Phylogeography of *Mycobacterium tuberculosis*

Mary B. O'Neill[1,2], Alex Zarley[3], Andrew Kitchen[4], Caitlin S. Pepperell[1]

[1]Departments of Medical Microbiology & Immunology and Medicine, University of Wisconsin-Madison, Madison, WI, USA
[2]Laboratory of Genetics, University of Wisconsin-Madison, Madison, WI, USA
[3]Department of Geography, University of Wisconsin-Madison, WI, USA
[4]Department of Anthropology, University of Iowa, Iowa City, IA 52242, USA

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

School of Medicine and Public Health
UNIVERSITY OF WISCONSIN–MADISON

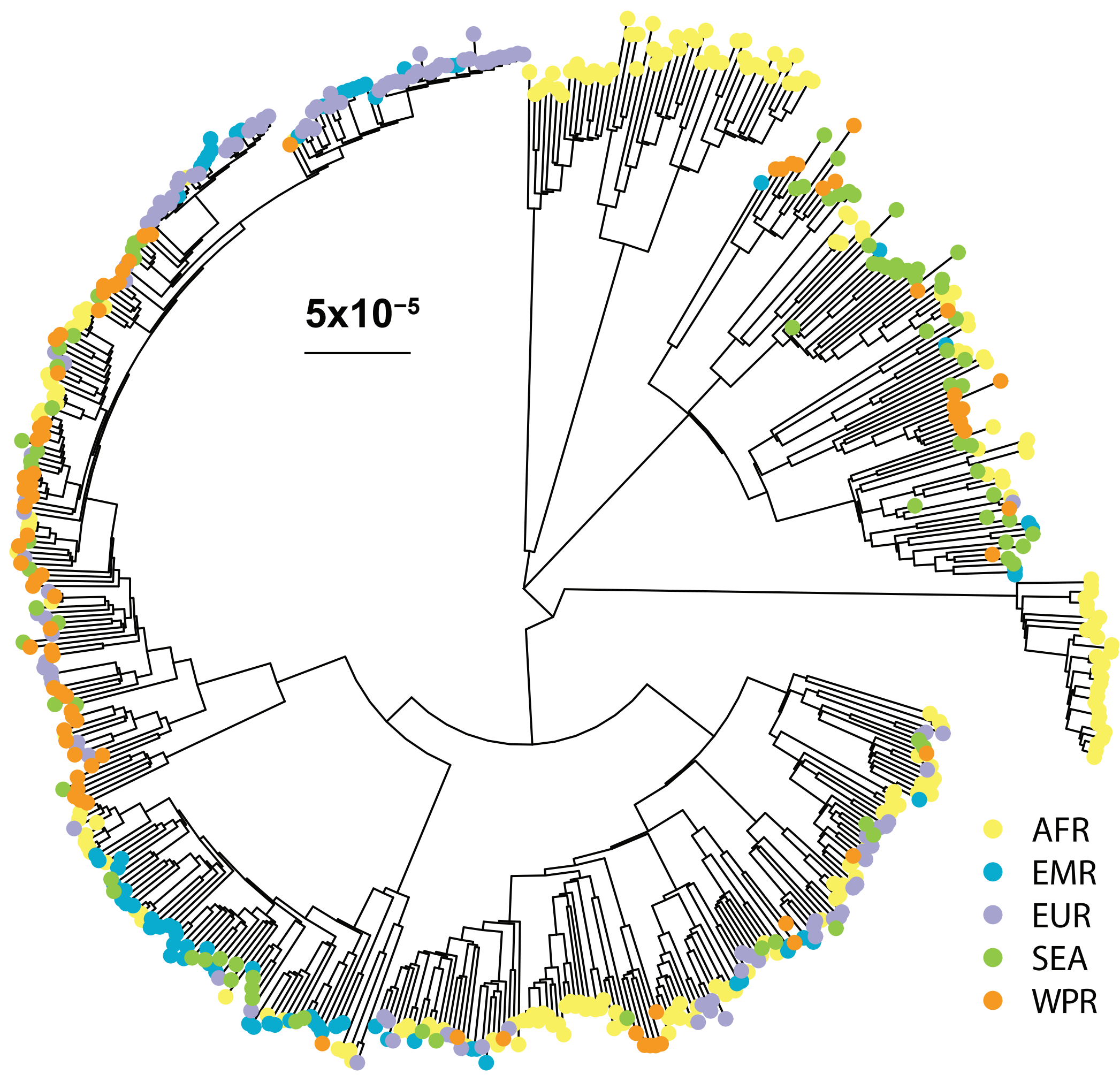
Introduction

Tuberculosis (TB) is caused by members of the *Mycobacterium tuberculosis* complex (MTBC), a group of closely related bacterial pathogens with varying host ranges. The most well-known member of the complex, *M. tuberculosis*, is the etiological agent of human TB and the leading cause of mortality due to a single infectious agent. The origin and history of TB remains contentious, with estimates of the time to the most recent common ancestor (TMRCA) of extant strains of the MTBC varying by orders of magnitude. Here, we present an overview of the phylogeographic diversity of *M. tuberculosis* and reconstruct historical patterns of TB migration.

Genetic Dataset

Sequence data of *M. tuberculosis* isolates from a collection of studies were assembled using our validated reference guided assembly (RGA) pipeline (https://github.com/tracysmith/RGAPepPipe) with H37Rv (NC_000962.3)[1] as the reference genome. Countries with large collections of sequenced genomes were sub-sampled. To ensure only high quality sequencing data were included, individual sequencing runs for which <80% of the H37Rv genome was covered by at least 20X coverage were discarded, as were runs for which <70% of the reads mapped to the reference. Pilon[2] was used to call variants; ambiguous calls and deletions were treated as missing data. High quality draft genome assemblies for which fastqs were not available were aligned to H37Rv with Mugsy[3]. Whole genome alignments were merged with the RGAs. Transposable elements, phage elements, and repetitive families of genes (PE, PPE, and PE-PGRS gene families) that are poorly resolved with short read sequencing were masked. Isolates with >20% missing sites were excluded from the dataset, resulting in a final alignment of 558 *M. tuberculosis* isolates from 52 countries. Single nucleotide polymorphisms (SNPs) with respect to H37Rv were extracted with snp-sites[4] resulting in 60,818 SNPs. Only sites where at least half of the isolates had confident data (i.e. non-missing) were included in the phylogenetic and phylogeographic models (3,838,249 bp, 60,787 SNPs).

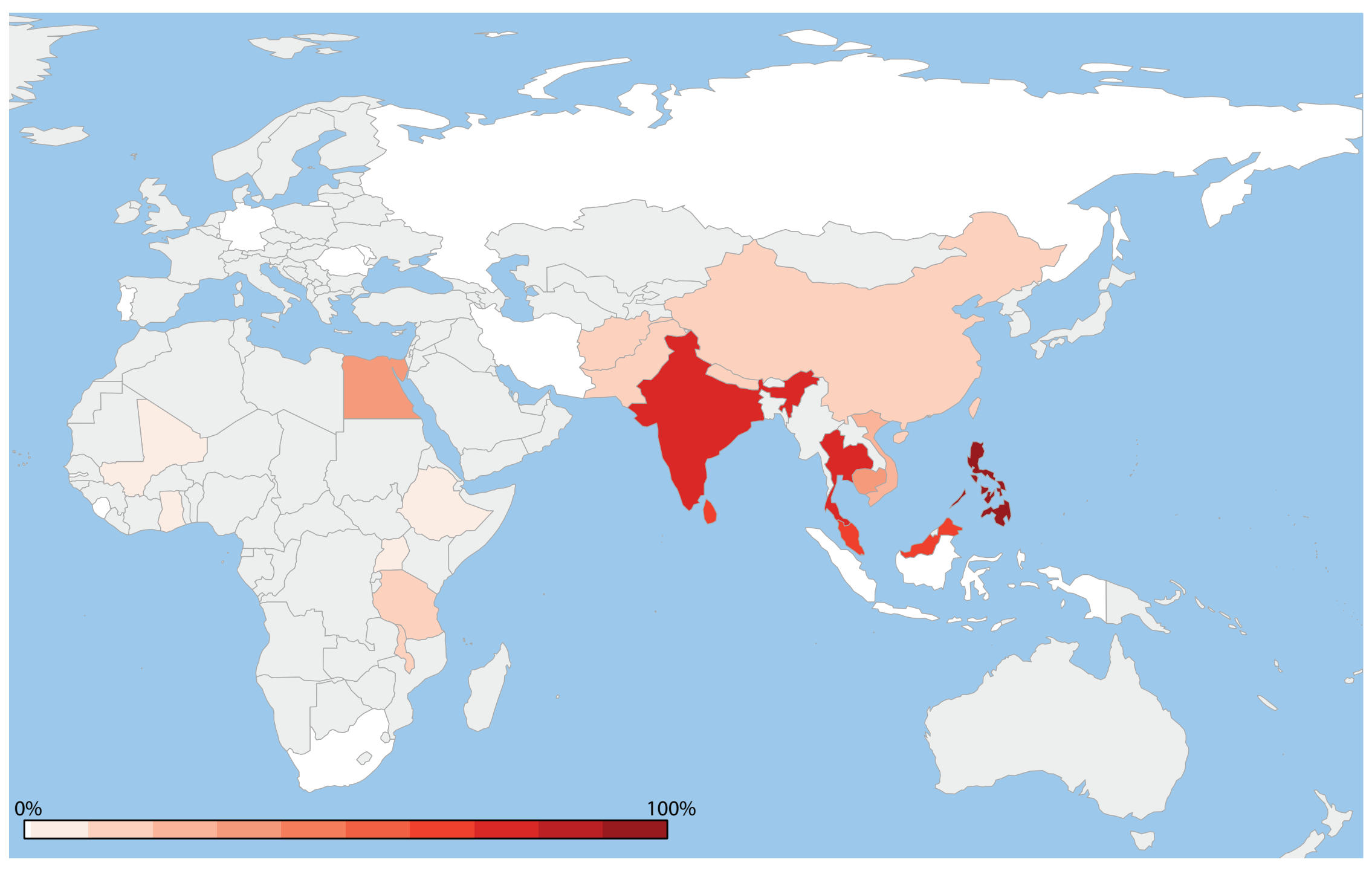
Phylogenetic Analyses



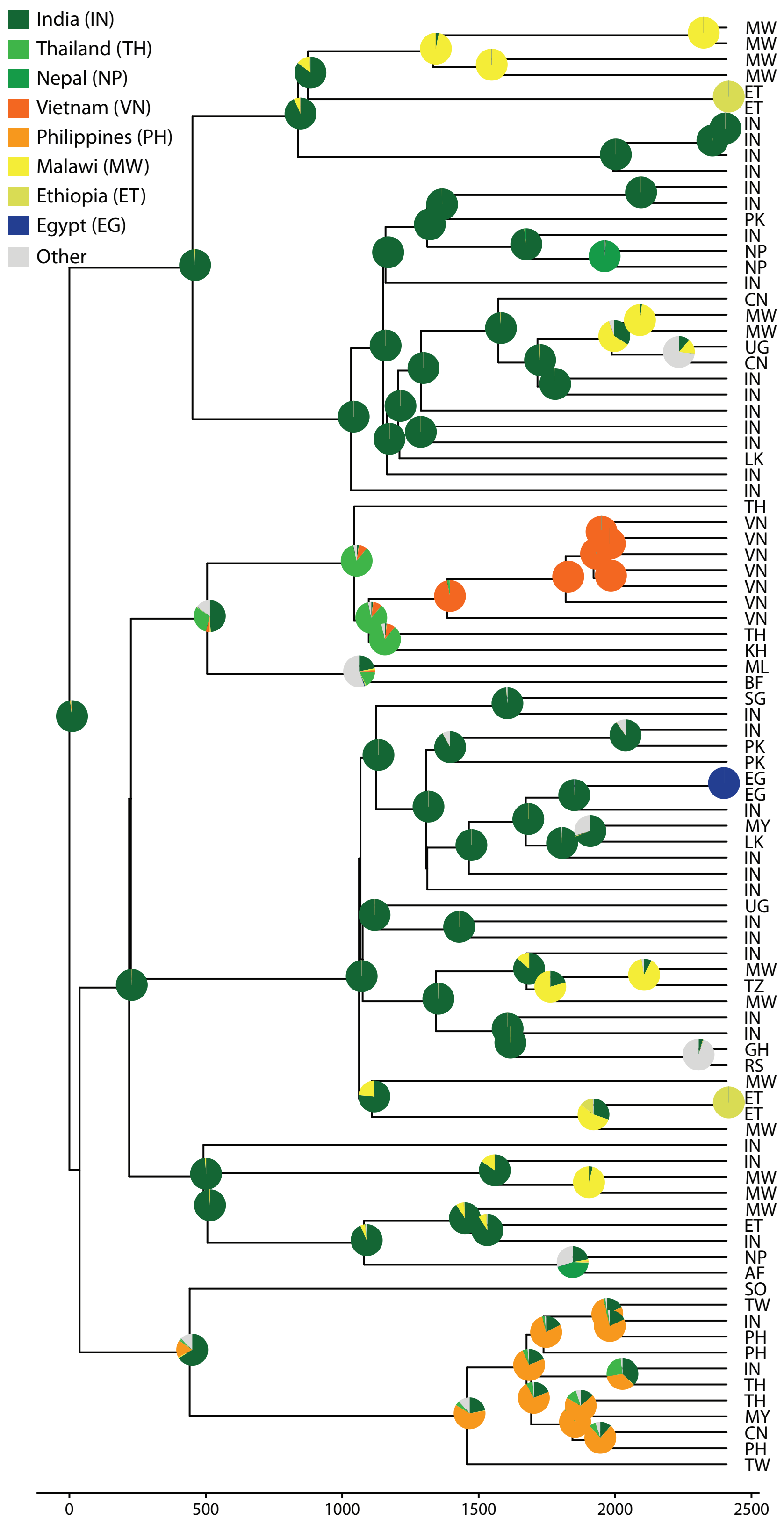
Maximum likelihood phylogeny of 'Old World' *Mycobacterium tuberculosis*. RAxML[5] analysis was performed using the General Time Reversible model of nucleotide substitution under the Gamma model of rate heterogeneity with 50 rapid bootstrap replicates. Tip color reflects the WHO geographic region of the isolate: Africa Region (AFR), Eastern Mediterranean Region (EMR), European Region (EUR), South-East Asia Region (SEA), and Western Pacific Region (WPR).

MBO is supported by the National Science Foundation Graduate Research Fellowship Program (DGE-1255259). CSP is supported by National Institutes of Health (R01AI113287).

Phylogeographic Analyses

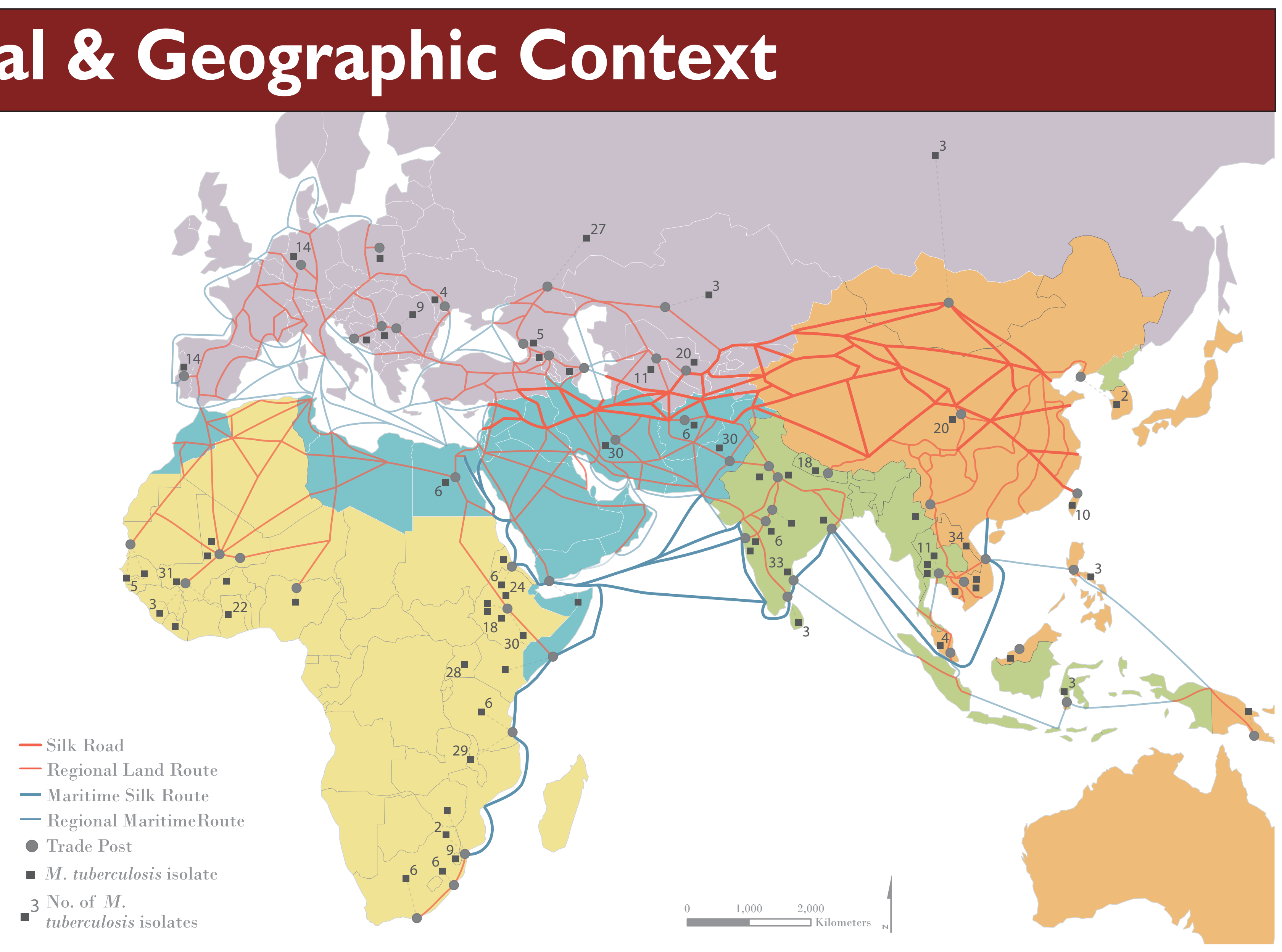


Frequency of lineage 1 in sampled countries. For each country with greater than one sample available, the percentage of isolates from that country belonging to lineage 1 are displayed. Grey countries were not sampled.



Phylogeographic reconstruction of lineage 1. BEAST[6] analysis was performed under a strict clock model with a skyline population model[7]. The data was calibrated using a rate of 5×10^{-8} substitutions/site/year.[8] Phylogeographic history was inferred using a previously described method where country of origin was modeled as a discrete trait[9]. The analysis was run with a chain of 100,000,000 states sampled every 10,000 states.

Historical & Geographic Context



***Mycobacterium tuberculosis* isolate locations and historic trade routes.** Geographic locations for each of the *M. tuberculosis* isolates were obtained from NCBI and/or the respective publications from which the isolates were first described. Data for all trade routes active throughout Europe, Africa, and Asia during 700 BCE - CE 1300 were compiled from the Old World Trade Routes Project (www.ciolek.com/owtrad.html). Trading posts, routes, and *M. tuberculosis* isolate locations were imported into ArcGIS 10.3 where isolates were assigned to the nearest trading post using the Generate Near Table tool. Projection is from World Mercator (WGS 1984) and Shapefiles are from Natural Earth (www.naturalearthdata.com). Country color reflects the WHO geographic region.

Ongoing Analyses

- Discrete phylogeography BEAST analyses with relaxed molecular clock
- Landscape-aware, continuous phylogeography BEAST analyses under strict and relaxed molecular clocks
- Correlation with described human admixture events

References

1. Cole, ST *et al. Nature* 393, 537–544 (1998); 2. Walker, BJ *et al. PLOS ONE* 9, e112963 (2014); 3. Angiuoli, SV & Salzberg, SL. Bioinformatics (2010); 4. Page, AJ *et al. Microb. Genomics* 2, (2016); 5. Stamatakis, A. *Bioinformatics* 30, 1312–1313 (2014); 6. Drummond, AJ *et al. Mol. Biol. Evol.* 29, 1969–1973 (2012); 7. Drummond, AJ *et al. Mol. Biol. Evol.* 22, 1185–1192 (2005); 8. Kay GL *et al. Nat. Commun.* 6, (2015); 9. Lemey, P *et al. PLoS Comput. Biol.* 5, (2009).