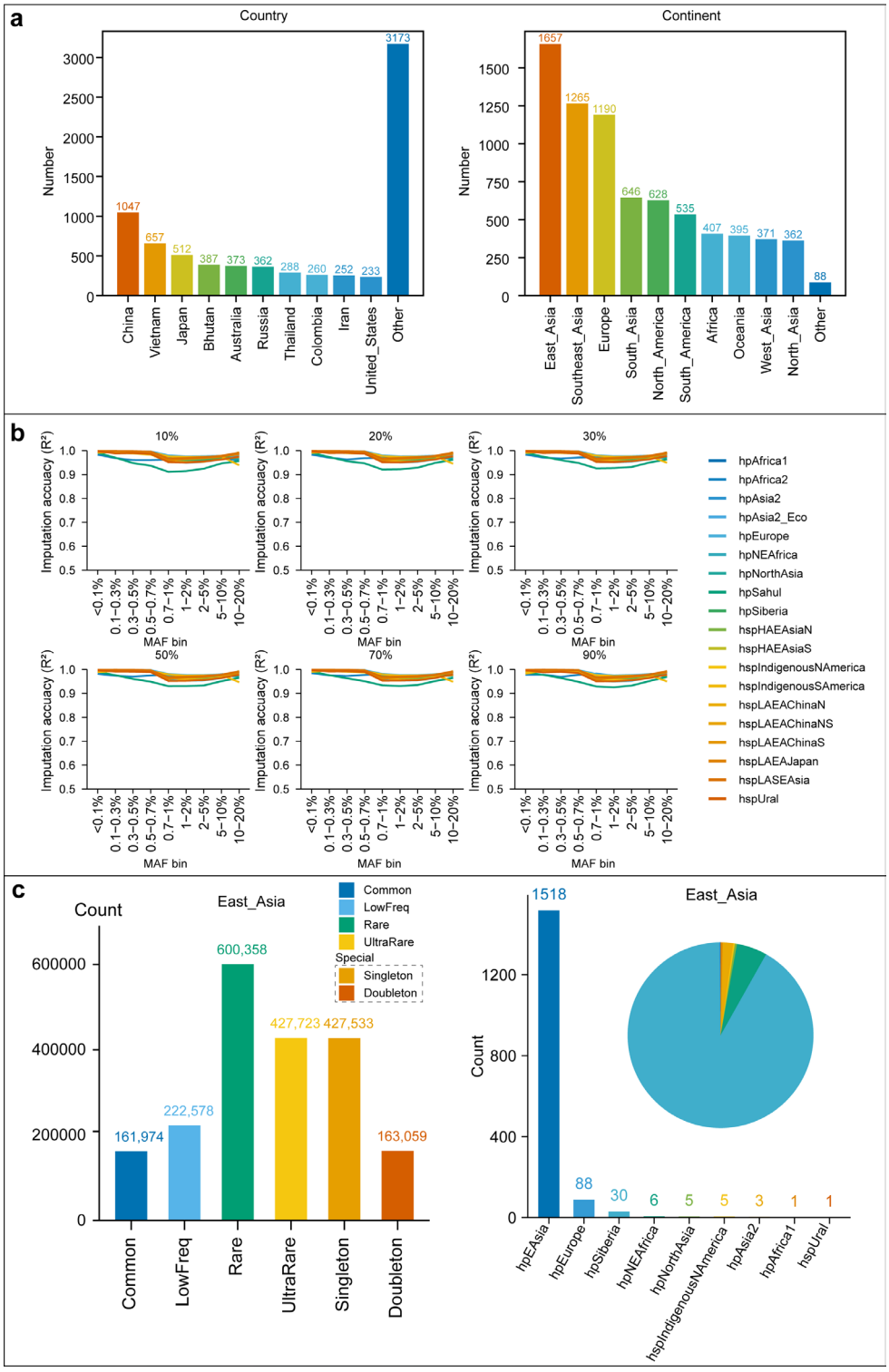
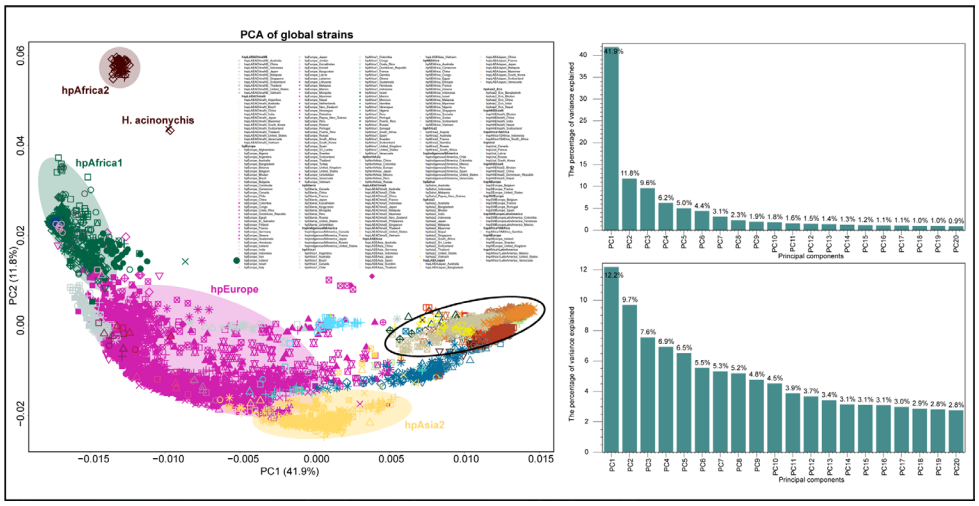
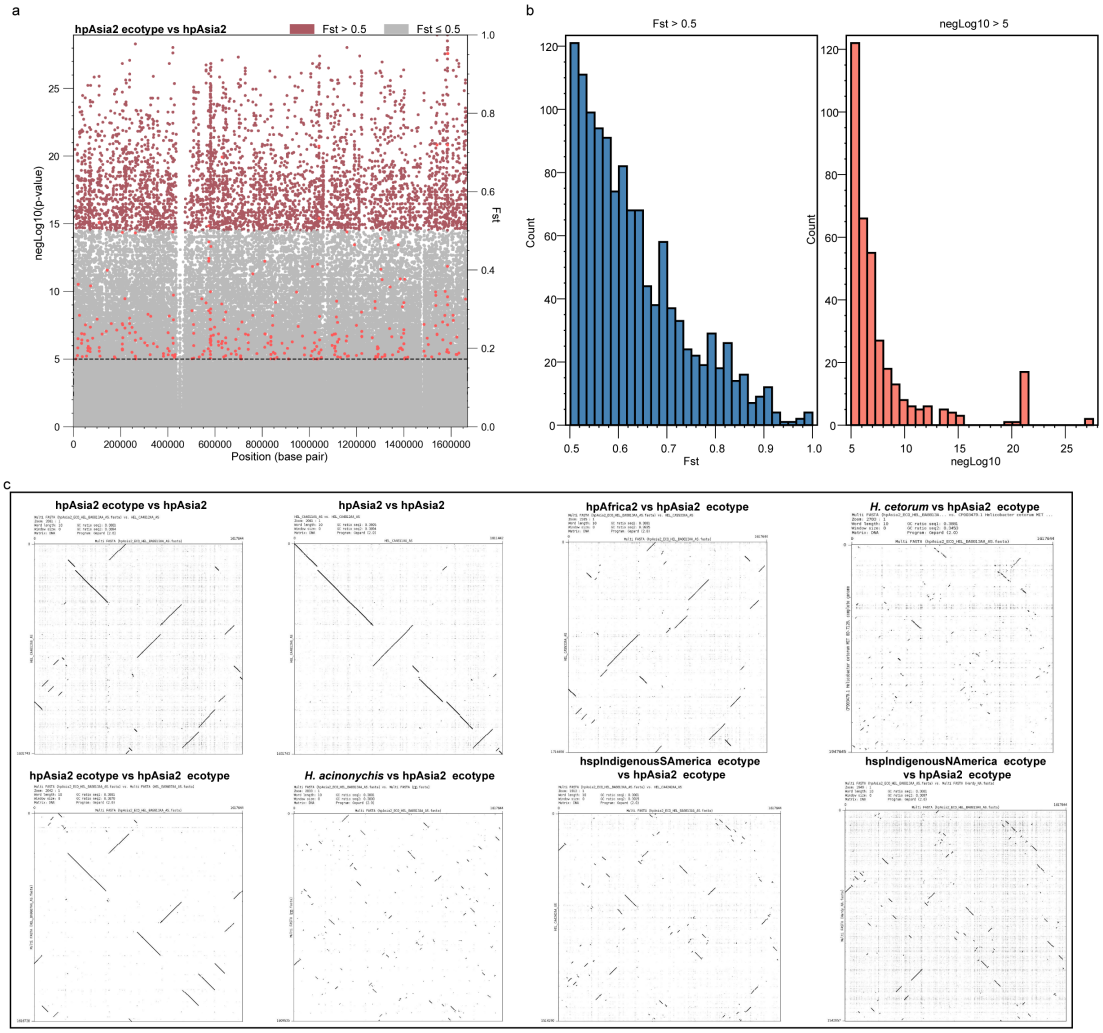
Extended Data Figs



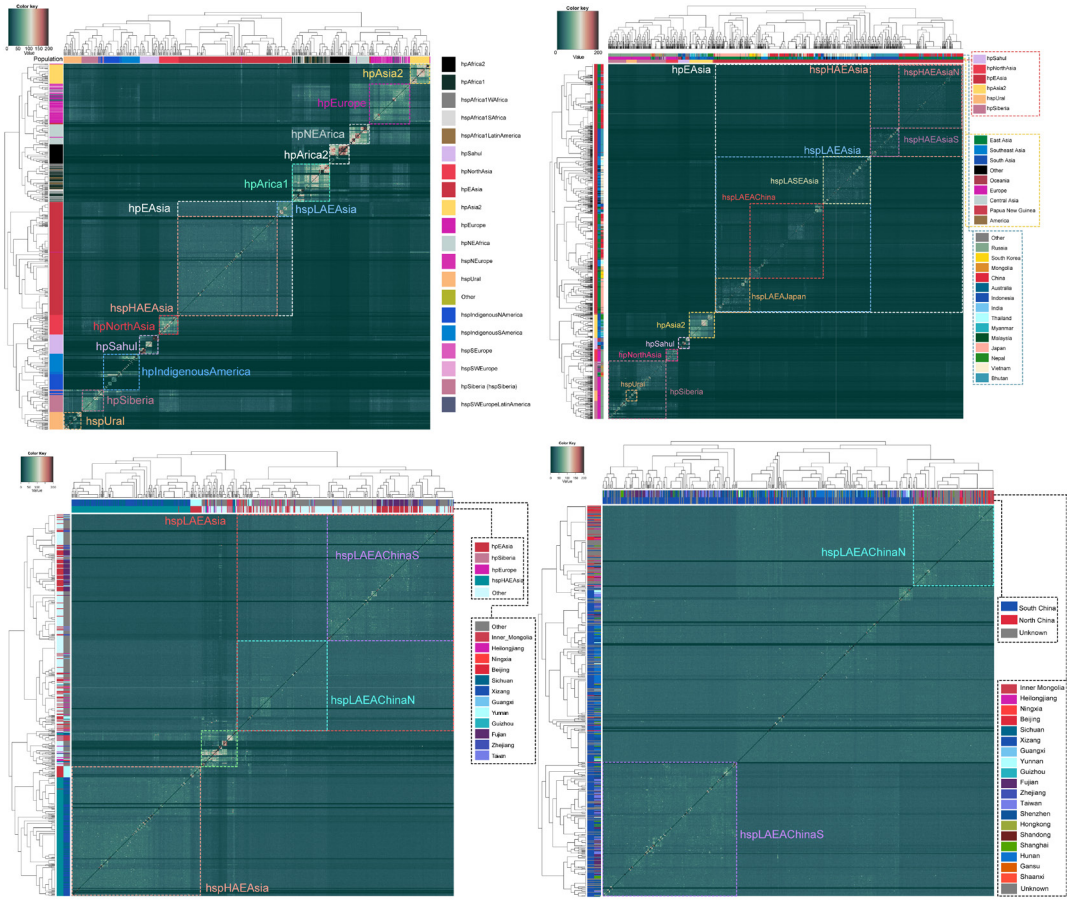
Extended Data Fig. 1| More information of sample distribution in the HPgnomAD database, variant discovery, and reference-panel performance evaluation. a, Bar chart of the number of strains collected per country and continent. **b,** Imputation accuracy across coverages, MAF bins, and populations. The x-axis denotes MAF bins, and the y-axis indicates empirical dosage R^2. **c,** Left: bar chart of variant counts; right: combined bar and pie charts showing population distribution in East Asia.



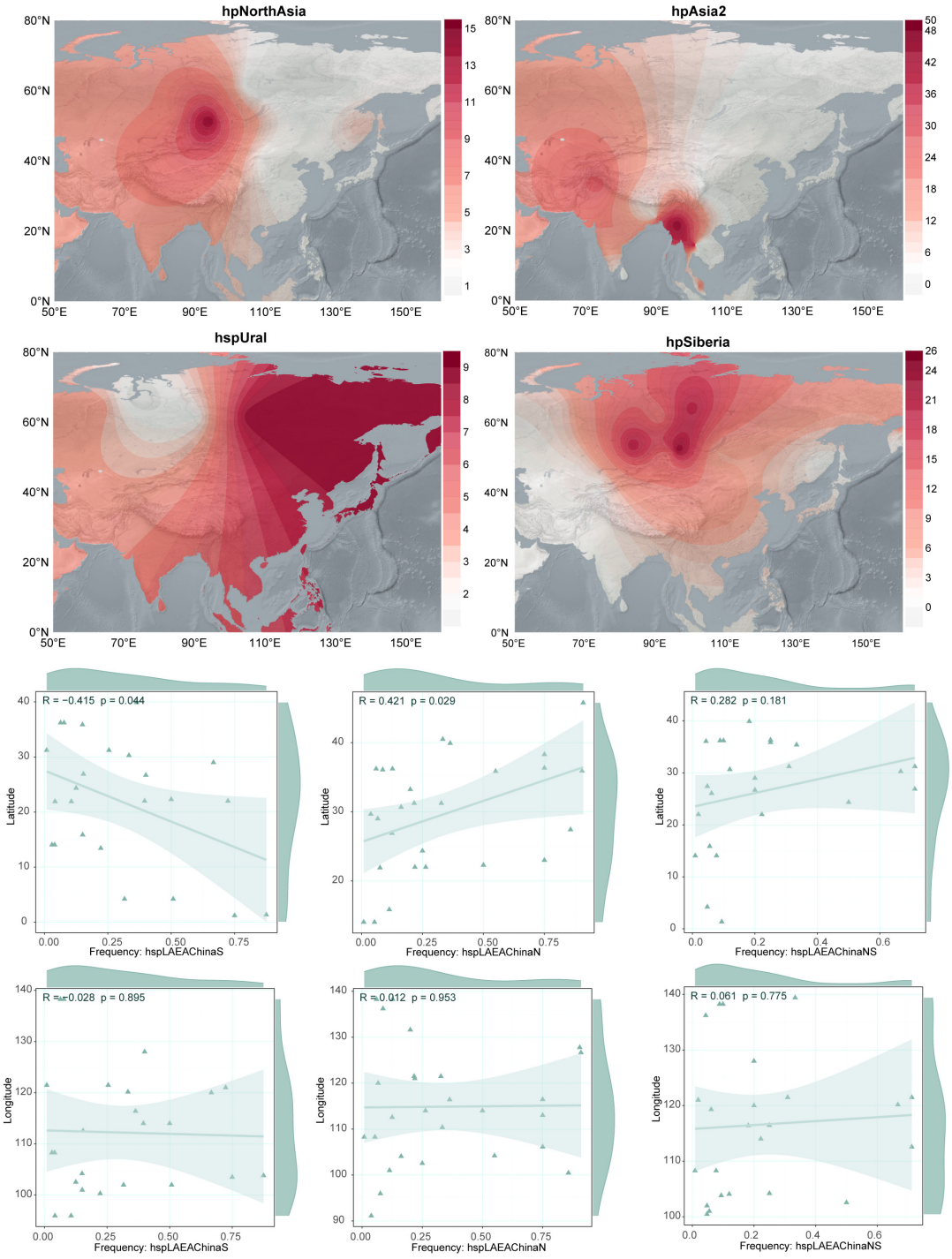
Extended Data Fig. 2| PCA analysis. Left, PCA of global strains, identical to **Fig. 2a** but colored by genetic background. Upper right, variance explained by the global PCA. Lower right, variance explained by the hpEAsia PCA.



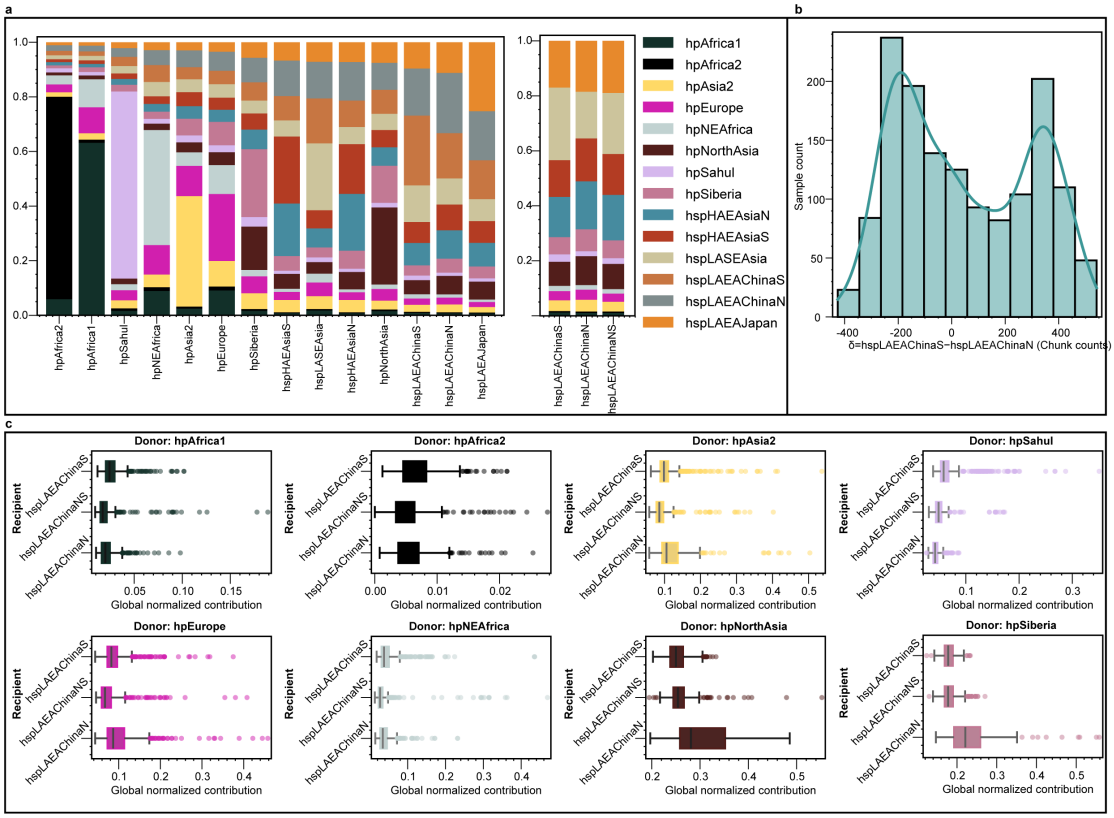
Extended Data Fig. 3| Genomic features of the hpAsia2 ecotype. a, Manhattan plot from a GWAS comparing the hpAsia2 ecotype with hpAsia2 strains. Each point represents a SNP; colors denote F_{ST} bins (gray and red). The horizontal line marks the genome-wide significance threshold ($-\log_{10}p = 5$), calculated with a Bayesian Wald test followed by Bonferroni correction (345,807 SNPs tested). **b,** Counts of differentiated loci. Bar charts show the number of sites with $F_{ST} > 0.5$ (left) and with GWAS significance ($-\log_{10}p > 5$; right). In total, 1,226 high F_{ST} loci and 359 GWAS-significant loci were detected, of which 36 satisfy both criteria. **c,** Whole-genome dot-plot comparisons within and between hpAsia2 ecotype genomes (genome pairs indicated above each panel). Continuous diagonals denote collinear genomes; breaks or multiple short diagonals reveal chromosomal rearrangements or differences in gene content, whereas long uninterrupted diagonals indicate highly similar genomes.



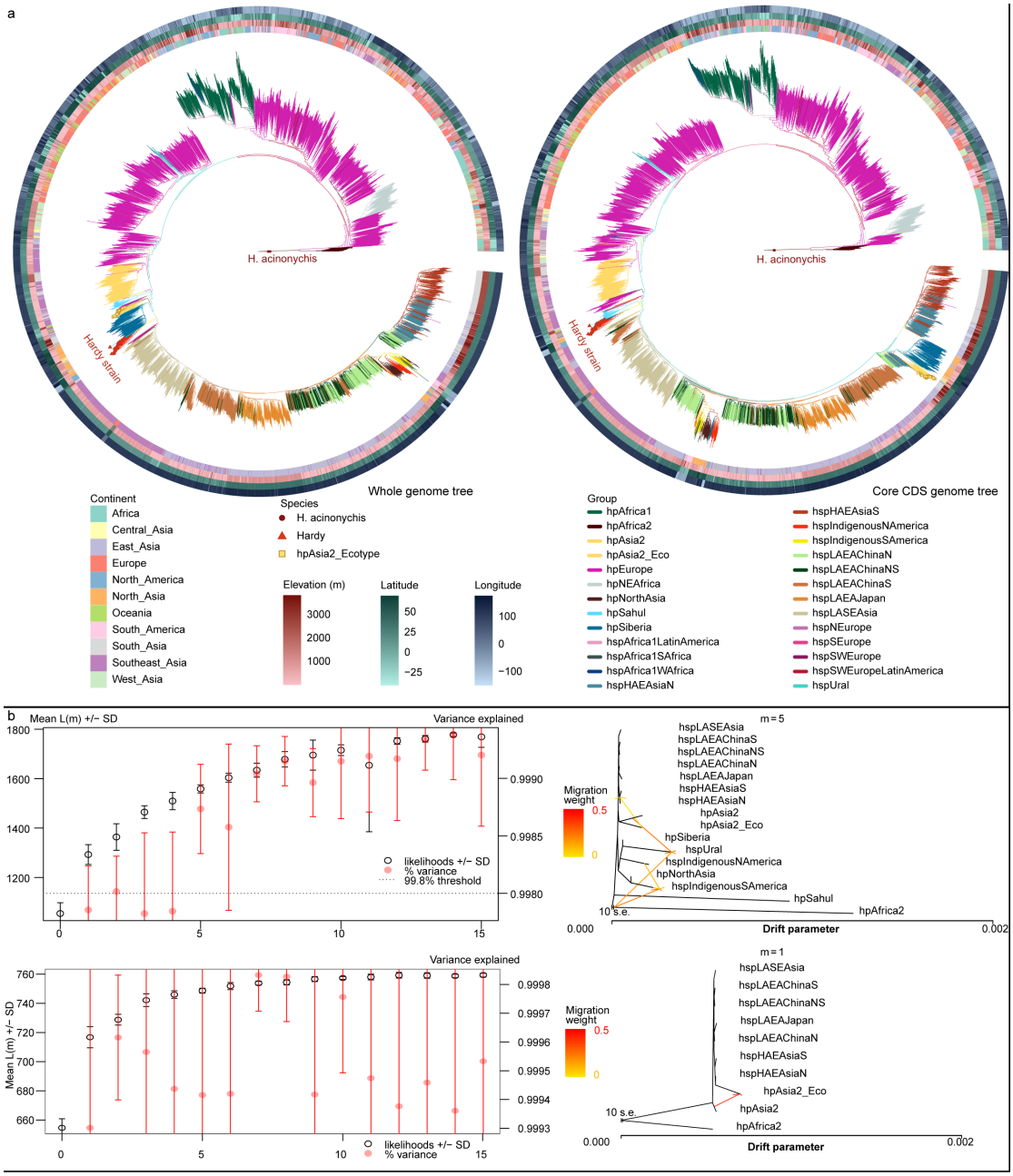
Extended Data Fig. 4| FineSTRUCTURE analysis of strains from global and hpEAsia perspectives. Strain labels defined in previous studies and geographic information are overlaid on the left dendrogram and at the top of the heatmap. Strains that have been redefined or newly defined in this study are highlighted with square markers and bold font. FineSTRUCTURE employs an in-silico chromosome painting algorithm that models each strain as a mosaic of its nearest neighbors, selected from among the other strains in the dataset. Each row in the heatmap represents the coancestry vector for a strain, reflecting the count of DNA segments contributed by each of the other strains. High coancestry between strains indicates that they share numerous genomic segments and likely originate from a common gene pool. Moreover, FineSTRUCTURE-based clustering is more sensitive to recent gene flow compared to traditional clustering methods based on genetic distances.



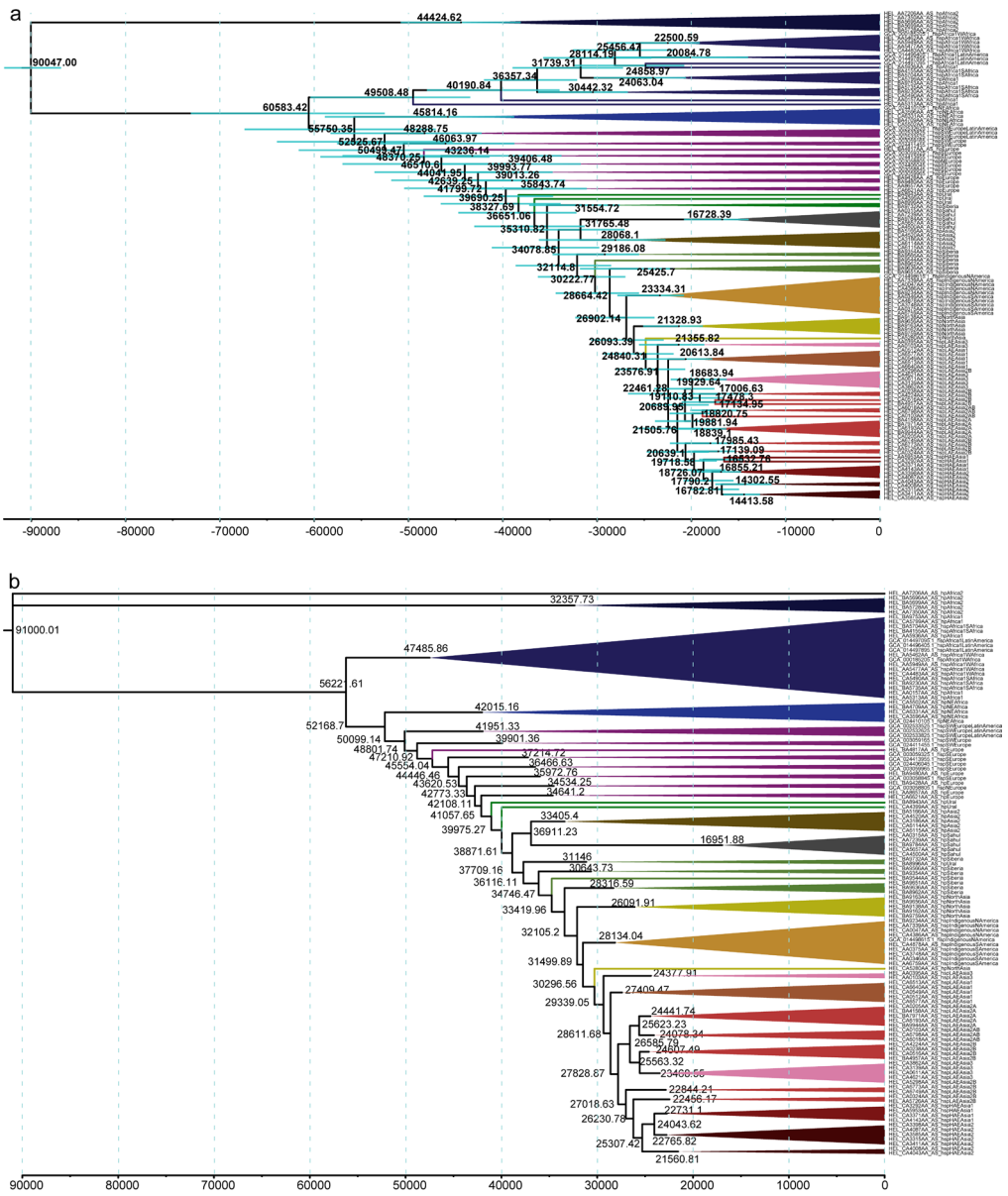
Extended Data Fig. 5| Geographic distribution of H. pylori strains across Asia and correlation analysis of hspLAEAChina. a, Spatial distribution patterns of different *H. pylori* strains in Asia are illustrated. Contour maps represent interpolated frequency distributions generated using the Kriging algorithm. **b,** Pearson correlation analysis between latitude, longitude, and frequency of hspLAEAChina strains.



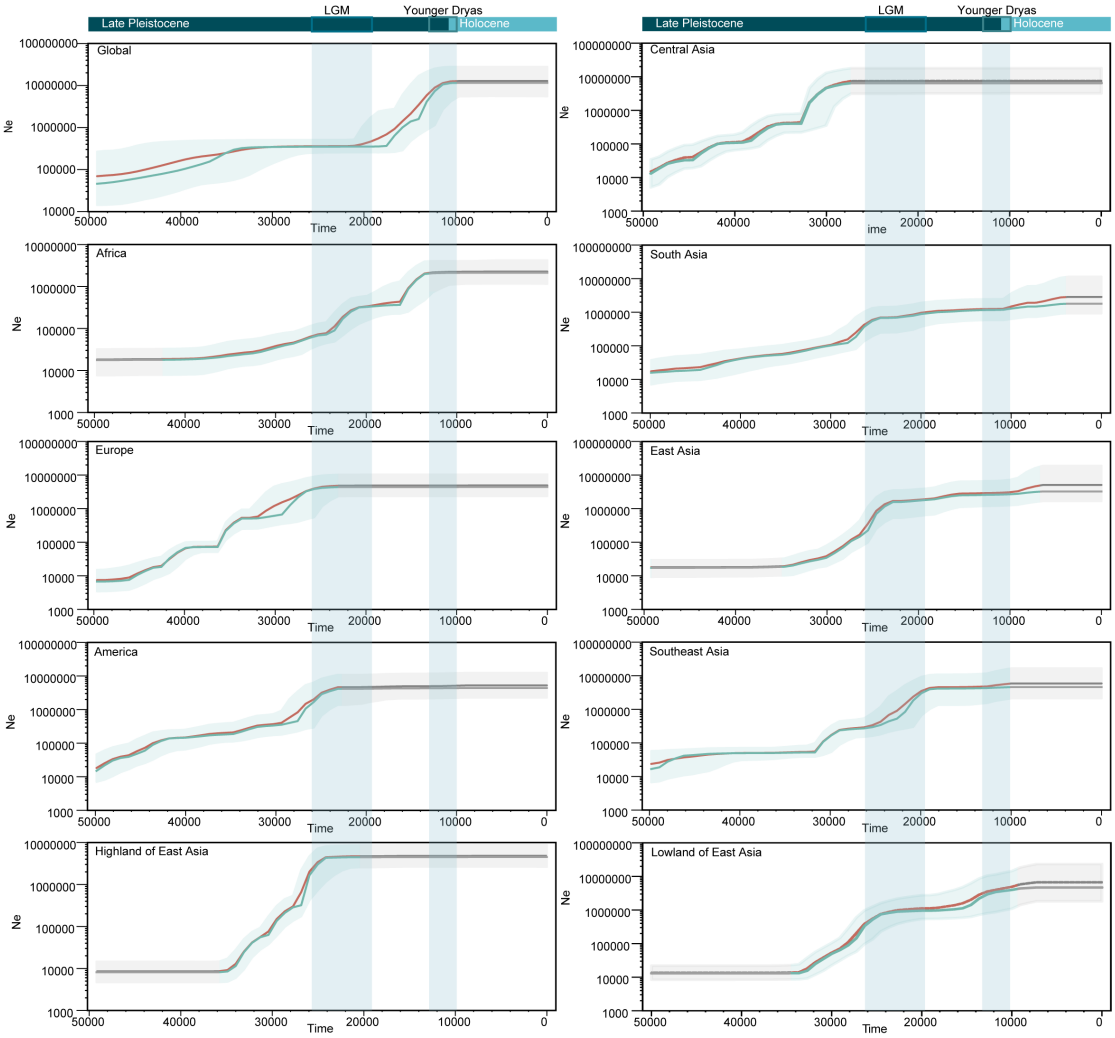
Extended Data Fig. 6| Average ancestry profiles of *H. pylori* in Asia. a, ChromoPainter results. The left panel includes the major global strains and all hpEAsia strains as recipients, while the right panel excludes hspLAEAChina from the donor set. **b,** Among the strains assigned to hspLAEAChina, the contributions from hspLAEAChinaS and hspLAEAChinaN differ. **c,** Full version of the box-and-whisker plots illustrating contributions from various donors across the three recipient categories.



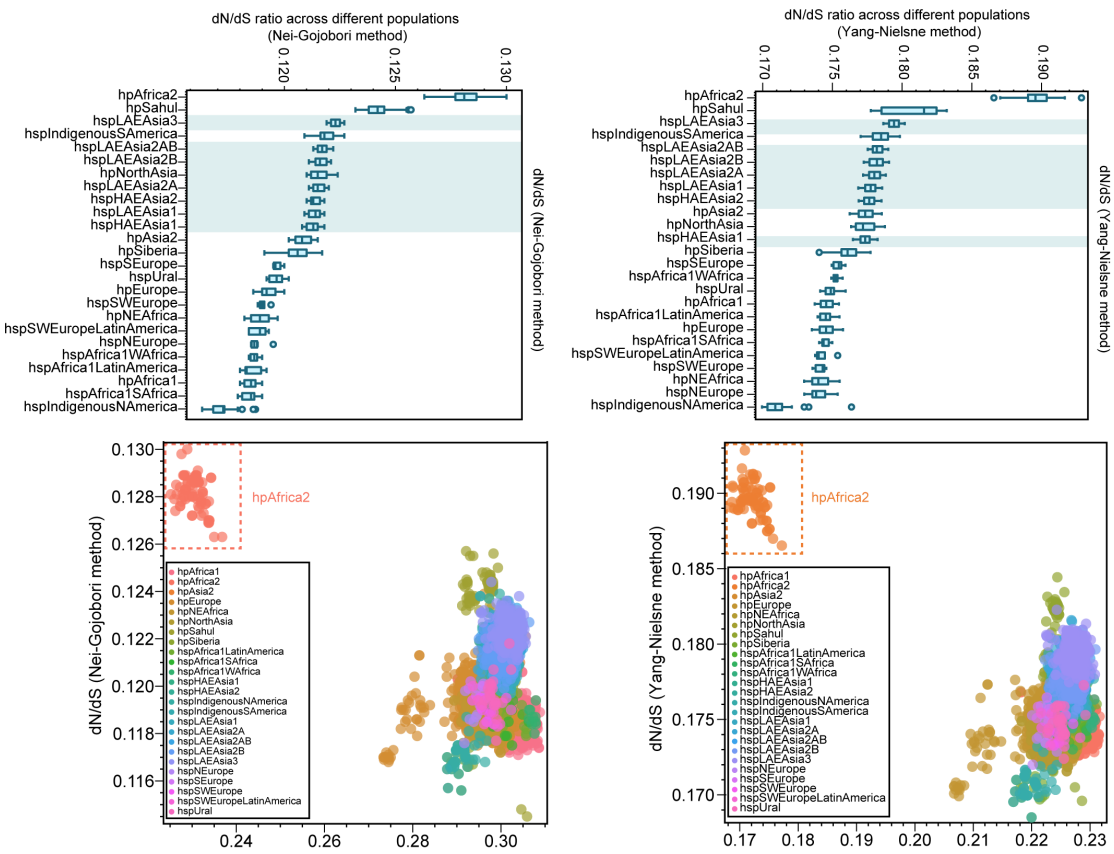
Extended Data Fig. 7| Phylogenetic trees and TreeMix result. a, Maximum-likelihood phylogenetic tree based on whole-genome sequences (left) and based on core CDS genome sequences (right). Branches are colored according to population. H. acinonychis serves as the outgroup, indicated by a red square at the tip of its branch. Hardy strains are marked by red triangle dots at the branch tips, and hpAsia2 ecotype are marked by yellow square at the tip of its branch. Concentric circles surrounding each tree represent, from innermost to outermost, the sampling location's classification by continent, elevation, latitude, and longitude. **b,** Treemix graphs for different populations. The top panel includes strains common across Asia; the bottom panel includes only strains common across East Asia. The optimal number of migration edges was chosen at the inflection point of maximal change in explained variance. The hpAfrica2 population was set as the outgroup. Arrows denote gene flow between branches, with arrow color indicating the migration-edge weight.



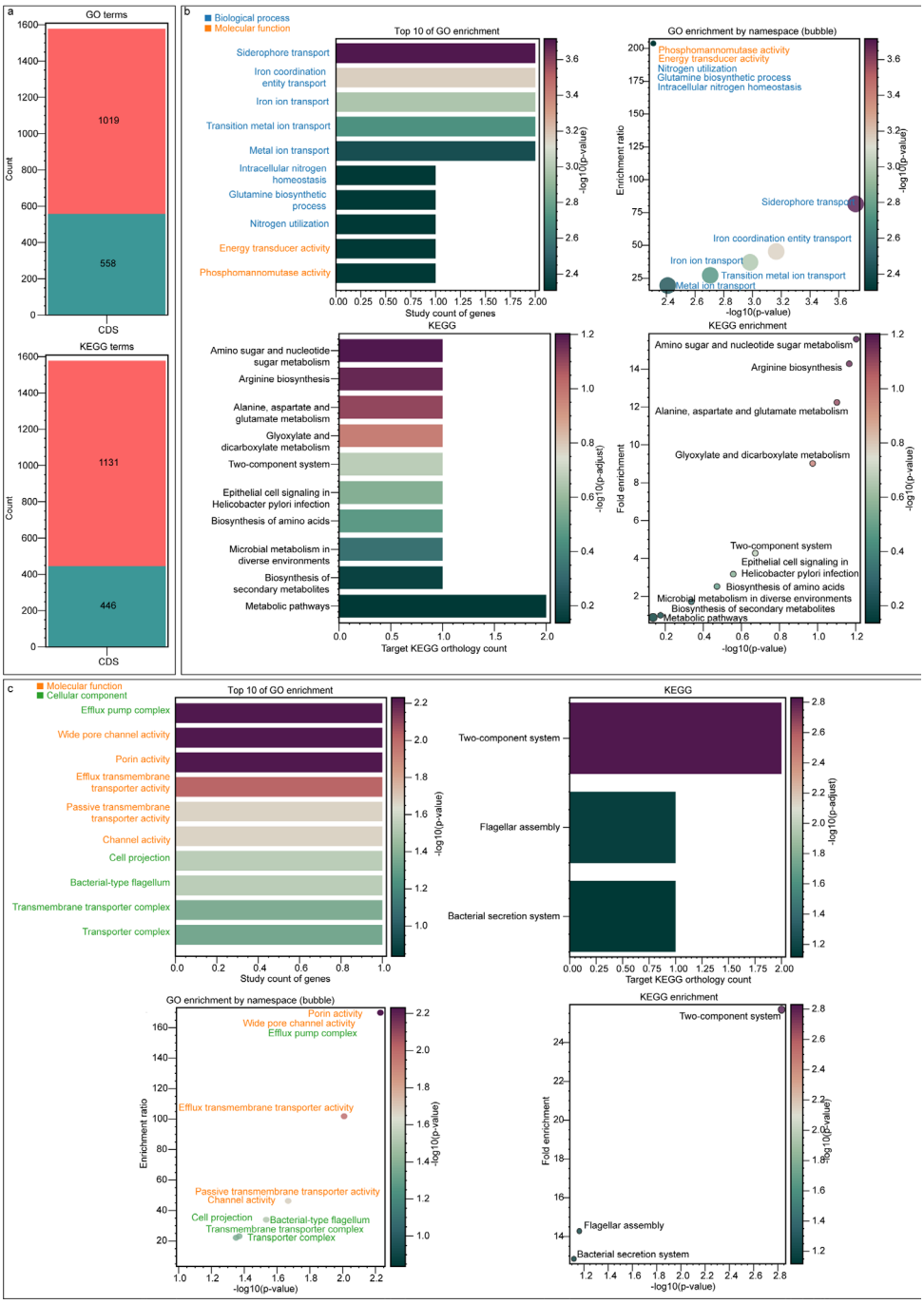
Extended Data Fig. 8| Time scaled phylogeny. a, Phylogenetic tree inferred using BEAST, presented here as the complete version of **Fig. 4c**. **b,** Phylogenetic tree inferred using PAML.



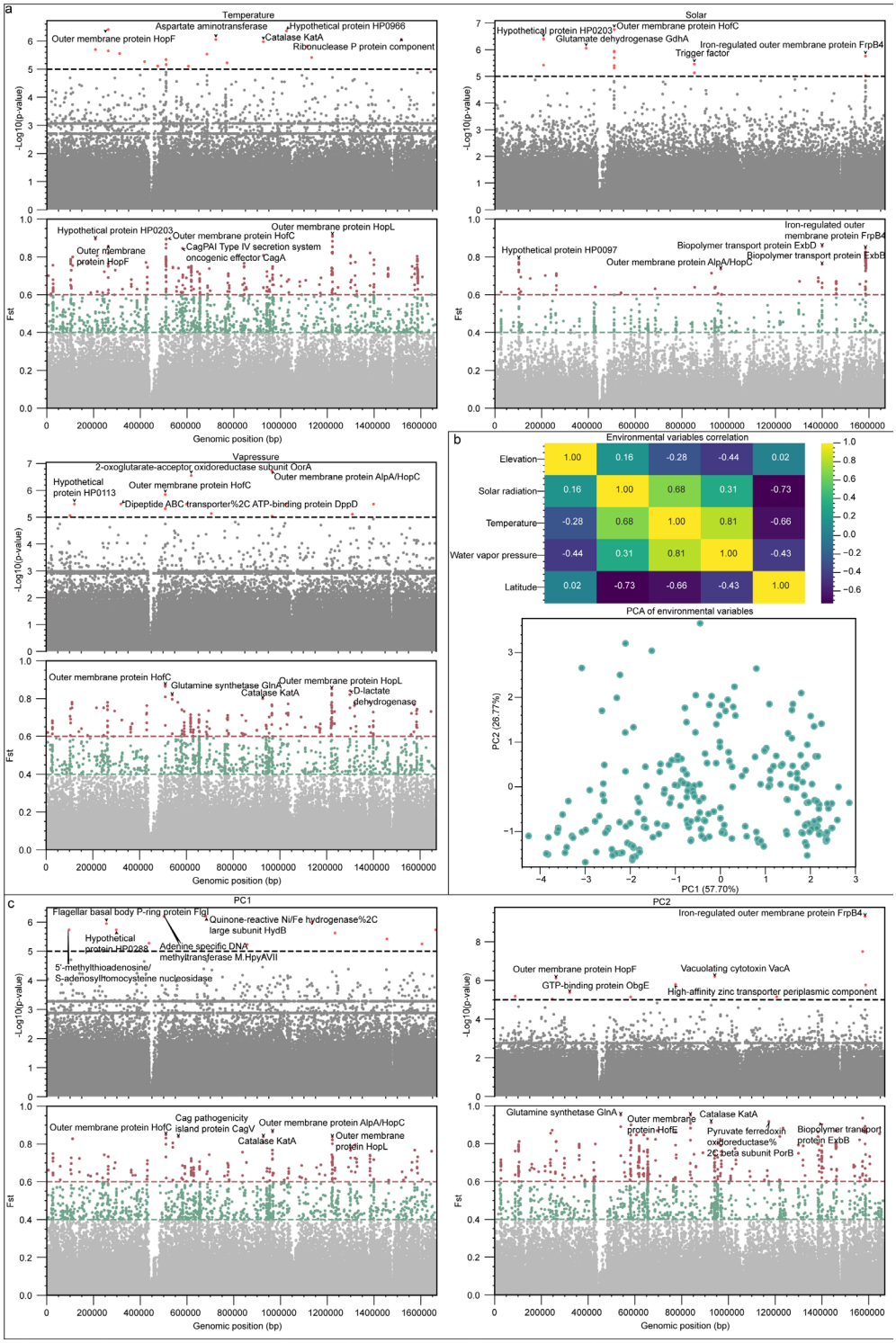
Extended Data Fig. 9| Bayesian skyline plots (BSPs) for each region. The x-axis represents time (in years before the present), and the y-axis denotes effective population size. The shaded area of each BSP curve indicates the minimum-to-maximum range, the green solid line represents the mean value, and the orange dashed line represents the median value. Major geological periods are marked at the top. Grey shaded intervals correspond to periods when the BSP curves plateau, suggesting that genetic data cannot reliably reconstruct population dynamics during those intervals.



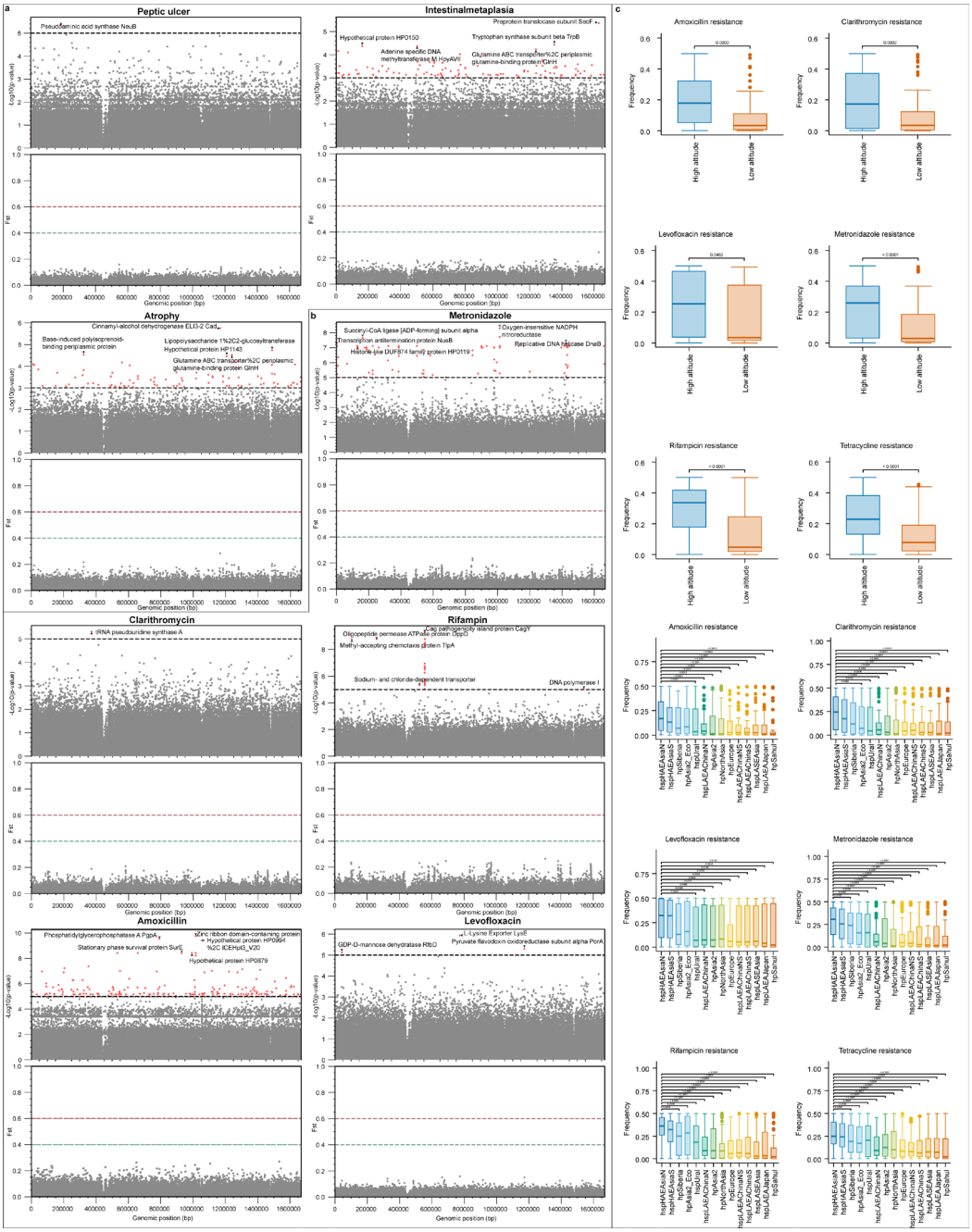
Extended Data Fig. 10| dN/dS estimates for all strains. dN/dS ratios were calculated using the Nei–Gojobori method (left panels) and the Yang–Nielsen method (right panels). The upper panels show box-and-whisker plots summarizing the distribution of *dN/dS* values for each population, whereas the lower panels depict scatter plots of individual gene estimates, with synonymous substitution rates (*dS*) on the x axis and corresponding *dN/dS* ratios on the y axis.



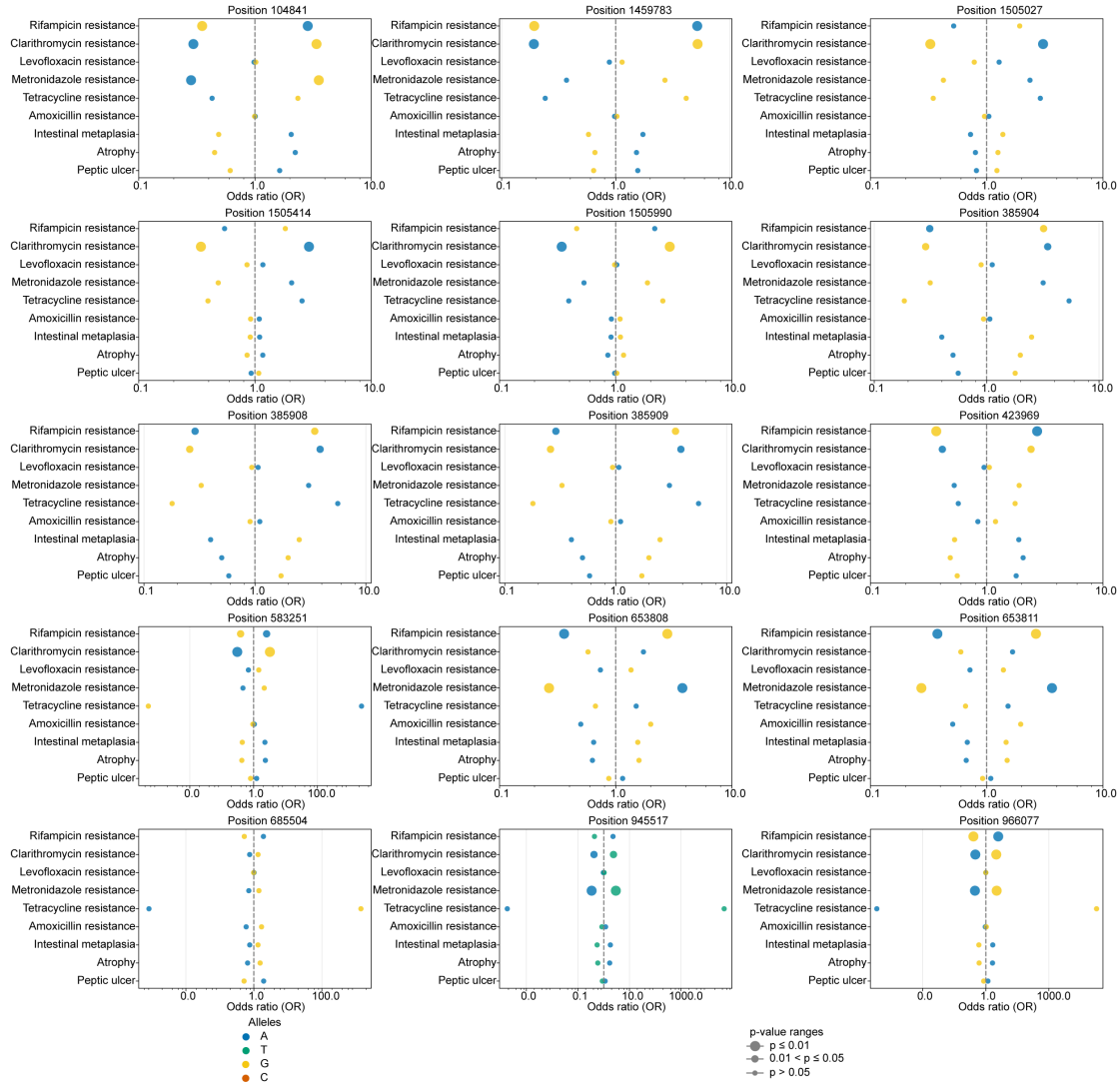
Extended Data Fig. 11| GO and KEGG enrichment analyses. a, Bar plot showing, for the 1,577 CDSs, the number of CDSs successfully annotated (red) versus not annotated (green) against GO and KEGG databases. **b,** GO-term and KEGG-pathway enrichment for genes differentiating high-altitude versus low-altitude strains. The target set comprises genes harboring SNPs that satisfy GWAS ($-\log_{10}p > 5$) or $F_{ST} > 0.6$ thresholds, and exhibit |Cliff's Δ| ≥ 0.35 in dN/dS comparisons; all other annotated genes serve as the background. Only the top 10 enriched terms are shown. **c,** GO-term and KEGG-pathway enrichment for genes differentiating northern versus southern East Asian strains, using the same target and background sets as in b. Only the top 10 enriched terms are shown.



Extended Data Fig. 12| GWAS and F_{ST} analyses for environmental variables. a, GWAS and F_{ST} analyses for temperature, solar radiation and vapor pressure. GWAS significance is indicated by a horizontal threshold at $-\log_{10}p = 5$. SNPs exceeding this threshold are colored red, those below grey. F_{ST} thresholds of 0.4 and 0.6 are marked; SNPs are colored grey ($F_{ST} < 0.4$), green ($0.4 \leqslant F_{ST} < 0.6$) or red (Fst $\geqslant$ 0.6). **b,** Correlation matrix and PCA of environmental variables. **c,** GWAS and F_{ST} analyses for PC1 and PC2 of environmental variables, applying the same significance and F_{ST} thresholds as above.



Extended Data Fig. 13| The GWAS and F_{ST} analyses of clinical phenotypes and frequency distribution of candidate loci. a, GWAS and F_{ST} analyses of host clinical traits. Panels show GWAS and F_{ST} results for peptic ulcer, intestinal metaplasia and gastric atrophy. Because the sample sizes for intestinal metaplasia and atrophy were limited (n = 133), a relaxed significance threshold of $-\log_{10}$ (*p*-value) = 3 was applied. **b,** GWAS and F_{ST} results for bacterial antibiotic-resistance phenotypes. Analyses are presented for metronidazole, clarithromycin, rifampicin, amoxicillin and levofloxacin. **c,** Boxplots showing the allele frequency distribution of candidate loci across populations. Candidate loci were those in the top 5% of each analysis based on empirical thresholds, intersected with the top 5% of loci from the high-versus-low altitude differentiation scan. Allele frequencies were computed for each locus in each population.



Extended Data Fig. 14 | Bubble plot of odds ratios for pleiotropic SNPs. The x-axis shows ORs on a log scale; the y-axis lists phenotypes. Bubble color indicates allele variant, and bubble size reflects Bonferroni-adjusted *p*-value categories ($p \leqslant 0.01$; $0.01 < p \leqslant 0.05$; $p > 0.05$). The dashed vertical line marks OR = 1. ORs and *p*-values were derived from logistic regression with Wald tests and Bonferroni correction. This panel presents the full loci including **Fig. 6**.