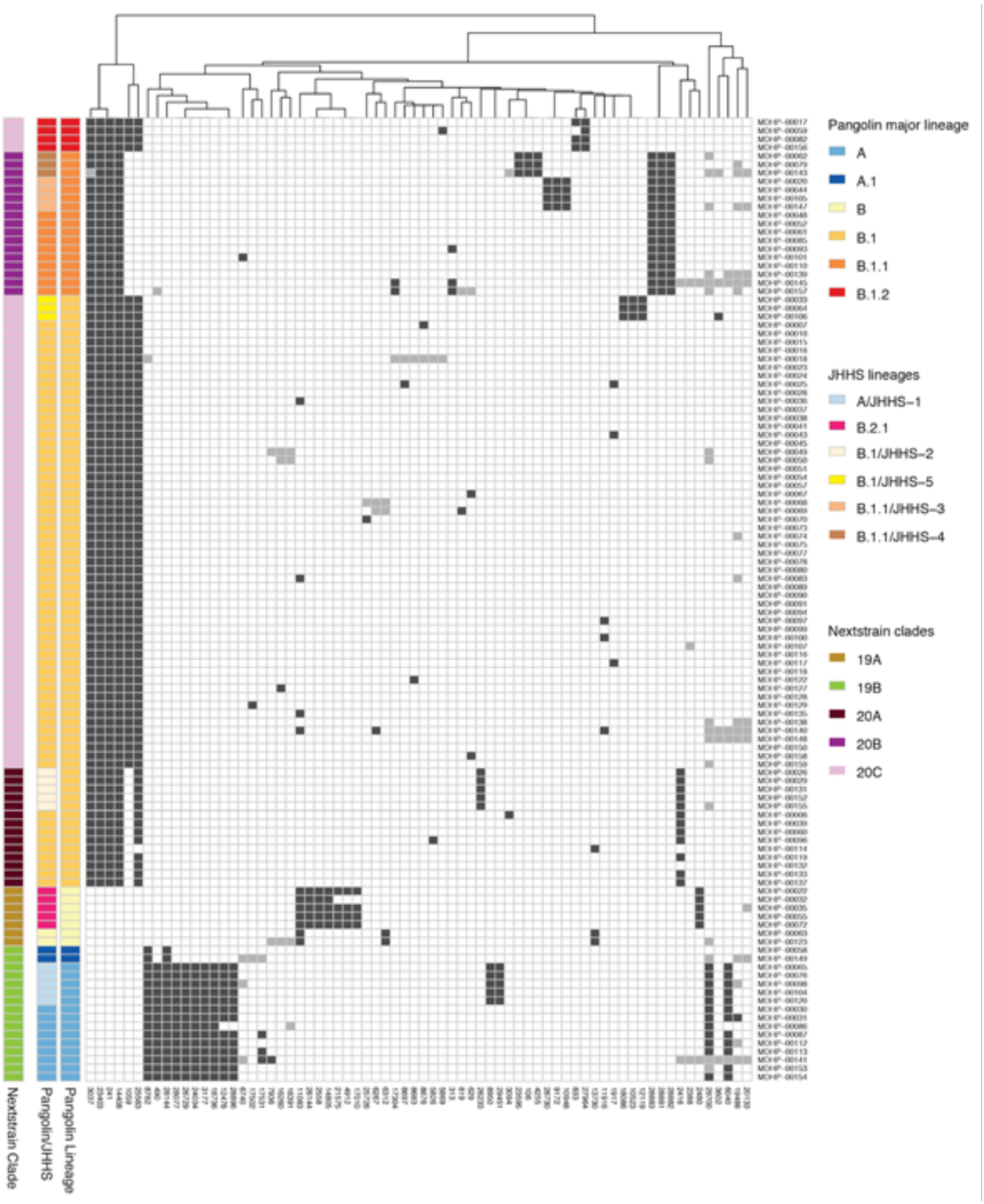
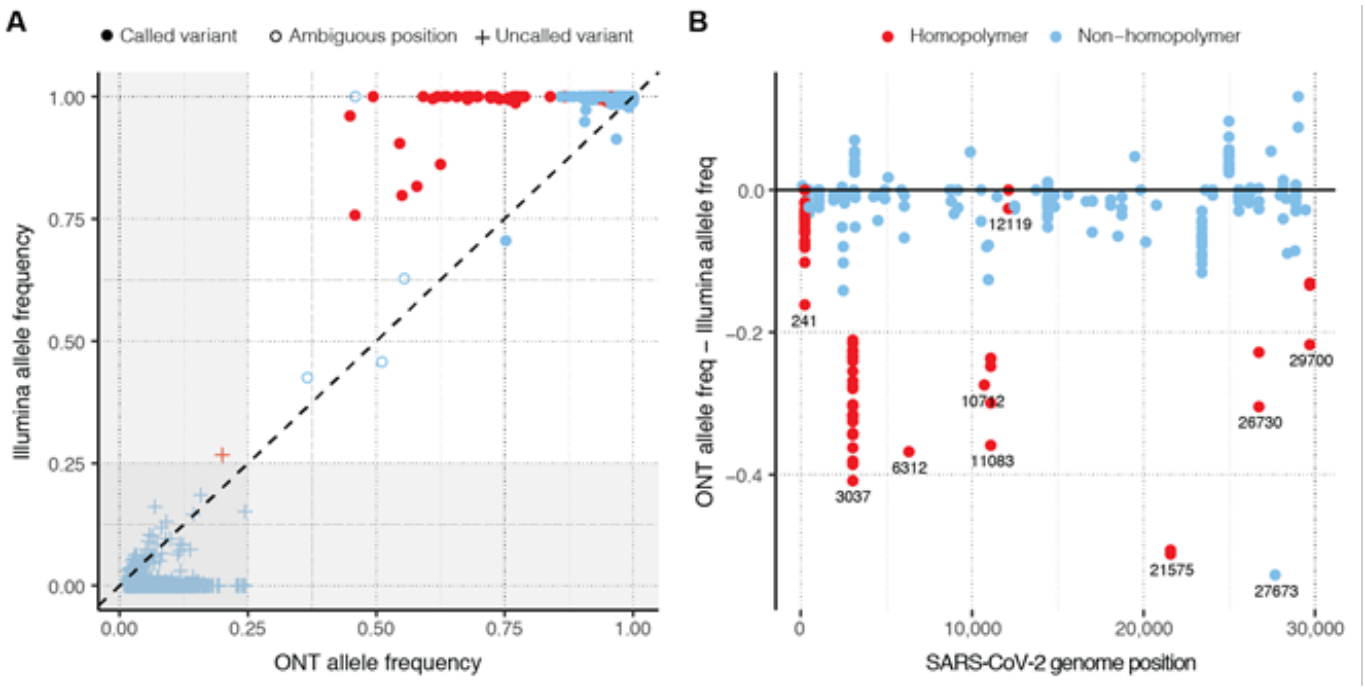
Supplemental Figures

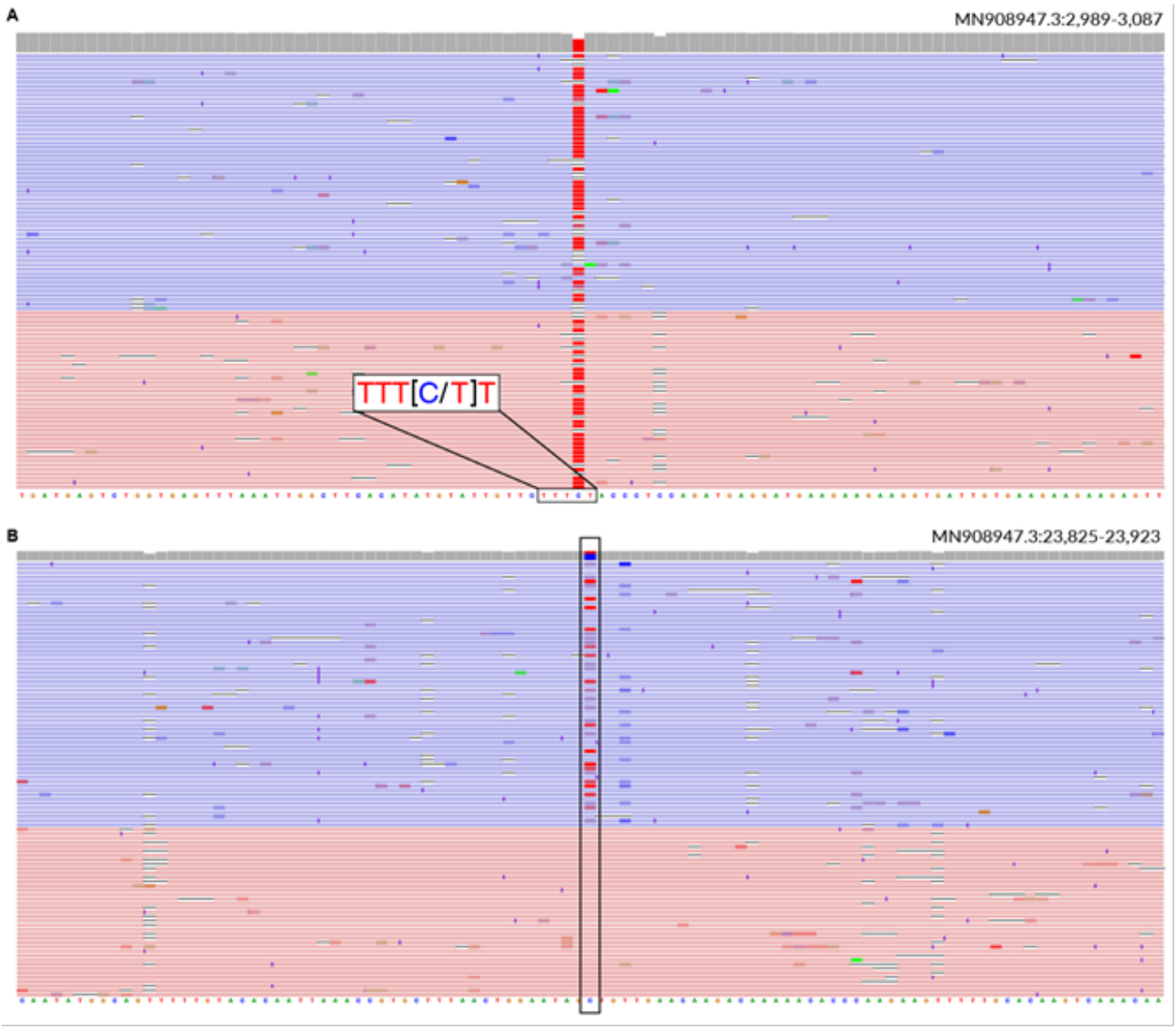
Supplemental Figure 1. Positive diagnoses by first three digits of zip code. Outlined regions correspond to areas sharing the first three digits of their zip code (Washington, DC outlined in black, all others grey). Each region is shaded by the number of patients with positive COVID-19 diagnoses in March 2020 (from the Johns Hopkins Health System) reporting home residence in that region. Regions with 1–5 positive diagnoses are shown as 5.



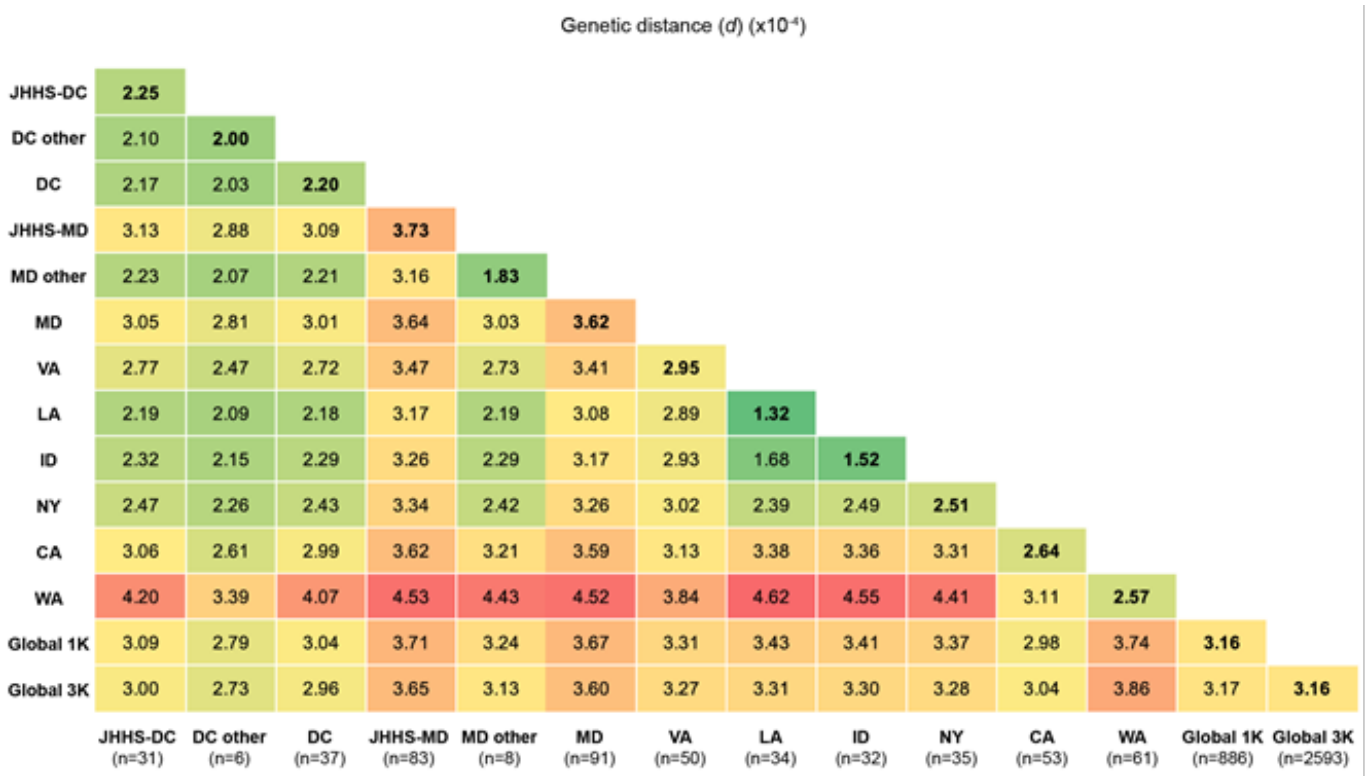
Supplemental Figure 2. Single nucleotide polymorphisms present in three or more JHHS sequences. Rows = samples sequenced as part of this study (114), sorted by Pangolin lineage. Columns = positions with consensus variants in three or more sequences. Black cells = variant present; grey cells = ambiguity present. Columns are clustered using complete linkage hierarchical clustering. JHHS lineages are sub-groups of pangolin lineages in which two or more variants are shared between three or more JHHS sequences. In one case (B.2.1), the JHHS sub-lineage is captured by a non-major pangolin lineage designation.



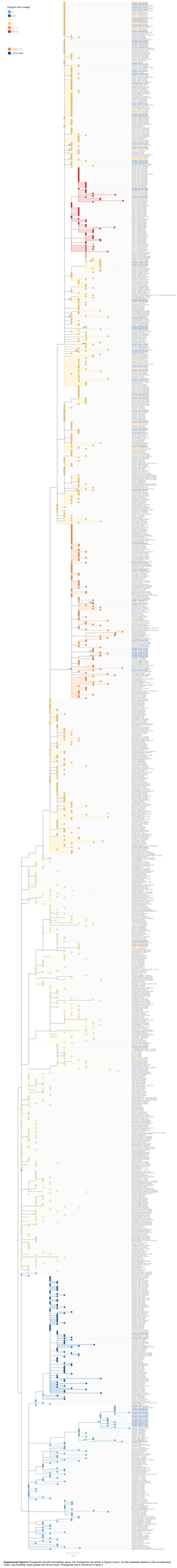
Supplemental Figure 3. Illumina and Oxford Nanopore allele frequency comparison. (**A**) Variant frequencies from tiled amplicons sequenced on the Oxford Nanopore (ONT) and Illumina platforms. All variants at >0.02 allele frequency are shown. Uncalled variants are those not present in the consensus sequences from these samples. Grey shaded regions represent frequency threshold (25%) used to automatically reject candidate variants. Almost all discrepancies in frequency occur within homopolymer regions (red symbols). (**B**) Difference between allele frequencies across the two sequencing platforms. Position 3,037 (within a T-homopolymer) accounts for most large discrepancies in frequency.



Supplemental Figure 4. Common Oxford Nanopore sequencing issues. (**A**) Example of a T-homopolymer region in the SARS-CoV-2 genome (sample MDHP-00028). A large number of reads have deletions, resulting in a mixed alternate allele frequency. (**B**) Example of strand bias in Oxford Nanopore sequencing reads (sample MDHP-00028). A mutation only occurs in minus strand reads.



Supplemental Figure 5. Evolutionary divergence in geographic regions. Pairwise matrix containing average pairwise genetic distances within (bolded in the diagonal) and between sequences from various geographic regions including: the District of Columbia (JHHS-DC + DC other), Maryland (JHHS-MD + MD other), Virginia (VA), Louisiana (LA), Idaho (ID), New York (NY), California (CA), Washington (WA), and two representative global subsamples (Global 1K and 3K). Green-to-red color scale highlights the most (red) and least (green) divergent within or between average distances.



Supplemental Figure 6. Phylogenetic tree with all bootstrap values >65. Phylogenetic tree shown in Figures 3 and 4, but with bootstraps labeled on their corresponding nodes. Only bootstrap values greater than 65 are shown. Phylogenetic tree is colored as in Figure 4.

Supplemental Tables

Supplemental Table 1. Aggregate JHHS diagnostic tests by date. Number of positive tests within the Johns Hopkins Health System on each day during March 2020.

Supplemental Table 2. Aggregate sample metadata. Clinical metadata for all tested samples within the Johns Hopkins Health System, for all 143 samples sequenced in this study, and for the 114 patient samples that produced complete genomes. Results are grouped into appropriate bins to protect patient privacy.

Supplemental Table 3. Sequencing metrics. Sample names and accession numbers (where relevant) for all 143 samples selected for sequencing alongside sequencing metrics. Clade assignments using the Pangolin and NextStrain nomenclature systems are also provided for each sample that resulted in a whole genome sequence. One sample (MDHP-00146) was submitted to public repositories prior to enforcing a strict threshold of at least 27,000 unambiguous nucleotides, and therefore has an accession number but was not included in this analysis.

Supplemental Table 4. Single nucleotide polymorphisms present in JHHS sequences. Single nucleotide polymorphisms (with annotations) present in the 114 complete genomes generated in this study, alongside the number of samples containing each variant. Sites with IUPAC ambiguity codes due to mixed allele frequencies are also included.

Supplemental Table 5. Accession numbers of samples used in phylogenetic analysis. Accession numbers and submitting lab information for all 1279 sequences included in our phylogenetic tree.

Supplemental Table 6. Accession numbers of samples used in genetic distance analysis. Geographic group assignments for all sequences used in genetic distance analysis, listed by accession number. This table includes accession numbers for sequences included in the Global 3K dataset but were not included in the phylogenetic tree (which is based on the Global 1K dataset).

Supplemental Table 7. Quartile values for pairwise genetic distances within each geographic region. Mean, median, min, max, Q1, and Q3 values for the pairwise genetic distance distributions of each geographic group used in our analysis.