Supplementary Information

Title: Hotspots of genetic change in *Yersinia pestis*

Authors: Yarong Wu[1#], Youquan Xin[2#], Xiaoyan Yang[2#], Kai Song[1], Qingwen Zhang[2], Haihong Zhao[2], Cunxiang Li[2], Yong Jin[2], Yan Guo[1], Yafang Tan[1], Yajun Song[1], Huaiyu Tian[3], Zhizhen Qi[2]*, Ruifu Yang[1]*, Yujun Cui[1]*

[1] State Key Laboratory of Pathogen and Biosecurity, Academy of Military Medical Sciences, Beijing, 100071, China

[2] Key Laboratory of National Health Commission on Plague Control and Prevention, Key Laboratory for Plague Prevention and Control of Qinghai Province, Qinghai Institute for Endemic Disease Prevention and Control, Xining, 811602, China

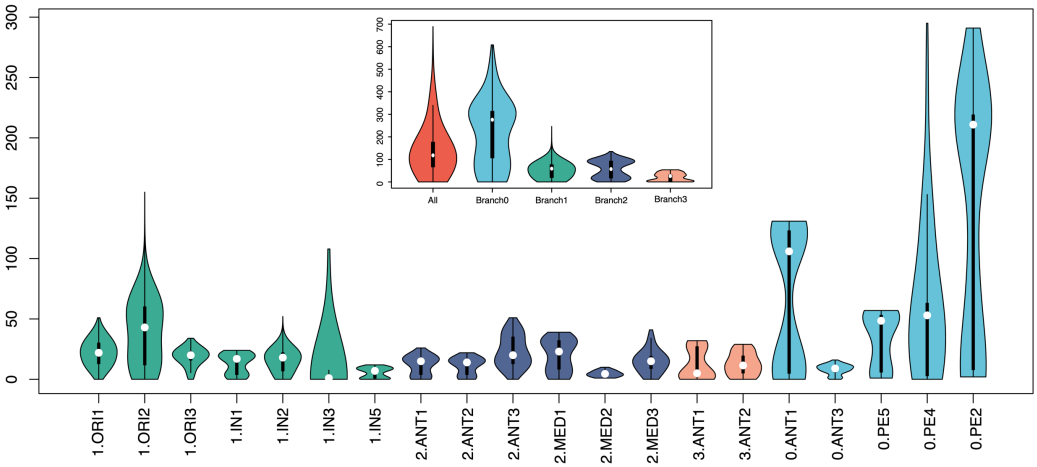
[3] State Key Laboratory of Remote Sensing Science, Center for Global Change and Public Health, Beijing Normal University, Beijing, 100875, China

[#] These authors contributed equally: Yarong Wu, Youquan Xin, Xiaoyan Yang

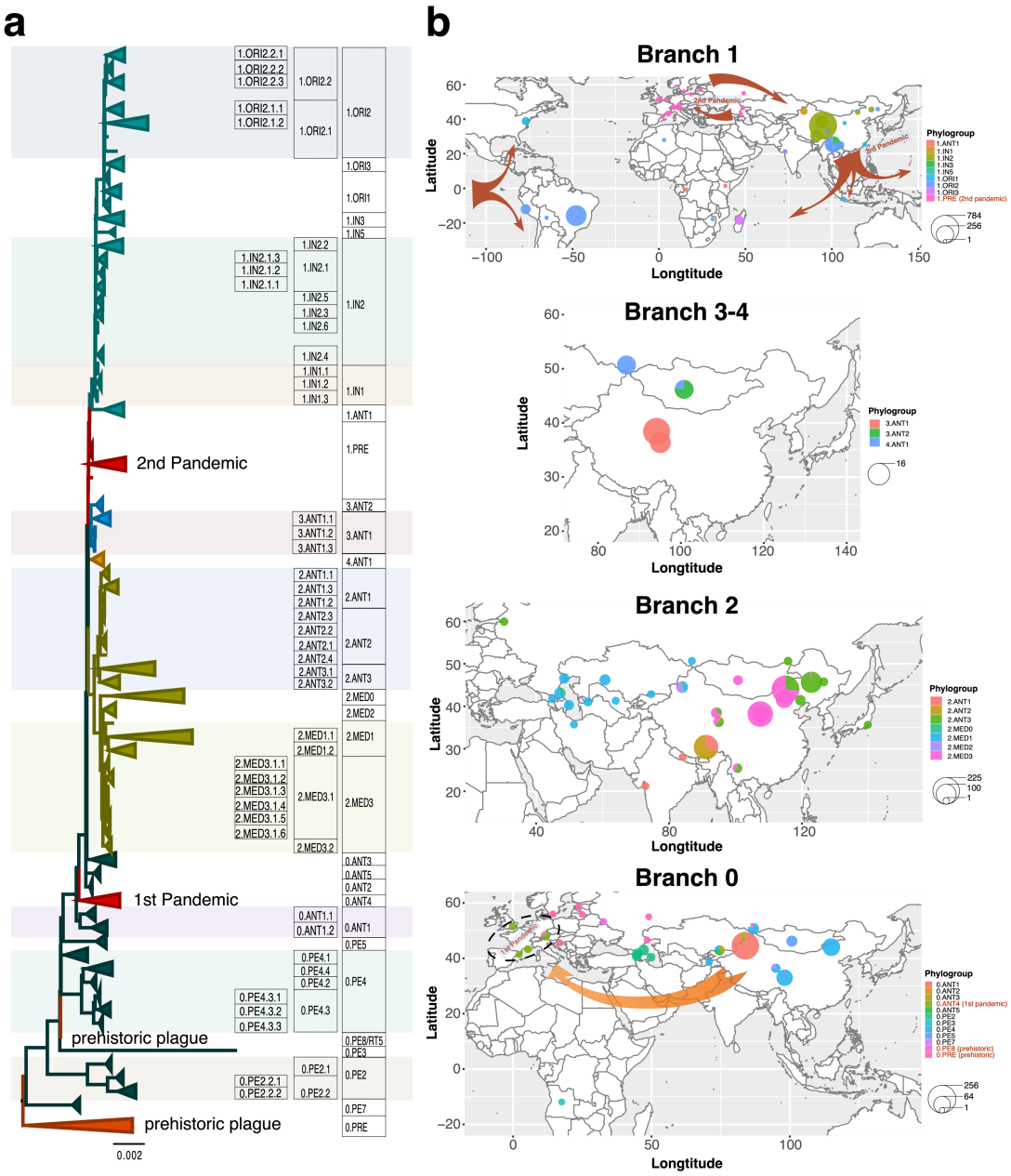
* Correspondence: cuiyujun.new@gmail.com (Yujun Cui); ruifuyang@gmail.com (Ruifu Yang); qzz7777@163.com (Zhizhen Qi)

Content

Supplementary Figures

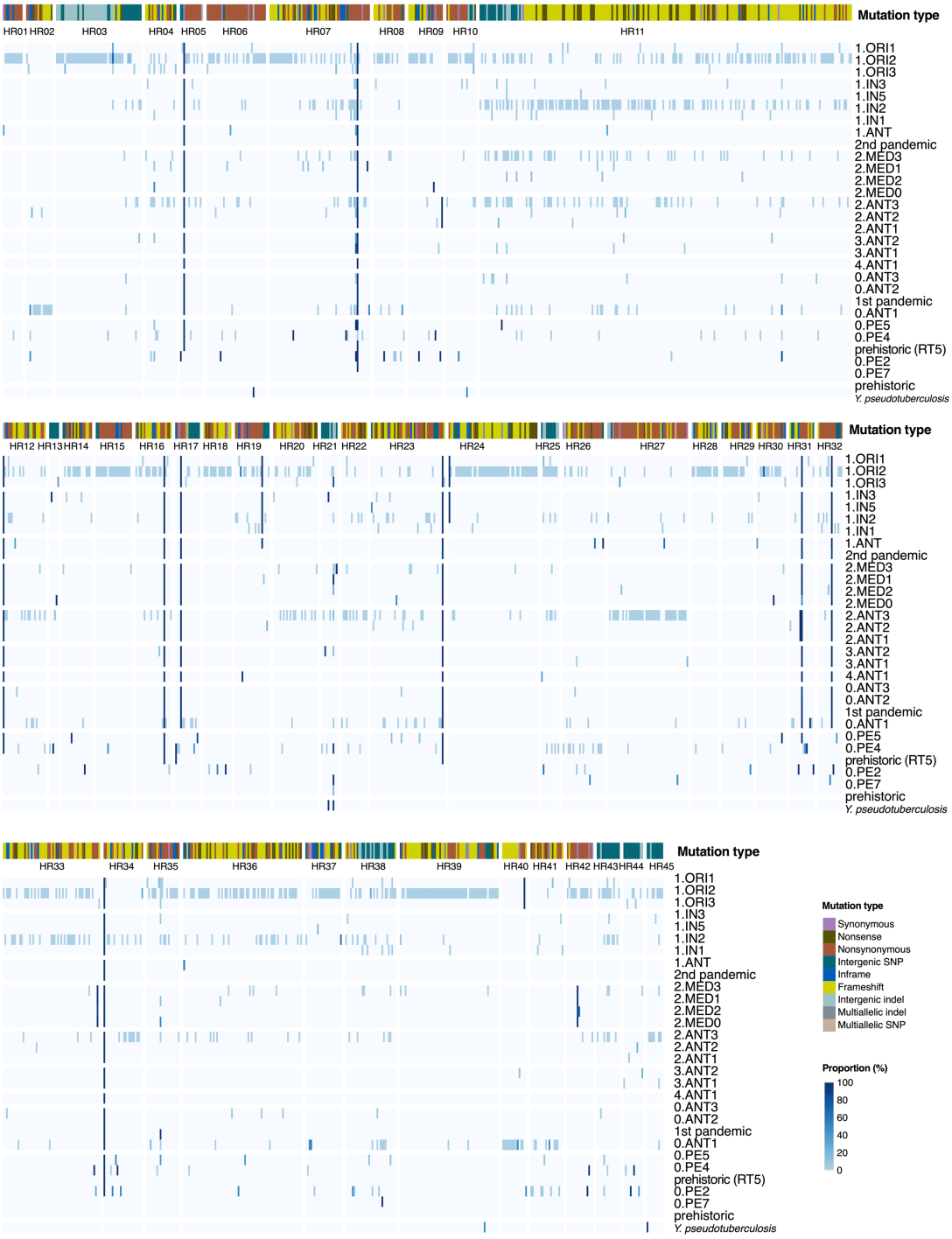


Supplementary Fig. 1 Pair-wise SNP distances for 3,318 *Y. pestis* strains. Branches and phylogroups with fewer than 5 genomes were excluded. In the violin plot, the white dot represents the median, the thick black bar shows the interquartile range (IQR), and the thin black line indicates the range of the data within 1.5× IQR. The plot width reflects data density, and the curved boundaries represent outliers beyond 1.5× IQR. Source data are provided as a Source Data file.

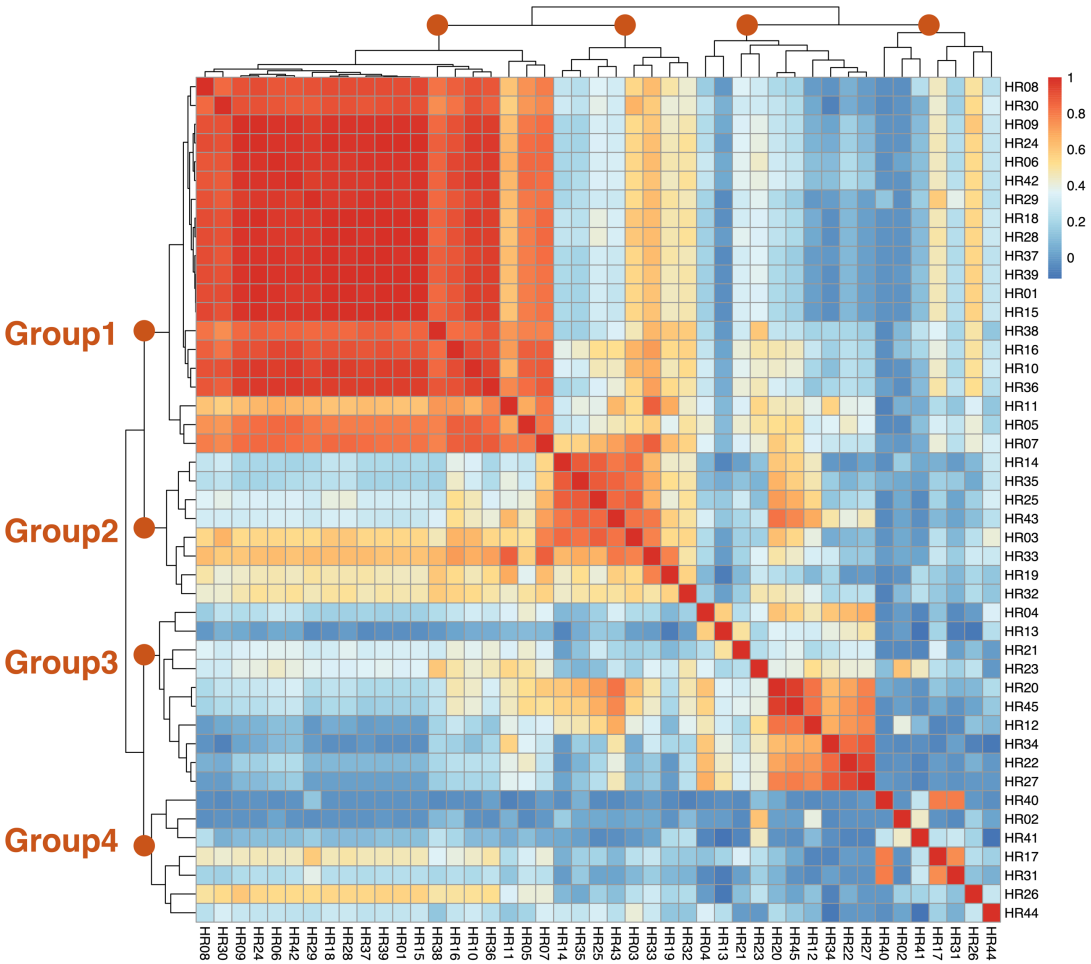


Supplementary Fig. 2 ***Y. pestis*** **three-level hierarchical nomenclature system and geographic distribution of five major branches. a,** A maximum likelihood (ML) tree derived from core genome SNPs (shared by ⩾95% strains) of 3,575 *Y. pestis* strains. **b,** Geographic distribution of five major branches, with arrows showing potential transmission routes of three historical pandemics. The maps were created using the ggplot2 and maps packages in R.

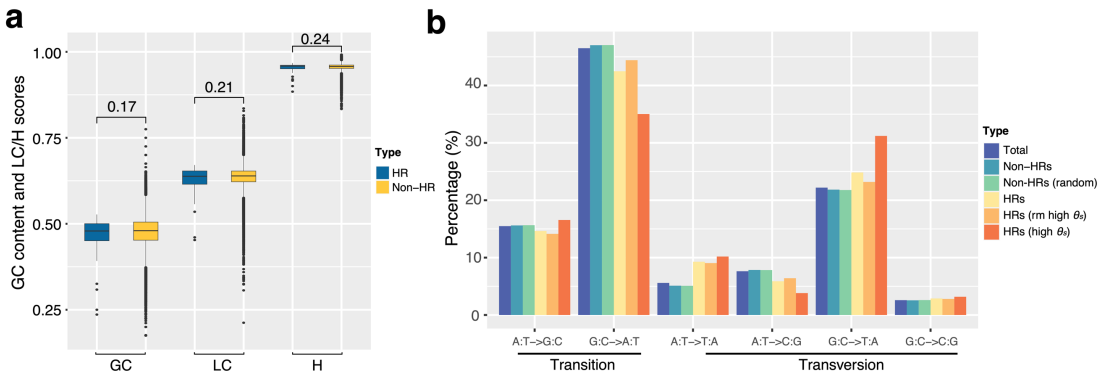
Supplementary Fig. 3 Genomic distribution of 45 variation hot regions (HRs). a, Five HRs confined to specific genes. Horizontal arrows represent genes within or adjacent to HRs, with gray shading marking HRs. Lines between arrows signify intergenic regions. Arrow direction indicates positive or negative strands. A consistent scale is applied to all 45 HRs, with arrow length proportional to gene length. Short vertical lines in various colors within genes and intergenic regions indicate different mutation types. **b,** Six HRs confined to intergenic regions. **c,** Thirty-four HRs targeted multigenic regions. Source data are provided as a Source Data file.



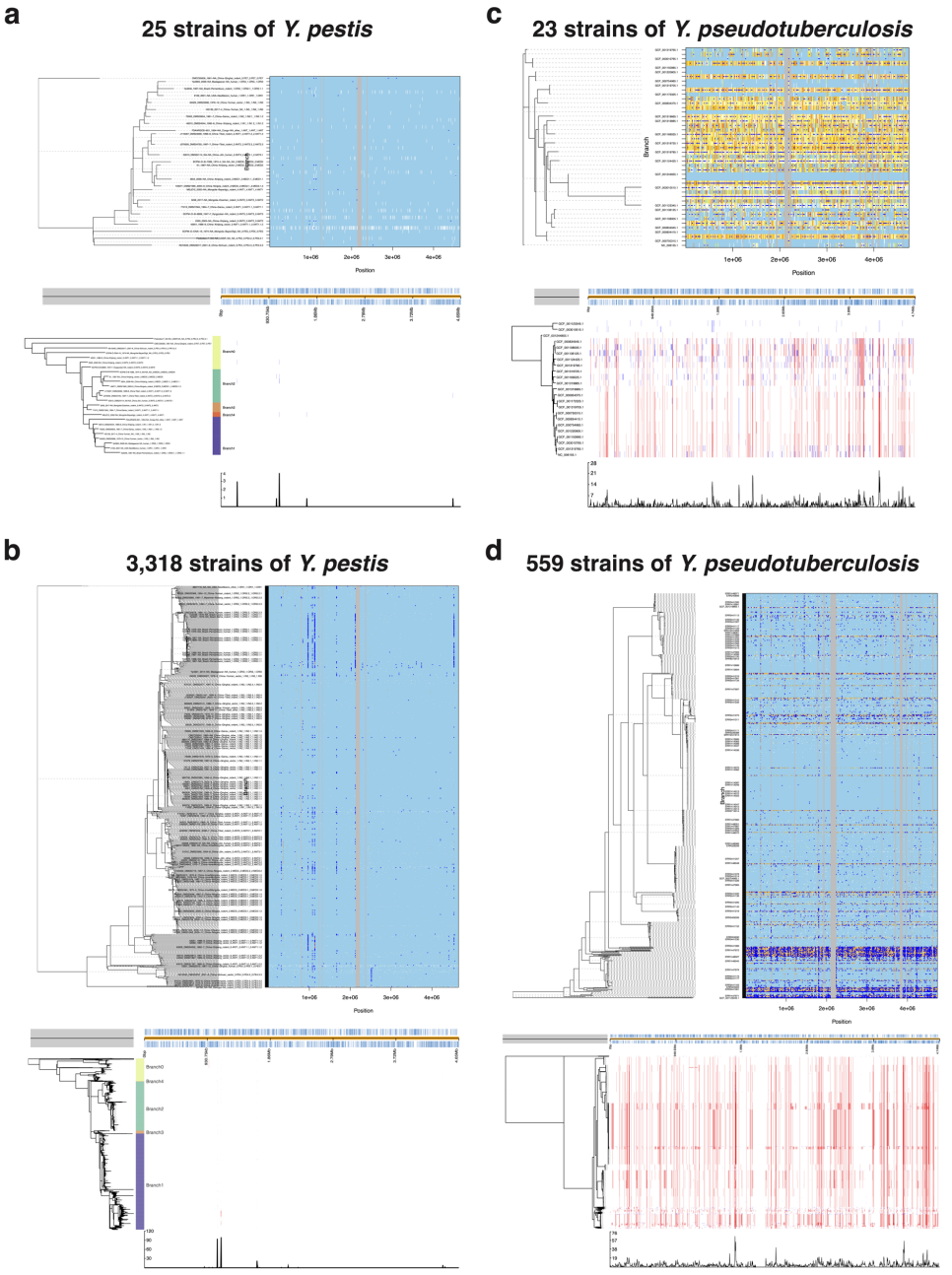
Supplementary Fig. 4 Distribution of mutation sites in 45 HRs across *Y. pestis* phylogroups. In the heatmap, each column corresponds to a mutation site, while each row represents a *Y. pestis* first-order phylogroup. The color intensity in the squares indicates the proportion of strains with mutations in each HR relative to the total number of strains in the respective group, with darker colors representing higher percentages. Colored vertical lines above indicate the corresponding mutation types. Source data are provided as a Source Data file.



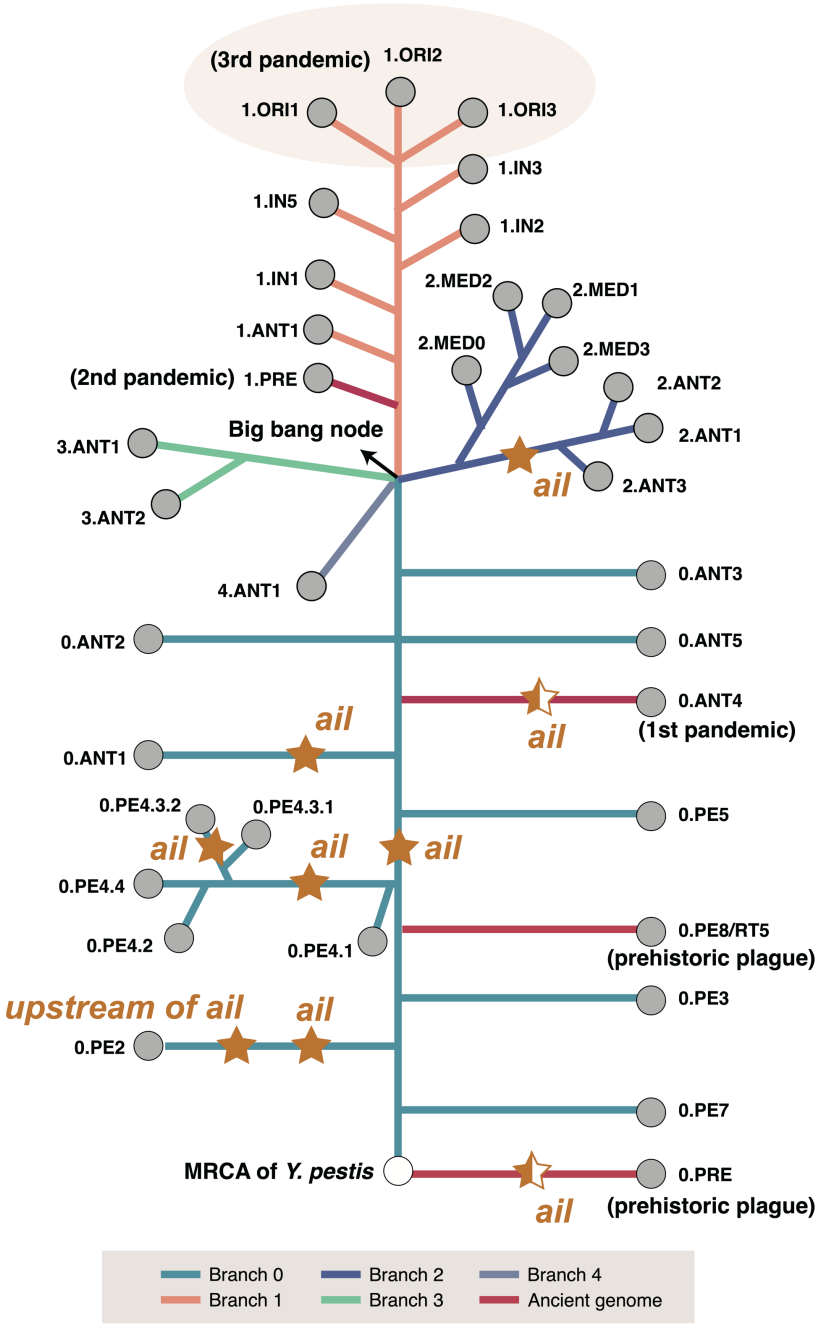
Supplementary Fig. 5 Non-random association analysis of 45 HRs. The association was measured using Pearson's correlation coefficient (*r*) based on the phylogroup proportion of the 45 HRs. Source data are provided as a Source Data file.



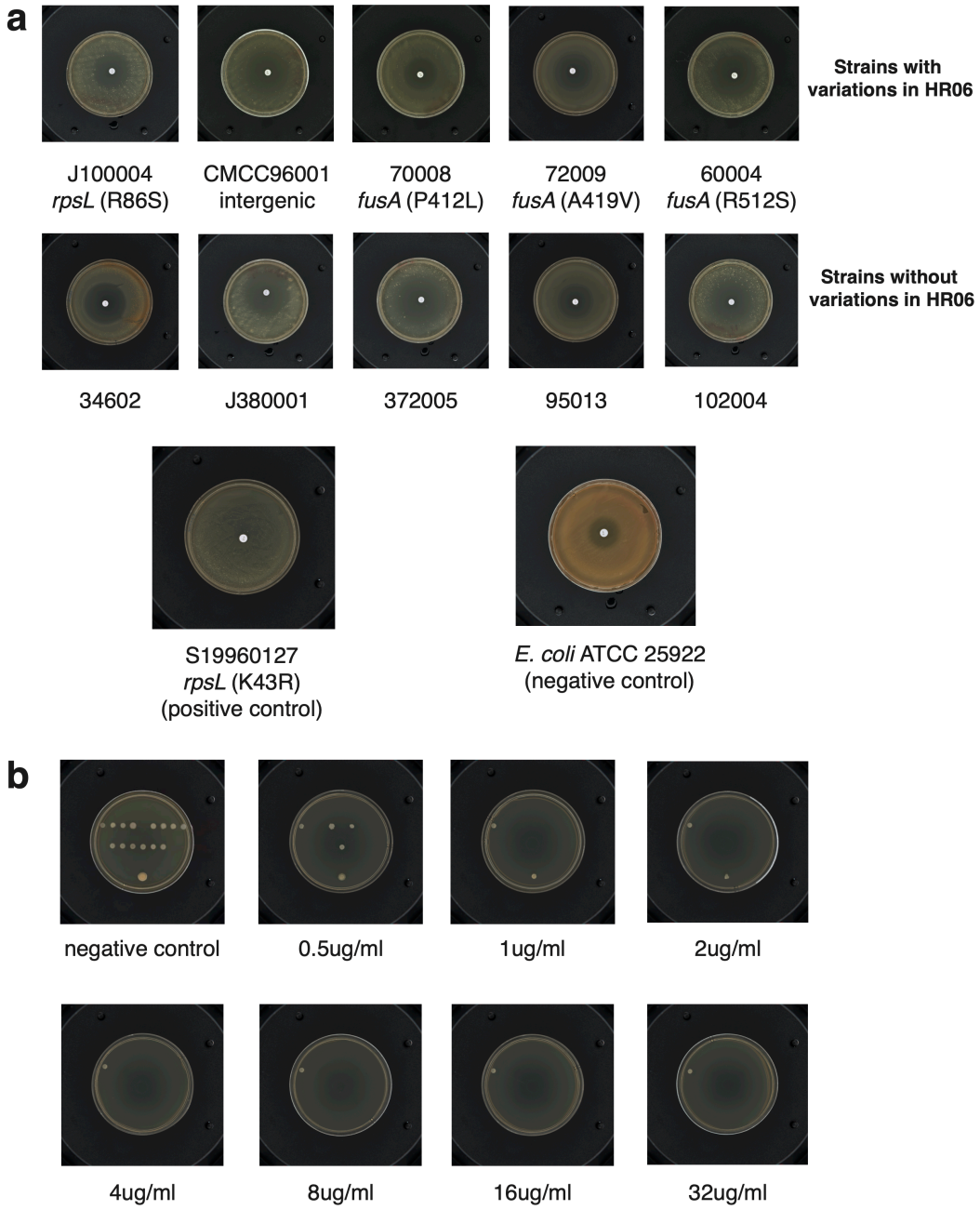
Supplementary Fig. 6 Comparison of genomic context and mutation spectrum between mutation hotspots and non-hotspots. a, Comparison of GC content, linguistic complexity (LC score), and Shannon entropy (H score) between hot regions (HRs) and non-HRs. Box plots display the median, Q1 and Q3 quartiles, whiskers at 1.5× IQR, and outliers as individual points. Welch's t-test *P*-values are annotated above. Source data are provided as a Source Data file. **b,** Distribution of mutation spectrum. The x-axis represents different types of transitions and transversions, while the y-axis indicates the proportion of each mutation type relative to the total number of mutations in each group. Different groups are represented by different colors. The "Non-HR (random)" group refers to the mean value derived from 1000 random samples taken across the genomic regions outside of the HRs, with each sample containing the same number of SNP sites as found in the HRs. The "HR (high θ_s)" group consists of the five HRs with elevated θ_s, while the "HR (rm high θ_s)" group includes all HRs except for these five. For each genomic site, SNPs with more than two alleles (i.e., non-biallelic SNPs) were counted multiple times based on changes relative to the reference sequence. Source data are provided as a Source Data file.


a
25 strains of Y. pestis
c
23 strains of Y. pseudotuberculosis
b
3,318 strains of Y. pestis
d
559 strains of Y. pseudotuberculosis
Position
Branch
e
Frequency
39872
2778
40
3
Number of mutations per strain in each HR
f
Linked mutations
Position relative to the length of HR (%)

Supplementary Fig. 7 Comparative recombination analysis in *Y. pestis* and *Y. pseudotuberculosis* genomes and physical linkage of variation sites within mutation hotspots. **a,** ClonalFrameML and Gubbins recombination analysis across 25 strains of *Y. pestis*; **b,** 3,318 strains of *Y. pestis*; **c,** 23 strains of *Y. pseudotuberculosis*; **d,** 559 strains of *Y. pseudotuberculosis*. Top panels show ClonalFrameML results, while bottom panels display Gubbins results**.** In ClonalFrameML, dark blue bars indicate recombination events, colored vertical bars represent branch substitutions, and grey areas are non-core regions. In Gubbins, red blocks indicate predicted recombinations on internal branches (shared among multiple strains due to common descent), and blue blocks represent recombination events unique to individual strains. **e,** Frequency distribution plot of mutations in each hotspot for individual genomes compared to the ancestral state. Source data are provided as a Source Data file. **f,** 31 groups of potential physically linked variations. The horizontal axis represents the relative position of variation sites within HRs, and the vertical axis shows the group index of physically linked variations within hotspots. Blue circles in the graph represent a single mutation event that may simultaneously introduce two variations in same strains, connected by gray lines; red circles represent a single mutation event that may simultaneously introduce three variations in same strains; yellow circles (connected by black lines) represent a single mutation event that may simultaneously introduce two variations, both of which are fixed in the same population. Source data are provided as a Source Data file.



Supplementary Fig. 8 Schematic representation of positive selection for *ail*-related fixed mutations within HR31. The branch lengths are arbitrary and represent only the phylogenetic topology. Five main branches (Branch 0–4) of *Y. pestis* are color-coded, with clades containing the ancient genomes highlighted in red. Filled pentagons mark seven *ail*-related mutations, including six within the *ail* gene and one upstream, fixed in the main branch and across multiple phylogroups. Partially filled pentagons indicate *ail* gene variations that occur in a subset of the ancient genomes.



Supplementary Fig. 9 Streptomycin susceptibility testing for *Y. pestis* strains. a, Assessment of drug resistance using the disk diffusion method. **b,** Evaluation of drug resistance by employing the agar dilution method with varying concentrations of streptomycin. Two groups of strains, with or without mutations in HR06, are tested, with streptomycin-resistant *Y. pestis* S19960127 and *Escherichia coli* ATCC 25922 as quality control strains. In the agar dilution method, strain order from top to bottom was as follows: HR06-mutated strains S19960127, J100004, CMCC96001, 42022, 95014, 70008, 72009, 60004 in the first row, non-HR06-mutated strains 34602, J380001, 372005, 95013, 102004, * (not included in this study) in the second row, and negative control *E. coli* ATCC 25922 in the third row.