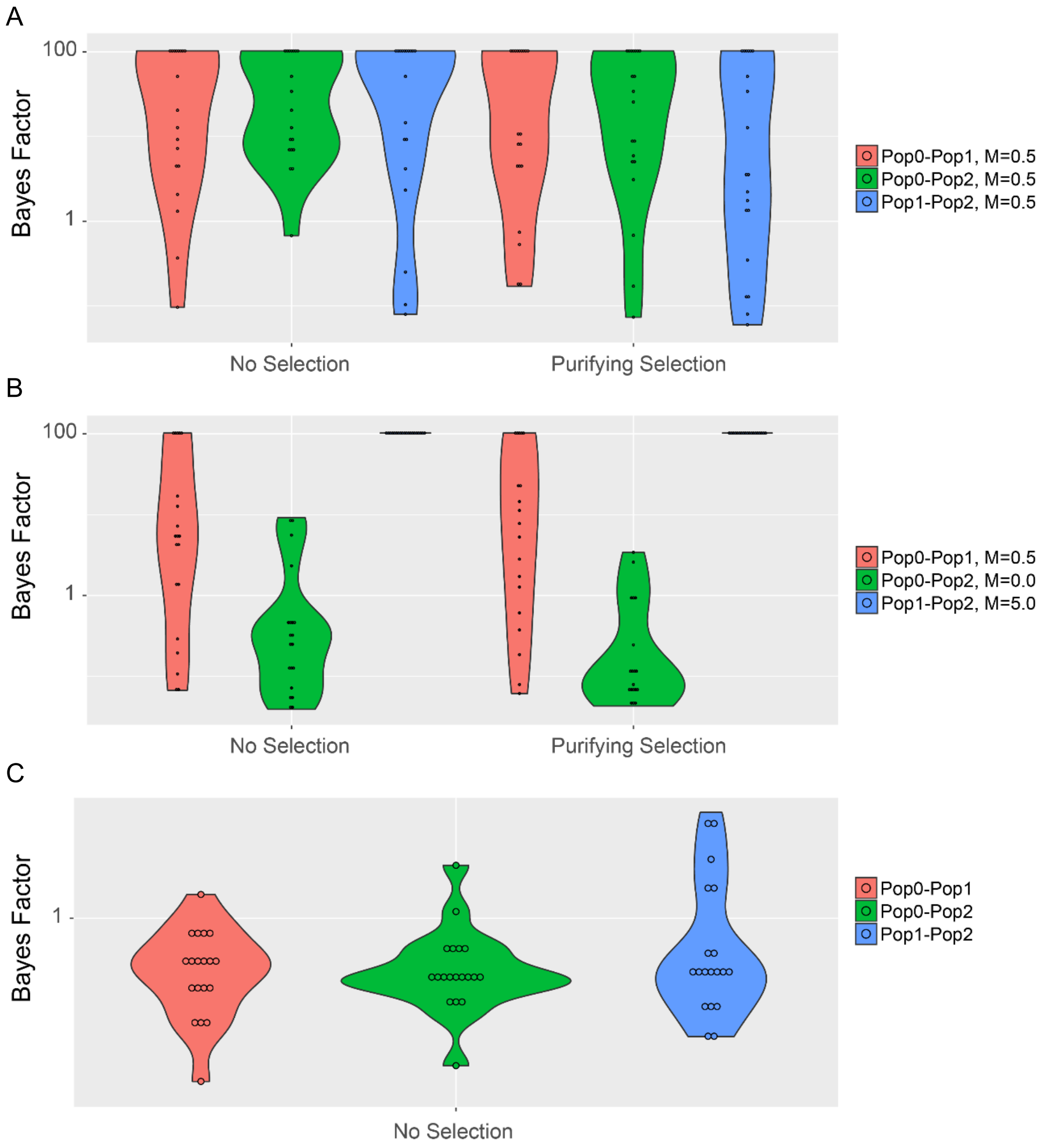

Fig. S14. Three Population Migration Simulation Rate Bayes Factors. Sequence alignments were simulated using the forward-in-time simulation package SFS_CODE (Hernandez 2008) and analyzed in BEAST (Drummond & Rambaut 2007). Alignments were simulated under (A) symmetric and (B) asymmetric migration regimes, as well as with and without purifying selection. The three possible pairwise migration rates (between populations 0 and 1 = Pop0-Pop1, between 0 and 2 = Pop0-Pop2, and between 1 and 2 = Pop1-Pop2) were all set to 0.5 (i.e., *M* = 0.5) in the symmetric migration regime. In the asymmetric regime, the alignments were simulated with different pairwise population migration rates (Pop0-Pop1 = 0.5, Pop0-Pop2 = 0.0, and Pop1-Pop2 = 5.0). Twenty alignments were simulated with and without purifying selection under each migration regime (total simulated alignments with migration = 80); (C) twenty alignments were also simulated under a scenario without pairwise migration (Pop0-Pop1 = Pop0-Pop2 = Pop1-Pop2 = 0.0) and without purifying selection. Bayes factor support for each pairwise population migration rate (i.e., Pop0-Pop1, Pop0-Pop2, and Pop1-Pop2 following the naming convention described in the legend) was estimated using BSSVS (Lemey et al. 2009) in the SpreaD3 package (Bielejec et al. 2016).

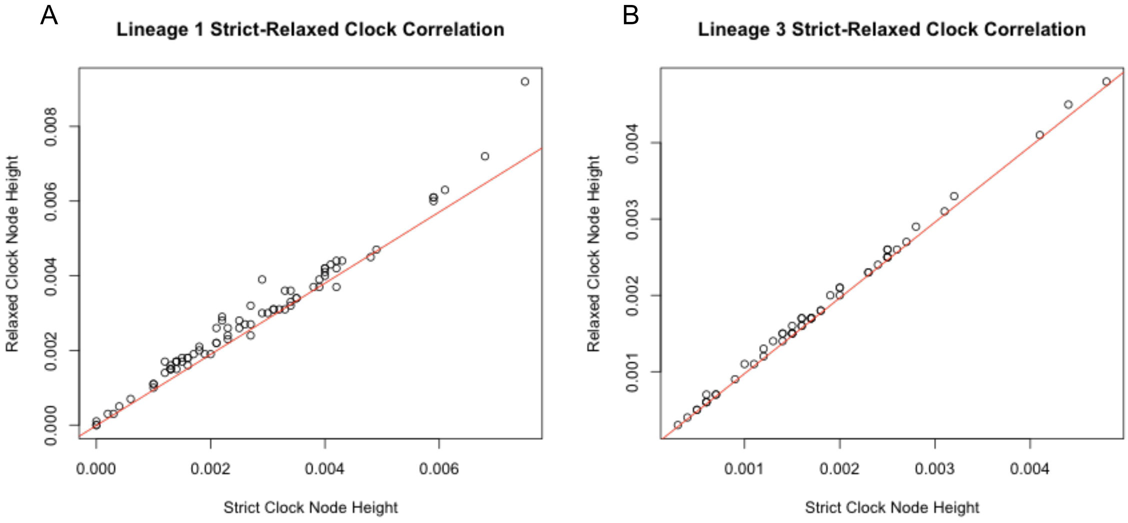

Fig. S15. Comparison of strict and relaxed clocks. (A) Correlation of Lineage 1 node heights estimated under strict and relaxed (uncorrelated lognormal clock). R^2 for the correlation is 0.9534. Analyses were performed in BEAST using the same conditions as described in the main text. (B) Correlation of Lineage 3 node heights estimated under strict and relaxed (uncorrelated lognormal clock). R^2 for the correlation is 0.9925. Analyses were performed in BEAST using the same conditions as described in the main text.

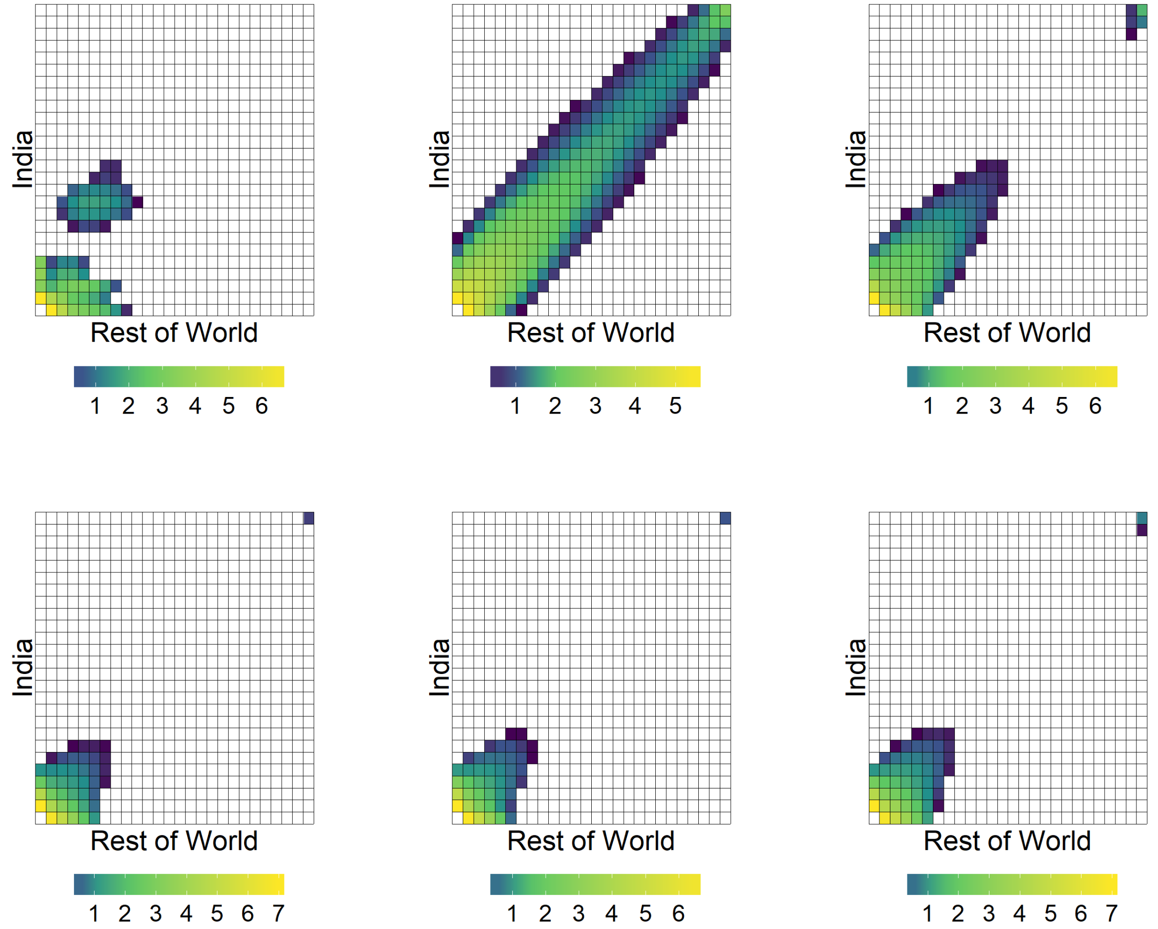

Fig. S16. Observed and inferred 2D SFS of India and the rest of the world (RoW). Heatmaps of 2D spectra colored by number of SNPs at each frequency in the population (log-transformed). SNP counts = 0 in white. From left to right, top to bottom: observed, no split and no migration, no migration, symmetric migration, unidirectional migration, and asymmetric migration.

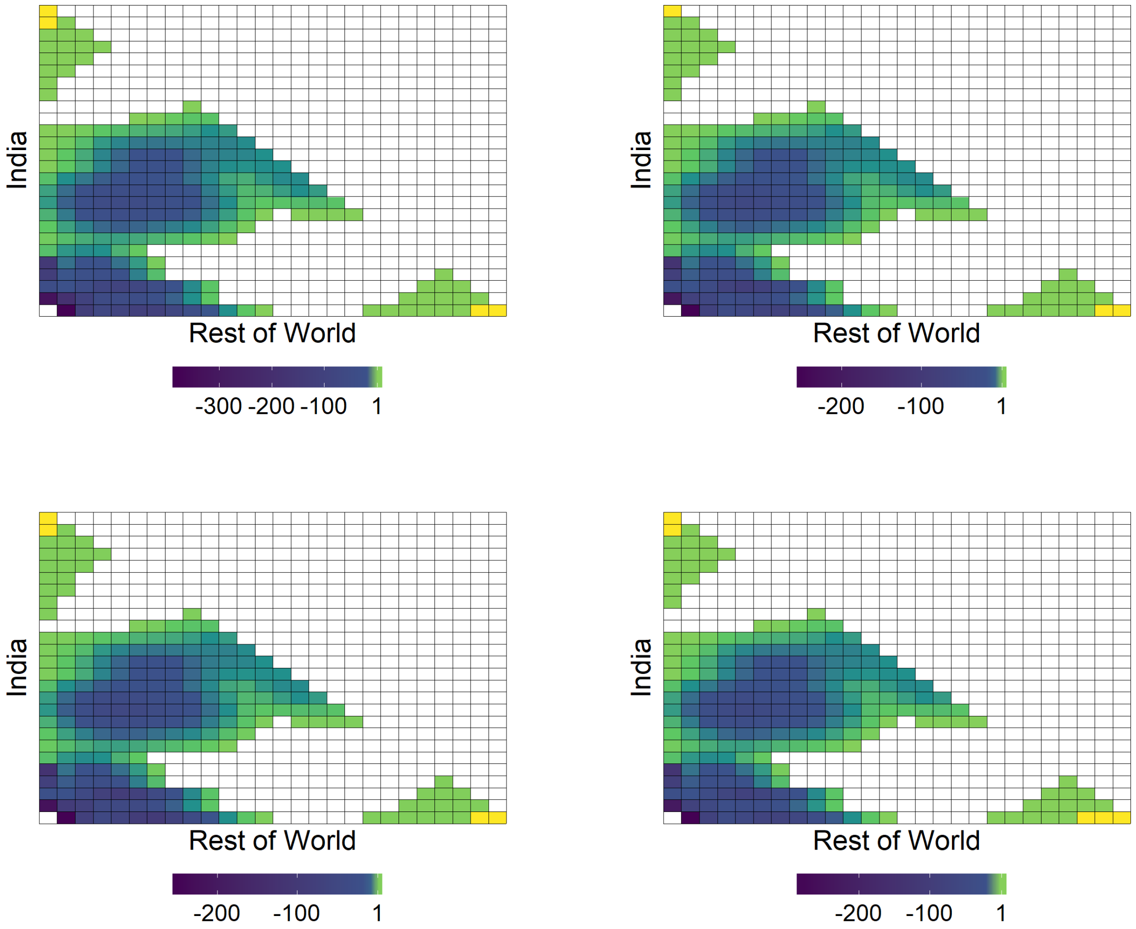

Fig. S17. Poisson residuals for best-fit migration models. Heatmaps of Poisson residuals between model and data for the 2D spectra calculated using ∂a∂I. Masked data colored in white. Positive or negative residuals indicate the model predicts too many or too few SNPs at a given frequency, respectively. From left to right, top to bottom: no migration, symmetric migration, unidirectional migration, and asymmetric migration.

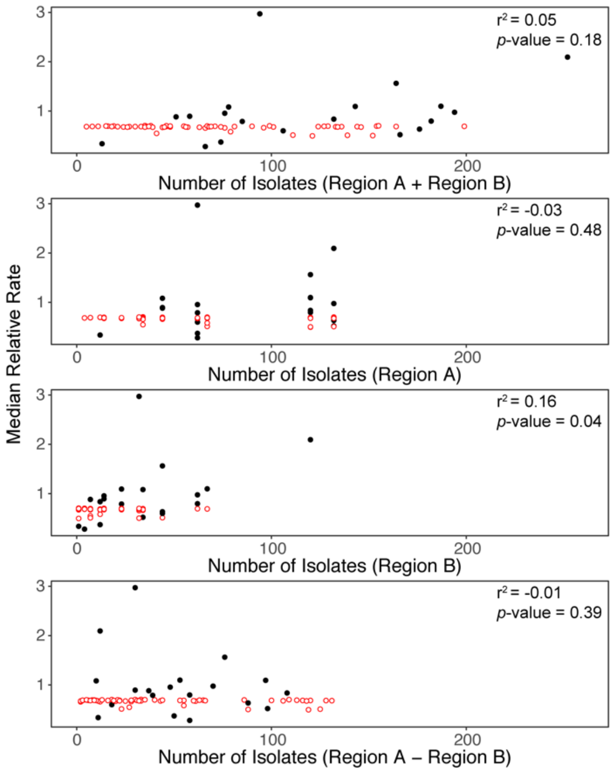

Fig. S18. Relationship between estimated migration rates and sample number. The median relative rate obtained from analysis of the Old World collection with BSSVS (*y*-axis) versus the number of isolates in the Old World collection, where Region A refers to the UN subregion with the most isolates in the comparison and Region B refers to the UN subregion with the least (*x*-axis). Adjusted r^2 and *p*-values for linear regressions of significant rates (Bayes factor > 5, black circles) are displayed. Open red circles correspond to rates for which there was low support (Bayes factor < 5).