PNAS
www.pnas.org

Supplementary Information for

Helicobacter pylori's historical journey through Siberia and the Americas.

Yoshan Moodley, Andrea Brunelli, Silvia Ghirotto, Andrey Klyubin, Ayas S. Maady, William Tyne, Zilia Y. Muñoz-Ramirez, Zhemin Zhou, Andrea Manica, Bodo Linz, Mark Achtman

#Corresponding author: Yoshan Moodley
Email: yoshan.moodley@univen.ac.za

This PDF file includes:

Supplementary Information Text (Methods and Materials)

Supplementary Figures (S1-S21)

Supplementary Tables (S1-S12)

SI References

Other supplementary materials for this manuscript include the following:

Supplementary Data sets S1 and S2 (excel format)

Supplementary Information Text

Materials and Methods

Strains

Esophagogastroduodenoscopy was performed with written informed consent during routine inspection in 2005-2006 of volunteers by a gastroenterologist (ASM), at government hospitals and clinics in Russia and Mongolia (Ethics certificate EA1/071/07, Charité, Berlin). Biopsies of the gastric mucosa were obtained from the antrum (and/or corpus) of the stomachs of individuals from 18 human populations representing 16 ethnic groups (Fig. S1). Location:ethnicities: North-western Siberia: Uralic-speaking Khanty and Nenet; Central Siberia: Turkic-speaking Tuvan and Tubalar and Mongolic-speaking Buryat and Mongolian; Northern Siberia: Tungusic-speaking Evenk and Turkic-speaking Yakut; Eastern Siberia: Tungusic-speaking Nanai, Ulchi and Orok; Beringia: Tungusic-speaking Even and Chukotko-Kamchatkan-speaking Koryak and Chukchi. Two other ethnicities, the Ket of the Yenisei Valley and the Nivkh of Sakhalin Island, spoke their own language isolate.

Gastric biopsies were added to PBS (phosphate buffered saline) solution, frozen immediately in liquid nitrogen, and kept at -80°C until they were grown on Pylori cultivation plates (bioMérieux, France) for 3-7 days at the Research Institute for Physico-Chemical Medicine, Moscow, according to [1,2]. DNA was extracted from cultures, grown from single colonies in BHI (brain-heart infusion) with 10% inactivated fetal bovine serum for 24 hours at 37 °C, using Wizard Genomic DNA Purification (Promega). Fragments of seven multilocus sequence typing (MLST) genes (*atpA*, *efp*, *mutY*, *ppa*, *trpC*, *ureI*, *yphC*) were amplified and sequenced as previously described [3,4]. We also sequenced draft genomes of 55 Siberian strains from 14 ethnic groups as well as 40 other representative genomes from other sources as described [5] to reconstruct the evolutionary history of *H. pylori* in the region.

Genetic Structure

We employed multiple methods to investigate the structure of genetic variation in our dataset of MLST sequences, which consisted of 1002 *H. pylori* strains from across Asia and the Americas, including the 396 Siberian strains isolated for this study. The MLST alignment thus contained 3406 nucleotide positions and 1952 polymorphic sites. First, we performed a discriminant analysis of principal components (DAPC) [6] on the MLST data. This method assesses the presence of clusters by optimizing the variation of allele frequencies between- and within-groups and returns the most highly supported subdivision according to Bayesian

information criteria (BIC). Because DAPC is a multivariate approach, and not model-based, it makes no assumptions about Hardy-Weinberg or linkage equilibrium. We assessed the number of clusters that is most supported for our *H. pylori* data set by employing the find.clusters function in adegenet 1.3–1 [7] comparing the results of 10 independent runs using a custom made R script. We then ran the DAPC analysis with 1,000,000 iterations checking the consistency of the inferred groups over 10 different runs. DAPC was performed on the entire dataset as well as exclusively on the hspIndigenousAmericas subgroup. We also conducted a Bayesian analysis of population structure on the MLST data using the model-based algorithm implemented by the software STRUCTURE [8]. We ran 100,000 iterations, discarding the first 10,000 as burn-in and testing 2 to 15 partitions (K) under the linkage model [9], replicating 5 runs for each value of K.

We also investigated genetic structure using a representative set of *H. pylori* whole genomes. First, we tested the consistency of MLST data by conducting five DAPC analyses each on Asian and American strains for which both MLST and genome data were available (79 strains). Both data sets were optimally clustered into the same three populations (Fig. S21) with only minor differences in individual assignments (Table S13).

We then continued to assess genomic structure of our full genome data set consisting of 94 genomes, 54 of which were isolated in Siberia (Table S2). To establish the general clonal structure of *H. pylori*, we first analysed a set of 40 genomes representing the global diversity of *H. pylori.* Then, to show how Siberian genomic variation partitioned within this global clonal structure, we then re-ran the analysis after including 54 genomes from Siberian strains isolated in this study. An hpAfrica2 strain (Khoisan03A) was used as the reference genome in all analyses to call single nucleotide polymorphisms (SNPs) that occurred in ≥95% of genomes in the alignment. Since *H. pylori* is highly recombinant[10], we were careful to first model and removed potentially recombinant sites from the genome alignment as these would violate the assumption of common ancestry and potentially blur the underlying clonal genomic structure. We used the iterative algorithm Gubbins, which scans the alignment searching for high density substitutions that would be typical of a recombination event[11]. We then used this recombination-free alignment to reconstruct maximum likelihood phylogenies using IQ-Tree2[12]. According to Bayesian and Akaike information criteria (BIC and AIC, respectively), the most likely nucleotide substitution model for both the 40-genome and 94-genome alignments was the Kimura-3-parameter model[13] (K3P), with ascertainment bias correction for SNP data (+ASC) and rate

heterogeneity modelled using the FreeRate model[14] with five rate categories (R5). We assessed branch support for both trees using an ultrafast bootstrap approximation with 1000 replicates.

We investigated the ancestry of our full genome data set using fineSTRUCTURE v 0.02 to define populations and sub-populations based on the similarity of the haplotype copying profiles obtained by an EM algorithm in ChromoPainter v.0.02[15]. Briefly, we performed the annotation of the genomes using Prokka v. 1.12 [16], the gff files were subsequently submitted to the Roary pan-genome pipeline v 3.12.0[17] using a blastp identity cut-off of 85% with the option not to split paralogs based on differential synteny. The core genome based on 1,084 genes was defined as genes present in > 95% of the genomes analysed and the core genome alignment (825,608 bp) was produced by Mafft[18]. Then, we conducted SNP calling for core genome alignment, and imputation for polymorphic sites with the frequency of missing data set to < 1% using BEAGLE v.3.3.2[19]. Finally, we ran core genome haplotype data (221,239 SNPs), using fineSTRUCTURE as described in [20]. Briefly, we set a constant recombination rate per base across the genome, with a normalization constant of 0.324, and ran the analysis to cluster strains based on the coancestry matrix with 200,000 Markov chain Monte Carlo iterations, discarding the first 100,000 iterations as burn-in. The results were visualized as a heat map with each cell indicating the proportion of DNA "chunks" a recipient receives from each donor using R[21]. To identify the proportion of ancestry of *H. pylori* isolates from hspSiberia1 and hspSiberia2, we designated Siberians isolates as recipients, and all other populations as potential donors. We then calculated the average proportion of ancestry from each population that is present in Siberian populations.

Demographic modelling

We attempted to reconstruct the evolutionary origins of the new subpopulations hspSiberia1, hspSiberia2, hspKet and hspAltai, and the migration into the Americas, by modelling their evolution within an ABC framework [22]. We defined Siberian populations based on observed genetic structure, regardless of the geographic origins of their human hosts. All evolutionary scenarios were based on *H. pylori*'s established split of the common ancestor of hpEastAsia and hpNorthAsia from hpAsia2 [23,24], followed by tree-like and admixture demographic scenarios (Fig. S7, Table S5). We performed four different ABC analyses, each considering the origins of one newly defined Siberian subpopulation. In the first two analyses, we estimated the models best accounting for the genetic variation found in hspSiberia1 and hspSiberia2 (Figs S10, S11, Table S6). In the third analysis we inferred the ancestry of hspKet by building alternative models taking account of the best topologies estimated in the first two comparisons (Fig. S12, Table S10). Since hspAltai evolved through divergence, rather than

admixture, we applied a tree-like model to time the split of this population from other hpNorthAsia strains (Table S12). We then used ABC to model are range of scenarios for the colonisation of the Americas by hspIndigenousAmericas bacteria (Table S14, Fig. S18), but defining populations based on geographic location, rather than population assignment.

The ABC framework allowed the assignment of posterior probabilities to alternative demographic models comparing summary statistics computed on the observed and simulated data sets. The simulated data were generated according to a specific demographic model (and a combination of parameter values) using coalescent theory and a mutational model. At each iteration, model parameters are drawn from prior distributions (Tables S5, S10, S12, S14), defined by our prior knowledge about the plausible values of demographic or evolutionary parameters. To reconcile coalescent generations with real time, we used the calibration of one year per generation previously determined from population divergence time estimates [23,24] using MLST data and mutation rates using whole genomes [25]. To generate the simulated datasets, we used the coalescent simulator fastsimcoal2 [26], within the software package ABCtoolbox [27], running 500,000 simulations for each tested model. We summarized the genetic data by calculating the following summary statistics: the number of haplotypes, the number of private polymorphic sites, Tajima's D, the mean number of pairwise differences within populations; the mean number of pairwise differences between populations and pairwise Fst. All the statistics were calculated with the arlsumstat software [28]. The simulations generating the summary statistics most similar to the observed ones, measured by mean Euclidean distance, were chosen to compute the posterior probability of each model using a weighted multinomial logistic regression (LR, Beaumont [29]). Under LR, the model is considered the categorically dependent variable in the simulations, while the summary statistics are the predictive variables. The regression is local around the vector of observed summary statistics and the probability of each model is finally evaluated at the point corresponding to the observed vector of summary statistics. Maximum likelihood was used to estimate the β coefficients of the regression model. To evaluate the stability of model posterior probabilities, we examined a range of thresholds by considering different numbers of retained simulations for LR (that is the 50,000, 125,000 or 250,000 best simulations). We checked the goodness of fit of our estimates using Principal Component Analysis, and estimated final model parameters using a locally weighted multivariate regression [22] on the 5,000 best-fitting simulations after a logtan transformation [30]. To evaluate the quality of the parameter estimation, we computed the coefficient of determination (R^2). As a general rule, an $R^2 < 0.10$ suggests that the summary statistics do not convey enough information about the posterior distribution of the estimated parameter[31].

Figure S1. Locations of 18 sampled Siberian populations (red) of 16 ethnicities across northern Eurasia. A further 36 populations (black) representing the total diversity of *Helicobacter pylori* in Asia were also included in our analyses. Tuvan (KZ): Kyzyl; Tuvan (TD): Todzha; Mongolia (UB): Ulan Bator; Mongolia (UG): Ulan Goom.

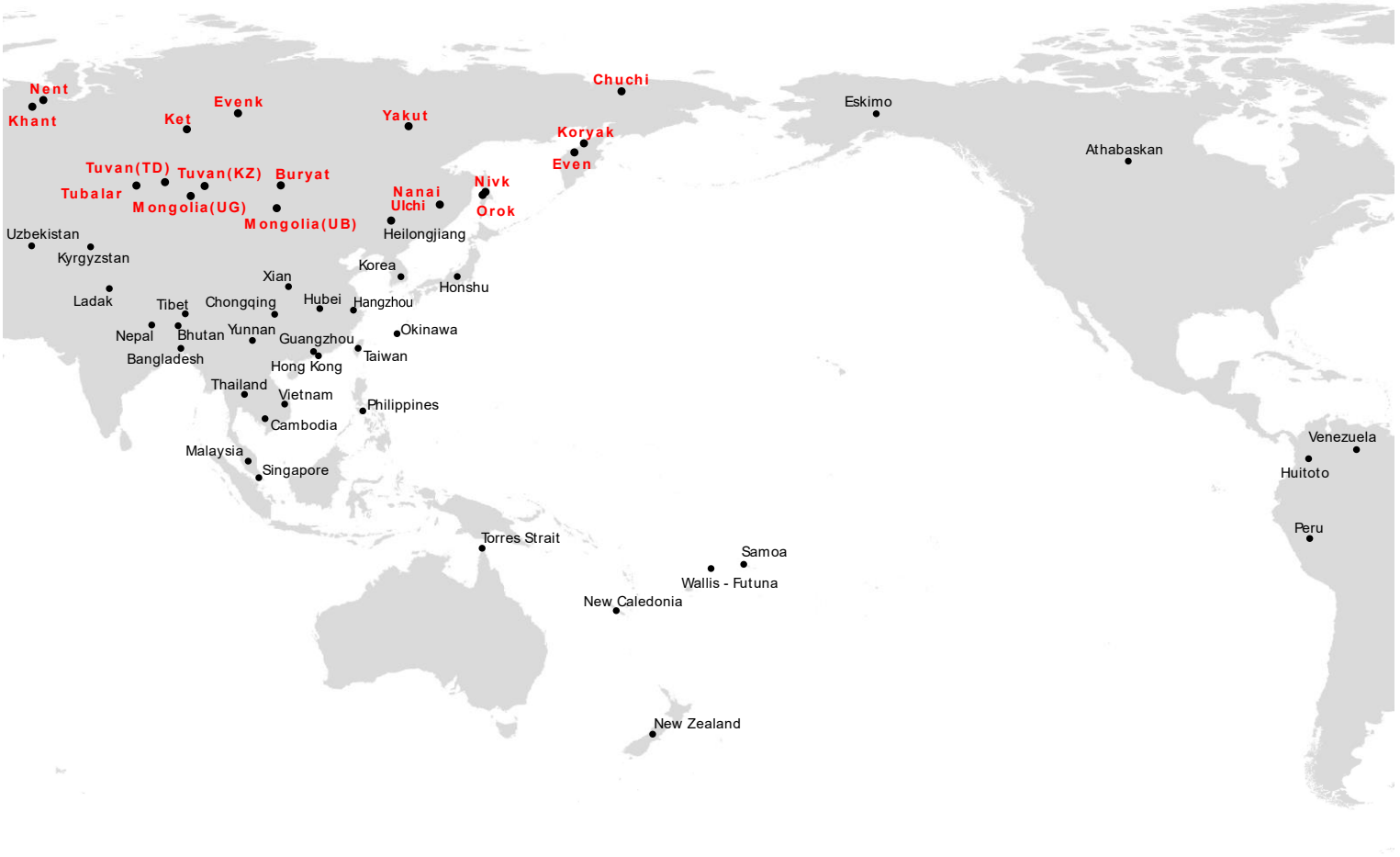

Figure S2. Results of discriminant analysis of principal components (DAPC) on the entire data set showing that 1002 *Helicobacter pylori* strains from 52 populations across eastern Eurasia and the Americas were divided optimally into 10 population clusters, based on the Bayesian information criterion (BIC).

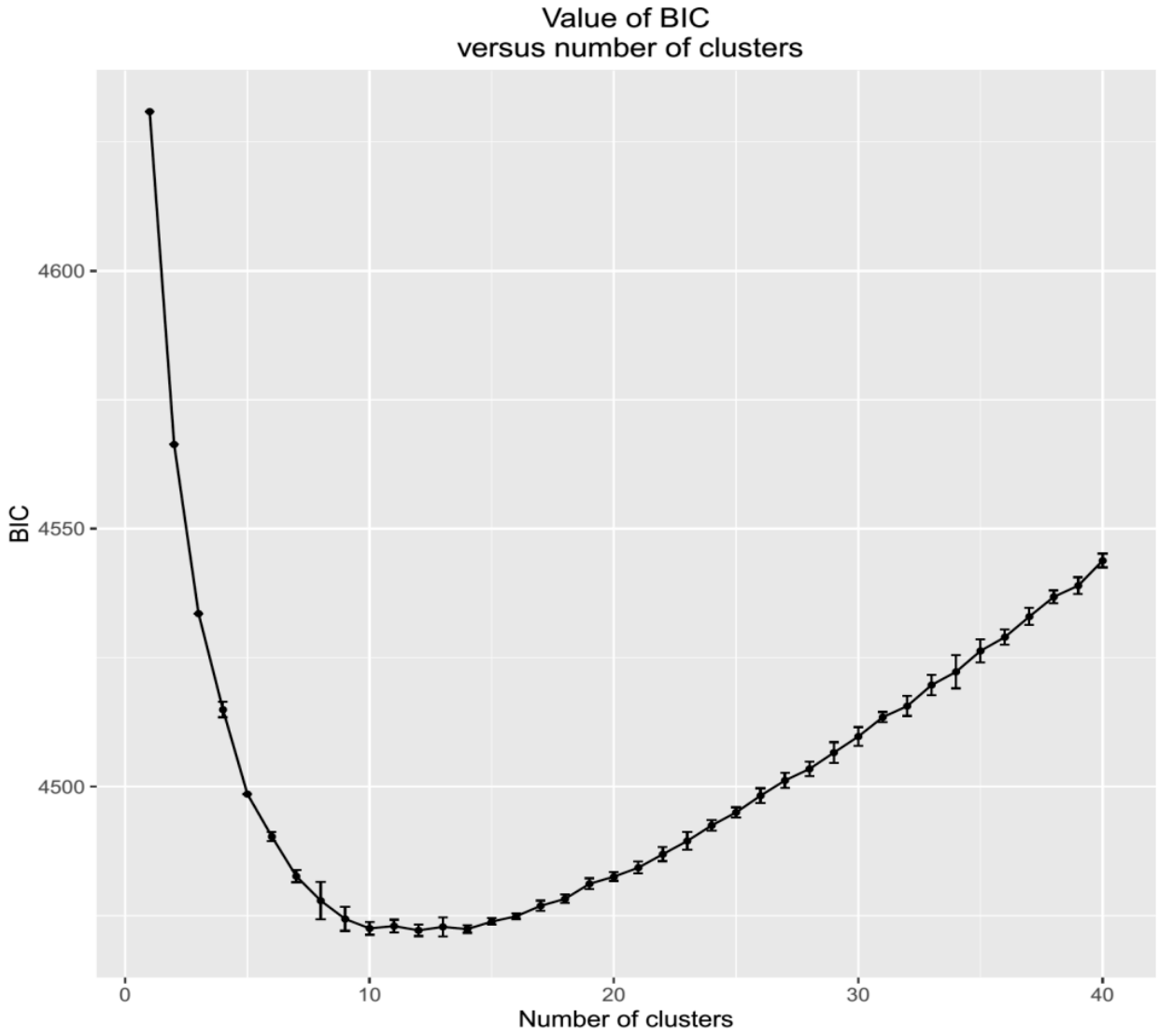

Figure S3. Structure plot for K=2-10 showing the distribution of genetic variation among *Helicobacter pylori* strains across eastern Eurasia and the Americas. The subpopulation colour key at K10 is given directly below the Structure plot. Clustering was consistent between Structure and DAPC analyes with the exception that hspKet was not identified through by Structure. Instead at K10, Structure reveal additional population clustering among Nepal and South-East Asia (here provisionally called hspNepal). Abbreviations: Tuvan (KZ), Kyzyl; Tuvan (TD), Todzha; Mongolia (UB), Ulan Bator; Mongolia (UG), Ulan Goom.

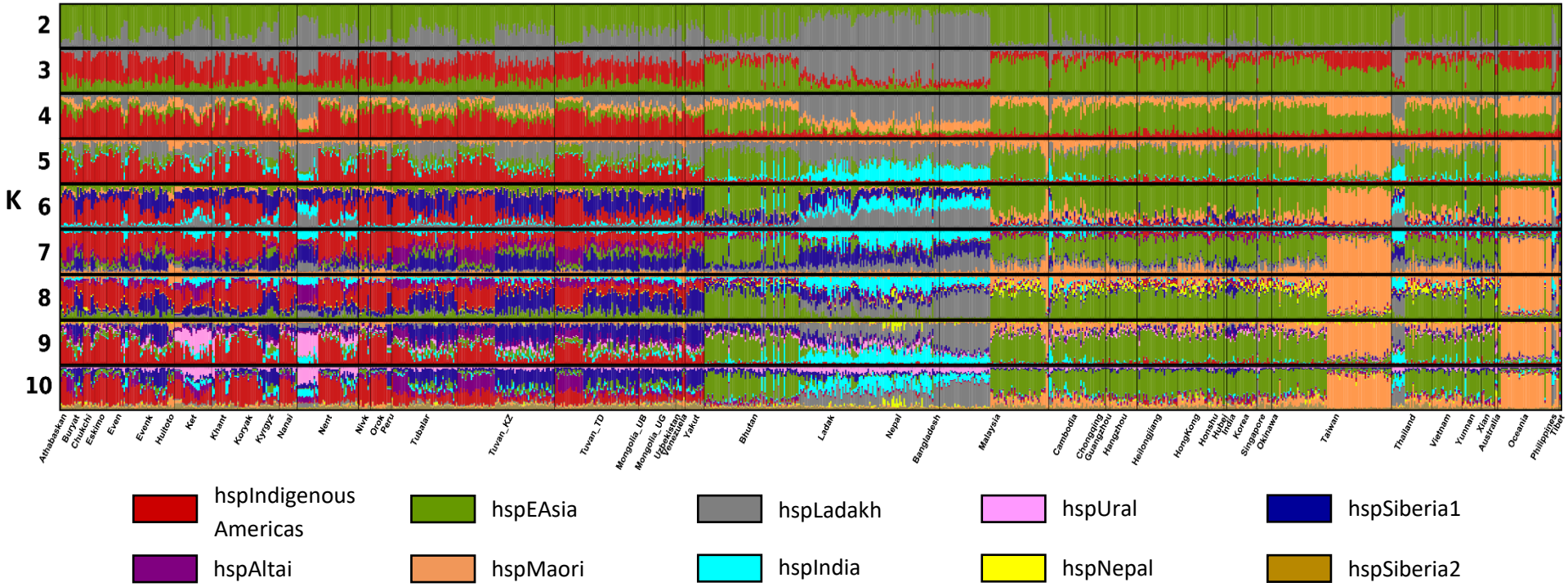

Figure S4. DAPC Scatterplot plotting discriminant functions 2 and 3 for *H. pylori* strains across eastern Eurasia and the Americas. Insets show the amount of PCA and DA variation retained for the analysis

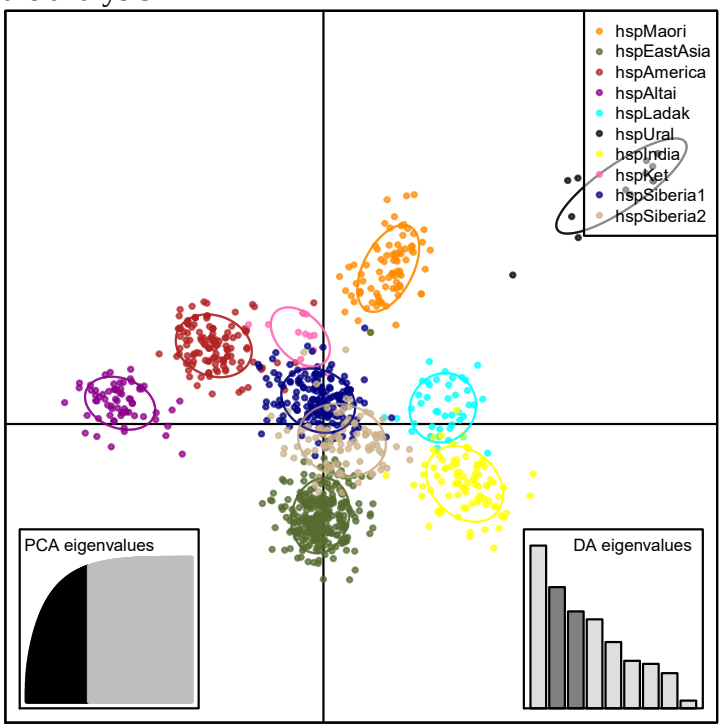

Figure S5. DAPC Scatterplot plotting discriminant functions 3 and 4 for *H. pylori* strains across eastern Eurasia and the Americas. Insets show the amount of PCA and DA variation retained for the analysis

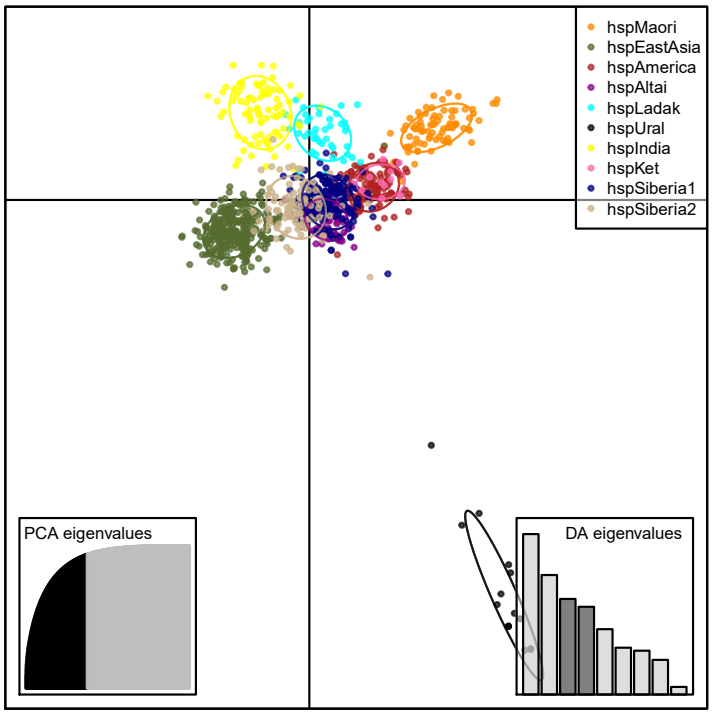

Figure S6. Global phylogenomic patterns of relatedness among *H. pylori* populations obtained through maximum likelihood analysis of genomic sites free of recombination using Gubbins[11]. Nodal bootstrap values were obtained using IQ-TREE[12] and all nodes with less than 95% support were collapsed for interpretation. **A.** The clonal structure of *H. pylori* using 40 non-admixed genomes from populations representing the total diversity of *Helicobacter pylori*. **B.** Phylogenomic structure of *H. pylori* after the addition of 54 newly sequenced Siberian genomes (denoted by stars), greatly increasing the diversity of this bacterium in Eurasia.

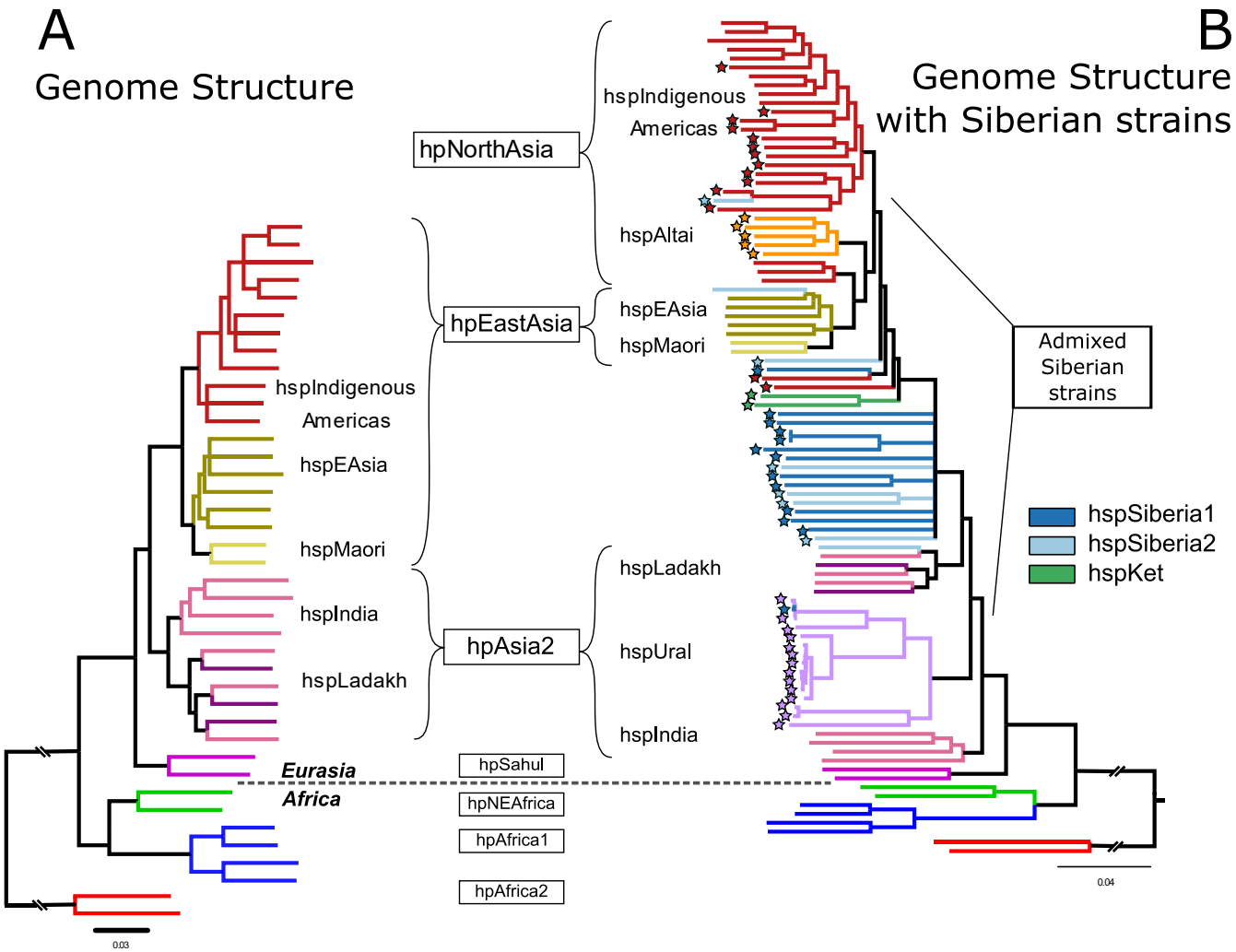

Figure S7. Tree-like (top) and admixture (bottom) models that were used to determine the origins of the newly defined populations hspSiberia1 and hspSiberia2 (unlabelled blue wedges). A2, hpAsia2; AM, hpNorthAsia; EA, hpEastAsia.

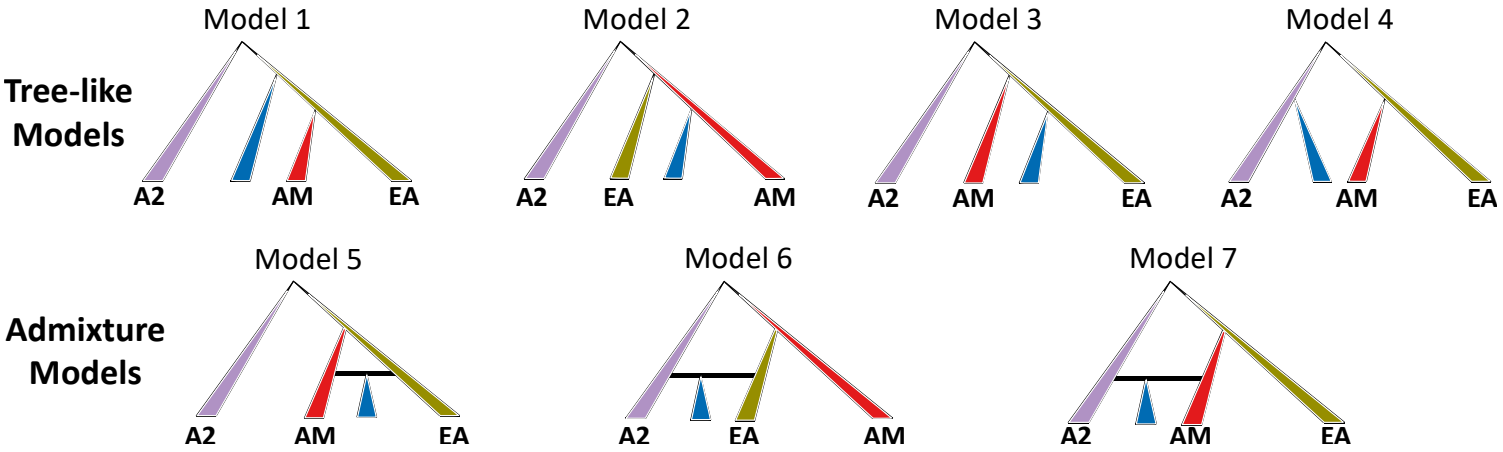

Figure S8. Model choice for the origin of hspSiberia1. Goodness of fit of the observed data to the simulate data was performed using principal component analysis of the best 3000 simulations for each model. Plots of principal components 1 and 2 are displayed below, showing that the simulated data were able to generate the observed variation. The orange dot represents the observed data. The best model was Model 7.

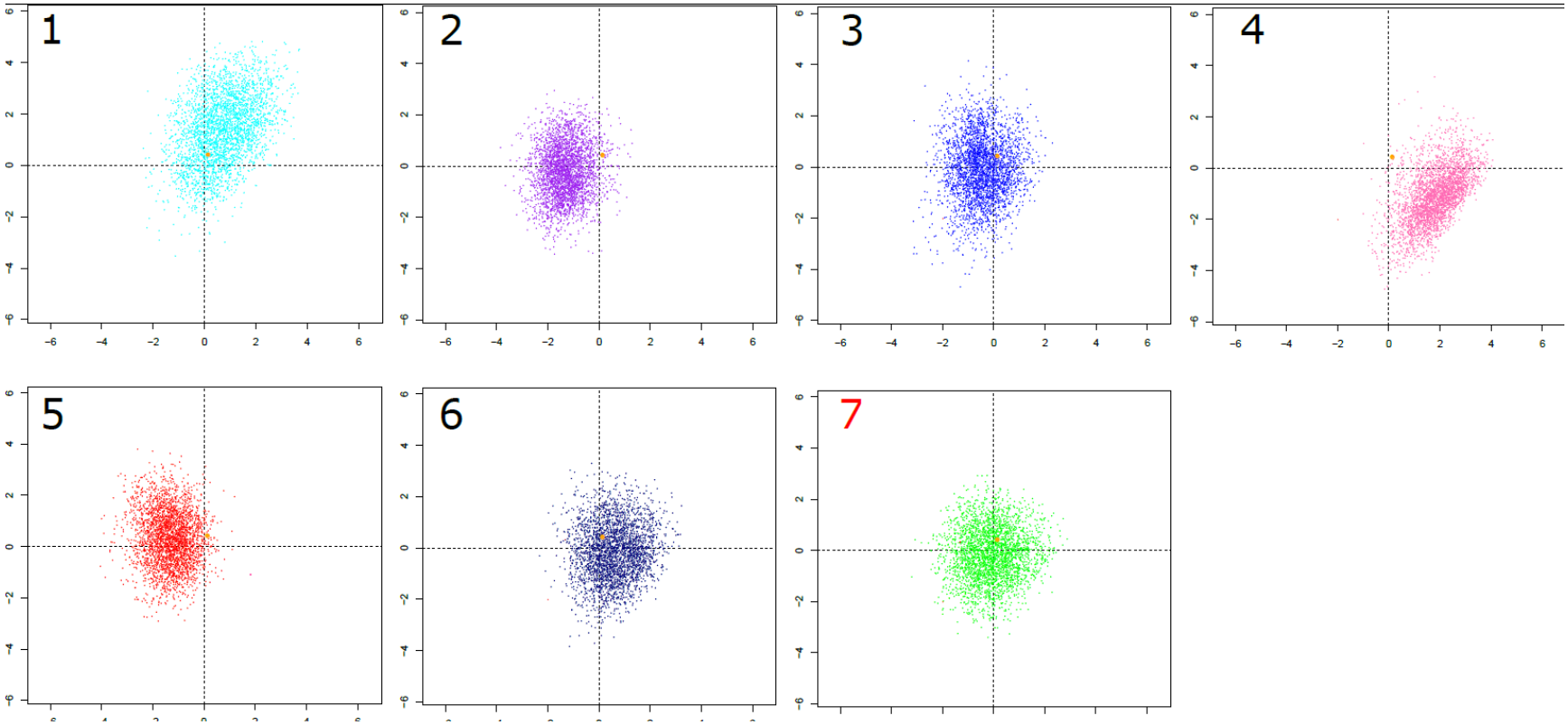

Figure S9. Model choice for the origin of hspSiberia2. Goodness of fit of the observed data to the simulated data was performed using principal component analysis of the best 3000 simulations for each model. Plots of principal components 1 and 2 are displayed below, showing that the simulated data were able to generate the observed variation. The orange dot represents the observed data. The best model was Model 6.

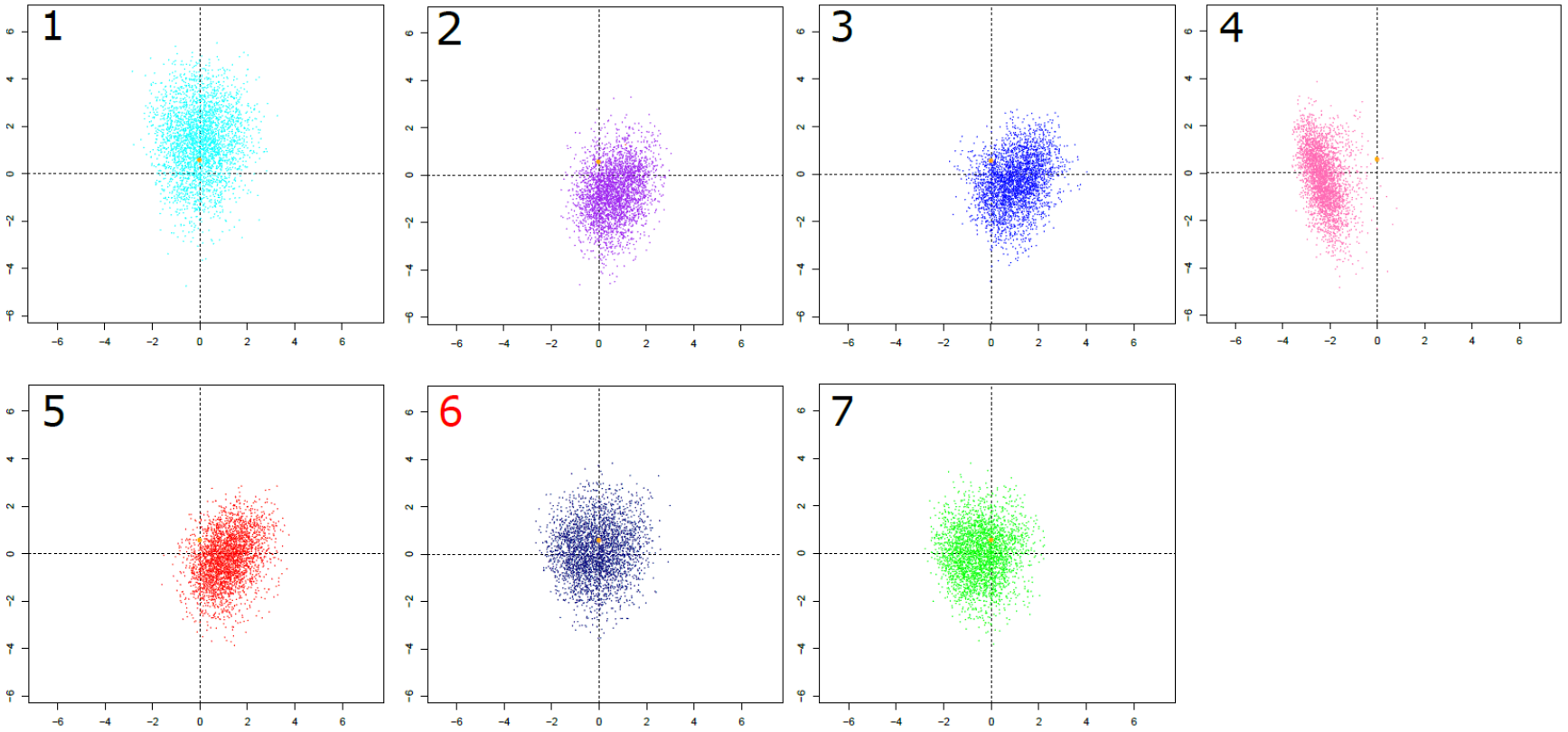

Figure S10. Posterior distributions of model parameters for the best model (Model 7) for the origin of hspSiberia1 based on 5,000 best-fitting simulations.

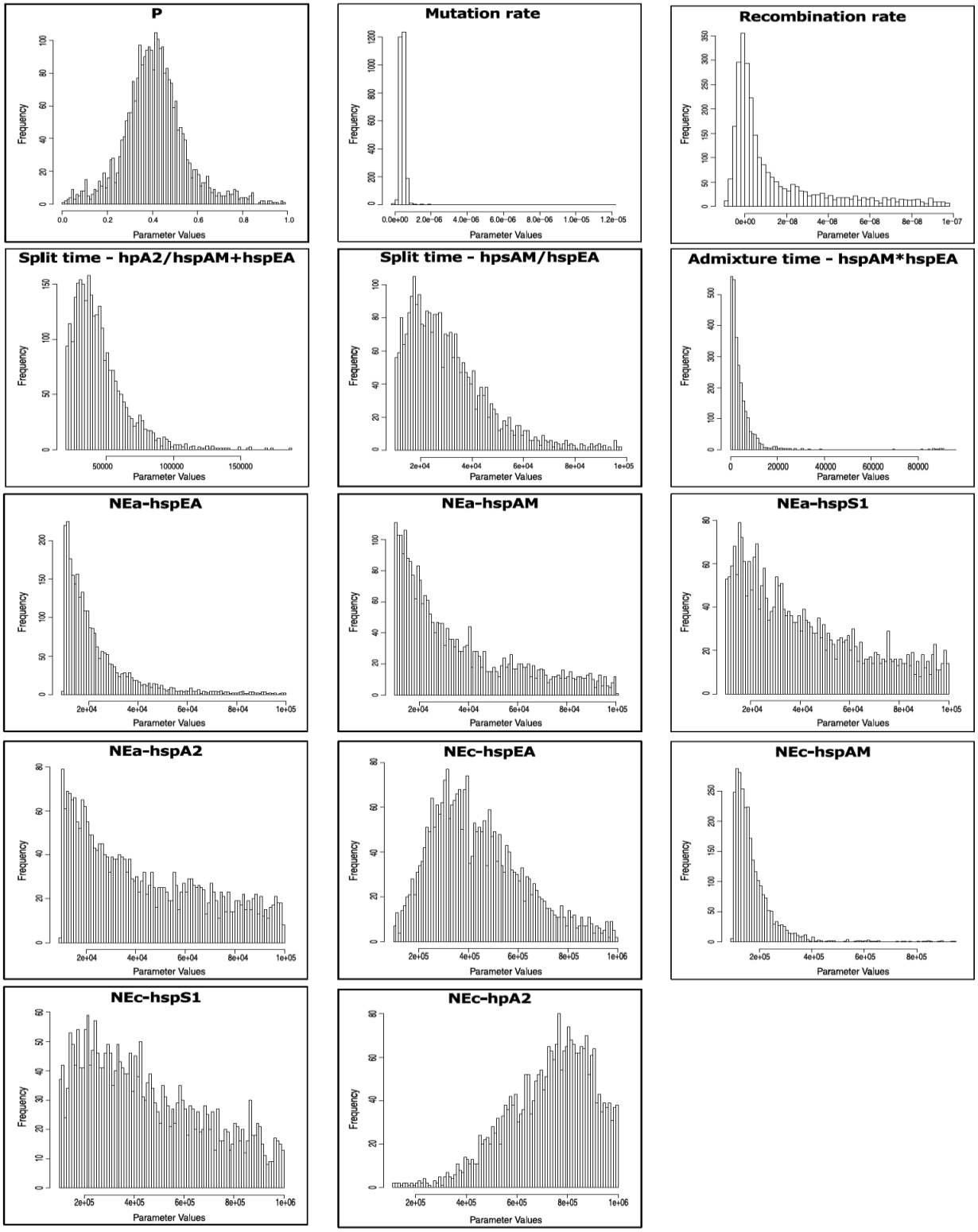

Figure S11. Posterior distributions of model parameters for the best model (Model 6) for the origin of hspSiberia2 based on 5,000 best-fitting simulations.

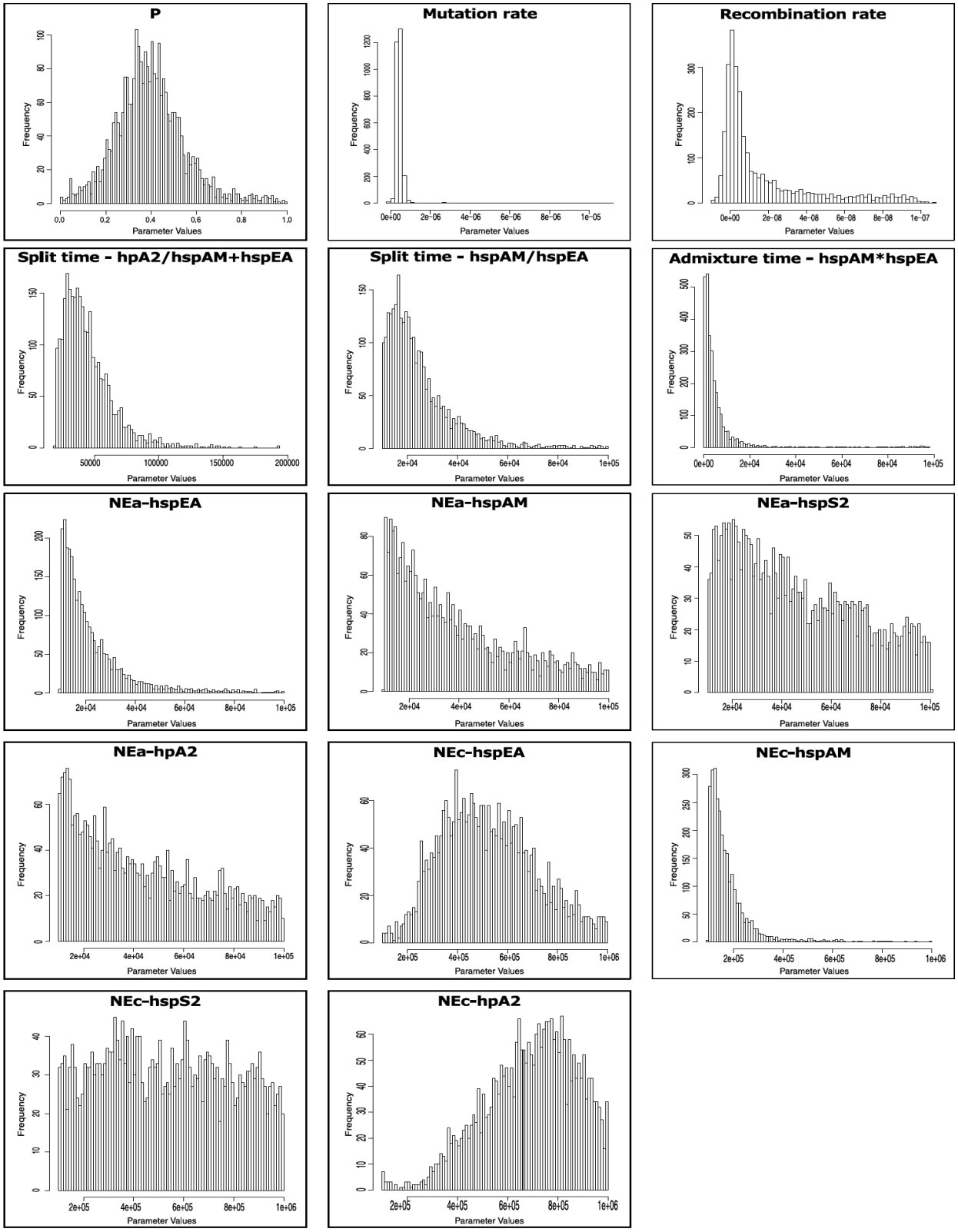

Figure S12. Admixture models inferring the origin of hspKet. This subset of models was designed by combining the best models from the previous analysis (Models 6 and 7), while allowing the evolution of hspKet (green-filled population) through admixture between populations. A2, hpAsia2; AM, hpNorthAsia; EA, hpEastAsia; S1, hspSiberia1; S2, hspSiberia2.

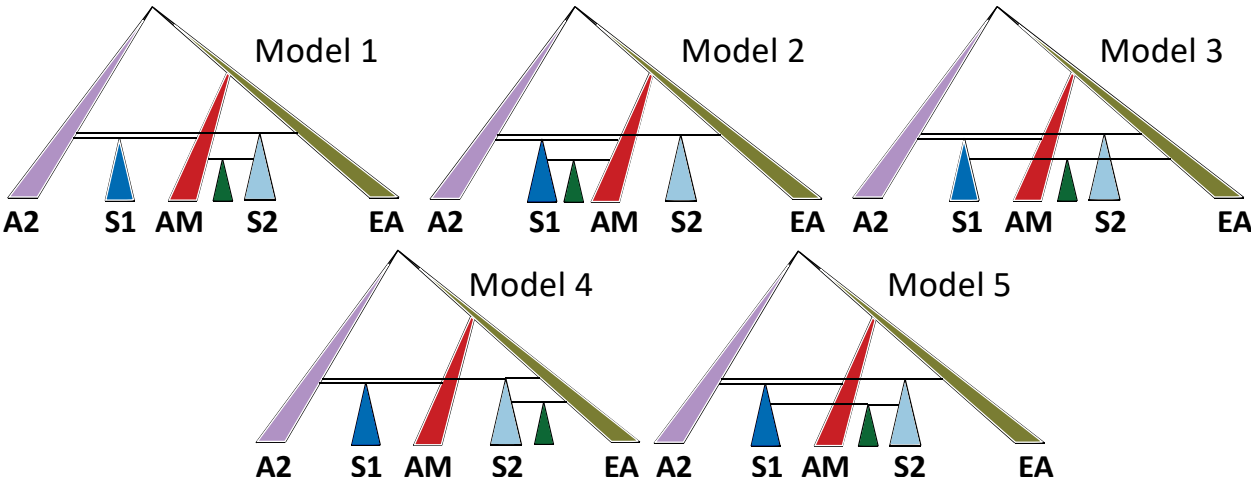

Figure S13. Model choice for the origin of hspKet. Goodness of fit of the observed data to the simulated data was performed using principal component analysis of the best 3000 simulations for each model. In this case, components 3 and 4 were most visually informative about the structure of genetic variation and they show that the simulated data were able to generate the observed variation. The orange dot represents the observed data. The best model was Model 1.

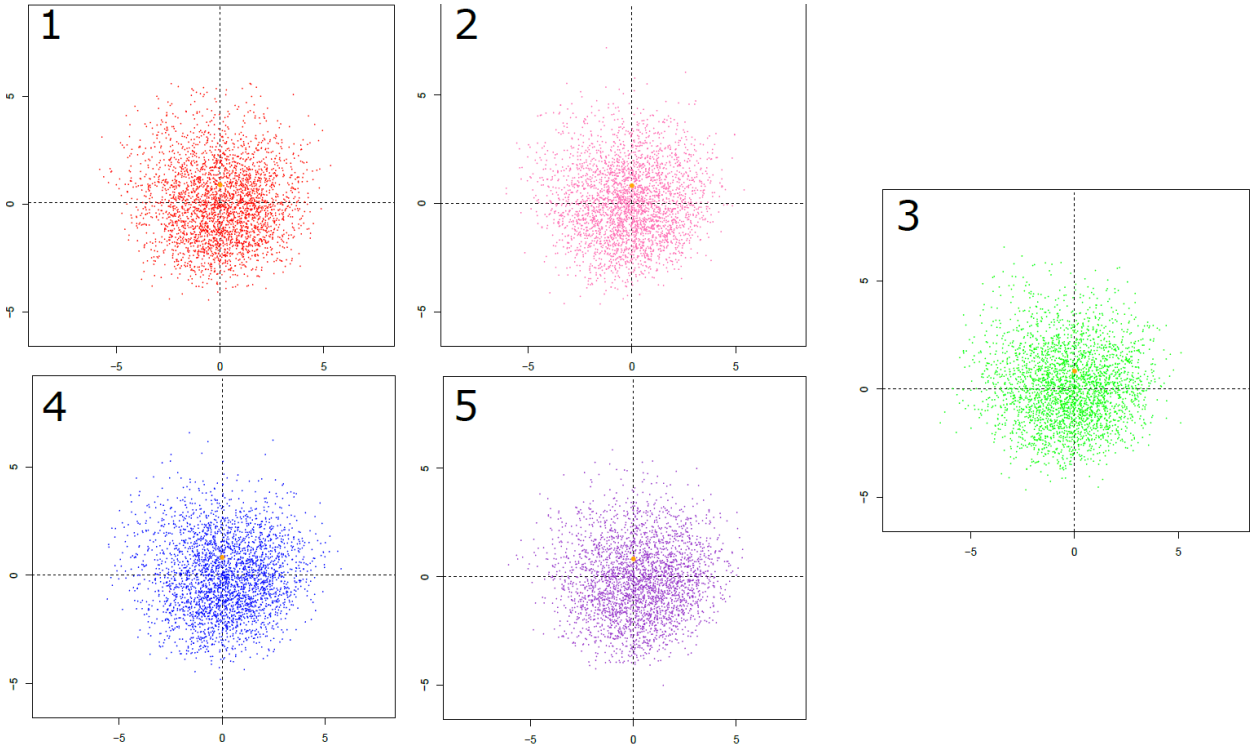

Figure S14. Posterior distributions of model parameters for the best model (Model 1) for the origin of hspKet based on 5,000 best-fitting simulations.

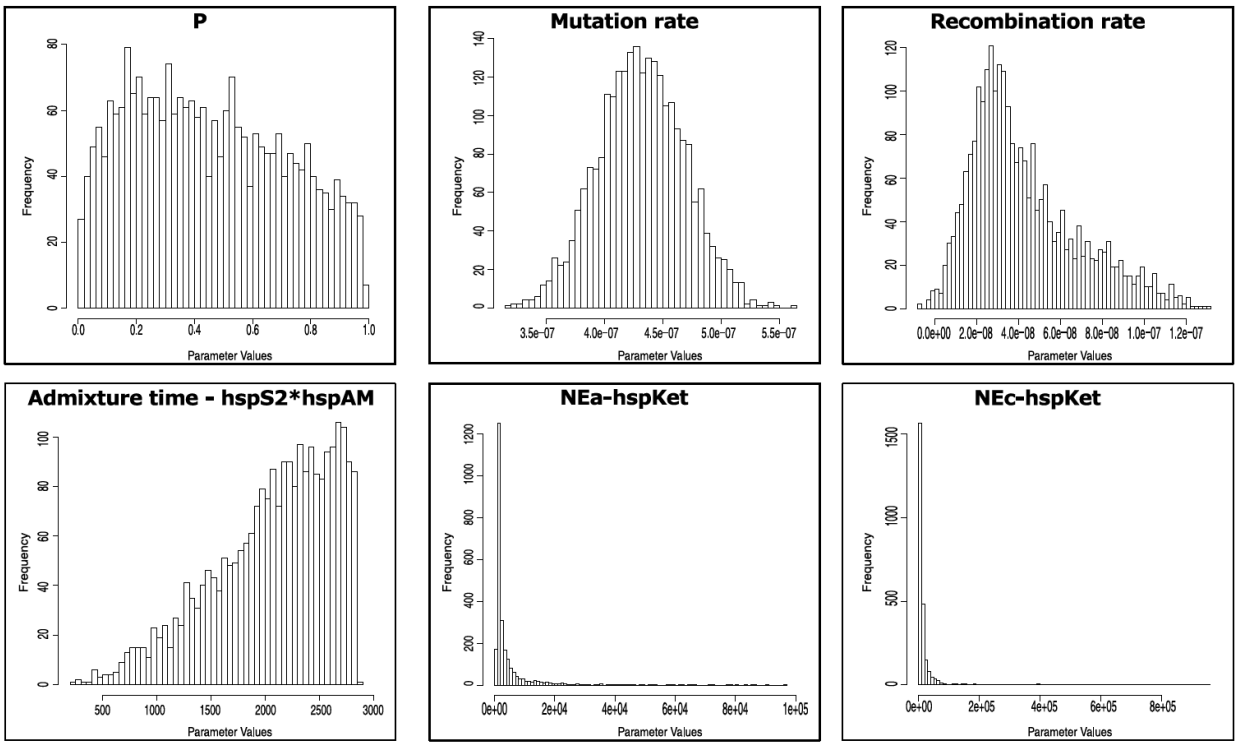

Figure S15. Tree-like model for the divergence of hspAltai. Goodness of fit of the observed data to the simulated data was performed using principal component analysis of the best 3000 simulations. A plot of principal components 1 and 2 is displayed below, showing that the simulated data were able to generate the observed variation. The orange dot represents the observed data.

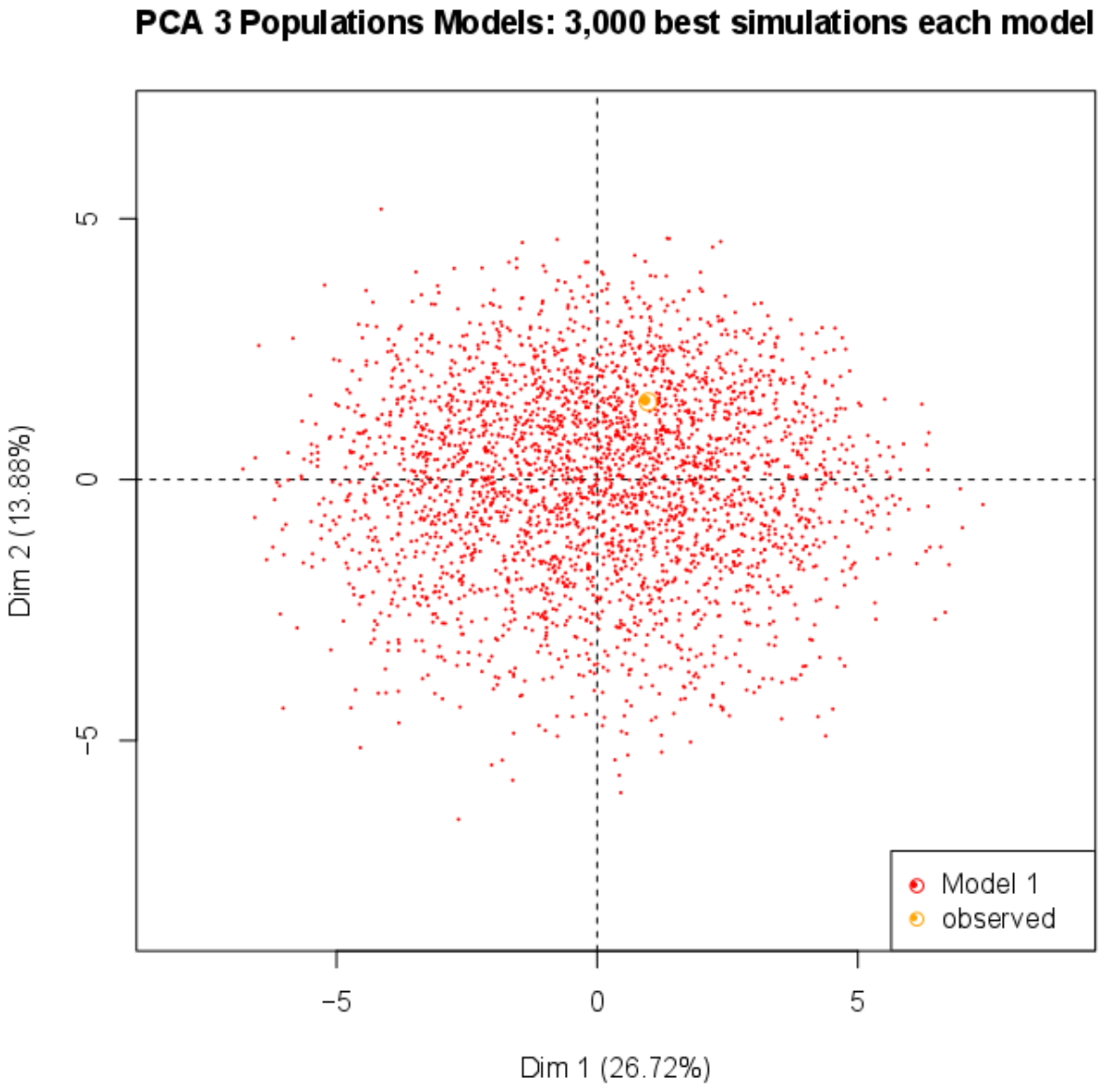

Figure S16. Posterior distributions of model parameters for the divergence of hspAltai from other hpNorthAsia strains, based on 5,000 best-fitting simulations.

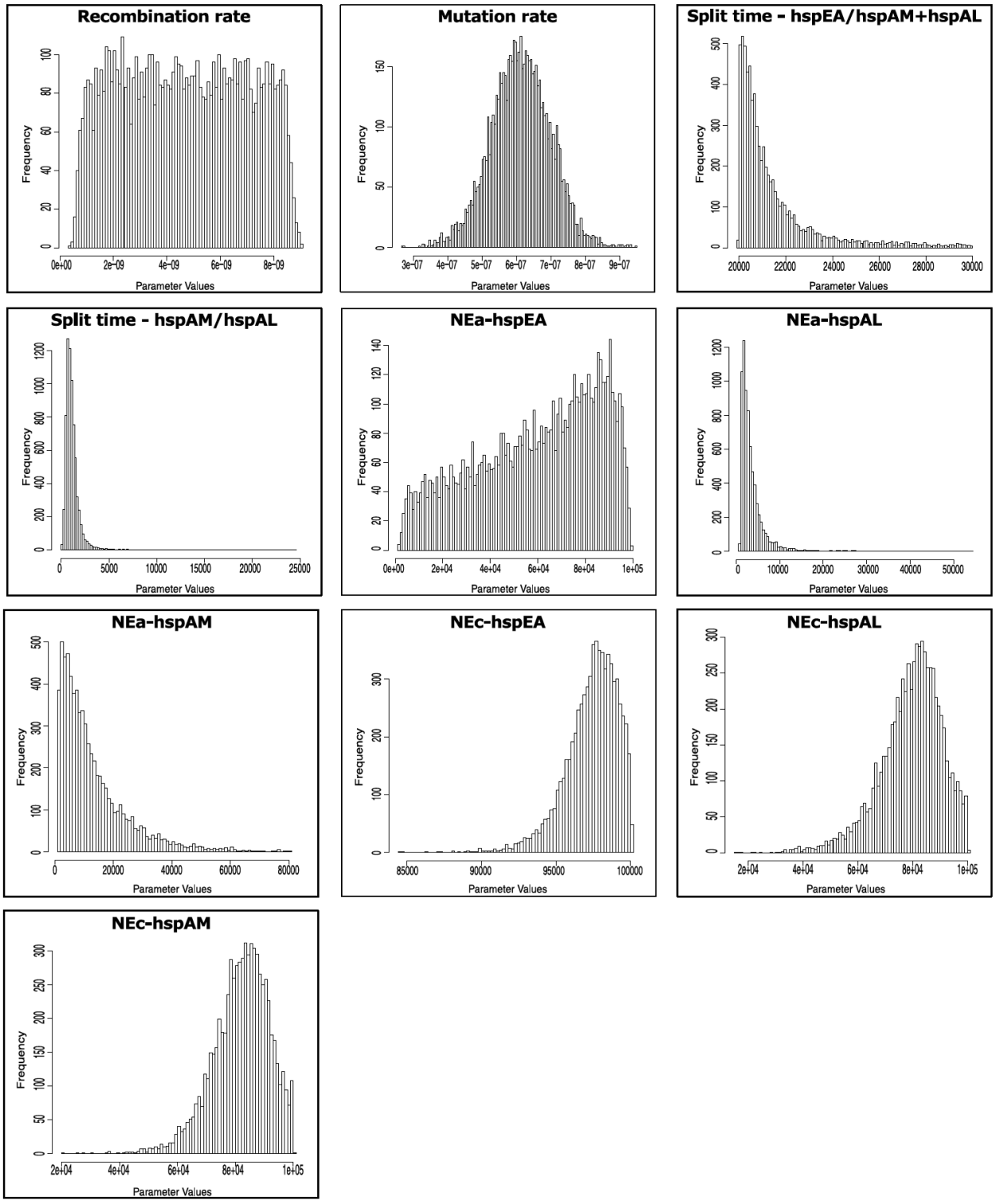

Figure S17. Results of discriminant analysis of principal components (DAPC) on the hspIndigenousAmericas data set. 123 *Helicobacter pylori* strains from 17 populations across eastern Eurasia and the Americas were consistently divided into four optimal sub-population clusters in ten independent runs of the analysis, based on the Bayesian information criterion (BIC).

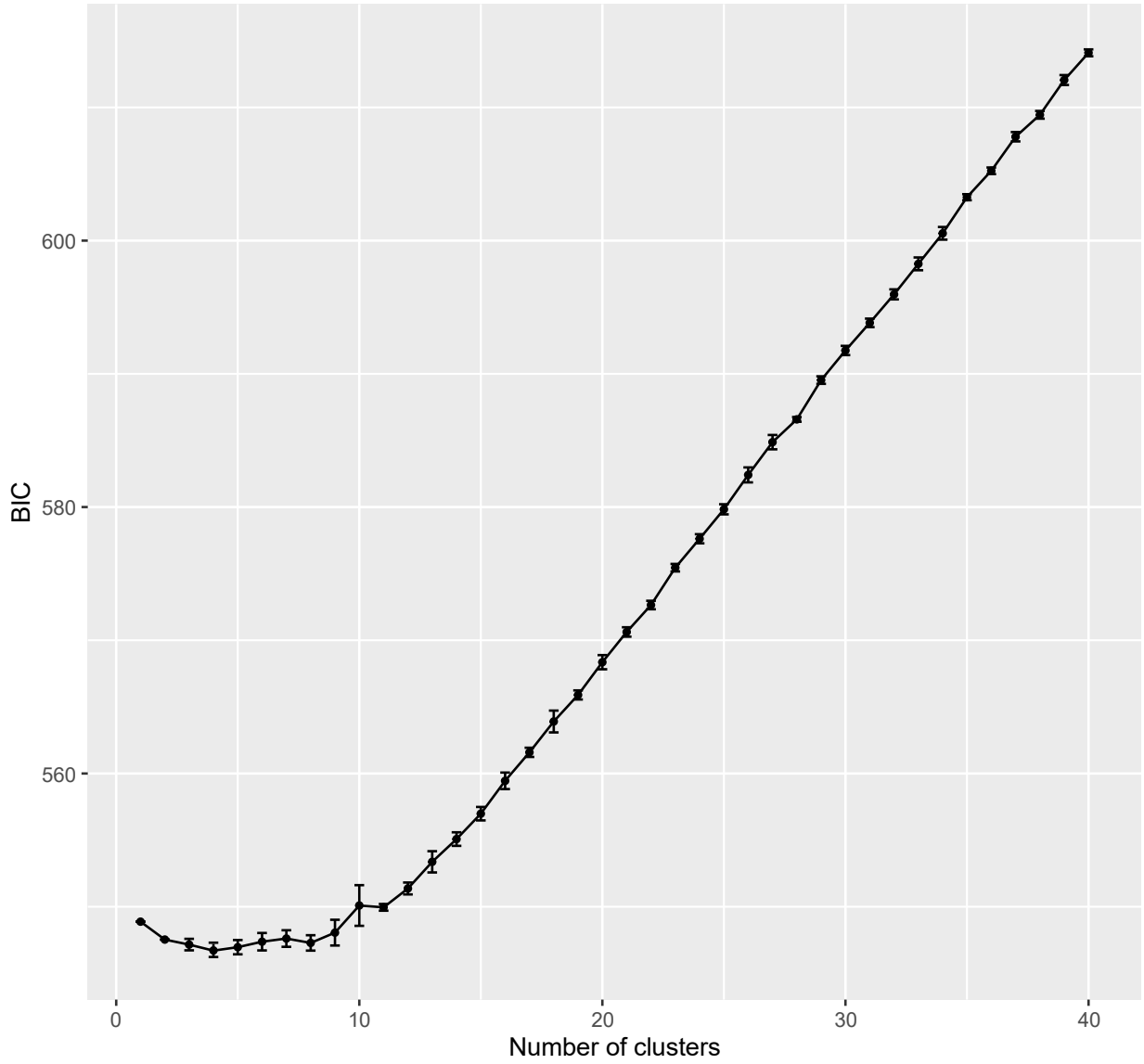

Figure S18. Tree-like and admixture models depicting the putative histories for *Helicobacter pylori*'s colonisation of the Americas. These models divide the subpopulation hspIndigenousAmericas into four geographic locations (NS, northern Siberia; ES, eastern Siberia; KC, Kamchatka; AM, America) using an east Asian population (hspEA) as outgroup.

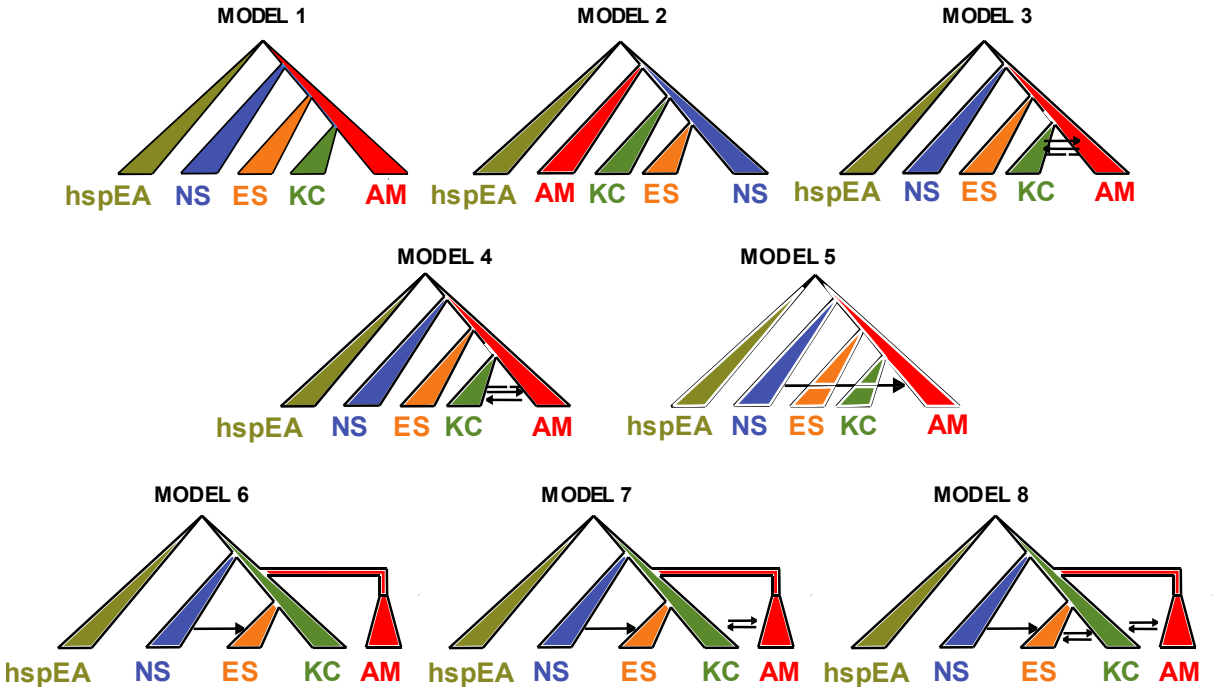

Figure S19. Model choice for the colonisation of the Americas by hspIndigenousAmericas. Goodness of fit of the observed data to the simulated data was performed using principal component analysis. Plots of principal components 1 and 2 are displayed below, showing that the simulated data were able to generate the observed variation. The orange dot represents the observed data. The best model was model 7, followed by Model 8.

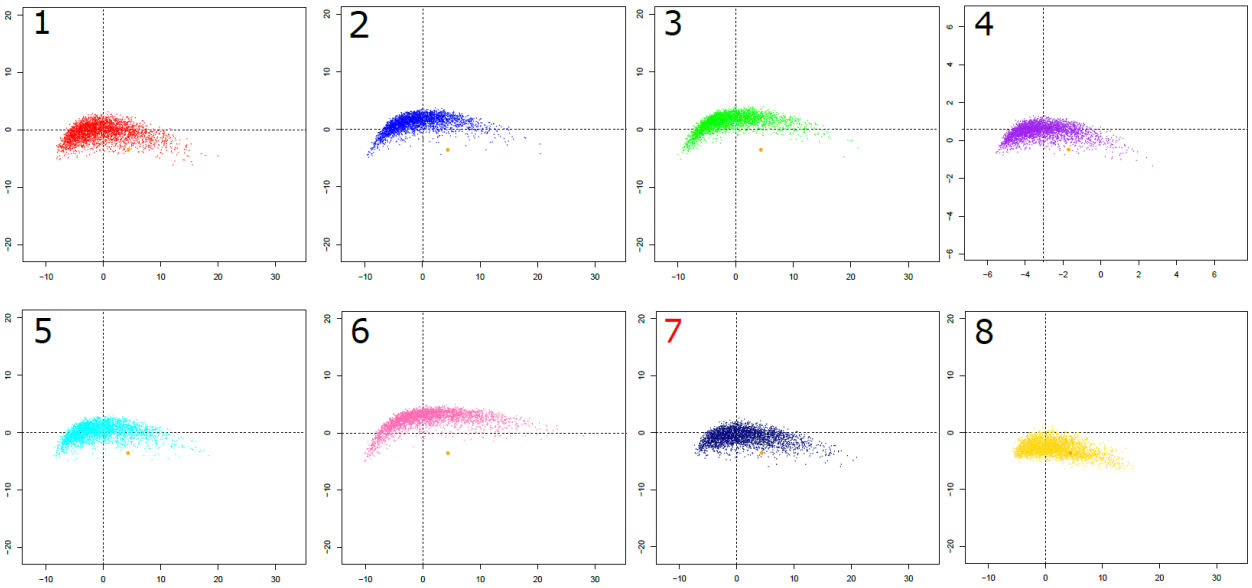

Figure S20. Posterior distributions of model parameters for the best model (Model 7) for the colonisation of Siberia and the Americas by hspIndigenousAmericas based on 5000 best-fitting simulations. This figure is continued on the following page.

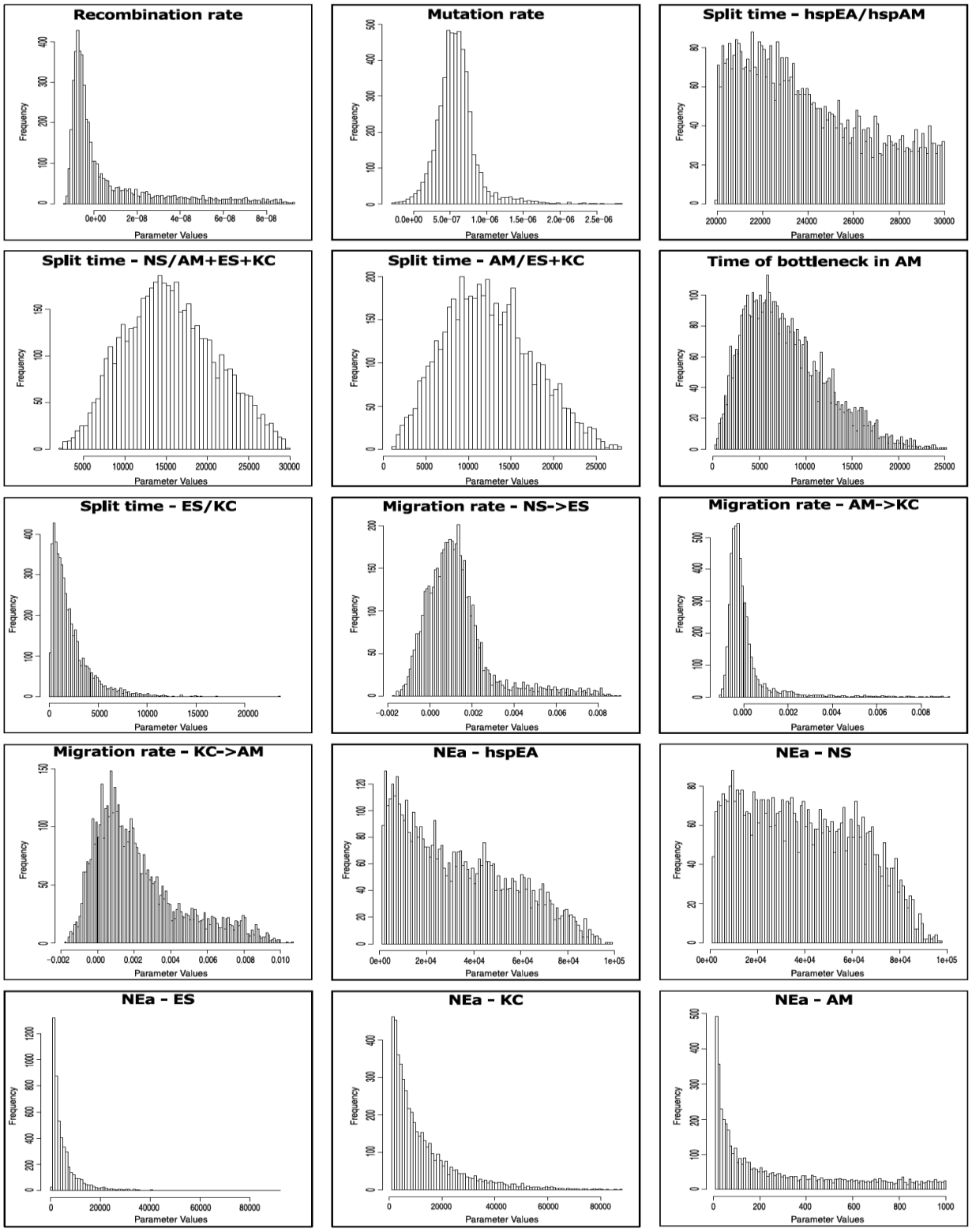

Figure S20. Continued. Posterior distributions of model parameters for the best model (Model 7) for the colonisation of Siberia and the Americas by hspIndigenousAmericas based on 5000 best-fitting simulations

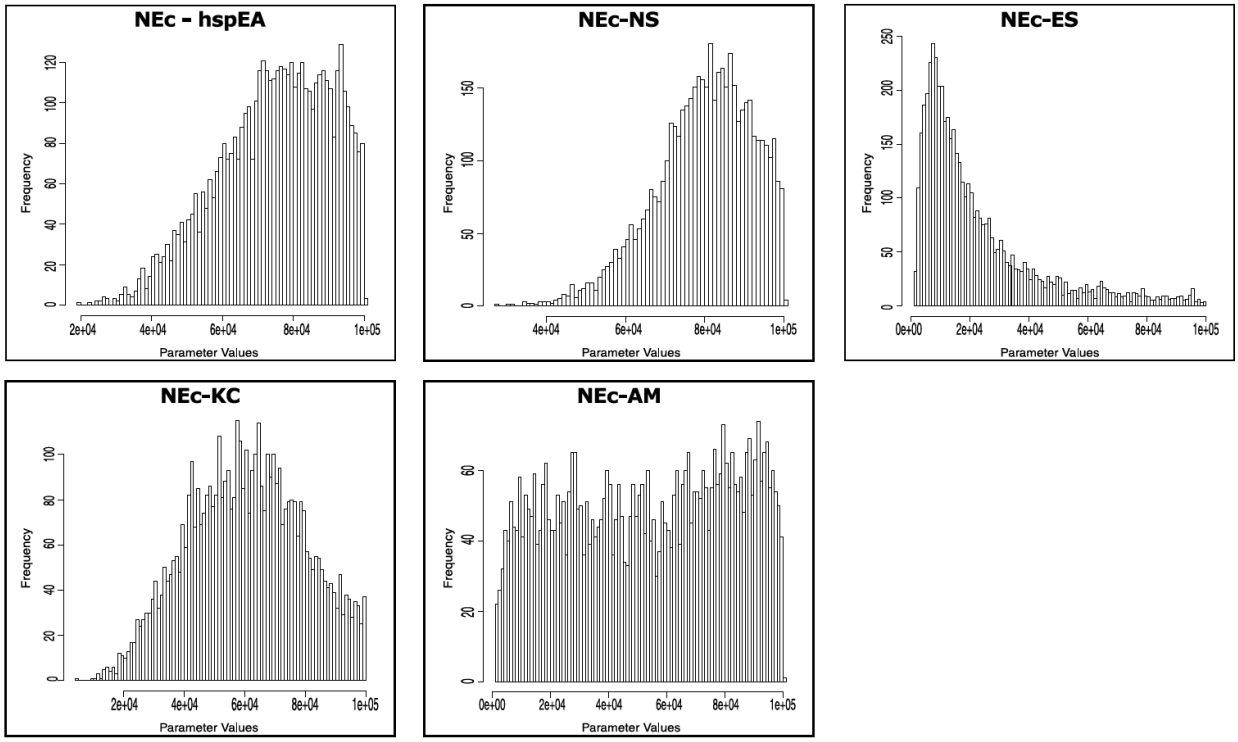

Figure S21. Head to head comparison of MLST and whole genome sequence data for the same 79 *H. pylori* strains from Asia and the Americas. A. Bayesian information criteria (BIC) plot summarising five DAPC runs fitting the MLST data in to 1-15 population clusters (K), with optimal K=3. B. Bayesian information criteria (BIC) plot summarising five DAPC runs fitting the genome data in to 1-15 population clusters (K), with optimal K=3. C. Scatterplot of discriminant functions 1 and 2 showing three distinct populations for *H. pylori* MLST data. D. Scatterplot of discriminant functions 1 and 2 showing three distinct populations for *H. pylori* genome data.

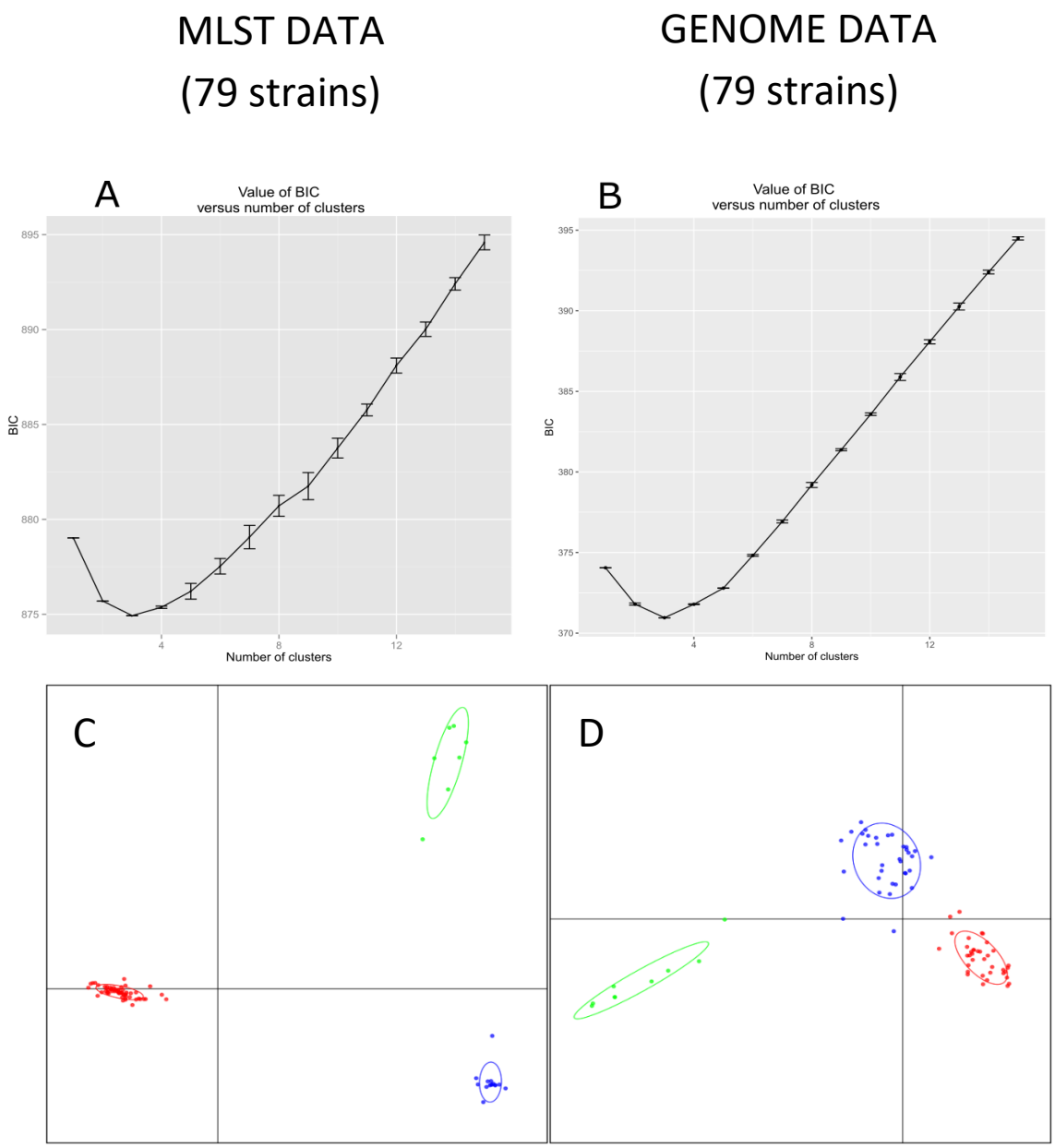

Table S1. Summary of Siberian ethnicities sampled for *Helicobacter pylori*.

Subregion in Siberia	**Ethnicity**	**Language family**	**Locality(ies)**	**Country**	**Year**	**Biopsies taken**	***H. pylori*+ Biopsies**	**European *H. pylori***	**Indigenous *H. pylori***
North-Western Siberia	Khant	Uralic	Muzhi, Shurishkari, Ovgort	Russia	2006	50	30	17	**13**
	Nenet	Uralic	Novi-port	Russia	2006	101	58	17	**41**
Central Siberia	Tuvan	Turkic	Kyzyl	Russia	2006	100	69	4	**65**
	Tuvan	Turkic	Todzha	Russia	2005	121	62	6	**56**
	Tubalar	Turkic	Altai	Russia	2005	79	43	0	**43**
	Buryat	Mongolic	Ulan Ude	Russia	2004	15	12	1	**11**
	Mongolian	Mongolic	Ulaanbatar	Mongolia	2004	80	11	2	**9**
	Mongolian	Mongolic	Ulaangom	Mongolia	2004	91	16	0	**16**
Northern Siberia	Ket	ISOLATE	Sulomai	Russia	2006	73	39	14	**25**
	Evenk	Tungusic	Yessei, Tura, Nidim, Tutongan, Chilinda	Russia	2006	79	38	11	**27**
	Yakut	Turkic	Yakutsk	Russia	2005	80	17	4	**13**
Eastern Siberia	Nanai	Tungusic	Troitckoe, Yandonga, Naichin, Arsenevo, Bulova, Uchta, Bogorodskoe	Russia	2006	40	21	9	**12**
	Ulchi	Tungusic	Bogorodskoe	Russia		12	5	5	**0**
	Orok	Tungusic	Sakhalin-Nogliki	Russia	2006	62	39	25	**14**
	Nivkh	ISOLATE	Sakhalin-Val	Russia	2006	32	17	9	**8**
Beringia	Even	Tungusic	Kamchatka- Esso-Anavgai	Russia	2006	43	32	18	**14**
	Koryak	Chukotko-Kamchatkan	Kamchatka-Palana	Russia	2006	47	35	14	**21**
	Chukchi	Chukotko-Kamchatkan	Bilibino-Keperveyem	Russia	2006	47	12	4	**8**
TOTALS						**1175**	**556**	**160**	**396**

Table S2. Proportions of ancestry of Siberian genomes belonging to hspSiberia1 and hspSiberia2 as determined by fineSTRUCTURE. The majority of Siberian ancestry derived from populations hpNorthAsia, hpAsia2 and hpEastAsia (red, bold text)

Strain name	Population	hpAfrica2	**hpNorthAsia**	hpAfrica1	**hpEastAsia**	hpNEAfrica	**hpAsia2**	hspKet	hpSahul
BURYAT19	hspSiberia1	0.00200339	0.53263592	0.00847747	0.10699452	0.00870383	0.28177805	0.04926414	0.01014261
BURYAT27	hspSiberia1	0.00227122	0.50955622	0.01081601	0.11267897	0.01264828	0.29767328	0.04197546	0.01238050
BURYAT49	hspSiberia1	0.00210493	0.52823952	0.00737572	0.10934907	0.00814839	0.27945614	0.05463337	0.01069278
Chukchi08	hspSiberia1	0.00181166	0.56451038	0.01130448	0.09541744	0.01061163	0.26183825	0.04463255	0.00987354
Even14	hspSiberia1	0.00178145	0.66650442	0.00927086	0.07990409	0.00752324	0.19237203	0.03680007	0.00584377
Evenky01	hspSiberia1	0.00154396	0.57675587	0.01039541	0.09771857	0.01090891	0.24972711	0.04448965	0.00846044
Evenky65	hspSiberia1	0.00300975	0.47910418	0.01945846	0.09213270	0.02538671	0.31967428	0.05055172	0.01068213
Evenky73	hspSiberia1	0.00164861	0.58637229	0.00928074	0.09834041	0.00962790	0.24319605	0.04332976	0.00820418
Khanty27	hspSiberia1	0.00004860	0.05940211	0.00060789	0.00396451	0.00027515	0.93155927	0.00359657	0.00054560
altai11	hspSiberia1	0.00207224	0.46983974	0.01171256	0.11771640	0.01295378	0.31809592	0.05573393	0.01187537
altai59	hspSiberia1	0.00197556	0.51796463	0.00796301	0.10569359	0.00808529	0.29617305	0.04867623	0.01346857
mong49	hspSiberia1	0.00174533	0.48511301	0.00983089	0.11756756	0.00959120	0.31915779	0.04731720	0.00967695
yak97	hspSiberia1	0.00181747	0.51754449	0.00842020	0.11149756	0.00705227	0.27583982	0.06558995	0.01223816
Khanty47	hspSiberia2	0.00082920	0.77242757	0.00315738	0.02006211	0.00341365	0.17397311	0.02286901	0.00326778
Nanai30	hspSiberia2	0.00128184	0.60005864	0.00707512	0.13704366	0.00594548	0.20553973	0.03358056	0.00947490
TUVAB15	hspSiberia2	0.00182802	0.49552021	0.00917331	0.11653393	0.00850671	0.30490384	0.05311093	0.01042299
Tuvac46	hspSiberia2	0.00246792	0.4845393	0.00949336	0.11720652	0.01105366	0.31017206	0.05314950	0.01191762
Tuvac80	hspSiberia2	0.00161871	0.52408067	0.00750231	0.11594997	0.00884579	0.28153928	0.05104410	0.00941909
mong44	hspSiberia2	0.00260527	0.46816329	0.01143128	0.12647345	0.01483060	0.31124282	0.05179275	0.01346047
Total Ancestry		**0.00181395**	**0.51780697**	**0.00909192**	**0.09906553**	**0.00969013**	**0.30810063**	**0.04484934**	**0.00958144**

Table S3. Details of prior distributions of all model parameters to infer the origin of hspSiberia1 and hspSiberia2 (seven models). These complement the *a priori* visual descriptions of each model in Fig. S4. NEa, Ancestral effective population size; NEc, current effective population size; T, time of population split; /, denotes a population split; +, denotes nested populations; *, denotes admixture between two populations. Rules for the timing of evolutionary events (T) were as follows: Tadm was always less than T1, which was less than T2, which was less than T3.

MODEL1

PARAMETER	DISTRIBUTION	MINIMUM	MAXIMUM
NEa-hpAsia2	uniform	10 000	1 000 000
NEa-hpNorthAsia	uniform	10 000	1 000 000
NEa-hspSiberia	uniform	10 000	1 000 000
NEa-hpEastAsia	uniform	10 000	1 000 000
NEc-hpAsia2	uniform	10 000	1 000 000
NEc-hpNorthAsia	uniform	10 000	1 000 000
NEc-hspSiberia	uniform	10 000	1 000 000
NEc-hpEastAsia	uniform	10 000	1 000 000
T1(hpNorthAsia/hpEastAsia)	uniform	1 000	100 000
T2(hspSiberia/hpNorthAsia+hpEastAsia)	uniform	1 000	100 000
T3(hpAsia2/hspSiberia+hpNorthAsia+hpEastAsia)	uniform	1 000	100 000
Mutation rate	uniform	1.00E-08	1.00E-05
Recombination rate	uniform	1.00E-10	1.00E-07

MODEL2

PARAMETER	DISTRIBUTION	MINIMUM	MAXIMUM
NEa-hpAsia2	uniform	10 000	1 000 000
NEa-hpNorthAsia	uniform	10 000	1 000 000
NEa-hspSiberia	uniform	10 000	1 000 000
NEa-hpEastAsia	uniform	10 000	1 000 000
NEc-hpAsia2	uniform	10 000	1 000 000
NEc-hpNorthAsia	uniform	10 000	1 000 000
NEc-hspSiberia	uniform	10 000	1 000 000
NEc-hpEastAsia	uniform	10 000	1 000 000
T1(hpNorthAsia/hspSiberia)	uniform	1 000	100 000
T2(hpEastAsia/hspSiberia+hpNorthAsia)	uniform	1 000	100 000
T3(hpAsia2/hpEastAsia+hspSiberia+hpNorthAsia)	uniform	1 000	100 000
Mutation rate	uniform	1.00E-08	1.00E-05
Recombination rate	uniform	1.00E-10	1.00E-07

MODEL3

PARAMETER	DISTRIBUTION	MINIMUM	MAXIMUM
NEa-hpAsia2	uniform	10 000	1 000 000
NEa-hpNorthAsia	uniform	10 000	1 000 000
NEa-hspSiberia	uniform	10 000	1 000 000
NEa-hpEastAsia	uniform	10 000	1 000 000
NEc-hpAsia2	uniform	10 000	1 000 000
NEc-hpNorthAsia	uniform	10 000	1 000 000
NEc-hspSiberia	uniform	10 000	1 000 000
NEc-hpEastAsia	uniform	10 000	1 000 000
T1(hpEastAsia/hspSiberia)	uniform	1 000	100 000
T2(hpNorthAsia/hpEastAsia+hspSiberia)	uniform	1 000	100 000
T3(hpAsia2/hpNorthAsia+hpEastAsia+hspSiberia)	uniform	1 000	100 000
Mutation rate	uniform	1.00E-08	1.00E-05
Recombination rate	uniform	1.00E-10	1.00E-07

MODEL4

PARAMETER	DISTRIBUTION	MINIMUM	MAXIMUM
NEa-hpAsia2	uniform	10 000	1 000 000
NEa-hpNorthAsia	uniform	10 000	1 000 000
NEa-hspSiberia	uniform	10 000	1 000 000
NEa-hpEastAsia	uniform	10 000	1 000 000
NEc-hpAsia2	uniform	10 000	1 000 000
NEc-hpNorthAsia	uniform	10 000	1 000 000
NEc-hspSiberia	uniform	10 000	1 000 000
NEc-hpEastAsia	uniform	10 000	1 000 000
T1(hpAsia2/hspSiberia)	uniform	1 000	100 000
T2(hpNorthAsia/hpEastAsia)	uniform	1 000	100 000
T3(hpAsia2+hspSiberia/hpNorthAsia+hpEastAsia)	uniform	1 000	100 000
Mutation rate	uniform	1.00E-08	1.00E-05
Recombination rate	uniform	1.00E-10	1.00E-07

MODEL5

PARAMETER	DISTRIBUTION	MINIMUM	MAXIMUM
NEa-hpAsia2	uniform	10 000	1 000 000
NEa-hpNorthAsia	uniform	10 000	1 000 000
NEa-hspSiberia	uniform	10 000	1 000 000
NEa-hpEastAsia	uniform	10 000	1 000 000
NEc-hpAsia2	uniform	10 000	1 000 000
NEc-hpNorthAsia	uniform	10 000	1 000 000

NEc-hspSiberia	uniform	10 000	1 000 000
NEc-hpEastAsia	uniform	10 000	1 000 000
Tadm(hpNorthAsia*hpEastAsia)	uniform	1 000	100 000
T1(hpNorthAsia/hpEastAsia)	uniform	1 000	100 000
T2(hpAsia2/hpNorthAsia+hpEastAsia)	uniform	1 000	100 000
Mutation rate	uniform	1.00E-08	1.00E-05
Recombination rate	uniform	1.00E-10	1.00E-07
P	uniform	0.000001	0.99999

MODEL6

PARAMETER	DISTRIBUTION	MINIMUM	MAXIMUM
NEa-hpAsia2	uniform	10 000	1 000 000
NEa-hpNorthAsia	uniform	10 000	1 000 000
NEa-hspSiberia	uniform	10 000	1 000 000
NEa-hpEastAsia	uniform	10 000	1 000 000
NEc-hpAsia2	uniform	10 000	1 000 000
NEc-hpNorthAsia	uniform	10 000	1 000 000
NEc-hspSiberia	uniform	10 000	1 000 000
NEc-hpEastAsia	uniform	10 000	1 000 000
Tadm(hpAsia2*hpEastAsia)	uniform	1 000	100 000
T1(hpNorthAsia/hpEastAsia)	uniform	1 000	100 000
T2(hpAsia2/hpNorthAsia+hpEastAsia)	uniform	1 000	100 000
Mutation rate	uniform	1.00E-08	1.00E-05
Recombination rate	uniform	1.00E-10	1.00E-07
P	uniform	0.000001	0.99999

MODEL7

PARAMETER	DISTRIBUTION	MINIMUM	MAXIMUM
NEa-hpAsia2	uniform	10 000	1 000 000
NEa-hpNorthAsia	uniform	10 000	1 000 000
NEa-hspSiberia	uniform	10 000	1 000 000
NEa-hpEastAsia	uniform	10 000	1 000 000
NEc-hpAsia2	uniform	10 000	1 000 000
NEc-hpNorthAsia	uniform	10 000	1 000 000
NEc-hspSiberia	uniform	10 000	1 000 000
NEc-hpEastAsia	uniform	10 000	1 000 000
Tadm(hpAsia2*hpNorthAsia)	uniform	1 000	100 000
T1(hpNorthAsia/hpEastAsia)	uniform	1 000	100 000
T2(hpAsia2/hpNorthAsia+hpEastAsia)	uniform	1 000	100 000
Mutation rate	uniform	1.00E-08	1.00E-05
Recombination rate	uniform	1.00E-10	1.00E-07
P	uniform	0.000001	0.99999

Table S4. Stability of model posterior probabilities under weighted multinomial logistic regression. The posterior probabilities for each model were calculated for three subsets (thresholds) of 10,000, 50,000 and 100,000 simulations. The posterior probabilities of the best model (highlighted in red text) must be high relative to other models and consistent across the three different thresholds.

A. Origin of hspSiberia1

	MODEL1	MODEL2	MODEL3	MODEL4	MODEL5	MODEL6	MODEL7
10,000	0.00	0.04	0.00	0.00	0.41	0.00	0.55
50,000	0.00	0.07	0.00	0.00	0.29	0.00	0.64
100,000	0.00	0.05	0.00	0.00	0.29	0.01	0.65

B. Origin of hspSiberia2

	MODEL1	MODEL2	MODEL3	MODEL4	MODEL5	MODEL6	MODEL7
10,000	0.00	0.00	0.00	0.00	0.52	0.48	0.03
50,000	0.00	0.00	0.00	0.00	0.29	0.68	0.03
100,000	0.00	0.00	0.00	0.00	0.29	0.68	0.03

C. Origin of hspKet

	MODEL1	MODEL2	MODEL3	MODEL4	MODEL5
50,000	0.46	0.12	0.09	0.10	0.22
75,000	0.44	0.14	0.09	0.10	0.22
100,000	0.42	0.16	0.09	0.10	0.23

D. Colonisation history of hspIndigenousAmericas in Siberia and the Americas.

	MODEL1	MODEL2	MODEL3	MODEL4	MODEL5	MODEL6	MODEL7	MODEL8
50,000	0.22	0.00	0.00	0.03	0.03	0.00	0.63	0.09
75,000	0.22	0.00	0.00	0.03	0.02	0.00	0.66	0.07
100,000	0.26	0.00	0.00	0.04	0.02	0.00	0.61	0.07

Table S5. Posterior parameter estimates for Model 7 and Model 6 explaining the evolutionary history of hspSiberia1 and hspSiberia2, respectively. This table complements the *a posteriori* visual descriptions of the best selected model in Fig. 3A/B. The S1 column shows parameters estimated using Model 7 for hspSiberia1. The S2 column shows parameters estimated using Model 6 for hspSiberia2. T, time of population split in generations or years; /, denotes a population split; +, denotes nested populations; *, denotes admixture between two populations; NEa, ancestral effective population size; NEc, current effective population size.

PARAMETER	MEDIAN		MODE		95% HPD-LOWERBOUND		95% HPD-UPPERBOUND		R.SQUARED	
	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
MUTATION RATE	4.12E-07	4.15E-07	4.00E-07	3.78E-07	2.05E-07	1.97E-07	6.74E-07	6.90E-07	0.902	0.901
RECOMBINATION RATE	3.44E-09	5.26E-09	6.83E-10	1.00E-10	1.00E-10	1.00E-10	9.02E-09	8.86E-09	0.007	0.007
T2 HPASIA2/HPNORTHASIA +HPEASTASIA)	40,270	40,327	32,889	32,115	20,000	20,000	82,250	82,484	0.471	0.468
T1 (HPNORTHASIA/HPEASTASIA)	27,054	21,104	17,633	15,941	10,000	10,000	62,385	53,393	0.489	0.487
TADM (HPASIA2*HPNORTHASIA)	2,630	2,933	929	959	100	100	13,418	18,054	0.505	0.520
NEA-HPEASTASIA	18,061	18,143	10,000	10,333	10,000	10,000	54,428	52,781	0.081	0.086
NEA-HPNORTHASIA	27,918	32,619	10,000	10,000	10,000	10,000	85,968	87,561	0.074	0.088
NEA-HPASIA2	37,134	38,752	10,000	10,000	10,000	10,000	91,305	90,972	0.049	0.049
NEa-hspSiberia1 (Model 7)	35,764	-	14,905	-	10,000	-	87,786	-	0.098	-
NEa-hspSiberia2 (Model 6)	-	42,121	-	20,153	-	10,000	-	91,953	-	0.088
NEC-HPEASTASIA	403,688	505,775	326,192	425,023	121,422	195,500	824,212	914,761	0.458	0.492
NEC-HPNORTHASIA	151,967	150,476	119,172	117,732	100,000	100,000	317,822	324,842	0.368	0.355
NEC-HPASIA2	753,215	710,769	794,960	768,137	416,832	364,896	1,000,000	1,000,000	0.438	0.445
NEc-hspSiberia1 (Model 7)	407,522	-	225,743	-	100,000	-	906,391	-	0.154	--
NEc-hspSiberia2 (Model 6)	-	527,863	-	332,223	-	100,000	-	944,734	-	0.116

Table S6. Details of prior distributions of all model parameters to infer the origin of hspKet (five models). These complement the *a priori* visual descriptions of each model in Fig. S9. NEa, ancestral effective population size; NEc, current effective population size; Tadm, time of admixture; *, denotes admixture between two populations.

MODEL1

PARAMETER	DISTRIBUTION	MINIMUM	MAXIMUM
NEa-hspKet	uniform	1 000	100 000
NEc-hspKet	uniform	1 000	100 000
Tadm(hpNorthAsia*hspSiberia2)	uniform	100	2 850
Mutation rate	uniform	1.00E-08	1.00E-05
Recombination rate	uniform	1.00E-10	1.00E-07
P	uniform	0.000001	0.99999

MODEL2

PARAMETER	DISTRIBUTION	MINIMUM	MAXIMUM
NEa-hspKet	uniform	1 000	100 000
NEc-hspKet	uniform	1 000	100 000
Tadm(hpNorthAsia*hspSiberia1)	uniform	100	2 850
Mutation rate	uniform	1.00E-08	1.00E-05
Recombination rate	uniform	1.00E-10	1.00E-07
P	uniform	0.000001	0.99999

MODEL3

PARAMETER	DISTRIBUTION	MINIMUM	MAXIMUM
NEa-hspKet	uniform	1 000	100 000
NEc-hspKet	uniform	1 000	100 000
Tadm(hspSiberia1*hpEastAsia)	uniform	100	2 850
Mutation rate	uniform	1.00E-08	1.00E-05
Recombination rate	uniform	1.00E-10	1.00E-07
P	uniform	0.000001	0.99999

MODEL4

PARAMETER	DISTRIBUTION	MINIMUM	MAXIMUM
NEa-hspKet	uniform	1 000	100 000
NEc-hspKet	uniform	1 000	100 000
Tadm(hspSiberia2*hpEastAsia)	uniform	100	2 850
Mutation rate	uniform	1.00E-08	1.00E-05
Recombination rate	uniform	1.00E-10	1.00E-07
P	uniform	0.000001	0.99999

MODEL5			
PARAMETER	DISTRIBUTION	MINIMUM	MAXIMUM
NEa-hspKet	uniform	1 000	100 000
NEc-hspKet	uniform	1 000	100 000
Tadm(hspSiberia1*hspSiberia2)	uniform	100	2 850
Mutation rate	uniform	1.00E-08	1.00E-05
Recombination rate	uniform	1.00E-10	1.00E-07
P	uniform	0.000001	0.99999

Table S7. Posterior parameter estimates for the best model (Model 1) explaining the evolutionary history of hspKet. This table complements the *a posteriori* visual descriptions of the best selected model in Fig. 3C. Tadm, time of admixture in generations or years; *, denotes admixture between two populations; NEa, ancestral effective population size; NEc, current effective population size;

PARAMETER	MEDIAN	MODE	95% HPD-LOWERBOUND	95% HPD-UPPERBOUND	R.SQUARED
P	0.425	0.217	1.00E-06	0.903	0.474
Mutation rate	4.31E-07	4.28E-07	3.60E-07	5.04E-07	0.980
Recombination rate	3.59E-08	2.78E-08	5.22E-09	8.95E-08	0.139
Tadm (hpNorthAsia*hspSiberia2)	2,165	2,668	1,017	2,850	0.166
NEa-hspKET	1,682	1,000	1,000	16,792	0.177
NEc-hspKET	7,323	3,398	1,000	62,445	0.424

Table S8. Prior distributions of model parameters to infer the divergence of hspAltai from hspIndigenousAmericas. NEa, ancestral effective population size; NEc, current effective population size; T, time of population split in generations or years; /, denotes a population split; +, denotes nested populations. T2 is assumed to always be greater than T1.

PARAMETER	DISTRIBUTION	MINIMUM	MAXIMUM
NEa-hpEastAsia	uniform	10 000	1 000 000
NEa-hspIndigenousAmericas	uniform	10 000	1 000 000
NEa-hspAltai	uniform	10 000	1 000 000
NEc-hpEastAsia	uniform	10 000	1 000 000
NEc-hspIndigenousAmericas	uniform	10 000	1 000 000
NEc-hspAltai	uniform	10 000	1 000 000
T1(hspIndigenousAmericas/hspAltai)	uniform	100	30 000
T2(hpEastAsia/hspIndigenousAmericas+hspAltai)	uniform	20 000	100 000
Mutation rate	uniform	1.00E-08	1.00E-05
Recombination rate	uniform	1.00E-10	1.00E-07

Table S9. Posterior parameter estimates for the model of evolutionary divergence of hspAltai from hspIndigenousAmericas strains. T, time of population split in generations or years; /, denotes a population split; +, denotes nested populations; NEa, ancestral effective population size; NEc, current effective population size.

PARAMETER	MEDIAN	MODE	95% HPD-LOWERBOUND	95% HPD-UPPERBOUND	R.SQUARED
Mutation rate	6.12E-07	6.10E-07	4.41E-07	7.80E-07	0.769
Recombination rate	4.65E-09	6.85E-09	1.00E-09	8.30E-09	0.001
T2 (hpEastAsia/ hspIndigenousAmericas+hspAltai	20,823	20,000	20,000	25,612	0.042
T1 (hspIndigenousAmericas/hspAltai)	983	752	199	2,328	0.695
NEa-hpEastAsia	63,923	84,842	12,069	100,000	0.417
NEa-hspAltai	2,606	1,663	1,000	8,040	0.407
NEa-hspIndigenousAmericas	9,472	3,644	1,000	36,564	0.403
NEc-hpEastAsia	97,642	97,762	94,139	100,000	0.843
NEc-hspAltai	80,900	83,683	58,673	100,000	0.768
NEc-hspIndigenousAmericas	82,938	84,911	64,465	100,000	0.827

Table S10. Prior distributions of model parameters to infer migration events into the Americas (hspIndigenousAmericas, eight models). These complement the *a priori* visual descriptions of each model in Fig. S15. For this analysis the aboriginal distribution of hspIndigenousAmericas in Eurasia and the Americas was divided into four geographic populations: NS, Northern Siberia; ES, Eastern Siberia; KC, Kamchatka-Chukotka; AM, America. Twenty hpEastAsia strains from Hong Kong were used as the outgroup population. NEa, ancestral effective population size; NEc, current effective population size; T, time of population split; T_Migration_STOP, time at which migration stops; T_Migration_START, time at which migration starts; STOP_BT_AM, time at which the American bottleneck stopped; /, denotes a population split; +, denotes nested populations; ->, denotes direction of migration. Rules for the timing of evolutionary events (T) were as follows: T_Migration_START/T_Migration_STOP was always less than T1, which was less than T2, which was less than T3, which was less than T4.

MODEL1

PARAMETER	DISTRIBUTION	MINIMUM	MAXIMUM
NEa-hpEastAsia	uniform	1 000	100 000
NEa-NS	uniform	1 000	100 000
NEa-ES	uniform	1 000	100 000
NEa-KC	uniform	1 000	100 000
NEa-AM	uniform	1 000	100 000
NEc-hpEastAsia	uniform	1 000	100 000
NEc-NS	uniform	1 000	100 000
NEc-ES	uniform	1 000	100 000
NEc-KC	uniform	1 000	100 000
NEc-AM	uniform	1 000	100 000
T1(AM/KC)	uniform	100	30 000
T2(ES/AM+KC)	uniform	100	30 000
T3(NS/AM+KC+ES)	uniform	100	30 000
T4(EA/NS+AM+KC+ES)	uniform	20 000	30 000
Mutation rate	uniform	1.00E-08	1.00E-05
Recombination rate	uniform	1.00E-10	1.00E-07

MODEL2

PARAMETER	DISTRIBUTION	MINIMUM	MAXIMUM
NEa-hpEastAsia	uniform	1 000	100 000
NEa-NS	uniform	1 000	100 000
NEa-ES	uniform	1 000	100 000
NEa-KC	uniform	1 000	100 000
NEa-AM	uniform	1 000	100 000
NEc-hpEastAsia	uniform	1 000	100 000
NEc-NS	uniform	1 000	100 000
NEc-ES	uniform	1 000	100 000

NEc-KC	uniform	1 000	100 000
NEc-AM	uniform	1 000	100 000
T1(NS/ES)	uniform	100	30 000
T2(KC/NS+ES)	uniform	100	30 000
T3(AM/KC+NS+ES)	uniform	100	30 000
T4(EA/AM+KC+NS+ES)	uniform	20 000	30 000
Mutation rate	uniform	1.00E-08	1.00E-05
Recombination rate	uniform	1.00E-10	1.00E-07

MODEL3

PARAMETER	DISTRIBUTION	MINIMUM	MAXIMUM
NEa-hpEastAsia	uniform	1 000	100 000
NEa-NS	uniform	1 000	100 000
NEa-ES	uniform	1 000	100 000
NEa-KC	uniform	1 000	100 000
NEa-AM	uniform	1 000	100 000
NEc-hpEastAsia	uniform	1 000	100 000
NEc-NS	uniform	1 000	100 000
NEc-ES	uniform	1 000	100 000
NEc-KC	uniform	1 000	100 000
NEc-AM	uniform	1 000	100 000
T_Migration_STOP	uniform	11 500	30 000
T1(AM/KC)	uniform	11 500	30 000
T2(ES/AM+KC)	uniform	11 500	30 000
T3(NS/AM+KC+ES)	uniform	11 500	30 000
T4(EA/NS+AM+KC+ES)	uniform	20 000	30 000
Migration KC->AM	uniform	1.00E-05	1.00E-02
Migration AM->KC	uniform	1.00E-05	1.00E-02
Mutation rate	uniform	1.00E-08	1.00E-05
Recombination rate	uniform	1.00E-10	1.00E-07

MODEL4

PARAMETER	DISTRIBUTION	MINIMUM	MAXIMUM
NEa-hpEastAsia	uniform	1 000	100 000
NEa-NS	uniform	1 000	100 000
NEa-ES	uniform	1 000	100 000
NEa-KC	uniform	1 000	100 000
NEa-AM	uniform	1 000	100 000

NEc-hpEastAsia	uniform	1 000	100 000
NEc-NS	uniform	1 000	100 000
NEc-ES	uniform	1 000	100 000
NEc-KC	uniform	1 000	100 000
NEc-AM	uniform	1 000	100 000
T_Migration_START	uniform	100	12 000
T1(AM/KC)	uniform	11 500	30 000
T2(ES/AM+KC)	uniform	11 500	30 000
T3(NS/AM+KC+ES)	uniform	11 500	30 000
T4(EA/NS+AM+KC+ES)	uniform	20 000	30 000
Migration KC->AM	uniform	1.00E-05	1.00E-02
Migration AM->KC	uniform	1.00E-05	1.00E-02
Mutation rate	uniform	1.00E-08	1.00E-05
Recombination rate	uniform	1.00E-10	1.00E-07

MODEL5

PARAMETER	DISTRIBUTION	MINIMUM	MAXIMUM
NEa-hpEastAsia	uniform	1 000	100 000
NEa-NS	uniform	1 000	100 000
NEa-ES	uniform	1 000	100 000
NEa-KC	uniform	1 000	100 000
NEa-AM	uniform	1 000	100 000
NEc-hpEastAsia	uniform	1 000	100 000
NEc-NS	uniform	1 000	100 000
NEc-ES	uniform	1 000	100 000
NEc-KC	uniform	1 000	100 000
NEc-AM	uniform	1 000	100 000
T1(AM/KC)	uniform	100	30 000
T2(ES/AM+KC)	uniform	100	30 000
T3(NS/AM+KC+ES)	uniform	100	30 000
T4(EA/NS+AM+KC+ES)	uniform	20 000	30 000
Migration NS->AM	uniform	1.00E-05	1.00E-02
Mutation rate	uniform	1.00E-08	1.00E-05
Recombination rate	uniform	1.00E-10	1.00E-07

MODEL6

PARAMETER	DISTRIBUTION	MINIMUM	MAXIMUM
NEa-hpEastAsia	uniform	1 000	100 000
NEa-NS	uniform	1 000	100 000

NEa-ES	uniform	1 000	100 000
NEa-KC	uniform	1 000	100 000
NEa-AM	uniform	1 000	100 000
NEc-hpEastAsia	uniform	1 000	100 000
NEc-NS	uniform	1 000	100 000
NEc-ES	uniform	1 000	100 000
NEc-KC	uniform	1 000	100 000
NEc-AM	uniform	1 000	100 000
T1(ES/KC)	uniform	100	30 000
T2(STOP_BT_AM)	uniform	100	30 000
T3(AM/KC+ES)	uniform	100	30 000
T4(NS/AM+KC+ES)	uniform	100	30 000
T5(EA/NS+AM+KC+ES)	uniform	20 000	30 000
Migration NS->ES	uniform	1.00E-05	1.00E-02
Mutation rate	uniform	1.00E-08	1.00E-05
Recombination rate	uniform	1.00E-10	1.00E-07

MODEL7

PARAMETER	DISTRIBUTION	MINIMUM	MAXIMUM
NEa-hpEastAsia	uniform	1 000	100 000
NEa-NS	uniform	1 000	100 000
NEa-ES	uniform	1 000	100 000
NEa-KC	uniform	1 000	100 000
NEa-AM	uniform	1 000	100 000
NEc-hpEastAsia	uniform	1 000	100 000
NEc-NS	uniform	1 000	100 000
NEc-ES	uniform	1 000	100 000
NEc-KC	uniform	1 000	100 000
NEc-AM	uniform	1 000	100 000
T1(ES/KC)	uniform	100	30 000
T2(STOP_BT_AM)	uniform	100	30 000
T3(AM/KC+ES)	uniform	100	30 000
T4(NS/AM+KC+ES)	uniform	100	30 000
T5(EA/NS+AM+KC+ES)	uniform	20 000	30 000
Migration NS->ES	uniform	1.00E-05	1.00E-02
Migration KC->AM	uniform	1.00E-05	1.00E-02
Migration AM->KC	uniform	1.00E-05	1.00E-02
Mutation rate	uniform	1.00E-08	1.00E-05
Recombination rate	uniform	1.00E-10	1.00E-07

MODEL8

PARAMETER	DISTRIBUTION	MINIMUM	MAXIMUM
NEa-hpEastAsia	uniform	1 000	100 000
NEa-NS	uniform	1 000	100 000
NEa-ES	uniform	1 000	100 000
NEa-KC	uniform	1 000	100 000
NEa-AM	uniform	1 000	100 000
NEc-hpEastAsia	uniform	1 000	100 000
NEc-NS	uniform	1 000	100 000
NEc-ES	uniform	1 000	100 000
NEc-KC	uniform	1 000	100 000
NEc-AM	uniform	1 000	100 000
T1(ES/KC)	uniform	100	30 000
T2(STOP_BT_AM)	uniform	100	30 000
T3(AM/KC+ES)	uniform	100	30 000
T4(NS/AM+KC+ES)	uniform	100	30 000
T5(EA/NS+AM+KC+ES)	uniform	20 000	30 000
Migration NS->ES	uniform	1.00E-05	1.00E-02
Migration KC->AM	uniform	1.00E-05	1.00E-02
Migration AM->KC	uniform	1.00E-05	1.00E-02
Migration ES->KC	uniform	1.00E-05	1.00E-02
Migration KC->ES	uniform	1.00E-05	1.00E-02
Mutation rate	uniform	1.00E-08	1.00E-05
Recombination rate	uniform	1.00E-10	1.00E-07

Table S11. Posterior parameter estimates for the best model (Model 7) explaining the colonisation history of hspIndigenousAmericas across Eurasia and into the Americas. This table complements the *a posteriori* visual description of the best selected model in Fig. 4C. For this analysis the aboriginal distribution of hspIndigenousAmericas in Eurasia and the Americas was divided into four geographic populations: NS, Northern Siberia; ES, Eastern Siberia; KC, Kamchatka-Chukotka; AM, America. Twenty hpEastAsia strains from Hong Kong were used as the outgroup population. NEa, ancestral effective population size; NEc, current effective population size; T, time in generations or years; STOP_BT_AM, time at which the American bottleneck stopped; /, denotes a population split; +, denotes nested populations; ->, denotes direction of migration.

PARAMETER	MEDIAN	MODE	95% HPD-LOWERBOUND	95% HPD-UPPERBOUND	R.SQUARED
Mutation rate	5.66E-07	5.70E-07	7.29E-08	1.04E-06	0.821
Recombination rate	1.00E-10	1.00E-10	1.00E-10	7.81E-08	0.001
T5 (EA/NS+AM+KC+ES)	23,451	21,200	20,000	28,523	0.015
T4 (NS/AM+KC+ES)	15,170	14,612	5,596	26,388	0.033
T3 (AM/KC+ES)	12,032	10,910	3,105	22,512	0.038
T2 (STOP_BT_AM)	7,120	5,064	680	16,849	0.071
T1 (ES/KC)	1,523	572	100	6,048	0.205
M1 Migration NS->ES	0.001	0.001	0.000	0.005	0.354
M2 Migration KC->AM	0.000	0.000	0.000	0.005	0.280
M3 Migration AM->KC	0.002	0.001	0.000	0.008	0.242
NEa-hpEastAsia	29,276	1,109	1,000	75,743	0.325
NEa-NS	37,175	13,188	1,000	77,921	0.347
NEa-ES	3,485	1,000	1,000	18,762	0.171
NEa-KC	8,537	1,000	1,000	40,832	0.152
NEa-AM	167	10	10	854	0.005
NEc-hpEastAsia	76,371	76,119	46,455	100,000	0.589
NEc-NS	81,071	82,911	57,495	100,000	0.645
NEc-ES	15,277	7,653	1,000	69,614	0.327
NEc-KC	60,449	61,960	28,505	98,861	0.461
NEc-AM	54,255	85,574	7,604	100,000	0.387

Table S12. Individual DAPC assignments for the head to head comparison of whole genome and MLST structure. Both data sets were optimally portioned into three populations, with minor differences in individual assignment.

	Genomes	**MLST**
Cluster 1 hpNorthAsia hpEastAsia	5 hspAltai 25 hspIndigenousAmericas 4 hspEastAsia 3 hspIndia 2 hspMaori 12 hspSiberia1 7 hspSiberia2 2 hspKet	5 hspAltai 27 hspIndigenousAmericas 4 hspEastAsia 2 hspMaori
Cluster 2 hpAsia2	4 hspIndia 2 hspLadak 1 hspMaori 2 hspSiberia1 1 hspSiberia2 2 hspUral	7 hspIndia 2 hspLadak 2 hspKet 1 hspMaori 14 hspSiberia1 8 hspSiberia2
Cluster 3 hspUral	6 hspUral 1 hspIndigenousAmericas	8 hspUral

Supplementary Information References

1 Momynaliev, K. *et al.* Population identification of Helicobacter pilory isolates from Russia. *Genetika* **41**, 1434 (2005).

2 Momynaliev, K., Smirnova, O., Kudryavtseva, L. & Govorun, V. Comparative genome analysis of Helicobacter pylori strains. *Molecular biology* **37**, 529-536 (2003).

3 Achtman, M. *et al.* Recombination and clonal groupings within *Helicobacter pylori* from different geographical regions. *Mol. Microbiol.* **32**, 459-470, doi:10.1046/j.1365-2958.1999.01382.x (1999).

4 Linz, B. *et al.* An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* **445**, 915-918, doi:10.1038/nature05562 (2007).

5 Zhou, Z. *et al.* The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. *Genome Res* **30**, 138-152, doi:10.1101/gr.251678.119 (2020).

6 Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* **11**, 94, doi:10.1186/1471-2156-11-94 (2010).

7 Jombart, T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403-1405 (2008).

8 Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155**, 945-959 (2000).

9 Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567-1587 (2003).

10 Suerbaum, S. *et al.* Free recombination within *Heliobacter pylori*. *Proceedings of the National Academy of Sciences* **95**, 12619-12624, doi:10.1073/pnas.95.21.12619 (1998).

11 Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15-e15 (2015).

12 Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution* **37**, 1530-1534 (2020).

13 Kimura, M. Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences USA* **78**, 454-458 (1981).

14 Yang, Z. A space-time process model for the evolution of DNA sequences. *Genetics* **139**, 993-1005 (1995).

15 Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet* **8**, e1002453 (2012).

16 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069 (2014).

17 Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691-3693 (2015).

18 Katoh, K., Kuma, K.-i., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511-518 (2005).

19 Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics* **84**, 210-223 (2009).

20 Yahara, K. *et al.* Chromosome painting in silico in a bacterial species reveals fine population structure. *Molecular biology and evolution* **30**, 1454-1464 (2013).

21 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2015).
22 Beaumont, M. A., Zhang, W. Y. & Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025-2035 (2002).
23 Moodley, Y. *et al.* Age of the Association between *Helicobacter pylori* and Man. *PLoS Pathog.* **8**, e1002693, doi:10.1371/journal.ppat.1002693 (2012).
24 Moodley, Y. *et al.* The peopling of the Pacific from a bacterial perspective. *Science* **323**, 527-530, doi:10.1126/science.1166083 (2009).
25 Montano, V. *et al.* Worldwide Population Structure, Long-Term Demography, and Local Adaptation of Helicobacter pylori. *Genetics* **200**, 947-963, doi:10.1534/genetics.115.176404 (2015).
26 Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. & Foll, M. Robust demographic inference from genomic and SNP data. *PLoS Genet.* **9** (2013).
27 Wegmann, D., Leuenberger, C., Neuenschwander, S. & Excoffier, L. ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinform.* **11**, 116, doi:10.1186/1471-2105-11-116 (2010).
28 Excoffier, L. & Lischer, H. E. L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Res.* **10**, 564-567, doi:10.1111/j.1755-0998.2010.02847.x (2010).
29 Beaumont, M. A. in *Simulation, Genetics, and Human Prehistory* (ed S. Matsumura, Forster, P. and Renfrew, C) pp. 135-154 (McDonald Institute Monographs, 2008).
30 Hamilton, G., Stoneking, M. & Excoffier, L. Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. *Proceedings of the National Academy of Sciences* **102**, 7476-7480 (2005).
31 Neuenschwander, S. *et al.* Colonization history of the Swiss Rhine basin by the bullhead (Cottus gobio): inference under a Bayesian spatially explicit framework. *Molecular Ecology* **17**, 757-772 (2008).