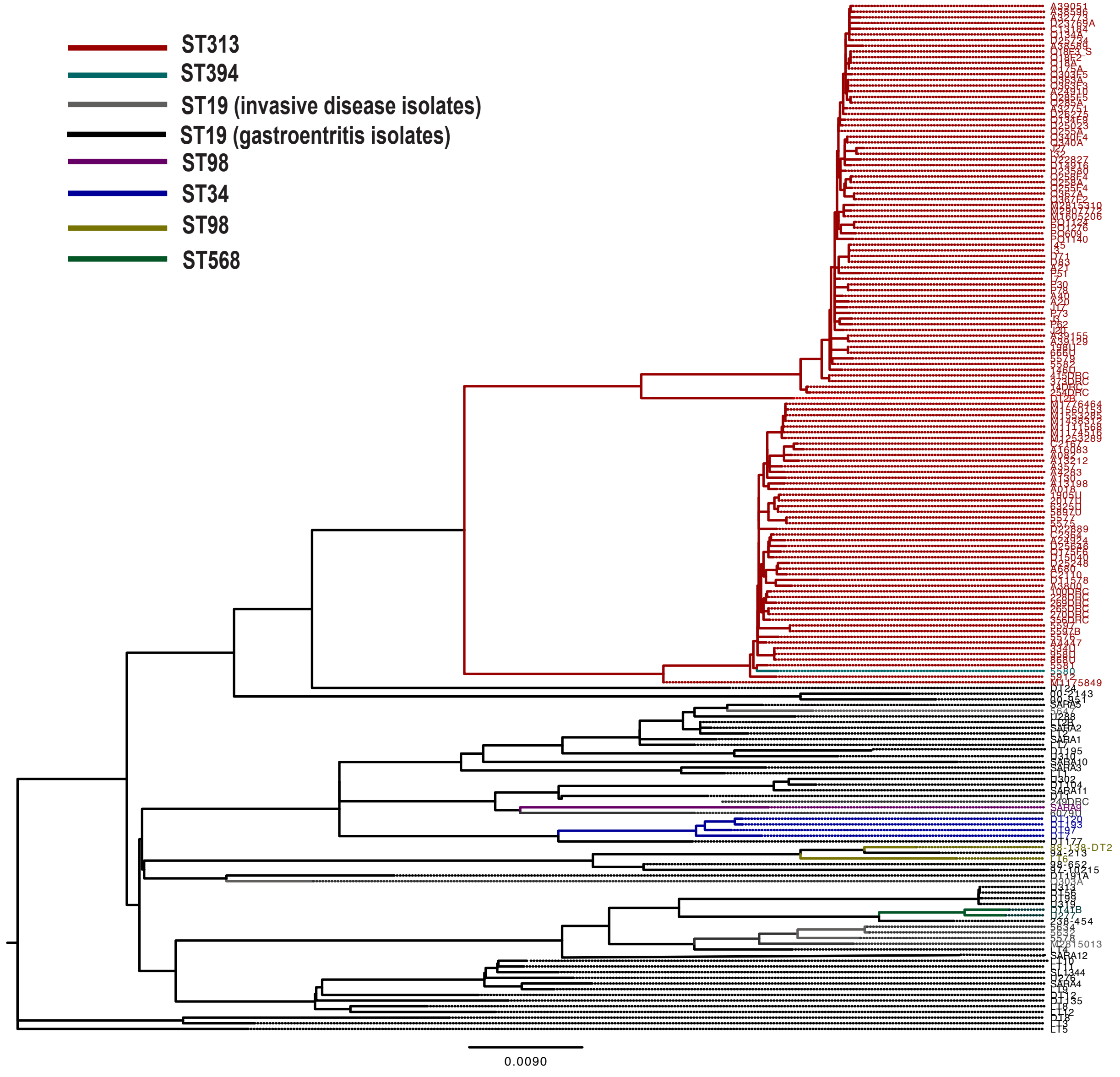
SUPPLEMENTARY INFORMATION

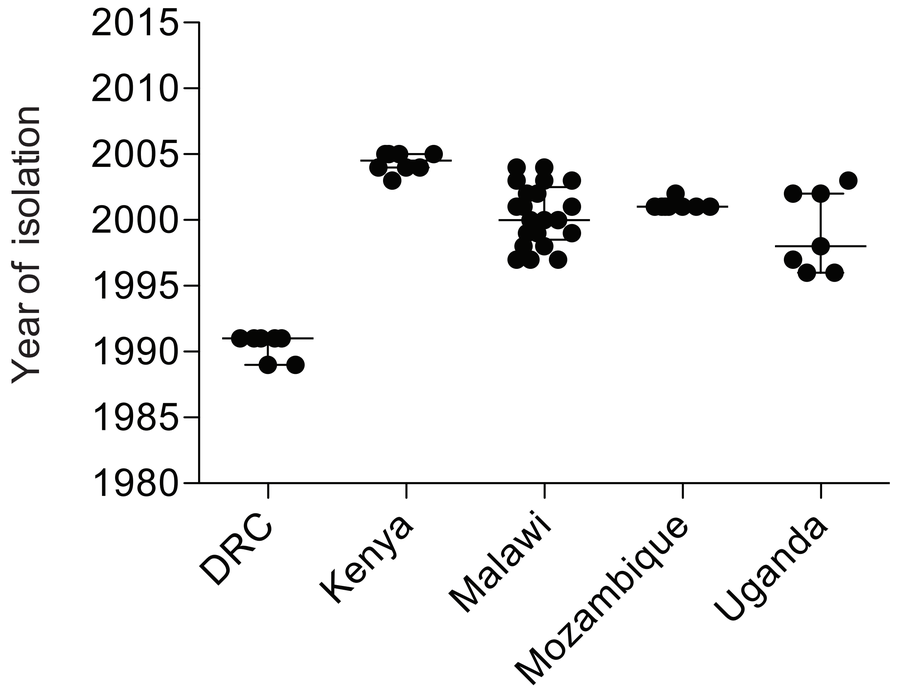
Intra-continental spread of human invasive *Salmonella* Typhimurium pathovariants in sub-Saharan Africa

Chinyere K. Okoro, Robert A. Kingsley, Thomas R. Connor, Simon R. Harris, Christopher M. Parry, Manar N Al-Mashhadani, Samuel Kariuki, Chisomo L. Msefula, Melita A. Gordon, Elizabeth de Pinna, John Wain, Robert S. Heyderman Stephen Obaro, Pedro L. Alonso, Inacio Mandomando, Calman A. MacLennan, Milagritos D. Tapia, Myron M. Levine, Sharon M. Tennant, Julian Parkhill, Gordon Dougan.



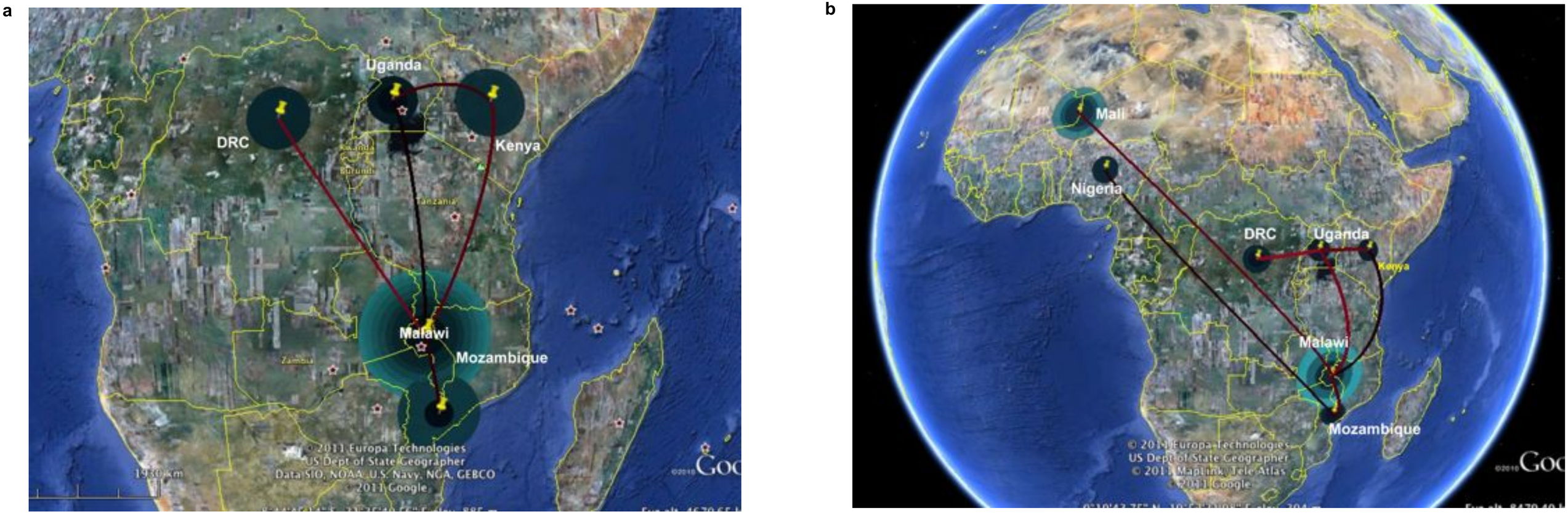
Supplementary figure 1. MLST distribution of *Salmonella* Typhimurium isolates used in the study. Scale bar represents number of SNPs per site.

a

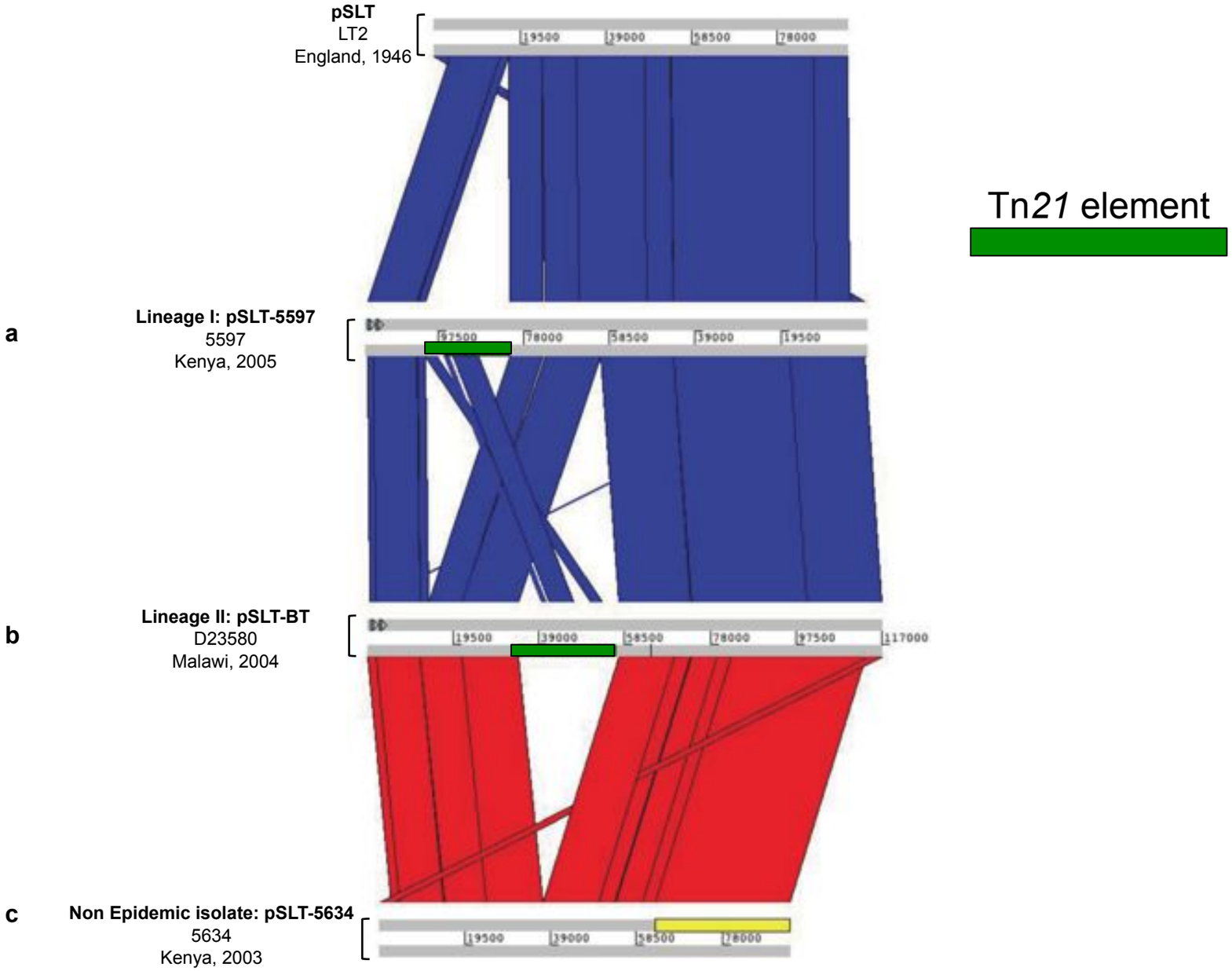


b

Supplementary figure 2. **Validation tests for the origin of lineage I. A**. Numbers of isolates from each sampled country plotted against the dates of isolation. Lines within the points in the graph show sample spread. Top line indicates the youngest isolates, bottom line, the oldest isolates and middle line indicates the median age of the group. **B**. Panel of 25 maximum clade credibility (MCC) trees showing phylogeographical reconstruction of lineage I, using 10 Malawian isolates selected randomly, 25 times. To the upper left of the trees are the root location state posterior probability distributions among the 5 different locations included in the analyses. Color-coding is the same as in Figure 2
.



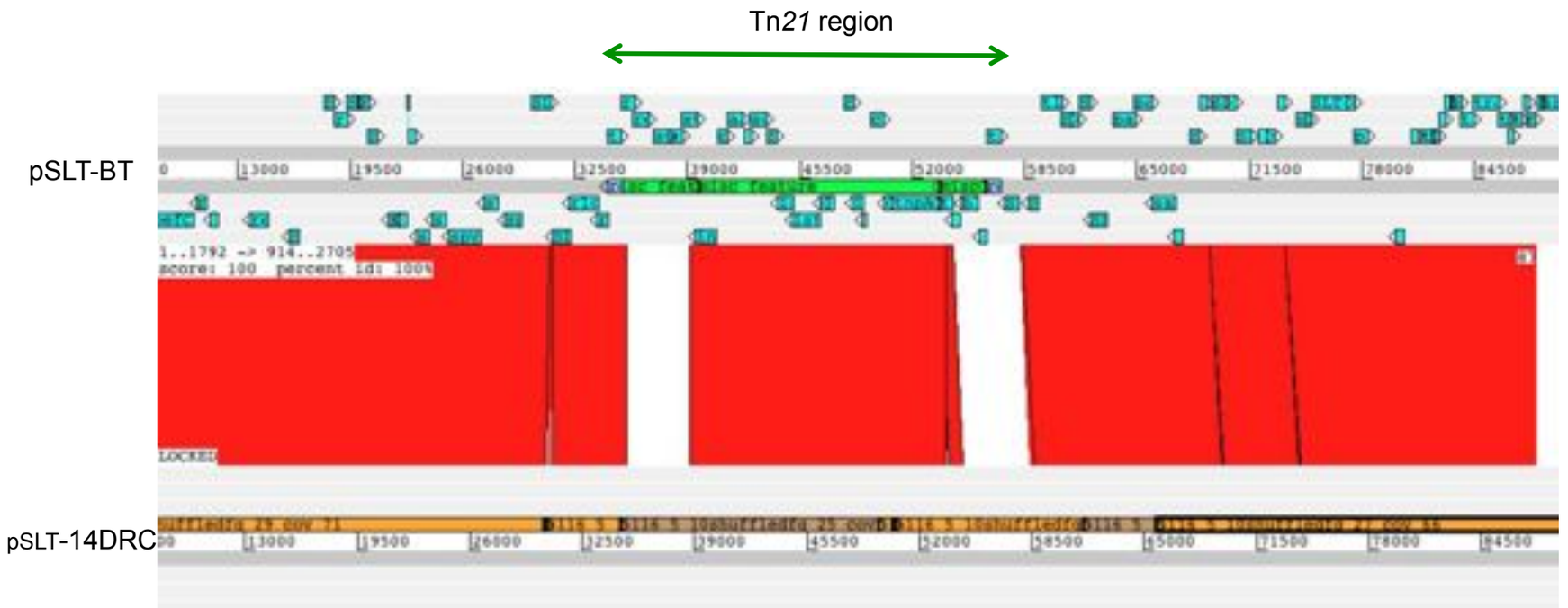
Supplementary figure 3: Geospatial dispersal of invasive *Salmonella* Typhimurium isolates in sub-Saharan Africa. Phylogeographical diffusion of lineages 1 (A) and 2 (B) across sub-Sahara Africa over time based on a discrete geospatial model with associated geographical coordinates. Discrete location states are annotated to tree nodes and branches that indicate location changes are represented on the map. Mapped objects are converted to the keyhole mark-up language (KML), using the SPREAD software and viewed in Google Earth. Colors of circular polygons are indicative of the number of lineages with the discrete state (location) at the given time point. Dark to light color gradients of polygons represent increasing number of lineages. The color gradient (red – dark red) shows ages (older – recent) of transmission lines.



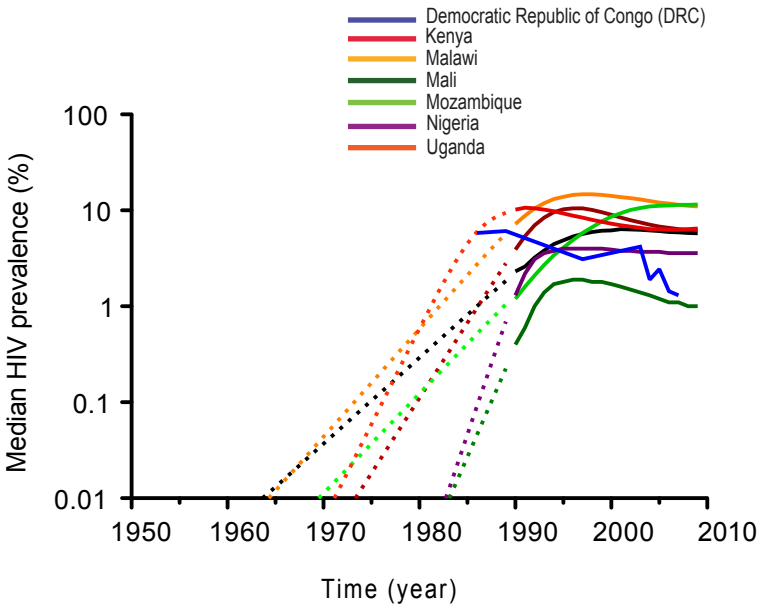
Supplementary figure 4. Comparison of Tn*21*-loci in representative isolates from different invasive lineages showing insertion of Tn*21* elements in two different loci in lineages I (A) and II (B) isolates and absence of the element in non-epidemic isolate (C), relative to the plasmid sequence of pSLT (top panel).

a

b



Supplementary figure 5. **Sequence comparisons of Tn*21* loci in the plasmids of 14DRC and D23580**. **A**. Percentage identity plots of multiple alignments of assembled plasmid sequences from invasive isolates D23580 (pSLT-BT), 14DRC(pSLT-14DRC) and 5597(pSLT-5597) and the plasmid of non-human invasive LT2 (pSLT). The alignment spans 100% of the nucleotides of the reference, pSLT-BT (top coordinate). The percentage identity of each of the three genomes being compared to pSLT-BT ranges from 50 – 100 (bottom to top). The dotted lines indicate sequences of Tn*21* elements on the plasmids of isolates associated with invasive disease and absence in strain pSLT. **B.** ACT comparison of Tn*21* loci in the plasmids of lineage II isolates 14DRC and D23580, showing nucleotide sequence match between the complete six-frame translations. The red blocks represent individual matches. Annotated genes in pSLT-BT are shown as blue with arrows direction representative of gene orientation.



Supplementary figure 6. **Percentage prevalence of HIV in sampled countries from 1960 – date**. HIV prevalence is defined here as the percentage of men and women between the ages of 15 and 49 who are HIV positive (UNAIDS, 2010). Dotted lines show predicted HIV prevalence values before monitoring and reportage for the different countries extrapolated backwards in time to 1960 (as previously described in Figure 2c). Prevalence data is plotted on a logarithmic scale to highlight changes in prevalence prior to 1990.

SUPPLEMENTARY NOTE

Sequence alignment and analyses of Tn*21* element and *cat* gene from 14DRC

We compared the assembled sequence of the Tn*21* elements from isolates 14DRC (DRC) and D23580 (Malawi) in order determine the extent of similarity of this element and to that found in lineage II clones. We used pSLT-BT (FN4320231), the plasmid sequence D23580 as a representative of this lineage. Sequence reads from 14DRC mapped[1] across the length of 70% (19/27) of the genes in pSLT-BT, thus indicating the presence of a similar but somewhat reduced element in this isolate. Pair-wise sequence alignment and visualisation of the draft assembly and complete sequence of Tn*21* elements in both isolates using the software programs Mugsy[2] and GmaJ[3], respectively, reveal a high level DNA sequence synteny between the regions under comparison in the two strains (Supplementary Fig. 5) This observation is consistent with our phylogenetic analyses which shows an acquisition of a Tn*21-like* element by isolates circulating around the same area and period very early on in the genealogy of the lineage. The presence of a *cat* gene on this 14DRC-Tn*21*-like element, with 62% sequence similarity to that in pSLT-BT also confirms our findings that the selective acquisition of these transposable elements and the *cat* gene by lineage II clones was precipitated by chloramphenicol usage.

SUPPLEMENTARY REFERENCES

1. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
2. Angiuoli, S.V. & Salzberg, S.L. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* **27**, 334-342 (2011).
3. Blanchette, M.*, et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**, 708-715 (2004).