Pros and cons of the genotyping approach

The hybrid approach we have used to reconstruct the phylogeographic history of *Y. pestis* has several limitations. Traces of strong selection that might be present in the genomes may have been eradicated by our exclusion of homoplasies, repetitive DNA and variable sequences. Indeed, the ratio of non-synonymous to synonymous mutations was near unity in the maximum parsimony genomes tree (**Fig. 1**; **Supplementary Table 14**), which indicates the absence of either purifying or diversifying selection in the core-genome that was studied here. The relative distances between populations along the main branches of the MSTree (**Fig. 2**) are accurate, because most of the tips in the tree terminate in genomic sequences. However, the tips of the tree are largely based on SNPs discovered from only 17 genomic sequences, resulting in phylogenetic discovery bias. As a result, the true diversity of populations in the MSTree is strongly underestimated, and the lengths of branches that are not on the path to a genomic sequence are much too short because they only reflect the diversity revealed by low resolution SNP discovery. Nevertheless, the genomes and minimal spanning trees share an identical branching order, which is fully parsimonious and reflects unidirectional, clonal evolution. These properties indicate that our genealogical reconstruction is very robust. The branches are defined by sequential fixed mutations, each of which only occurred once prior to multiplication and spread.

Mutation rate in nature. Historical records show that an initial plague epidemic reached Madagascar in 1898 and that a second wave began in 1921 [11]. We calculated mutation rates per year in nature based on the frequencies of SNPs discovered by dHPLC with 81 isolates from Madagascar of known dates of isolation (Madagascar isolate 1494 was excluded because its date of isolation was unknown). Mutation rates were inferred using a full maximum likelihood model which assumes that after introduction in Madagascar, demographic expansion was strong enough to result in perfect star genealogies, i.e. without any coalescent events. Indeed, almost all mutations discovered by dHPLC were singleton mutations (**Supplementary Fig. 5**), which strongly supports this assumption. The likelihood of the model for each locus is given by the binomial probability of the number of mutations observed on all strains given the sum of the genealogical branch lengths for all strains (i.e. date of strain collection - date of expansion start) and the mutation rate per locus and per year. The point multilocus mutation rate estimate (mutation rate per nucleotide per year) and its 95% confidence interval were inferred by considering the product of the above-described likelihood function for all loci, under the assumption that all loci have a defined number of sites, mutations are independent and occur with a single constant mutation rate. For this analysis, 382 loci from coding regions spanning a total of ~167 kb were screened in 81 Malagasy strains, assuming a demographic expansion that started either in 1898 or 1921. Comparable results were obtained for both dates (**Supplementary Table 2**). The 81 isolates fell into two distinct groups in a minimal spanning tree (**Supplementary Fig. 5**) and therefore the same calculations were performed for each of the two subgroups, again with comparable results. The calculations were performed using a procedure written in R[48], which is available at http://research.ucc.ie/NG1/index.html as are other supplementary Figures, Tables and R scripts.

Estimates of divergence times. The divergence times for each node of a genome tree for 15 *Y. pestis* genomes (excluding FV-1 and Angola) plus the outgroup *Y. pseudotuberculosis* IP32953 were estimated using BEAST 1.5.3 [22]. The polymorphisms analyzed were derived from all coding regions that were present in all genomes, except for repetitive regions. These analyses were restricted to coding regions because mutation rates had been exclusively calculated on mutations within coding regions (**Supplementary Table 2**).

Divergence times were calculated using the geometric means from three replicate BEAST runs for both a strict and a relaxed clock, each of which was calculated for two different substitution rates (**Supplementary Table 2B**). A constant population size was assumed and the HKY substitution model was chosen for the simulations because of the limited variability within *Y. pestis*. The final runs had a chain length of 50,000,000 steps, which sufficed for reaching stationarity. Intermediate runs of 20-50 million steps were also performed in order to adjust the tuning and weighting parameters, as suggested in the BEAST output files.

For the strict clock, whose dating estimates are shown in **Fig. 1**, the substitution rates used were the upper and lower bounds of the 95% confidence interval of six independent estimates of the mutation rate in Madagascar (**Supplementary Table 2A**). For the (uncorrelated log-normal) relaxed clock, the highest and lowest of the six mutation rates were used. Ages were then estimated from the corresponding lower 95% HPD and upper 95% HPD estimates, respectively, that were generated by BEAST. These runs were based on default parameters, except that the prior for ucld.stdev was a normal distribution centred on a mean of 0.75, which was estimated on the basis of an initial run without a defined mutation rate. These calculations were designed to result in a broad range of estimated dates in order to ensure that they encompassed the true dates. Both sets of estimates overlapped and we chose the estimates from the strict clock analyses for **Fig. 1** because a strict clock makes fewer assumptions.

Molecular clocks during epidemics.

Many RNA viruses are thought to have evolved in the last centuries[49] according to nucleotide sequences from archival virus strains that were isolated over several decades. Particular clones of *Staphylococcus aureus*[50,51] also have mutation rates that are fast enough to allow the deduction of their evolutionary histories on the basis of dated archival isolates, and the clock rates for *Buchnera*[52] and *Helicobacter pylori*[53] are only slightly slower. However, the endemic clock rate for *Y. pestis* of $2.3 \bullet 10^{-8}$ to $2.9 \bullet 10^{-9}$ is 10fold (*Buchnera*) to 1500fold (*S. aureus*) slower than for these more rapidly mutating bacteria. Dating evolutionary events in *Y. pestis* and many other bacteria will therefore probably continue to depend on correlations with independent historical events, such as are described here.

It is worth noting that only very few SNPs mark the microevolution of the 1.ORI1 radiation. Two SNPs mark the first node isolated from India, Hawaii and California, possibly reflecting limited prior microevolution in China. Seven other SNPs seem to have become fixed within the U.S.A. and the genome of isolate CO92 at the end of this radiation possesses four additional, strain-specific SNPs. These observations argue for neutral clock-like evolution with time rather than for bursts of adaptation or accelerated mutation rates. However, in those cases where historical records were correlated with dispersion events, the known dates of spread are near the lower end of the evolutionary date estimates, or later. The 3rd pandemic began to spread ~150 ya whereas we estimated that 1.ORI evolved >212 ya. 1.ANT evolved >628 ya, whereas we invoke voyages by Zheng He approximately 580 ya as the route whereby those bacteria were transmitted to Africa. Genotypes from the Black Death in Europe dating to the mid-14th century (~610 ya)[14] are as old as the split between branches 1 and 2, which was estimated as >728 ya. If this trend were consistent, then 2.MED should have first reached Western Asia in the last 235 years, and the oldest branches of *Y. pestis* began their spread only slightly more than 2,500 ya. But if our mutation rates are accurate, why should historical dates slightly postdate our youngest dating estimates based on a fixed molecular clock rate?

The timing for the accumulation of SNPs in certain lineages also raises questions about a fixed molecular clock. Ten strain-specific SNPs were found in each of the genomes of two strains from Madagascar. This number of SNPs is compatible with the endemic mutation rate in Madagascar because those two strains were isolated in 1995 and 2005, respectively. However, the endemic mutation rate is too low to readily account for the seven SNPs that differentiate Madagascar nodes in the blue cluster from their direct ancestors in Israel and India because those SNPs must have become fixed by 1926, the date of isolation of EV76 (**Supplementary Fig. 5**). Under the assumption of a strict molecular clock, the probability of such rapid accumulation of seven SNPs is below 2%, even for our highest plausible substitution rate (see below). The same problem applies to seven other country-specific SNPs that distinguish Madagascar nodes from their descendents in Turkey, a transmission which had probably happened by the 1930's. These observations suggest that clock rates might accelerate during epidemic spread, which might also account for our observation that historical transmissions slightly post-dated our age estimates.

Transiently accelerated accumulation of SNPs might possibly reflect a rapid, initial phase of selective adaptation to new hosts and vectors after spread to a new area, similar to experimental evolution with *Escherichia coli* in the laboratory over 40,000 generations[54]. However, the patterns of the SNPs specific to Madagascar or Turkey (country-specific SNPs) argue against positive selection. In the laboratory experiments, all SNPs were non-synonymous and clustered in a limited number of genes. In contrast, 5/14 country-specific SNPs in 1.ORI3 were synonymous, and each of the eight missense mutations was in a distinct housekeeping or hypothetical gene. Similar patterns apply to almost all other SNPs described here, and D_N/D_S ratios were not markedly different from 1.0 for the entire dataset or for individual branches of the tree (**Supplementary Table 14**). Similarly, only 36/1152 SNPs were homoplasies (**Supplementary Table 12**) and only one of 3,349 genes (*aspA*) contained multiple

SNPs, which are hallmarks of selection. Thus, the rapid accumulation of SNPs during geographical spread is unlikely to reflect positive selection.

A second possibility was that increased mutation rates might arise due to the transient existence of mutators[55]. Indeed, 20 strain-specific SNPs were detected by dHPLC in strain IP619, part of radiation iv to South Africa (**Fig. 2**), whereas no more than 3 strain-specific SNPs were found by dHPLC for other isolates. It seemed possible that IP619 is a mutator. Spontaneous mutants occurred in strain CO92 at an average frequency of 1.2×10^{-8} for resistance to nalidixic acid (Nal^R) and 6.4×10^{-9} for resistance to rifampicin (Rif^R) in 10 parallel cultures inoculated from single colonies. Spontaneous mutants occurred in IP619 at an 83fold to 188fold higher rate (Nal^R: 1.0×10^{-6}; Rif^R: 1.2×10^{-6}), confirming that it is a mutator. Similarly, 708 strain-specific coding SNPs were found in the Angola genome (**Supplementary Table 5**), many more than in any other genome. However, Angola is not a mutator, nor did the Angola genome[27] or any others contain any signatures of epistatic reversion of mutations in genes encoding mismatch repair.

A third possibility was that neutral demographic processes, such as bottlenecks or rapid expansion during epidemics, could account for the rapid accumulation of SNPs during outbreaks. However, both a simple cumulative binomial distribution and the explicit simulation of serial outbreaks described below agreed that seven SNPs are unlikely to have accumulated by 1926 (**Supplementary Figs. 8 - 10**), even when we used the highest plausible estimate of the endemic mutation rate. We raise the possibility of an alternative mechanism that would elevate the mutation rate per unit time, namely if transmission between hosts were more rapid during outbreaks than during endemic sylvatic disease. Elevated transmission rates would effectively result in a larger number of cell divisions per unit time without changes in the mutation rate per bacterial generation. Transiently elevated transmission rate during epidemics might account for the rapid accumulation of multiple SNPs seen during geographic transmissions in some outbreaks, which in turn could result in a slight overestimate of TMRCA's based on the endemic mutation rate.

Simulating the genetic evolution of strains during outbreaks. We used extensive simulations to follow the genetic polymorphism of strains in different demographic contexts. To this end, we developed a forward-time, individual-based simulation system, implemented by the functions *haploPop* and *haploPopDiv* in the *adegenet* package[56] written in R[48]. Scripts reproducing our simulations are provided in **Supplementary R Scripts for simulations in Figs. S8-10**. This new simulation tool offers considerable flexibility in the specification of genetics and demography, including complex epidemic scenarios in a metapopulation context. Reproduction and death of the strains, carrying capacity, and seeding of new populations can be considered as fixed parameters or random variables, in which case any probability distribution can be used. While the basic implementation of this simulation system (*haploPop*) returns haplotype(s) at the final generation, *haploPopDiv* also records different measures of genetic differentiation at each step. Outputs of *haploPop* can also be used as inputs for a new batch of simulations with different sets of parameters, thus allowing for alternation between different demographies.

As a proof of principle, we first ran simulations of single outbreaks caused by an organism with a genome of four million base pairs and a mutation rate of 10^{-7} per site per generation. We assumed the complete absence of genetic recombination, and all simulations were initiated with 100 genetically invariant strains. We first assessed the dynamics of genetic differentiation in a control population at equilibrium for mutations and/or drift. To this end, we simulated a population of strains with constant growth rate (R_0=1.1) and a carrying capacity of 100 susceptible hosts (K), renewed at each generation, which is equated here with successive host to host transmissions. These simple population dynamics are in effect equivalent to a Wright-Fisher model and may correspond to the stable transmission of *Y. pestis* in a reservoir population of rodents. This control simulation was run for 10,000 generations, of which the first 1,000 generations were discarded to ensure that equilibrium had been reached. After each generation, we recorded the pair-wise genetic distances (in numbers of mutations) within a sample of 50 strains, as well as the distribution of allele frequencies. Differences with time were calculated as the differences in average genetic distance between random pairs of samples that were taken 10 generations apart. This operation was replicated 1,000 times to obtain a reference distribution.

Secondly, we simulated non-equilibrium outbreaks using different fixed growth rates (R_0=1.1, 1.5, and 2.0), with 50 replicates for each value of R_0. This population dynamic modelling was designed as an SIR model: we used a fixed number of susceptible hosts (*K*=100,000), which were removed from the population after infection. As before, after each generation we monitored the average genetic distance between pairs of strains, as well as the distributions of allele frequencies. Because outbreak data largely reflect the late phase of pathogen expansion, the difference in pair-wise genetic distances were compared between the generation with the largest number of pathogenic strains, which corresponds to the peak of the outbreak, and the sample taken 10 generations previously.

Our simulations clearly confirm prior analyses[57-59] that an excess of genetic differentiation between strains accumulates during rapid population expansions. In the population at mutation/drift equilibrium, the mean genetic distance between strains taken 10 generations apart does not differ significantly from zero (Student *t* test: t_{eq}=-0.09, p=0.54). In contrast, the mean distance at the peak of the outbreak is substantially greater than zero in all simulated outbreaks (Student *t* tests: $t_{\mathrm{R0=1.1}}$=54.22, $p<2.2\mathrm{x}10^{-16}$; $t_{\mathrm{R0=1.5}}$=80.83, $p<2.2\mathrm{x}10^{-16}$; $t_{\mathrm{R0=2}}$=154.23, $p<2.2\mathrm{x}10^{-16}$; **Supplementary Fig. 8**). Interestingly, pairwise genetic distances in the population at equilibrium show a very large variance, unlike outbreak data. Comparable variability in equilibrium populations was also observed in control simulations performed with

Easypop[60], and represents a hallmark of the stochasticity inherent to single locus dynamics in equilibrium populations.

The distributions of allele frequencies observed during outbreaks also differ dramatically from the equilibrium population (**Supplementary Fig. 9**). There is a massive excess of rare variants segregating during outbreaks, and no derived mutation reaches high frequency or fixation. This shift in allele frequencies is driven by the inefficacy of genetic drift in exponentially growing populations. However, while the previous simulations provide a proof of principle for the excess diversity which can be generated by epidemic outbreaks, they fail to reproduce the pattern observed in *Y. pestis*. For instance in Madagascar, we observed both considerable segregating diversity, as well as a sizeable number of fixed private variants, i.e. seven polymorphisms were fixed in Madagascar and absent from the hypothetical source populations.

In the next step, we ran simulations that were parameterised with estimates from *Y. pestis*. We used the same 4 Mb genome size but with an intermediate point mutation rate of 1.25 x 10^{-8} per nucleotide per year. We considered an average serial generation time (the time between two successive infections by the same strain) of five days, thus leading to a per generation mutation rate of 1.7 x 10^{-10}. We allowed the strains to evolve in epidemics of susceptible rodent host populations ranging from 10,000 to 100 million individuals and intrinsic growth rates (R_0) ranging between 1.01 and 2.0. Sampling was performed during the peak of the outbreak to reflect the overspill to human hosts. Using these mutation rate estimates, we were unable to reproduce the large genetic diversity observed in the empirical data, even when considering immense host populations sizes (up to 100 million). The simulations also did not lead to the fixation of SNPs that were not present in the hypothetical source population. We reasoned that the empirical distribution of mutations observed in *Y.*

pestis could not result from a single epidemic, but instead required the dynamics of serial outbreaks.

Therefore, we finally considered a model of sequential outbreaks and specifically explored the parameter combination that would lead to patterns compatible with the ones observed in Madagascar. As plague is seasonal in Madagascar[11] and elsewhere, we simulated 80 successive outbreaks representing the number of years between the initial arrival of *Y. pestis* and the average sampling dates of the isolates. We considered a rat population, the primary susceptible rodent host in Madagascar[11], between 10,000 and 1 million individuals, and intrinsic growth rates between 1.1 and 2.0. An additional key parameter in the serial outbreak model is the effective number of strains contributing to the following outbreak (N), the bottleneck size, for which we considered values between 10 and 100. We ran extensive simulations for all parameter combinations recovering both the number of segregating SNPs and the number of fixed mutations that were absent in the source.

Depending on the parameter value combination, we recovered very different patterns of genetic diversity and resulting tree topologies. It proved straightforward to delineate a broad parameter space leading to tree topologies qualitatively similar to the one observed for the Madagascar isolates. In particular, we could reproduce the sizeable diversity in terms of pair-wise differences between isolates. Interestingly, we also regularly observed two dominant strains separated by one or two SNPs, as also found in the empirical data. However, there is one striking feature of the tree for isolates from Madagascar that we were systematically unable to reproduce in our simulations, namely the accumulations of seven new fixed mutations by 1926, which is at most 27 years after importation. Even when we applied the upper bound of the mutation rate estimated from the maximum likelihood analysis (2.3×10^{-8} mutations per site per year; **Supplementary Table 2**), we never recorded such a high fixation rate within 27 years (**Supplementary Fig. 10**). Our simulations confirm that while the demographic dynamics of epidemic outbreaks can generate considerable transient genetic diversity, they do not affect the fixation probability of mutations on their own. This observation is in line with classical results in population genetics, which predict that the fixation rate, in contrast to genetic diversity, is independent of fluctuations in population sizes[61].

Additional Citations

48 R Development Core Team. R: A language and environment for statistical computing. 2004.

49 Holmes, E. C. Evolutionary history and phylogeography of human viruses. *Annu.Rev Microbiol.* **62**, 307-328 (2008).

50 Harris, S. R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469-474 (2010).

51 Nubel, U. *et al.* A timescale for evolution, population expansion, and spatial spread of an emerging clone of methicillin-resistant *Staphylococcus aureus*. *PLoS Pathog* **6**, e1000855 (2010).

52 Moran, N. A., McLaughlin, H. J., & Sorek, R. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science* **323**, 379-382 (2009).

53 Morelli, G. *et al.* Microevolution of *Helicobacter pylori* during prolonged infection of single hosts and within families. *PLoS Genet* **6**, e1001036 (2010).

54 Barrick, J. E. *et al.* Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* **461**, 1243-1247 (2009).

55 Giraud, A. *et al.* Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut. *Science* **291**, 2606-2608 (2001).

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

57 Rogers, A. R. & Harpending, H. Population growth makes waves in the distribution of pairwise genetic differences. *Mol Biol Evol* **9**, 552-569 (1992).

58 Slatkin, M. & Hudson, R. R. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**, 555-562 (1991).

59 Holmes, E. C., Nee, S., Rambaut, A., Garnett, G. P., & Harvey, P. H. Revealing the history of infectious disease epidemics through phylogenetic trees. *Philos.Trans.R.Soc.Lond B Biol Sci* **349**, 33-40 (1995).

60 Balloux, F. EASYPOP (version 1.7): a computer program for population genetics simulations. *J Hered.* **92**, 301-302 (2001).

61 Kimura, M. *The neutral theory of molecular evolution*Cambridge University Press, Cambridge (1983).