Supplementary Methods:

Properties of strains chosen for genome sequencing. We compared the genomes of 149 isolates of Paratyphi A from diverse sources and dates (Dataset S1); including seven from Genbank, and 142 which were sequenced for this study (Dataset S9). According to the public MLST website (http://mlst.warwick.ac.uk) (3), all 54 Paratyphi isolates are assigned to one of eight STs within a tight cluster of closely related STs that is designated eBG11. eBG11 also includes six Sendai isolates and one isolate with no designated serovar, for a total of 61 isolates. STs 85 (42 isolates) and 129 (13 isolates) contain multiple isolates whereas the other STs only contain one isolate each. The sole lineage G genome belonged to ST479, all 45 lineage A genomes were in ST129, and all other genomes were in ST85.

Dataset S9 also lists the susceptibility or resistance to antimicrobials for all but one genome. These were based on minimum inhibitory concentrations (MIC). MICs for nalidixic acid, trimethoprim, chloramphenicol and tetracycline were determined by E-test strips (AB BIODISK, Solna, Sweden), using bacterial growth on Mueller-Hinton agar, and the results were interpreted according to the guidelines of the European Committee on Antimicrobial Susceptibility Testing (EUCAST, Version 3.1; http://www.eucast.org). *Escherichia coli* ATCC 25922 was used as an antibiotic-sensitive control strain. For published genomes, we used the metadata associated with the genomic sequence.

DNA preparation, sequencing, *de novo* Assembly and SNP calls. DNA was prepared from 5 ml overnight cultures using Jetflex Genomic DNA Purification Kits (Genomed, Germany) according to the manufacturer's instructions. Whole genome sequencing was performed using an Illumina HiSeq 2000 on 300 bp paired-end libraries in 96-fold multiplexes, using the parameters and with the results in Dataset S9.

For all 142 new sequences, we *de novo* assembled contigs from short reads with SOAPdenovo (4), and intra-scaffold gaps were filled using GapCloser v1.10 (SOAP package) as described in Zhou *et al*. (5). BOWTIE 2 (6) was then used to remap the reads to the assembled scaffolds. SAMtools and BCFtools (7) were used to validate the quality of each called base in the assemblies, filtering sites with quality <20, read coverage < 5, or >20% inconsistencies of base calls in overlapping reads with the base called in the assembly process. We applied the same pipeline to five public genomes (8), for which short reads were available.

Core genome assignment. The assembled genomes were aligned to a reference genome for Paratyphi A (strain ATCC 9150), using the MUMMER NUCMER module (9) with the '--mum' parameter. Regions that were absent in any of the 149 genomes were removed. In cases where multiple regions aligned to a common position within the reference, we retained the aligned region with the highest similarity. This filtering process was carried out with the delta-filter utility in the MUMMER (9) package, followed by a dedicated PERL script, MUMer_filter (https://sourceforge.net/projects/paratyphia/files/MUMmer_filter/). Three criteria were used to identify other repeat regions, which were subsequently also removed from the core genome. These included: 1) multiple segments of >50 bp dispersed throughout ATCC 9150 with a BLASTn hit of >94% identity when ATCC 9150 was blasted against itself; 2) VNTRs identified by Tandem Repeat Finder 4.04; 3) CRISPR1 and CRISPR2 regions. After removing repeat regions, the core genome comprised 4.07 Mb (Dataset S10), with 4,584 SNPs among the 149 genomes (Dataset S11).

Small insertions and deletions of up to 29 bp (indels) were identified in the core genome by using BOWTIE 2 with default parameters to map reads to the reference genome. The Samtools-formatted output was analyzed with DINDEL (10), which is one of the two most accurate down-stream analyzers of such data (11), and the best analyzer for assemblies with low coverage.

Genealogy and mapping of genetic variations. A maximum parsimony genealogy was calculated with MEGA 5 (12) on the basis of all 4,584 non-repetitive, core SNPs. The root of the genealogy was determined by identifying the consensus ancestral nucleotide variant in 12 outgroup genomes from serovars Choleraesuis (AE017220), Dublin (CP001144), Enteriditis (AM933172), Gallinarum (AM933173), Heidelberg (CP001120), Newport (CP001113), Agona (CP001138), Paratyphi B (CP000886), Paratyphi C (CP000857), Schwarzengrund (CP001127), Typhi (AE014613) and Typhimurium (AE006468) (Dataset S11). For nucleotides which varied between the twelve outgroup genomes, the ancestral nucleotide was only scored when that nucleotide variant was found at least twice and was present in at least 2/3 of outgroup genomes that possessed an orthologous region. Of the 4,508 core SNPs that met these criteria, 4,482 (99.4%) support the root position in the maximum parsimony tree that is presented here, which is between lineage G and the other lineages.

The genealogies of mutations in the core genome and genomic islands were estimated by mapping the changes of characters onto individual branches of the maximum parsimony genealogy. These analyses were performed using the maximum likelihood approach (13) that is implemented in the function "ACE" (Ancestral Character Estimation) in R package "APE" (14). All mutations and genomic islands that were identified as occurring on at least two independent branches by ACE were scored as homoplastic.

Identification of recombinant segments (RecHMM). CLONALFRAME (15) has been widely used in comparative genomics to identify recombinant segments among limited numbers of genomes (16), but according to our experience it is incapable of performing genomic comparisons with >40 genomes in less than several weeks of computing time, and did not converge reproducibly with large data sets. An alternative iterative method to identify significant clusters of SNPs has been described (17), but was not publicly available at the time these analyses were performed. A third algorithm for these purposes, STARRINIGHTS (18), has also been described but we had difficulties in implementing it. We therefore developed a novel program in R based on a Hidden Markov Model, RecHMM (Recombination by Hidden Markov Model), which recognizes significant clusters of SNPs by a similar algorithm to CLONALFRAME, but can analyze at least 149 genomes overnight. RecHMM is faster than CLONALFRAME because it uses a fixed topology whereas CLONALFRAME calculates its own topology. For Paratyphi A, we used a maximum parsimony topology, which is justifiable because only 0.6% of SNPs in the core genome were homoplastic. The RecHMM R-script is available for public download at https://sourceforge.net/projects/paratyphia/files/RecHMM/, which also provides a full description of its algorithm.

Brief description of RecHMM. The Markov structure for each branch of a genealogy in RecHMM is depicted in Fig. S6a. Each site in the core genome is assigned to one of two hidden states, 'non-recombinant' and 'recombinant', with characteristic densities of nucleotide variants designated *m* and *v*, respectively. The individual nucleotides at each site in the core genome result from a sequence of emissions from these hidden states, which can result in variant SNPs and indels. The assignment of each site to a state is purely dependent on the properties of the

preceding site as modified by the recombination rate per site, r, and the inverse of the average recombination tract length, δ, except that the first site has an initial parameter, r'.

The algorithm assumes that recombinant nucleotides are imported from a single outgroup donor, which results in constant v and δ for all branches. Parameters m and r are proportional to the length of each branch, l, and vary between branches. ρ and θ are the average recombination and mutation rates over the entire genealogy. Then for each branch, $m=\theta l/2$ and $r=\rho l/2$.

We applied a modified, iterative E-M (expectation–maximization) algorithm to estimate the maximum likelihood of the four global parameters of this Hidden Markov Model for the Paratyphi A dataset as ρ=286.4, θ=5002.1, v=0.09, δ=0.15. In the E step, we used the forward-backward algorithm to calculate all possible samples from the Markov model in each branch, as is typical of E-M algorithms. In the M step, our algorithm summarizes the results for all branches, thus providing an estimate of the overall likelihood of the genealogy. That estimate was then used to maximize expectations for the whole genealogy by testing new parameter values. Nucleotides were initially assigned to the recombinant state if they fell between the closest pairs of SNPs/indels with likelihood values for the recombinant state that surpassed the values at each of 10 arbitrary cut-off values drawn from [0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9] of all SNPs/indels. We then iterated the E-M algorithm until successive likelihoods differed by <0.00001. Of the ten arbitrary cut-off values, the cut-off of 0.01 had the greatest likelihood, and was used to calculate the likelihood of recombination for each nucleotide. Based on the frequencies of false positive and false negative calls within the simulations described below, we then scored recombinant stretches as stretches of sites with a likelihood of recombination of ≥0.5 containing at least one site with likelihood ≥0.8.

Evaluation of RecHMM by simulations. In order to compare the accuracy of RecHMM relative to ClonalFrame, and to deduce optimal parameters for scoring the lengths of recombinant stretches, we used both methods to analyze the results from 350 coalescent simulations with SimMLST (19) of random genealogies of 100 sequences of 10 kb with an average recombination stretch of 236 bps (20). The simulations consisted of 10 independent runs each for 35 combinations of various rates of recombination (ρ) [50, 100, 200, 300, 400, 600, 800] and mutation (θ) [100, 200, 300, 400, 500]. From each simulation, we analyzed the shortest sub-tree of 10 sequences whose branch length was approximately 0.1 of the branch length of the global genealogy. This strategy was intended to mimic a closely related subset of sequences with the same ρ and θ as the entire dataset, but which recombined with other sequences which were 10 times as divergent.

Because RecHMM does not estimate phylogenies, the clonal phylogenies generated by SimMLST were provided as a fixed topology to both ClonalFrame and RecHMM. RecHMM was run ten times per simulation in order to determine which yielded the highest likelihood among the initial cut-offs described above. ClonalFrame was run twice per simulation, using the parameter "-x 200000 -y 300000 -z 20 -T". The pairs of ClonalFrame runs always converged, and were merged for final analyses. The two programs were then compared for their individual estimates of two global parameters used by SimMLST, namely ρ/θ, the relative rate of recombination and mutation, and δ, the inverse of the average length of recombinant stretches. Our results confirmed previous observation (20) that ClonalFrame slightly under-estimates ρ/θ whereas RecHMM yielded estimates that were quite similar to those used by SimMLST (Figs. S7a, b). Our results also show that δ tends to be over-estimated by up to 1.5fold by

CLONALFRAME and by up to twofold by RecHMM (Figs. S7c, d). Thus, recombinant stretches calculated by RecHMM might be expected to be up to twofold shorter than the true stretches.

We also attempted to calculate the frequencies for both methods of false negative and false positive assignments to recombination for individual SNPs. Unfortunately, SIMMLST only records the multiple local trees that resulted from recombination, and does not provide details of the individual recombination events. The calculation of a so-called ancestral recombination graph (ARG) which reconstructs the history of such events is NP-hard (21), preventing the calculation of unambiguous ARGs. Approximate solutions to identify the single, shortest recombination pathway have been proposed, including PRUNIER (22), which is more accurate than two alternatives, but are not always correct because local trees can result from multiple pathways of equal probability. We therefore modified the algorithm used by PRUNIER to compare up to 30 unique pathways of recombination for each local tree, thereby maximizing the global likelihood of correct reconstruction of recombination events at the genomic level. This somewhat *ad hoc* approach is not necessarily ideal for reconstructing recombination pathways, but is better than the available alternatives. The ExpandTrees R-script for this approach is available for public download at https://sourceforge.net/projects/paratyphia/files/ExpandTree/, which also provides a full description of its algorithm.

Our reconstructed recombinational pathways were then used to compare the frequencies of false negative and false positive calls of recombinant SNPS by CLONALFRAME and RecHMM (Fig. S7e) in stretches of SNPs with recombinant likelihoods of ≥0.5. Both methods yielded frequencies of <20% false negative or false positive errors regardless of the likelihood cut-off used to initiate the recognition of recombinant SNPs within recombinant tracts. RecHMM yielded an average of 10% false positive and false negative calls when recombinant stretches were initiated with SNPs with a recombination likelihood of 0.8. CLONALFRAME yielded comparable false positive rates but had a slightly higher false negative rate. The sum of the false positive plus false negative error rates for RecHMM depended to a certain extent on the global parameters used in the simulations (Fig. S7f), ranging from <10% in simulations with high θ and low ρ to 42% in simulations with low θ and high ρ.

Designations for nodes, branches and lineages. Each terminal node in the maximum parsimony genealogy contained only one isolate, and received the same designation as the isolate. Internal nodes were designated as N001- N149, increasing sequentially from the root to the tips (Fig. S1). Branches were designated by the nodes they connect, e.g. N001.N002 and N149.ATCC-9150 (Dataset S11). The same designations also apply to the maximum likelihood genealogy because it shares the same topology. Branch designations A-G were assigned on the basis of their distinct clustering within the maximum likelihood tree (Fig. S2).

Bayesian estimates of age, transmissions and population fluctuation. Calculation of the root to tip distances *versus* dates of isolation indicated a linear relationship ($R^2 = 0.4$; Fig. S8). We estimated the population history of Paratyphi A with the Bayesian algorithms in BEAST v1.8.0 (23). The input consisted of a continuous core genome alignment consisting of all 4,525 SNPs in the non-repetitive, non-recombinant core genome, supplemented by the numbers of invariant A, C, T and G nucleotides. The dates of isolation of each strain were included in the Bayesian model as tip dates, as were discrete traits consisting of 11 countries from which at least four strains had been isolated and five continents/super regions for the other isolates (Dataset S1 and S9). We included continents/super-regions for isolates from 28 countries with only 1-3 isolates because their countries of origin offered little information, and would have dramatically

increased the complexity of the Bayesian model. Initial comparisons showed that the root positions of all maximum clade credibility trees differed from the root indicated by outgroup analysis, which would reduce the dating accuracy of the MRCA. We therefore assigned all isolates from lineages A-F to a single monophyletic clade in order to ensure that the root was between lineage G and the other lineages. In the initial phase, five independent Markov Chain Monte Carlo (MCMC) analyses were run for each of sixteen different combinations of clock rate and population models (Dataset S2) for 20 million states, with sampling every 1,000 iterations. Two combinations (Log-normal clock rate with either the GMRF or the exponential growth population model) did not initiate, and were excluded from further analyses. The initial 10-30% of samples from the beginning of each run were treated as burn in because they had significantly lower likelihoods than subsequent samples. The remaining samples from each model were combined with LogCombiner v1.8.0. The Bayesian model with a relaxed exponential clock rate and a Bayesian Skyline population size yielded the highest Bayes factor of all combinations, and was used to compare four discrete phylogeographic models (24): 1) symmetric (equal two-way) transmissions, 2) symmetric transmissions with BSSVS (Inferring social networks with Bayesian Stochastic Search Variable Selection), 3) asymmetric (including unequal) transmissions, 4) asymmetric transmissions with BSSVS. We also calculated 5) asymmetric transmissions with no transmission from Western Europe and 6) asymmetric transmissions with BSSVS and no transmission from Western Europe. Five independent runs of 40 million iterations were performed for each discrete phylogeographic model, with sampling every 1,000 iterations. The highest Bayes factor was obtained for asymmetric transmissions with BSSVS. However, according to this model, France was the source for multiple historical transmissions. This observation probably reflects a sample bias because many of our isolates were from the Institut Pasteur collection which has a predominance of isolates from France. Furthermore, transmissions from France contradicts the epidemiological record according to which Paratyphi A in Western Europe have been rare since its first description in 1898 (25). We therefore used the maximum clade credibility tree from the model with asymmetric transmissions with BSSVS in which transmissions from Western Europe were prohibited even though the likelihood of these results were slightly less likely (decreased ln(Bayes factor) of 1.1fold) than for the best model. Dataset S3 presents the estimated dates of the root nodes (N001) and all other nodes from the best model.

The maximum clade credibility tree (Fig. 1) was summarizes from 176K trees remaining after the removal of burn-in with TreeAnnotator, and visualized with FigTree. The topology of branches whose posterior probability was > 50% was concordant with the topologies found by both maximum parsimony and maximum likelihood. Fig. 1 indicates transmissions for internal nodes where ≥80% of the sampled trees supported one continent/super region or country which differed from the closest ancestral super region or country with ≥80% support.

We also estimated the population history of Paratyphi A, including its age, the mutation rate and effective population size with BEAST 2 (26), which incorporates fewer population models and phylogeographic models than BEAST 1.8. All analyses were performed with a fixed topology based on the maximum likelihood genealogy (Fig. S2). The best model (highest Bayes factor) for this data was a Generalized Time-Reversible (GTR) model with four different categories of rate heterogeneities plus an estimated proportion of invariable sites, and this was used for all subsequent analyses. A comparison of nine combinations of different clock and population models (Dataset S2) showed that the model based on a constant mutation rate and a Bayesian-Skyline population size was much worse than all others (relative Bayes factor $<10^{-135}$), and it was discarded. (We also discarded a tenth model, the BEAST 2 Birth-Death Skyline model,

because its estimate of TMRCA decreased discontinuously in steps of 50 years during long runs, rather than converging on a reproducible estimate). The highest Bayes factor was obtained with a relaxed exponential clock rate and a Bayesian Skyline population model, and the Bayes factors of the seven other models were worse by ratios of between 3.1 and 5.6 x 10^{10} (Dataset S2). Dataset S2 also presents the mean clock rates for all eight models.

Temporal mapping of genetic variations. To estimate the dates of changes in nucleotides and genomic content (Datasets S11, 12), we mapped each such event to the branches of the 176K BEAST 1.8 trees sampled from the preferred model (asymmetric, BSSVS and no transmission from Western Europe), except within lineage G, which has no sub-branches. We assigned an equal probability for each year between our estimated dates of the nodes flanking branches in which mutations and changes had occurred. The years before isolation in Figs. 2, 3E, 3F and S4 are summaries of those estimates over all trees.

Temporal mapping of transmissions. To estimate the dates of transmissions (Fig. 1b, Dataset S3), we analyzed the 176K trees sampled from the best model. We assigned an equal probability for each year between our estimated dates of the nodes flanking branches in which transmissions had occurred, and summarized the probability estimates over all trees (Fig. S3).

Gene prediction and annotation within the pan-genome. The contents of the pan-genome were calculated as described (5). Briefly, CDSs in all assemblies were predicted *ab initio* in Glimmer 3.02 (27) and aligned with all 149 genomes to identify High Scoring Matches (HSMs). HSMs which covered ≥60% of the query sequences and did not map to the edges of contigs were retained as homologs of the query CDS, as were HSMs at the edges of contigs which covered ≥40% of the query sequences. Overlapping HSMs were then merged into groups of homologs. When an HSM in a homologous group was present in the reference genome (ATCC 9150), that CDS was chosen as a representative of the group, or the longest ATCC 9150 CDS when multiple CDS homologs existed. Otherwise, the longest homologous CDS in the CDS dataset was chosen. All other CDSs were removed as redundant, resulting in a pan-genome of CDSs. These CDSs were translated and used to search the non-redundant UniRef100 dataset (28) in UniProtKB and the eggNOG 3.0 database (29) using BLASTp. Putative gene annotations and functions were assigned to CDSs using the criteria of E value <1e-05 and identity ≥30%. Small CDSs predicted by Glimmer and other HMM based gene prediction methods have been reported to be unreliable (30). We therefore removed from the pan-genome all CDSs of <100 amino acids that lacked a clear match or a clear functional designation, leaving a total of 4,938 pan-genomic CDSs for the reconstruction of genomic islands.

Reconstruction and annotation of genomic Islands. Genomic islands (Dataset S12) were identified as described (5). Briefly, scaffolds in assemblies were connected according to their relationships with the reference genome ATCC 9150, and the remapping of paired-end reads. Inter-scaffold breakpoints induced by repetitive regions were fully resolved for repetitive regions which belonged to different genomic islands, but only partially resolved by walking through as many unique bases as possible for multiple repetitive regions within a genomic island. We accepted genomic annotations for the functions of genomic islands within ATCC 9150, and assigned functional designations to other genomic islands based on the best hits in BLASTn searches of the RefSeq non-redundant nucleotide database in GenBank.

Annotation of genes related to insertion sequences. All predicted CDSs in the pan genome were screened against the IS finder database (31) using both BLASTn and BLASTp.

CDSs were assigned to IS elements if they had a best hit with ≥80% similarity and ≥ 60% coverage with BLASTn, or, when those criteria were not matched, with ≥60% similarity and ≥60% coverage with BLASTp.

Regions of nucleotides under selection (DHMM). Regions of clustered SNPs/indels arise under diversifying selection. CDSs containing such regions can be identified by programs such as PAML (32) on the basis of their unusual *d*N/*d*S ratios. However, this approach is restricted to coding regions, which excludes sites within non-coding and intergenic regions even though they are known to contain regulatory elements, including sRNA, and their binding sites. We therefore developed an alternative method by implementing a second HMM-based model to identify clustered SNPs/indels throughout the core genome, thus examining both coding and non-coding regions. This model was implemented in DHMM (Density by Hidden Markov Model), which is an R script available online at https://sourceforge.net/projects/paratyphia/files/DHMM/, including details on the algorithms it uses.

Briefly, we treat the core genome as a serial sampling from *k* hidden states whose emissions are responsible for all changes at the nucleotide level. Genomic regions that are assigned to the same state would be expected to have undergone similar patterns of selection, while those sampled from different states should have undergone different evolutionary scenarios.

The Markov structure in DHMM for k=3 is illustrated in Fig. S6b. The observed numbers of independent SNPs/indels at each site based on a maximum parsimony tree are determined by two groups of four parameters for synonymous SNPs, non-synonymous SNPs, SNPs in non-coding regions and SNPs in indels. For each hidden state i, there are four emission parameters ($p_{i,S}$ $p_{i,NS}$ $p_{i,NC}$ and $p_{i,indel}$) and a set of four other parameters ($c_{i,S}$ $c_{i,NS}$ $c_{i,NC}$ and $c_{i,indel}$) to reflect the differing potential for nucleotide variants according to the location of the nucleotide site, which is affected by the boundaries of CDSs and the nucleotide position within codons. With these eight parameters, DHMM calculates the conditional probability of the distribution of nucleotide variation based on a Poisson distribution. The Hidden Markov structure also includes two other essential parameters, the matrix U, consisting of the probabilities of each of the k hidden states for the first nucleotide, and the transition matrix T, which contains the transition probabilities from each state to each of the hidden states for the next site.

These parameters suffice to calculate maximum likelihood estimates from HMMs by the E-M algorithm for any given value of k. We therefore used the Akaike Information Criterion (AIC) as well as the Bayesian Information Criterion (BIC) to compare the results with k values ranging from one to six for each data set. We analyzed 100 initial random parameter combinations for each data set by iterating the E-M algorithm until the difference of probabilities between successive iterations was <0.00001. For Paratyphi A, the strongest BIC support was for the model with k=3 while AIC most strongly supported k=4 (Dataset S13). After extracting the per-site composition with the forward-backward algorithm, both models yielded very similar assignments of individual sites to the different states, except that the four states preferred by AIC assigned homoplastic indels to a different state than homoplastic SNPs whereas most of these were merged in one of the three states preferred by BIC. >98.5% of the sites in the core genome were assigned to state 1, which we interpret as representing invariant sites as well as mutations that are evolving neutrally. In order to identify clusters of SNPs/indels in the other states, we mapped stretches containing sites with a probability of state 1 ≤0.25 flanked by nucleotides with

a probability of state 1 ≥0.5. The ends and locations of these regions were almost identical in both models, and we arbitrarily chose k=3 for further analyses because it has fewer states than k=4.

DHMM results with published data sets. According to a χ^2 based method (33), most SNPs that have arisen recently in the genetically monomorphic *S. enterica* serovar Agona (5) (TMRCA ~80 years) and *Yersinia pestis* (33) (TMRCA ~3,000 years) have not (yet) been subjected to Darwinian (diversifying) selection. In contrast, the *cag* genomic island in *Helicobacter pylori* (34) (TMRCA: >50k years) contains numerous SNPs that show traces of Darwinian selection according to PAML (35). We used these three data sets involving different species and different time scales as test objects for DHMM (Fig. S9). Our comparisons using the same approach as for Paratyphi A indicated that k=1 was most likely for the non-recombinant, core genome of serovar Agona (Fig. S9b), which confirms previous conclusions that its core genome did not contain a detectable proportion of SNPs and indels which had undergone Darwinian selection. The optimal DHMM model for *Y. pestis* was k=2 (Fig. S9a). This model assigned all 28 homoplastic sites (33) to state 2, and all other sites to state 1. Homoplastic sites can arise by Darwinian selection (36), recombination or mutational hotspots, but recombination and hotspots are very rare in *Y. pestis*, and this result indicates that DHMM can identify rare homoplasies that might be under Darwinian selection.

Mutations in *H. pylori* are highly homoplastic, which has been attributed to very frequent recombination (37) or mutational hotspots (38). The frequency of homoplasies is so high that it is very difficult to reconstruct genealogies without first removing a large proportion of the data (39). In order to compare DHMM with PAML, we therefore applied DHMM to *cag* sequences presented as a linear array of invariant and polymorphic sites. The best model from DHMM analysis on the *H. pylori cag* PAI included five different states according to BIC (Fig. S9c). States 1, 3 and 5 had dN/dS ratios between 0.03 and 0.29 (Dataset S14), which may represent different levels of purifying selection. These three states included 31% of all non-synonymous mutations, 84% of all synonymous mutations and 58% of mutations in non-coding regions (Dataset S14). States 2 and 4 had dN/dS ratios of 0.97 and 1.12, respectively, and contained higher numbers of non-synonymous mutations (69%) but lower numbers of synonymous mutations (16%) and mutations in non-coding regions (42%). The dN/dS ratios of mutations in these two states were >4 fold higher than the average dN/dS ratio of 0.24 for the whole *cag* island. Furthermore, the proportion of sites under selection in each gene correlated fairly strongly (R^2=0.55) with the published results from PAML on these genes (34) (Fig. S9d). Thus, DHMM and PAML yield similar results for genes with high diversity that seem to have undergone a certain degree of diversifying selection, except that DHMM also provides data for non-coding regions and can be used to detect homoplasies in genetically monomorphic bacteria.

We also tested the Paratyphi A genomes with the χ^2 based method (33). This test identified 31 CDSs which contained significant clusters of non-synonymous SNPs (Fig. S10), 17 of which overlapped with the 76 regions identified by DHMM which recognizes multiple types of mutations rather than focussing exclusively on non-synonymous mutations within CDSs.

References

1. McClelland M et al. (2004) Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of *Salmonella enterica* that cause typhoid. *Nature Genet* 36:1268-1274.

2. Holt KE et al. (2009) Pseudogene accumulation in the evolutionary histories of *Salmonella enterica* serovars Paratyphi A and Typhi. *BMC. Genomics* 10:36.

3. Achtman M et al. (2012) Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. *PLoS Pathog* 8:e1002776.

4. Li R et al. (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* 20:265-272.

5. Zhou Z et al. (2013) Neutral genomic microevolution of a recently emerged pathogen, *Salmonella enterica* serovar Agona. *PLoS Genet* 9:e1003471.

6. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357-359.

7. Li H et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 25:2078-2079.

8. Liang W et al. (2012) Pan-genomic analysis provides insights into the genomic variation and evolution of *Salmonella* Paratyphi A. *PLoS ONE* 7:e45346.

9. Kurtz S et al. (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5:R12.

10. Albers CA et al. (2011) Dindel: accurate indel calls from short-read data. *Genome Res* 21:961-973.

11. Neuman JA, Isakov O, Shomron N (2013) Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection. *Brief. Bioinform.* 14:46-55.

12. Tamura K et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28:2731-2739.

13. Pagel M (1994) Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London Series B-Biological Sciences* 255:37-45.

14. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics.* 20:289-290.

15. Didelot X, Falush D (2007) Inference of bacterial microevolution using multilocus sequence data. *Genetics* 175:1251-1266.

16. Dingle KE et al. (2014) Evolutionary history of the *Clostridium difficile* Pathogenicity Locus. *Genome Biol Evol* 6:36-52.

17. Croucher NJ et al. (2011) Rapid pneumococcal evolution in response to clinical interventions. *Science* 331:430-434.

18. Shapiro BJ et al. (2012) Population genomics of early events in the ecological differentiation of bacteria. *Science* 336:48-51.

19. Didelot X, Lawson DJ, Falush D (2009) SimMLST: simulation of multi-locus sequence typing data under a neutral model. *Bioinformatics.* 25:1442-1444.

20. Didelot X, Lawson DJ, Darling A, Falush D (2010) Inference of homologous recombination in bacteria using whole genome sequences. *Genetics* 186:1435-1449.

21. Wang L, Zhang K, Zhang L (2001) Perfect phylogenetic networks with recombination. *J Comput. Biol* 8:69-78.

22. Abby SS, Tannier E, Gouy M, Daubin V (2010) Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. *BMC Bioinformatics* 11:324.

23. Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 29:1969-1973.

24. Lemey P, Rambaut A, Drummond AJ, Suchard MA (2009) Bayesian phylogeography finds its roots. *PLoS Comput Biol* 5:e1000520.

25. Bainbridge FA (1912) The Milroy Lectures On Paratyphoid Fever and Meat Poisoning. *Lancet* 179:705-709.

26. Bouckaert R et al. (2014) BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput Biol* 10:e1003537.

27. Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics.* 23:673-679.

28. Suzek BE et al. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics.* 23:1282-1288.

29. Powell S et al. (2012) eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res* 40:D284-D289.

30. Hyatt D et al. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119.

31. Siguier P et al. (2006) ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* 34:D32-D36.

32. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586-1591.

33. Cui Y et al. (2013) Historical variations in mutation rate in an epidemic pathogen, *Yersinia pestis*. *Proc Natl Acad Sci USA* 110:577-582.

34. Olbermann P et al. (2010) A global overview of the genetic and functional diversity in the *Helicobacter pylori cag* pathogenicity island. *PLoS Genet* 6:e1001069.

35. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13:555-556.

36. Roumagnac P et al. (2006) Evolutionary history of *Salmonella* Typhi. *Science* 314:1301-1304.

37. Suerbaum S et al. (1998) Free recombination within *Helicobacter pylori*. *Proc Natl Acad Sci USA* 95:12619-12624.

38. Meinersmann RJ, Romero-Gallo J, Blaser MJ (2008) Rate heterogeneity in the evolution of *Helicobacter pylori* and the behavior of homoplastic sites. *Infect Genet Evol* 8:593-602.

39. Moodley Y et al. (2009) The peopling of the Pacific from a bacterial perspective. *Science* 323:527-530.

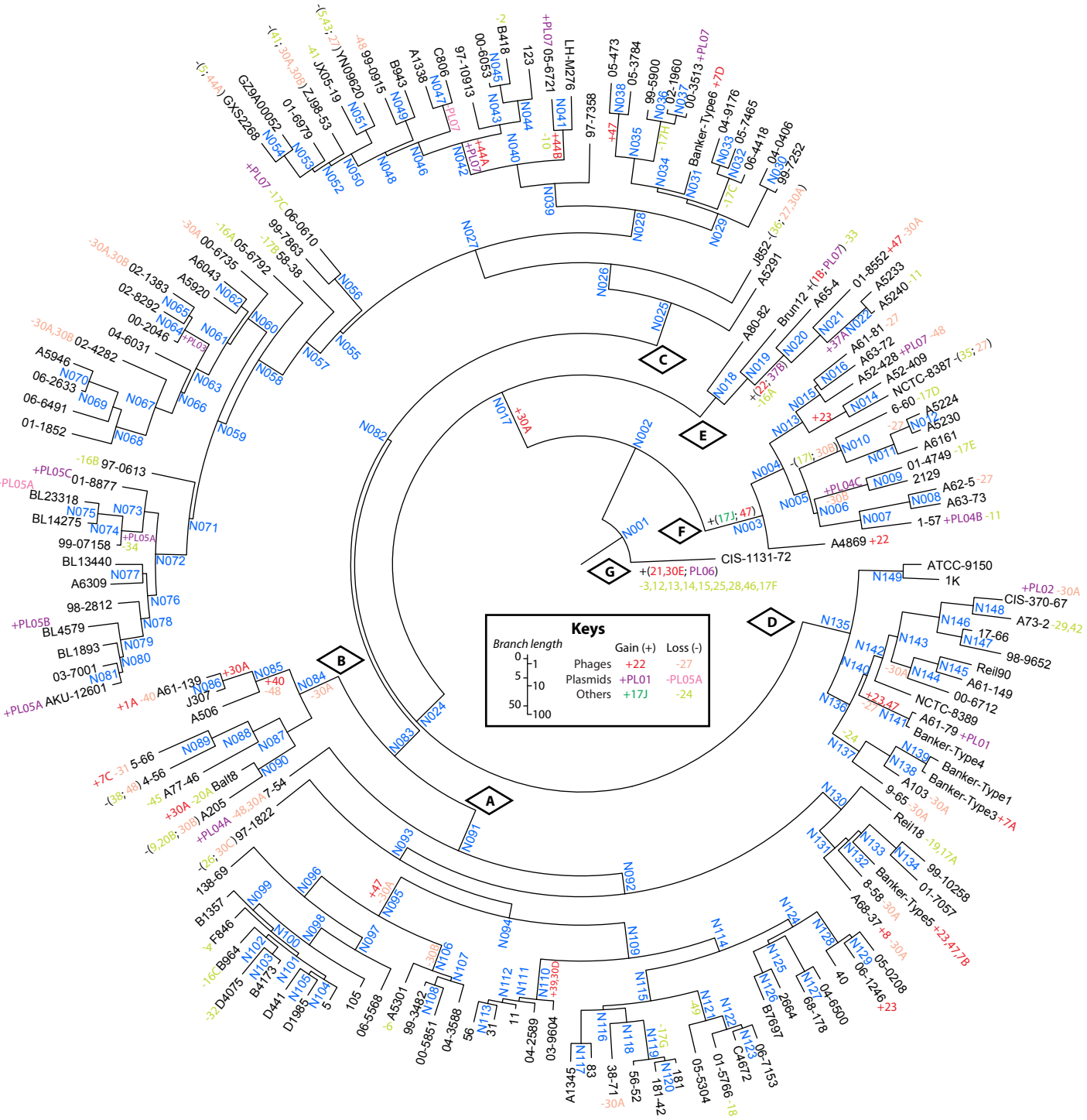

Fig. S1. Circular maximum parsimony tree based on 4,584 SNPs. The branch lengths were logarithm transformed for readier visual comparisons. The names of terminal nodes are black whereas internal nodes are in blue. Other colors indicate variations of genomic islands or plasmids in the accessory genome.

Fig. S2. Maximum likelihood genealogy derived from 4,525 SNPs in the non-recombinant, non-repetitive core genome. The genealogy was estimated using the GTR+4+I model in RAxML. Lineages are indicated by grey polygons, with filled circles for each genome, color coded according to geographic source. Note that the genetic distance between lineage G and other lineages is not drawn to scale.

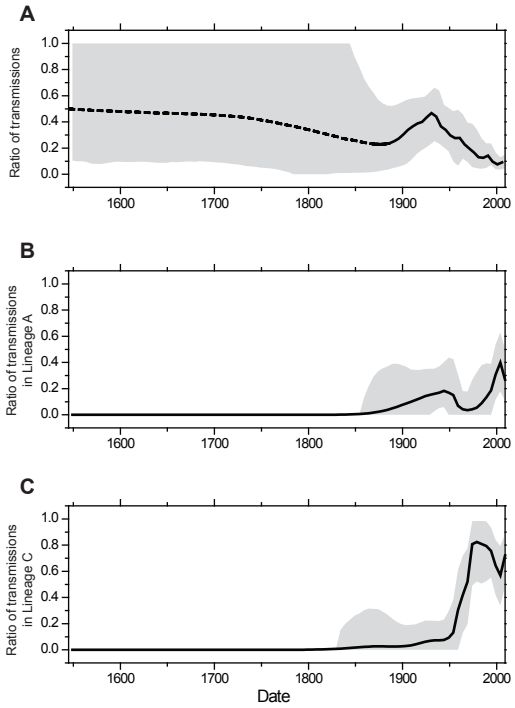

Fig. S3. Temporal mapping of transmissions. In all parts, solid black lines with 95% lineage confidence intervals (grey) indicate average ratios over all trees. (A) Frequencies of transmissions relative to number of branches. The dashed line before late 1900s indicates suspicious average value with wide CI95%; (B-C) Frequencies of transmissions lineage lineages A and C relative to all lineages.

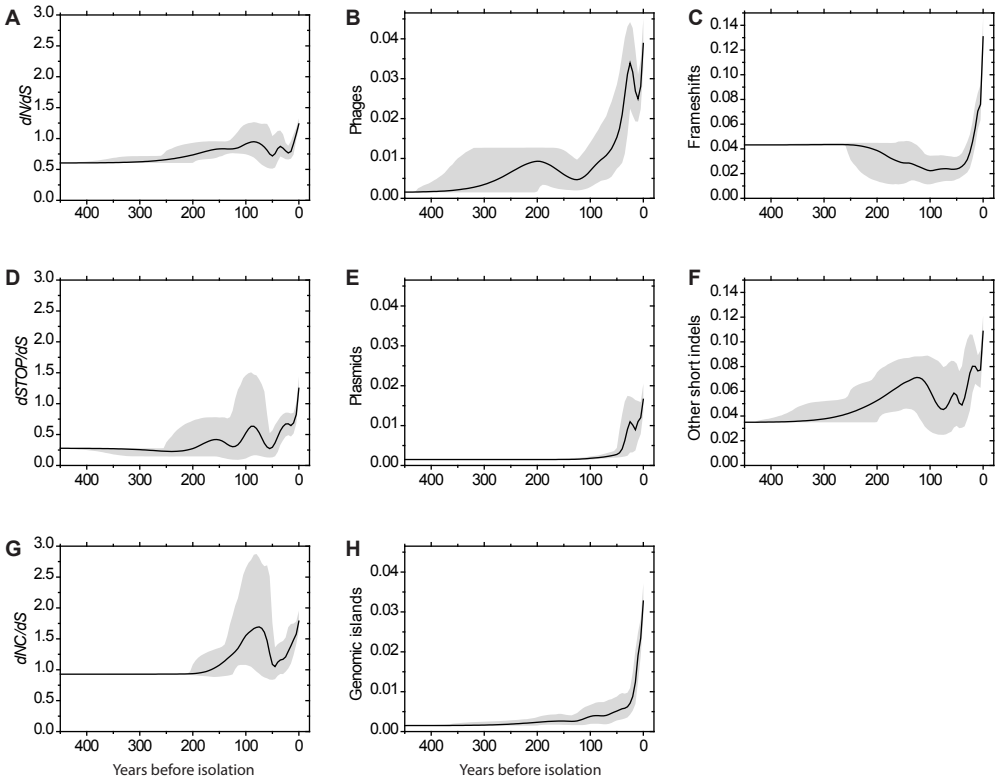

Fig. S4. Temporal mapping of SNPs (A, D and G), large indels (B, E and H) and short indels (C and F) with 95% of confidence intervals in grey. Black curves are average ratios that are also shown in Fig. 2 A-C.

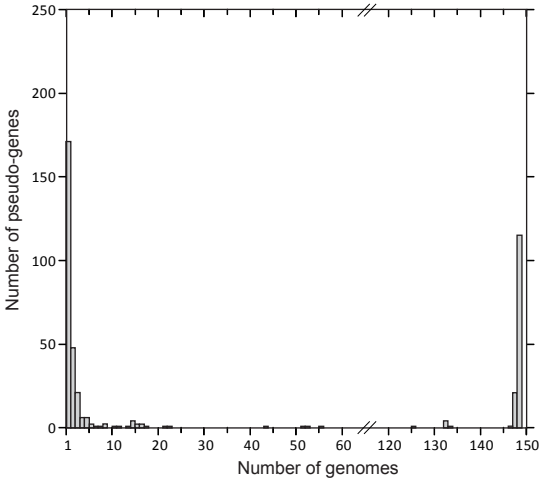

Fig. S5. Histogram of number of pseudo-genes according to the number of genomes in which they were present.

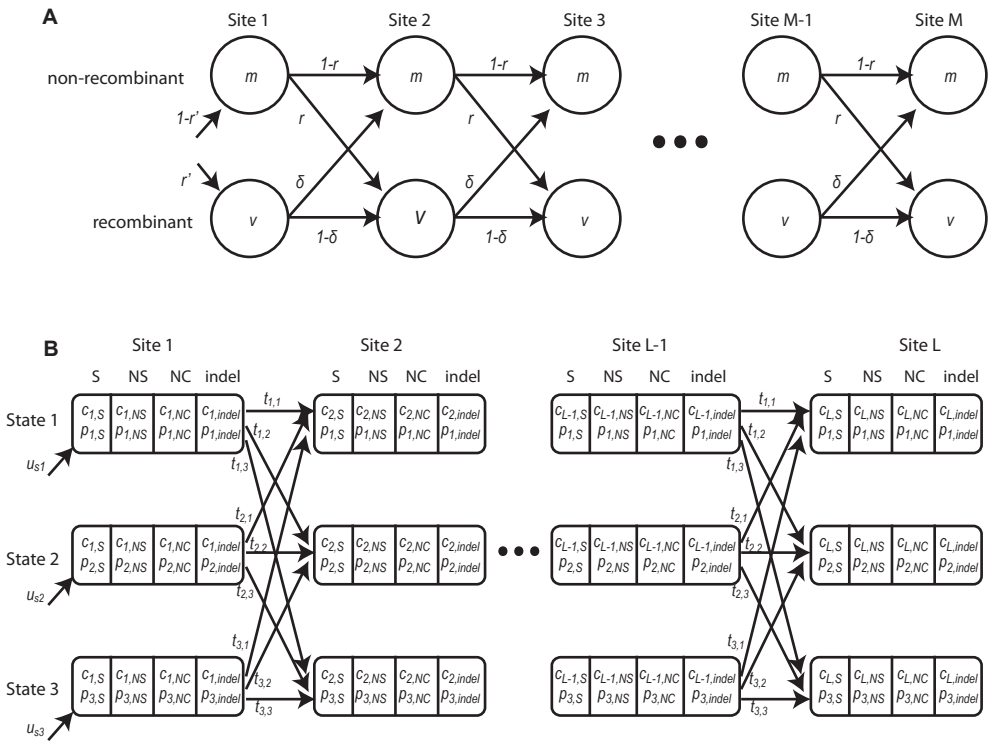

Fig. S6. Illustrations of HMM structures in RecHMM (A) and DHMM with three hidden states (B). The columns represent the first and last sites within the core genome of Paratyphi A, whereas the rows represent different hidden states, with the parameters that are estimated indicated for each site (A, circles; B. rounded rectangles) and for each transition (arrows).

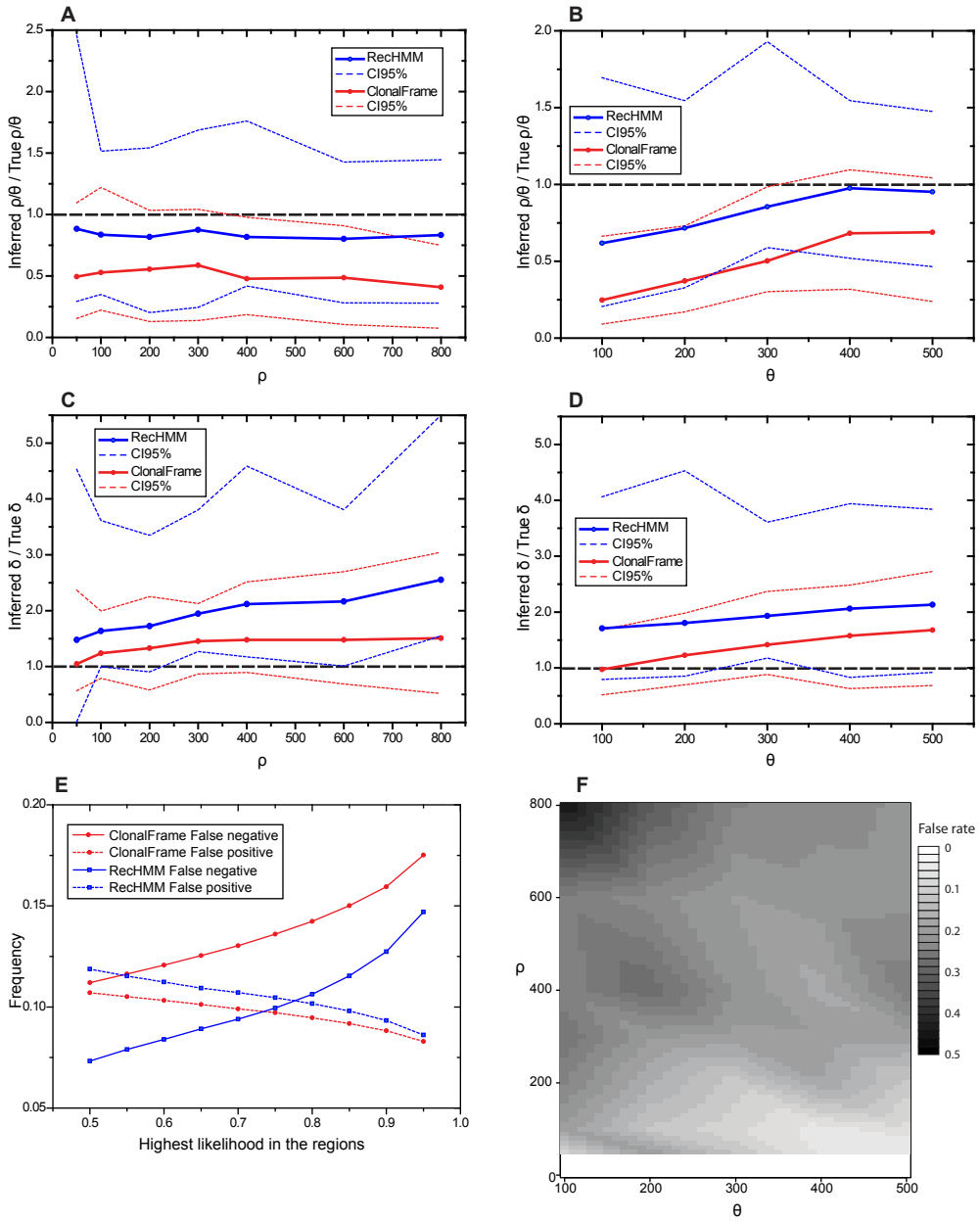

Fig. S7. Comparison of CLONALFRAME (red) and RecHMM (blue) values of recombination rate (ρ) mutational rate (θ) and average recombination tract length (δ) in simulated data. A-D) Ratio of Inferred ρ/θ (A-B) and δ (C-D) with each circle representing average results (thick lines) and their 95% confidence intervals (dashed line) for 10 simulations. E) Frequencies of false positive and false negative errors in assignment of SNPs to recombinational/non-recombinational tracts. F) Interpolated heatmap (function “interp” in R package “akima”) of the sum of false negative and false positive error frequencies by RecHMM for different combinations of θ and ρ.

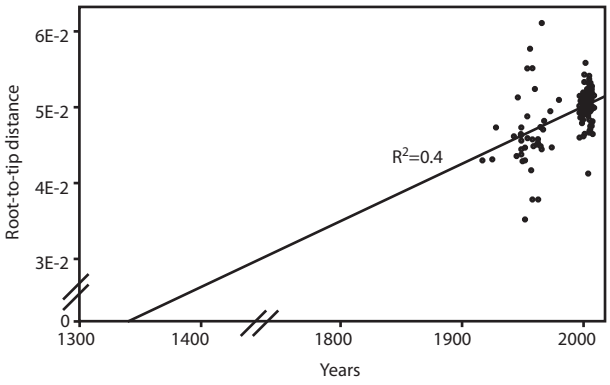

Fig. S8. Root-to-tip genetic distances to the MRCA of 4,525 SNPs *vs* dates of bacterial isolation of 149 isolates of Paratyphi A (circles).

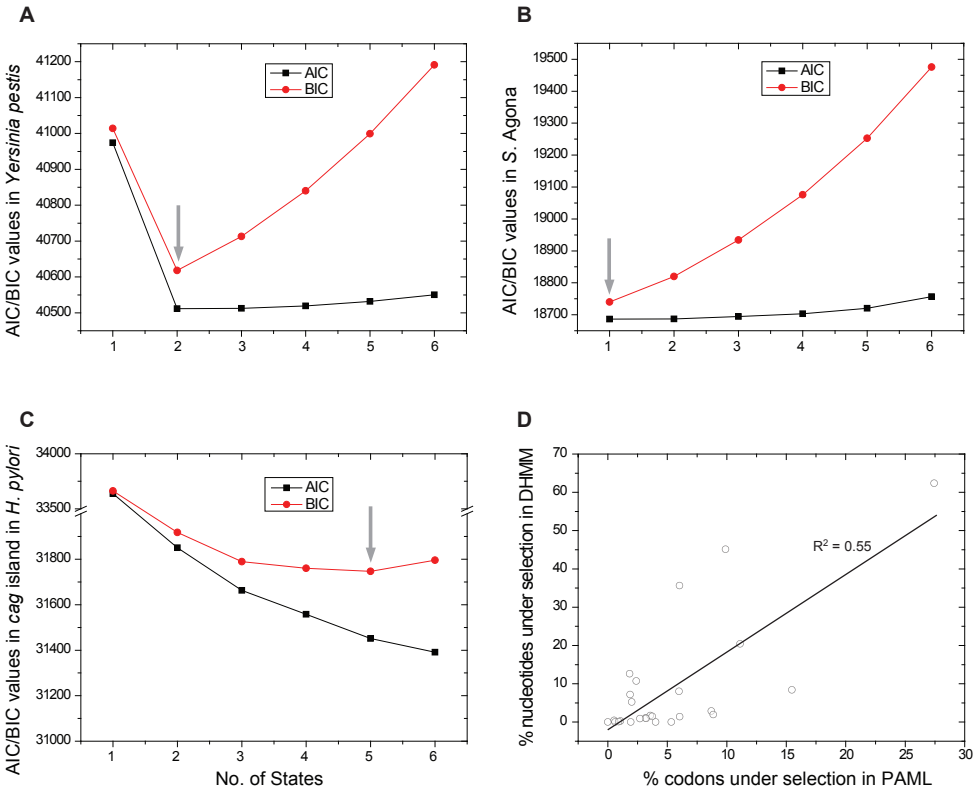

Fig. S9. Numbers of DHMM states in different bacterial data sets and divergent selection inferred by DHMM vs PAML. (A-C) AIC(Akaike information criterion; black) and BIC (Bayesian information criterion, red) *vs* number of states for *Y. pestis* (A), Agona (B) or *H. pylori* (C). Grey arrows highlight the numbers of states with the lowest BIC values. (D) Number of codons under selection according to PAML *vs* number of nucleotides under selection according to DHMM in *H. pylori*. Each circle represents a distinct genes in the the *cag* island.

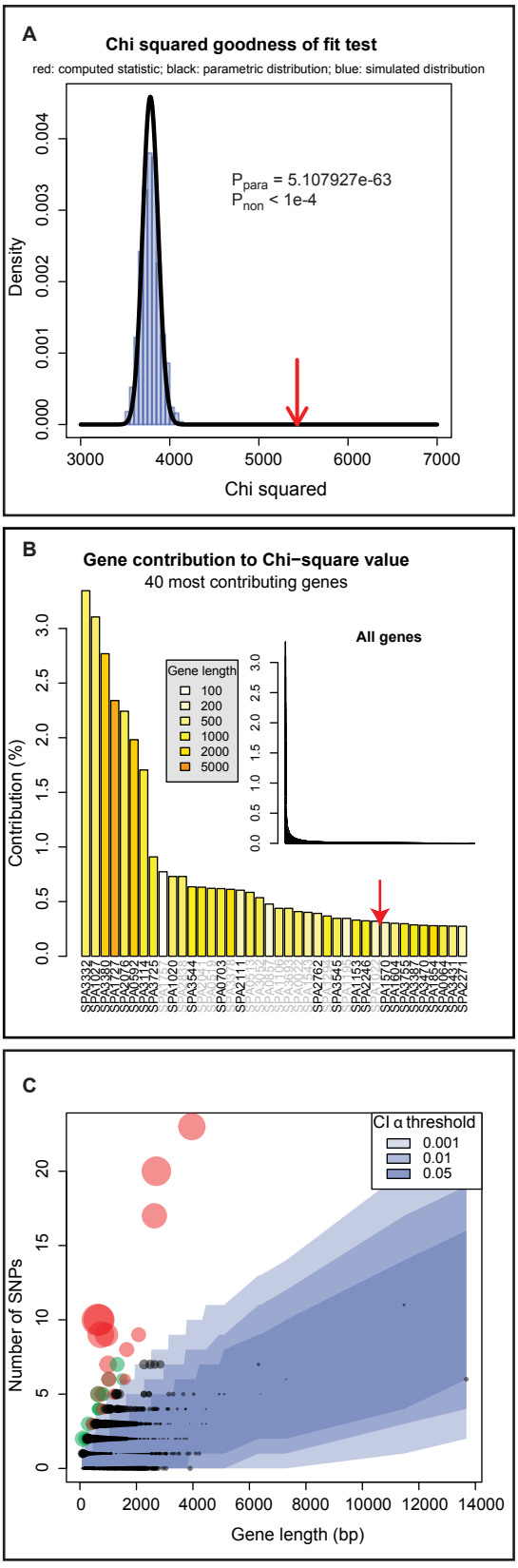

Fig. S10. Statistical tests of neutrality of non-synonymous SNPs within the CDSs in the non-recombinant, non-repetitive core genome. A) Parametric and non-parametric χ^2 goodness-of-fit test of the densities of non-synonymous mutations in genes. Red arrow: observed χ^2 from the core genome; Black curve: parametric distribution; Blue histograms: simulated non-parametric distribution. B) Contribution of the 40 genes which most contribute to the χ^2 values. Significant outliers are to the left of the red arrow. Gene lengths are indicated by histogram colors and names at the bottom of the figure are grey when those genes were not identified by DHMM. C) Numbers of non-synonymous mutations per gene in the non-recombinant, non-repetitive core genome as a function of gene length. Each gene is represented by a black circle, whose size is proportional to the contribution shown in part B. Shades of blue indicate different α-thresholds (0.05, 0.01, 0.001) of the confidence intervals of the theoretical expectations, where 0.05 indicate CI 95%. The 31 identified outliers are indicated in red when they were also identified by DHMM, otherwise in green.