doi:10.1038/nature10549

Materials and Methods

Mortality records. Figure S1 is based on data collected and published by Cohn, 2003[1] (Fig. 7.33). The City of London did not publish bills of mortality in the 14th century. Cohn counted the numbers of last wills and testaments submitted to the Court of Husting in London each month in 1348 and 1349. While only relatively prosperous individuals would have had last wills and testaments, the temporal pattern of deaths indicated by these monthly counts is likely to be strongly correlated with the full epidemic curve for the population of London as a whole.

Samples. Dental material from the four individuals was harvested from the East Smithfield collection at the Museum of London as previously described[2]. The East Smithfield cemetery of London, England (Figure S2) is one of the few excavated medieval catastrophe burials in Europe, and historical documentation clearly indicates purchase of this land in late 1348 or early 1349 specifically for interment of victims of the pestilence[3,4], who had quickly exhausted available plots in local perish cemeteries. Shortly after the Black Death, the Abbey of St. Mary Graces was established on this land in 1350. The plot later housed the nineteenth-century Royal Mint for the city of London, hence the skeletal collection has been equivocally referred to as the "Royal Mint collection" in published work.

Extraction and screening. All manipulations of samples and DNA extracts were performed in facilities specifically dedicated to the extraction of ancient DNA with no prior exposure to sources of *Y. pestis* DNA, and following stringent protocols to limit laboratory contamination[5]. Pulverisation of dentin from both roots and molars was performed as previously described[2]. DNA extraction from dental material was performed by a protein kinase mediated lysis, followed by purification via either a modified phenol chloroform method as previously described[2] (McMaster) or a guanidinium-silica method[6] (MPI). A minimum of one negative extraction blank per 8 samples was carried along throughout.

Library preparation and indexing. 75µl of each DNA extract, extraction blank control or water library blank control (a minimum of one per every 10 libraries), was used for library preparation following the protocol of Meyer and Kircher (2010)[7] with modifications for ancient DNA[8]. For extracts purified via phenol chloroform, supernatant and pellet fractions were combined, and 75µl of the resultant pool was used in the immortalisation procedure. For all libraries, the blunt-ending step was preceded by treatment with uracil-DNA-glycosylase as well as Endonuclease VIII[9] to remove deaminated cytosines that contribute to C > T damage motifs in ancient DNA[10]. The chemistry of this procedure was scaled to accommodate a larger volume of extract. Post library preparation, each library was amplified with two 5'-tailed 'indexed' PCR primers, adding sample-specific indexes to both library adapters in this process[8]. Individual PCR reactions contained a maximum of $5x10^8$ library template molecules, and were amplified in 100µl reactions with the following thermal profile: 12 minute initial denaturation at 95°C, 10 cycles consisting of a 30 second denaturation at 95°C, a 30 second annealing at 58°C, and a 45 second elongation at 72°C, ultimately followed by at 10 minute final elongation at 72°C.

Array design. We used Agilent one million (1M) custom features arrays to capture *Yersinia pestis* genomic DNA libraries. We designed two 1M arrays by tiling 60 bp probes every three bases along the complete 4.6 Mb *Y. pestis* strain CO92 genome (NC_003143), the 70Kb pCD1 (NC_003131) and 96Kb pMT1 (NC_003134) plasmids. We included additional 60 bp probes every 1 base to get extra coverage of 933 single nucleotide polymorphisms (SNPs) that have been used to performed phylogenetic analysis of different *Y. pestis* isolates[11], and eight *Y. pestis* genomic regions that contain the virulence genes *hms* (hemin storage locus, YPO1951- YPO1954), *T3SS* (Type III secretion system, YPO0255 - YPO0273), an enhancing factor (YPO0339), the putative insecticide enhancing factors (YPO3678, YPO3681), and the putative insecticide toxin (YPO2312, YPO2380, YPO3673, YPO3674). The array design pipeline for mammalian genomes we used removes probes that contain repetitive elements based on 15-mer frequency counts[12]. First a table of frequencies for every 15-mer found in a given genome (sense and antisense) is calculated. Then, for each probe, the genomic frequency values of all of the probe's 15-mers are averaged. If this average genomic frequency of the 15-mers in one probe is above a given threshold, the probe is removed. A threshold of 100 has been used to design arrays based on mammalian genomes[12, 13]. We applied here the same methodology to remove probes that contain repetitive elements. Using 100 as a threshold, no probes were discarded. The highest frequency for a 15-mer in the *Y. pestis* genome was 66. The first array we designed (array A) contained probes covering the first 2.5 Mb of the *Y. pestis* chromosome, the second array (array B) contained the last 2.1 Mb of the chromosome, the plasmid pCD1, the plasmid pMT1, several virulence genes, and all SNPs taken from Morelli et al. 2010[11].

Array capture. Capture was performed following methods previously described[17] where all amplification reactions were performed with Phusion Hi-Fidelity DNA polymerase (New England Biosystems) in 100µl reactions with the following chemistry: 50µl Phusion High Fidelity Master Mix, 400 nM each p5 – p7 bridge amplification primer, and 10µl of template. Thermal profile consisted of a 3 minute initial denaturation at 95°C, between 5 to 10 cycles of a 30 second denaturation at 95°C, 30 second annealing at 60°C, and a 45 second extension time at 72°C, and a final elongation at 72°C for 5 minutes. PCR products were spin column purified over MinElute and quantitated via an Agilent 2100 bioanalyzer following manufacturer's protocols. The products were captured on two copies of each array A and array B in parallel, four arrays in total. After the first round of capture the 490µl eluate from the two identical arrays were pooled and PCR amplified in 20 independent PCRs in 100µl reactions containing 50µl PhusionTM High-Fidelity Master Mix, 1µM of each p5 – p7 bridge amplification primer and 50 µl library template and a total of 20 cycles. PCR products were spin column purified and quantified via an Agilent 2100 bioanalyzer. A second round of array capture was performed using the same conditions as the first round, only this time each of the two library pools were put on single arrays (array A or array B). The eluted 490µl from each array was subsequently amplified in 10 reactions using the amplification conditions described above.

Sequencing. The enriched libraries were pooled in equimolar concentrations and sequenced on the *Illumina Genome Analyzer IIx* platform using 2 x 76 + 7 +7 cycles on a single sequencing lane according to

the manufacturer's instructions for multiplex sequencing (FC-104-400x v4 sequencing chemistry and PE-203-4001 cluster generation kit v4). The manufacturer's protocol was followed except that an indexed control PhiX 174 library (index 5'- TTGCCGC-3') was spiked into each lane yielding a fraction of 2-3% control reads in all lanes of the run. The sequencing data were analyzed starting from QSEQ sequence files and CIF intensity files from the Illumina Genome Analyzer RTA 1.6 software. The raw reads were aligned to the PhiX 174 reference sequence to obtain a training data set for the base caller Ibis[15]. Raw sequences called by Ibis 1.1.1 were filtered for the 'AATCTTC' index as described[15]. The paired-end reads were subjected to a fusion process (including removal of adapter sequences and adaptor dimers) by requiring at least an 11nt overlap between the two reads. In the overlapping sequence, quality scores were combined and the base with the highest base quality score was called. Only sequences merged in this way were used for further analysis. A small number of larger molecules (longer than 141nt) had lower sequence quality, and were thus discarded. The total number of merged reads for individual samples and the total pool can be found in Table S1b.

Mapping, and calculation of average fragment length and genome coverage.

Mapping. Merged reads were aligned with BWA[16] to the CO92 *Y.pestis* reference genome (NC_003143), the 70Kb pCD1 (NC_003131), the 96Kb pMT1 (NC_003134), and the 9.6Kb pPCP1 (NC_003132.1) CO92 plasmids using default parameters. Mapping alignments were converted to SAM/BAM format[17] with BWA's *samse* command. Accurate mapping into different genomic regions was complicated by the presence of a 1956-bp transposase that is present in one copy in the pCD1 and the pPCP1, and in multiple copies on the pMT1 and the chromosome. One copy of this transposase was removed from the reference plasmid sequences prior to mapping, though its multiple copies in the chromosome and pMT contributed to inflated coverage estimates for these genetic components prior to additional filtering (Table S1b). The pooled dataset of all enriched extracts generated a total of 20,512,847 merged reads, of which 13,180,582 (64.26%) mapped to the CO92 reference chromosome. Multiple amplifications of the DNA library and enriched fractions performed as part of the capture procedure made it necessary to remove all identical molecules from the merged reads such that each starting library template was considered only once in genome assembly. This poses certain limitations when working with data from a bacterial population, since accuracy in identifying a unique template sequence requires properly distinguishing true genetic variation in the population, which can accrue rapidly even within a single host during pandemic episodes, from artefacts introduced via PCR misincorporations and Illumina sequencing. Duplicate molecules were removed using *rmdup*. This method clusters all reads together that have identical start and end coordinates in the mapping alignment, and uses the read with the highest mapping quality as representative of the cluster. This process, therefore, reduced the error signal resulting from amplification and sequencing, and also reduced the number of reads mapping to the transposase of the chromosome and pMT. The total number of mapping fragments before and after duplicate removal was calculated with *flagstat*[17] (Table S1b). Duplicate removal resulted in a total of

2,812,240 (13.71%) chromosomal reads, and filtering with Q>30 reduced this number to 2,366,647 (11.53%) high quality chromosomal reads. The average fragment length and length distribution was calculated using the SAM format loaded into *Galaxy*[18, 19]. Using the *pileup* function of *samtools*[17] all BAM files were converted into *pileup* files to calculate average coverage over the genome by dividing the total number of nucleotides mapping to the CO92 genome by its length using a mapping quality of Q>30. Average coverage was estimated at varying fold coverage and with and without mapping quality filtering (Table S1a, S1b, and S1c). To determine if we have sequenced the enriched library to exhaustion we estimated how often we have sequenced each original fragment for array A and array B separately. Average coverage of the target chromosomal region for each array was calculated using a mapping quality filter of Q>30 for the total reads both before and after duplicate removal (*rmdup*). Array A gave an average coverage of 66.09 before and 36.46 after duplicate removal, and array B gave 34.87 before and 22.41 after. Thus for array A, each original fragment was seen on average approximately 1.8 times whereas for array B each fragment was seen on average 1.5 times. This indicates that array A was sequenced to a greater depth than array B. The differences in sequencing depth for the arrays could be potentially explained by differences in enrichment efficiencies or experimental variation. Regardless of different sequencing depths, each array covered a comparable number of sites for the target region at a minimum of 1-fold coverage, indicating that further sequencing would not contribute to additional information on sites not currently covered (Figure 1b, main manuscript).

Genome architecture and contig assembly. Using the read dataset for individual 8291 aligned to the CO92 chromosome with BWA[16], all regions with a minimum mapping quality of 25 were extracted. A reference-guided sequence assembly was performed in Velvet version 1.1.03[20] using the extracted reference sequences supplied to the Columbus extension for reference-aided assembly in addition to all mapped and unmapped reads for individual 8291 in separate channels. A hash length of k = 21 was identified as producing an optimal assembly after iterations using several values for k. NCBI BLAST 2.2.21 was used for local similarity searches. The reference-guided assembly produced 130,556 contigs with an N50 of 288 bp. When filtered for contigs 500bp or larger 2134 contigs remained totalling 4,013,009 bp. This consisted of 30,959 unmapped reads. To identify contigs that potentially contained novel architecture distinct from CO92 and stemming from unmapped reads, all contigs were queried against the CO92 chromosome using BLAST[21] with an E-value cutoff of 1e-50. Twenty-nine were contigs that had no hits to CO92: six had 100% identity to human mtDNA, 8 matched mobile elements that were common to several species of bacteria, and 16 had no or low significant similarity to sequences in the NCBI database. A potential variant was considered for further examination if the contig sequence produced multiple hits to the CO92 chromosome. Any such contig would have segments that occur at a distance in the CO92 chromosome that the BLAST algorithm would be unable to extend through. Ideally these segments would be non-overlapping and would occur at a distance of more than one read length (54bp). Each contig was queried against other NCBI *Yersinia* sequences to check if the sequence proposed by the contig was observed in other species in the genus. Of the 20 such contigs identified (Table S2), half were either eliminated as mapping errors or were removed due to their location within

known repetitive regions. The remaining 10 contigs provided indications of altered structure in the ancient organism as compared to the reference CO92 genome. The contig numbers, length, description of the change relative to CO92, and the number of *Yersinia* strains in which each structural variant was observed are recorded in Tables S2a and S2b. An example is provided in Figure 2 in the main manuscript using contig number 6149, which suggests a breakpoint joining sequence in the *pbpC* pseudogene, which is intact in ancestral *Yersinia* strains, but spans a gap of 231 kb in the CO92 genome. The mapping alignment shows both joined ends occurring next to regions of high coverage (insertion element IS285), though a read depth of 20 is depicted in Figure 2 owing to space restrictions and resolution. Only short overlaps of several bases are observed across the breakpoint in the mapping alignment and no read is aligned across its centre at that position. Unmapped reads have spanned this breakpoint producing a natural overlap joining these two ends in reference-guided assembly. This specific orientation is shared only with *Y. pseudotuberculosis*, Microtus, Pestoides, and B42003004, which are ancestral to all currently circulating *Y. pestis* strains commonly associated with human infection. Separate orientations exist for this region in branch 1 (*pestis* Z176003, *pestis* D182038, *pestis* D106004, CO92, Antiqua) and branch 2 (Medievalis str. Harbin 35, Nepal516, KIM) strains, indicating that rearrangements in the *pbpC* pseudogene occurred as independent events on different lineages. This ancestral character supports the authenticity of our sequence data because these contigs demonstrate marked differences from modern *Y. pestis*. Each proposed breakpoint position was examined in the 8291 mapping alignment as well as a mapping alignment for each contig to determine if 1) the reconstruction was supported by multiple reads, 2) it did not occur in a repetitive region of CO92, and 3) was not produced by mis-assembly of a collapsed repeat. Additionally, the mapping alignment needed to indicate that the breakpoints occur in relatively poorly mapped regions of CO92 such as that described for contig 6149 above. This process revealed that 10 of the 20 contigs resulted from improper assembly (Table S2b).

Genomic analysis. To identify positions that differ from the CO92 *Y.pestis* reference genome, pileup files generated from BAM files were filtered for positions with a mapping score of Q>30, a major base frequency of at least 95% and at least a 5 fold coverage using *Galaxy*[18, 19]. Using these conditions 97 nucleotide positions were found to be different between the CO92 reference genome and the East Smithfield individuals 11972, 8291, 8134, and 94 positions between 6330 and CO92 (Table S3a). All 97 (or 94) nucleotides were found to be in the ancestral state in the East Smithfield individuals when compared to *Y. pseudotuberculosis*, thus not a single unique derived position was found in the ancient bacterial genome. Furthermore, all of these positions are known variants in extant *Y. pestis* strains[11]. The same strategy was used for reference comparison against the three plasmids. The average coverage of the plasmids was found to be similar for both the pMT1 and the pCD1. The PCP1 plasmid has an estimated copy number of 186 within *Y.pestis*[22] and was not put on the designed capture arrays since it was sequenced at high coverage in a previous work[2]. The estimated low coverage of less than one fold was, therefore, expected. For the pMT1 plasmid 2 differences to CO92 were found, and 4 differences for the pCD1. As with the chromosomal positions, all differences from the reference sequences in the

plasmids were ancestral in the East Smithfield individuals. Differences from the reference sequence are recorded in Table S3a and S3b.

Variant positions were identified by comparing base call frequency data from all high-quality reads for each position with individuals treated separately. A position was considered polymorphic if at least two different nucleotides were recorded in reads with a minimum of 5-fold coverage, permitting 0.9675 confidence. All polymorphic positions are shown in Table S5.

Heat plots were generated for the CO92 genome and for the plasmid genomes to show the relationship between read depth and GC content. In each case the number of sites that have a given number of reads per site and a given GC content (averaged over the 30 neighbouring bases) is colour encoded. The heat plots (Figure S6) show that most sites are covered by a comparatively small number of reads and have intermediate GC content. There is a small skew toward less coverage when the GC content is low. This skew becomes more apparent in the plasmids.

Phylogenetic analysis. We used 1694 of the 1748 positions that were previously identified as variant in modern *Y. pestis* genomes[11], since a multiple alignment of 17 modern genomes as well as the outgroup *Y. pseudotuberculosis* using MAUVE[23] revealed only 1694 of them to be variant in extant *Y. pestis*. The genome sequence of the Angola strain (NC_010159) was excluded from phylogenetic analysis due to its high divergence compared to the *Y. pseudotuberculosis* outgroup[11]. Strain NZ_AAUB01000 was excluded from further analysis due to poor sequence quality. The 27 previously unreported variant positions discovered in our analysis that are derived in the CO92 reference sequence were not included in the 1694 variant position since they could not be confirmed via HPLC[11] and could potentially represent sequencing artefacts in the publically available reference sequence (NC_003143). All 1694 positions were called from the four East Smithfield individuals (Table S4) using *pileup* files and filtered for a minimum coverage of 5-fold and mapping quality of Q>30 using Galaxy[18, 19]. A 95% majority base filter was not used in this step. Using the software SplitsTree4[24] median networks were calculated for all 1694 positions from the 18 modern genomes as well as a subset excluding the Branch 0 genomes[11], and compared to all four East Smithfield individuals. We found that three of the four East Smithfield individuals clustered together (8124, 8291, 11972) whereas one individual (6330) fell closer to other Branch 1 strains (Figure 3, main manuscript). Phylogenetic trees using the Maximum Parsimony (MP) and the Neighbour Joining (NJ) algorithm implemented in MEGA 4.1[25] confirmed a distinct position of individual 6330 compared to the others (Figure S8a and S8b). Thus individual 6330 was excluded for further analysis and individuals 8124, 8291, and 11972 were pooled to get a higher number of informative sites (Table S4, pool 3). The resulting MP and NJ trees are well resolved based on 1000 bootstrap replicates with bootstrap values >90 (Figure S8c and S8d).

Divergence times. To estimate the divergence times of all *Y.pestis* strains the Bayesian algorithm implemented in BEAST v1.5.3 was used. The pool (8124, 8291, 11972) consensus was analyzed using a relaxed uncorrelated log-normal molecular clock as well as the strict clock option[26]. As a tip calibration point we set the prior distribution age of the East Smithfield sample to 1348 - 1350 AD as well as the

isolation years for the modern strains as described elsewhere[11]. For each analysis, we used a model that assumes a constant population size across the phylogeny and ran 50,000,000 generations of the Markov Chain Monte Carlo with the first 5,000,000 generations discarded as burn-in. We chose the HKY sequence evolution model. We performed two BEAST runs and merged both using Logcombiner[26]. The estimated mutation rate was 1.96E-08 (HPD 95% lower 1.7E-08, HPD 95% higher 2.2E-08) mutations per site and per year over the whole genome, using a strict molecular clock[26] . This estimate is similar to previously determined mutation rates for *Y.pestis*[11]. A consensus tree of all 90,000 trees was inferred using TreeAnnotater V.1.4.8. As observed for the NJ and MP tree all nodes were found to be well resolved in the Bayesian tree with posterior probabilities of >0.96 (Figure S9). The estimated coalescence times for the strict and relaxed molecular clock are shown in Figure S9.

Identifying synonymous vs non-synonymous changes.

Comparisons with related genomes. Descriptions of genomic regions for the nucleotide differences were determined using the current annotation of C092 and its plasmids. The location of the difference is given in the *mpileup* file and this was determined to be either synonymous or non-synonymous given the estimated mapping position and the location within the annotation.

A number of differences in coding regions for proteins involved in various aspects of pathogen virulence were observed between the ES strain and both *Y. pestis* Microtus and *Y. pestis* CO92 (Tables S3 and S6d). Although the specific role of these differences in virulence has not been assessed, the genes in which they appear are well known virulence factors.

Supplementary figures:

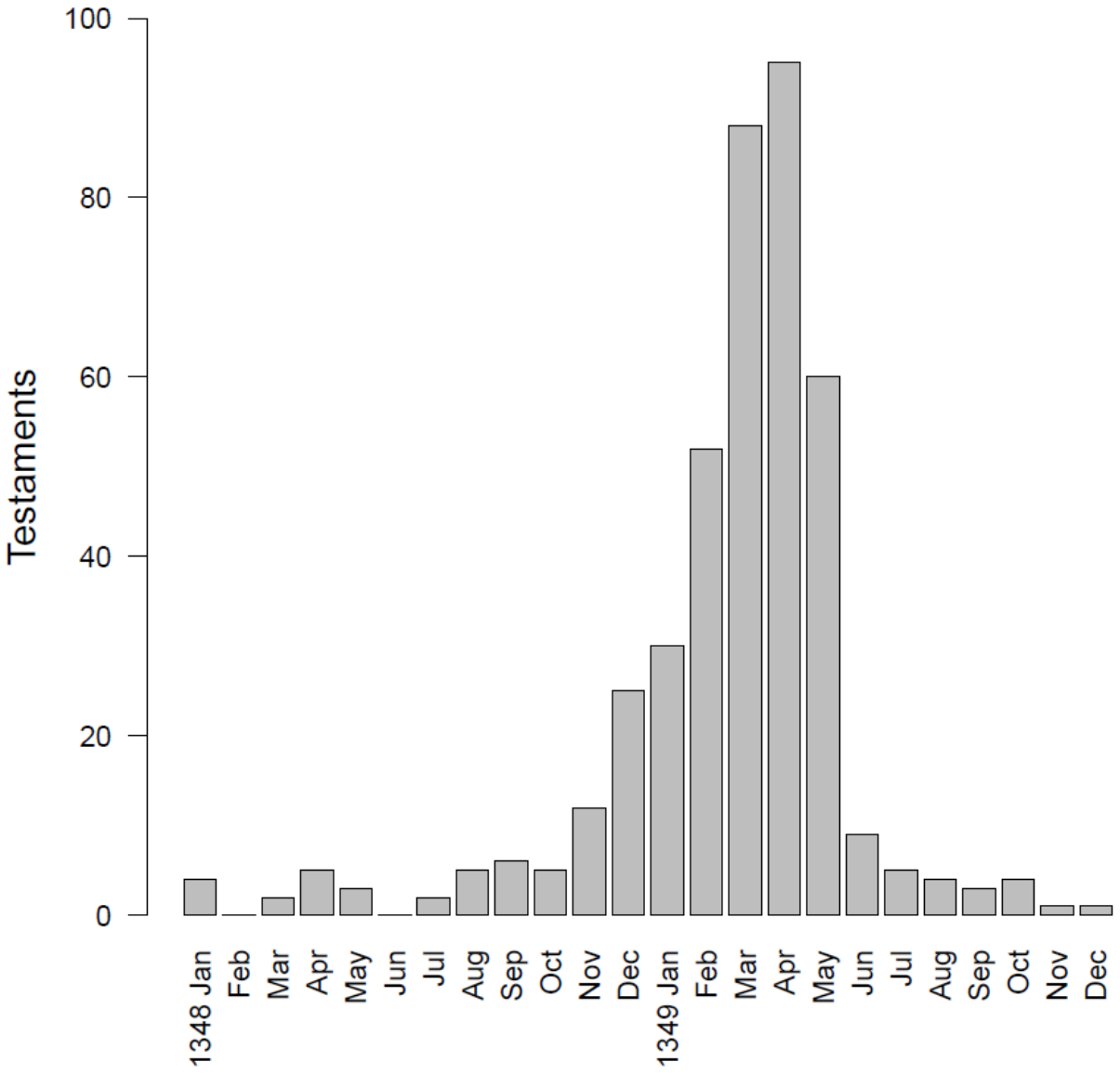

Figure S1 – Monthly numbers of Last Wills and Testaments submitted to the Court of Husting in London in 1348 and 1349, as counted by Cohn, 2003[1]. Deaths reported in this way would have represented a small sample of all deaths in London, but the structure of the epidemic curve is clear nevertheless.

LONDON, 1593. By JOHN NORDEN.

Figure S2 – John Norden's map of London in 1593. East Smithfield is indicated on the far right of the map in red, just north of the river.

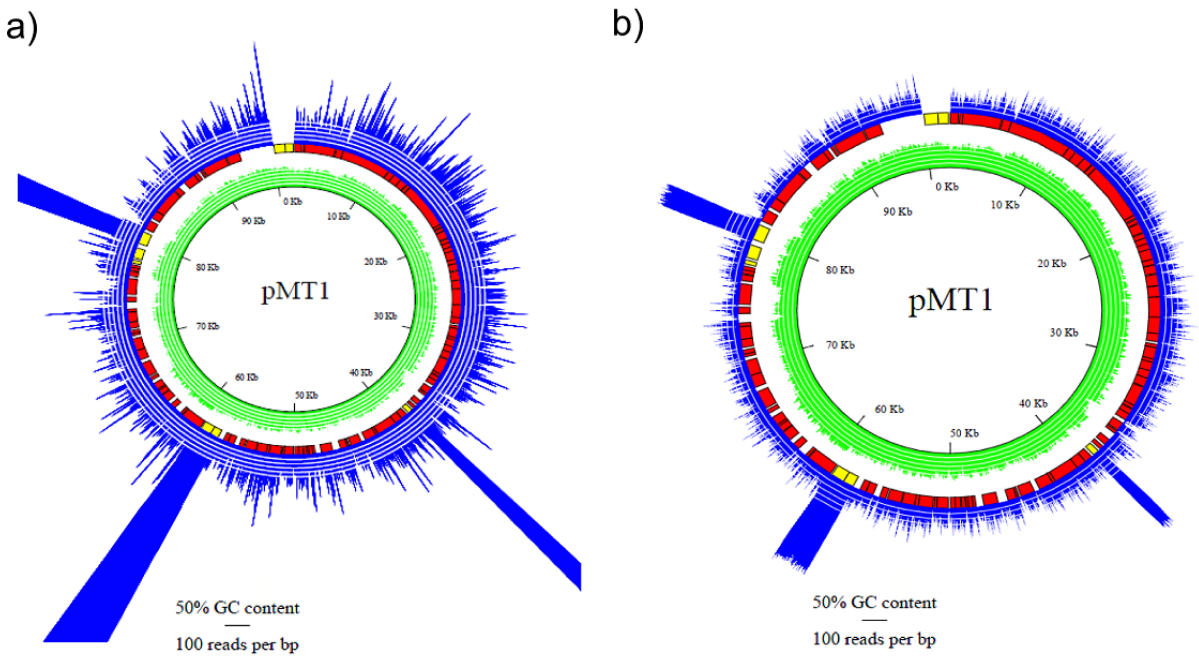

Figure S3 – Effect of duplicate removal via *rmdup*. The coverage plot in a) shows the mapping of all fragments against the pMT1 reference (NC_003134) prior to duplicate removal and before filtering for mapping quality. Coverage is shown in blue, GC frequency for the chromosome in green. Scale lines correspond to 10, 20, 30, 40, 50 fold coverage and to 0.1, 0.2, 0.3, 0.4, 0.5 GC content. Red corresponds to coding regions, yellow to mobile elements. A large tranposase corresponding to positions 1 – 1962 was removed, though three additional copies of this transposase in the pMT have contributed to three areas of dense coverage. The coverage plot in b) shows the effect of duplicate removal, which greatly reduces the number of fragments mapping to the pMT reference, most notable in coverage density in the three remaining transposases.

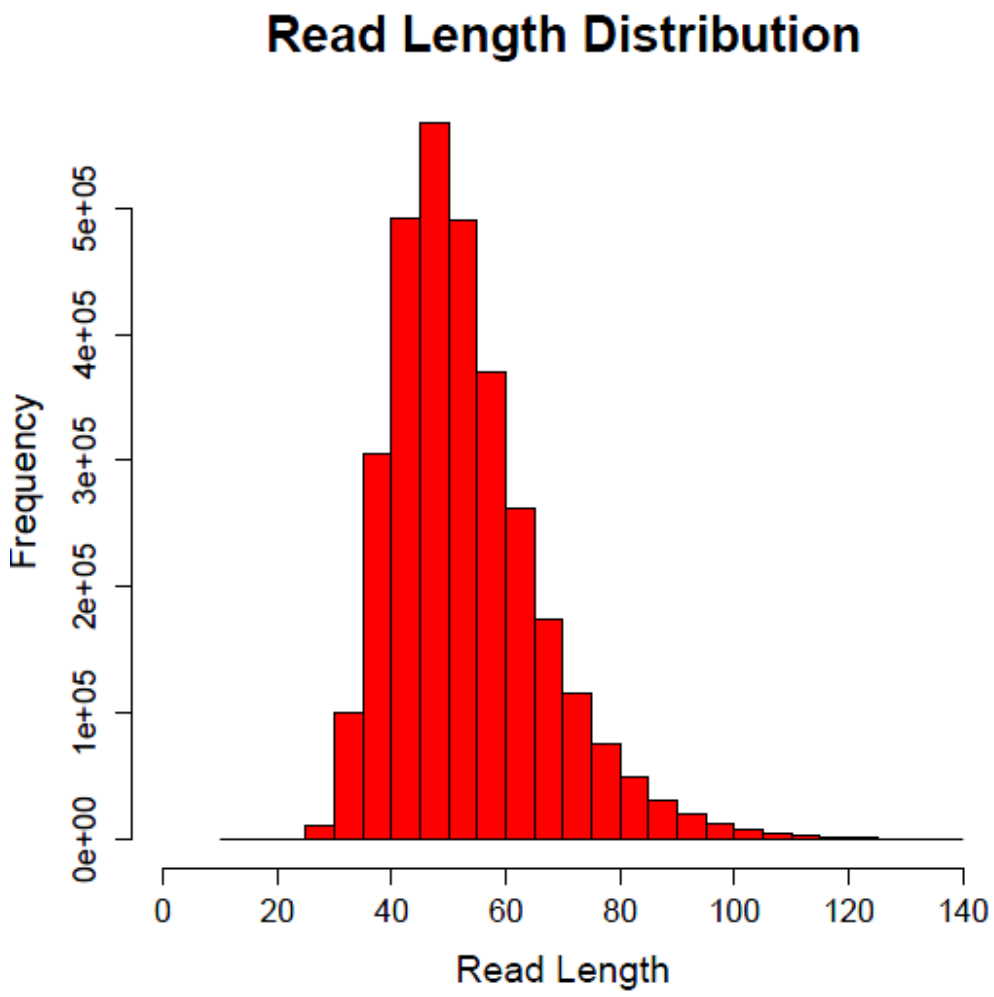

Figure S4 – Fragment size distribution of reads post duplicate removal (*rmdup*) and mapping filtering (Q>30) from a pooled collection of all libraries.

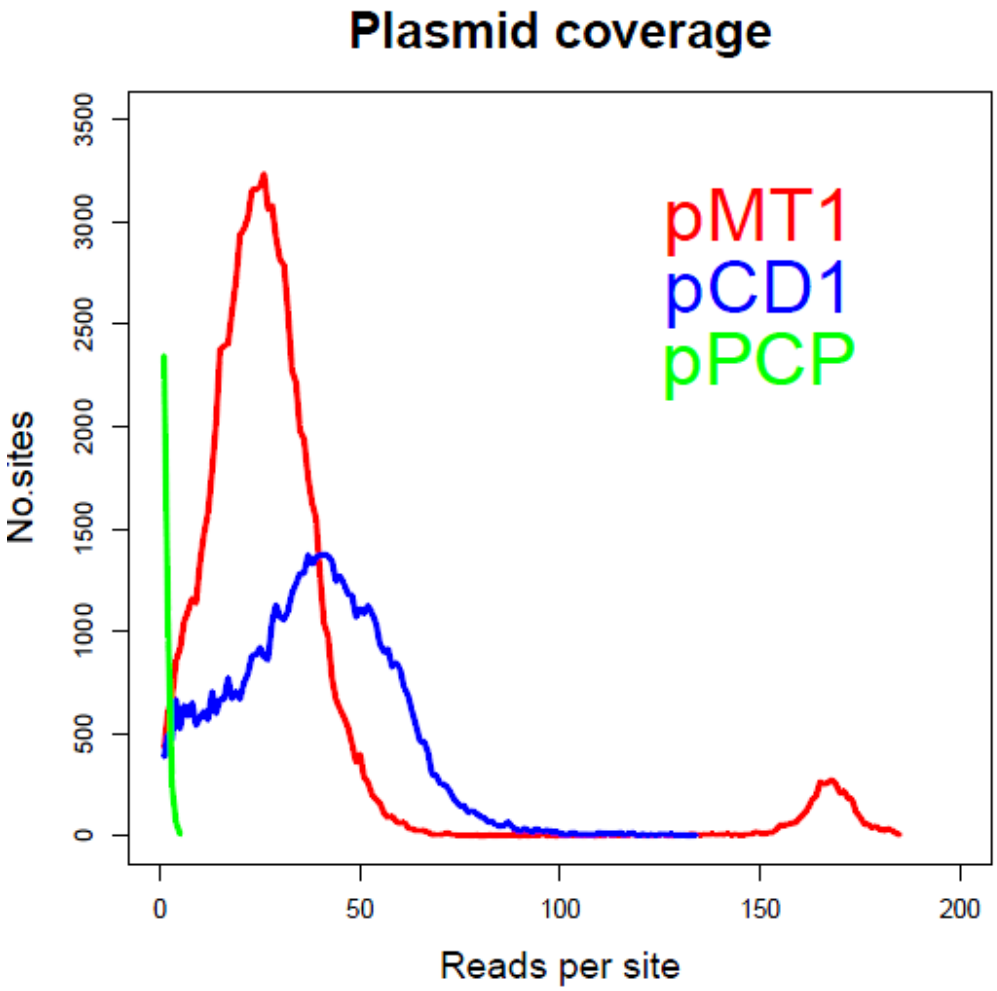

Figure S5 – Coverage plot for East Smithfield *Y. pestis* plasmids

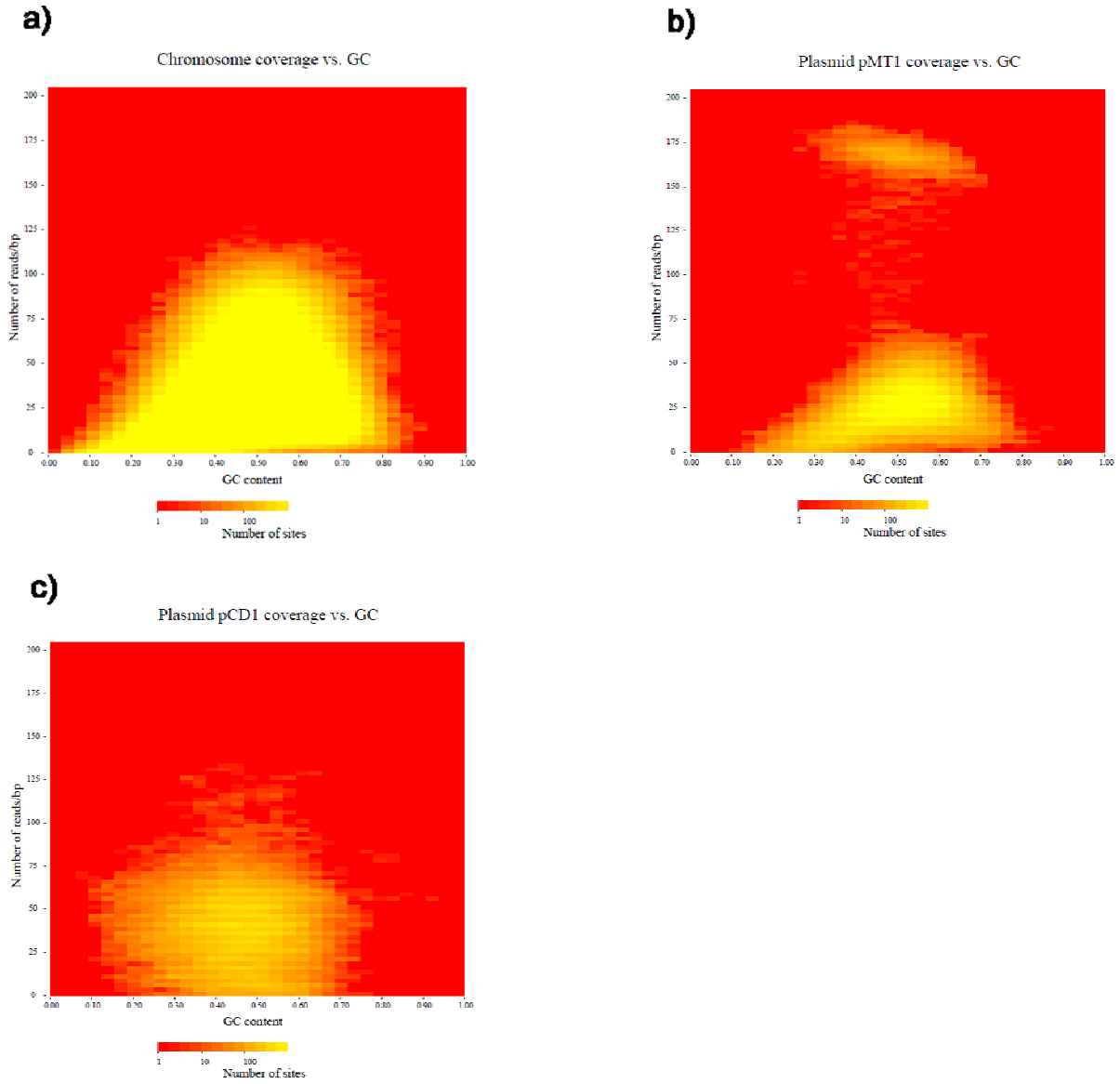

Figure S6 – Heat plots showing relationship of GC content on coverage for a) the chromosome, b) the pMT plasmid, and c) the pCD1 plasmid. The number of sites is shown as a function of the number of reads per site and the GC content averaged over the 30 neighbouring bases. Most sites are covered by fewer than 30 reads and have intermediate GC content. There is a small skew toward less coverage when the GC content is low.

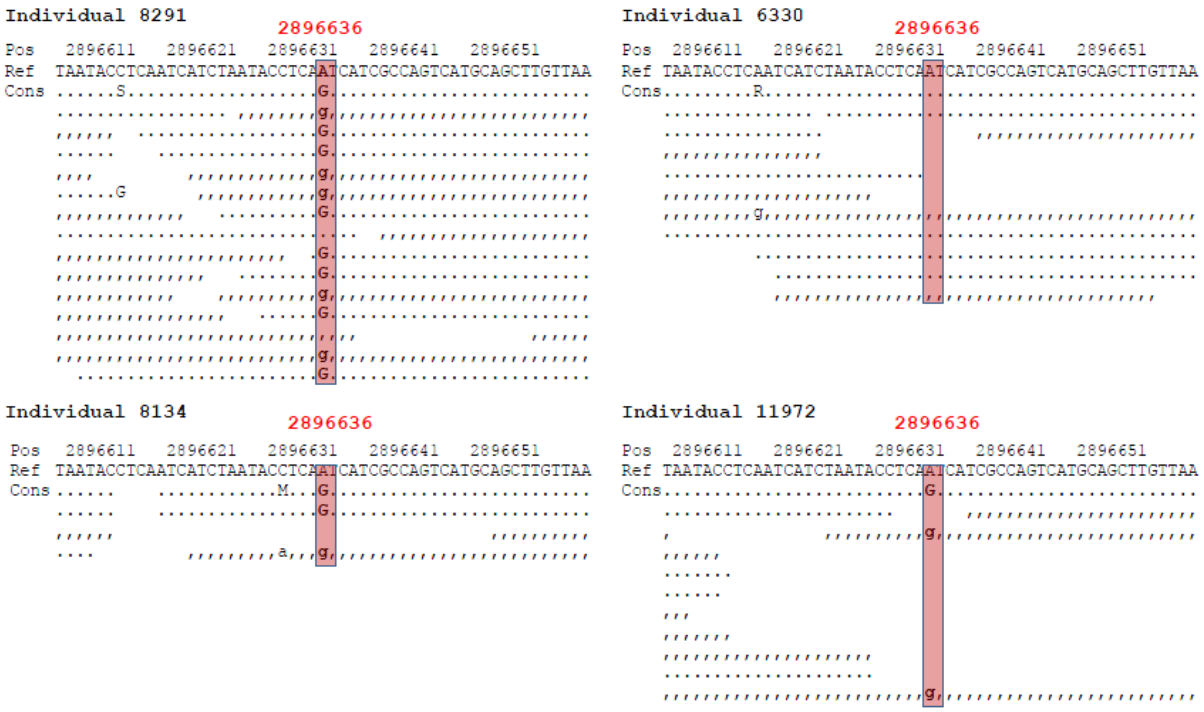

Figure S7 – Alignment showing base frequencies for position 2896636 in all four individuals. The position is polymorphic in individual 8291 and fixed derived in 6330. Individuals 11972 and 8124 show the ancestral state, though coverage is too low to ascertain whether the position is polymorphic with 0.9675 confidence.

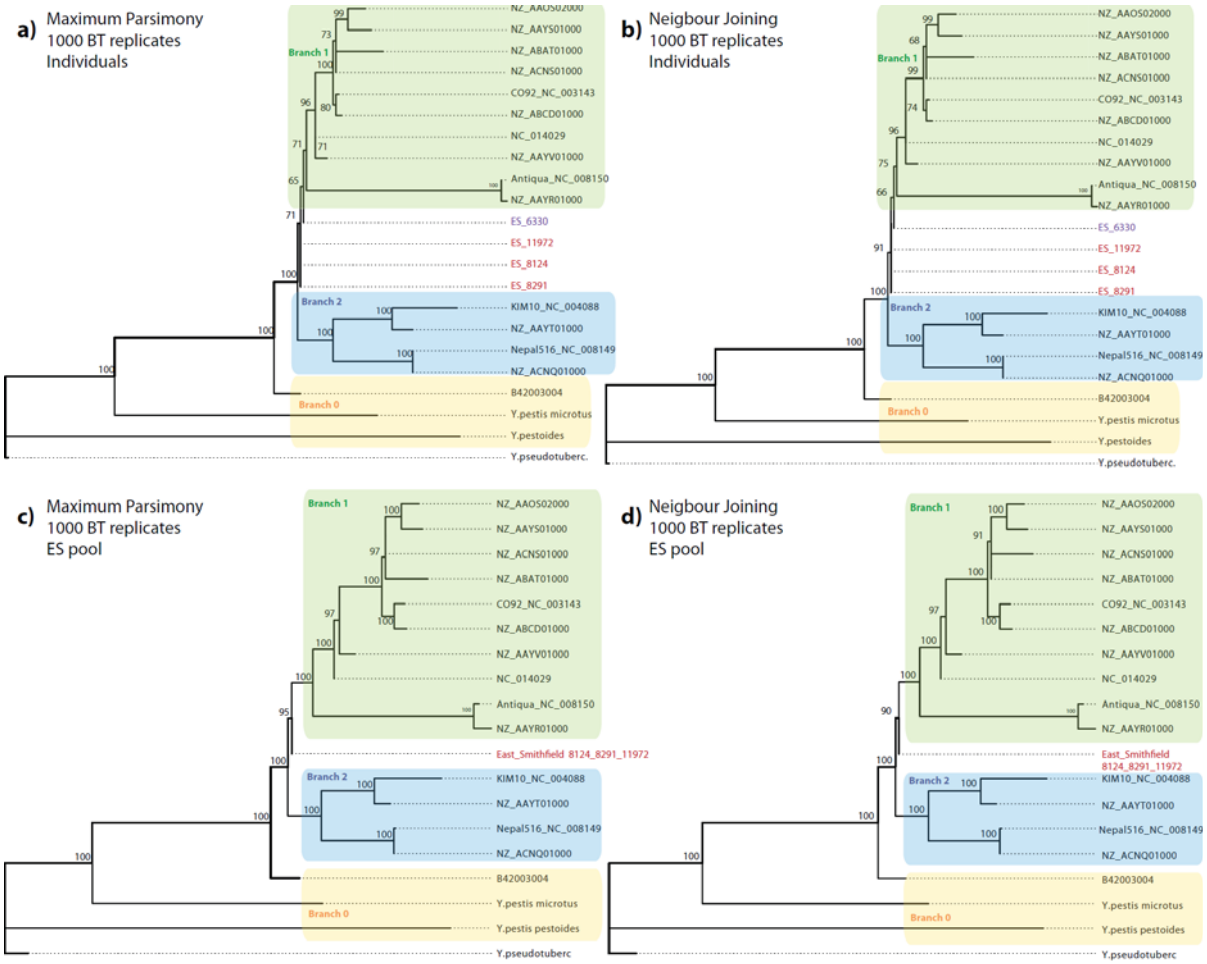

Figure S8 – Phylogenetic trees based on 1694 positions known to be variable within modern *Y. pestis* strains, reconstructed using Maximum Parsimony (a and c) and Neighbour Joining (b and d) algorithms implemented in MEGA 4.1[25]. For a) and b) the East Smithfield individuals are kept separate; for c) and d) individuals 8124, 8291, and 11972 are pooled together.

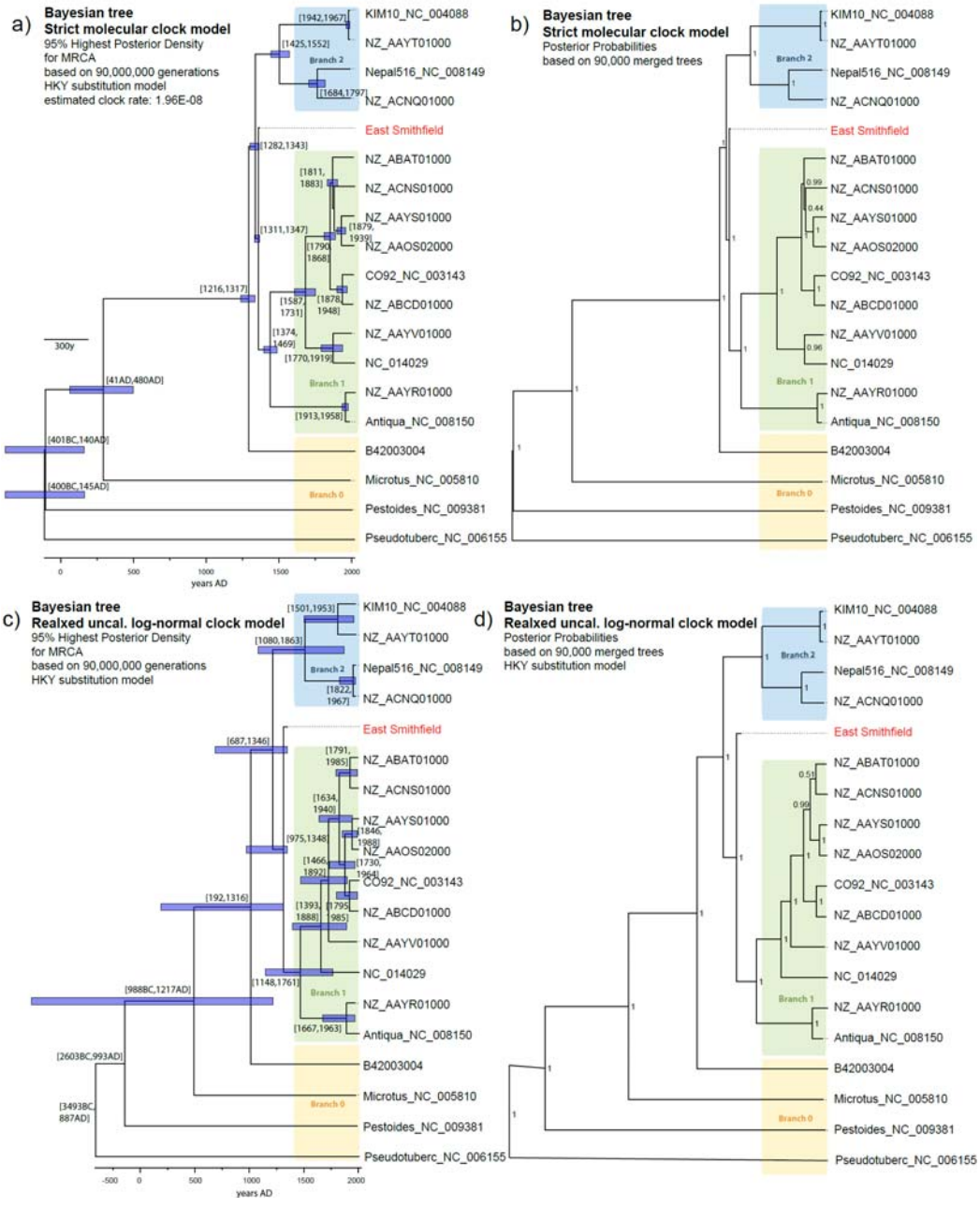

Figure S9 – Phylogenetic trees reconstructed using a Bayesian approach with BEAST and two times 50,000,000 generations. Each tree is derived from a consensus of 90,000 merged trees[26]. Trees a) and b) used a strict molecular clock; trees c) and d) used a relaxed uncorrelated log-normal clock. Trees a) and c) show divergence dates on each node; b) and d) show posterior probabilities for each node. Blue bars show 95% Highest Posterior Density for the divergence date of the different *Y.pestis* strains.

Supplementary tables:

Table S1 – Estimates of East Smithfield sequence coverage.

Table S1a – Fold coverage estimates showing effect of duplicate removal (*rmdup*) and mapping filtering.

fold chromosomal coverage	post duplicate removal	post duplicate removal and mapping filtering Q>30
1 fold	99.25%	93.48%
2 fold	98.62%	92.85%
3 fold	97.98%	92.22%
4 fold	97.29%	91.54%
5 fold	96.55%	90.80%
10 fold	91.50%	85.76%
30 fold	48.30%	42.67%

Table S1b – description of all mapped reads

[Included as a separate file]

S1c – Fold coverage estimates presented as average coverage per site.

individual	Chromosome	pMT1	pCD1	pcp1
ES_11972	6.6	10.1	7.9	0.3
ES_6330	3.5	7.6	5.4	0.1
ES_8124	2.2	7.5	2.7	0
ES_8291	14.8	19.9	19.4	0.2
Pool	28.2	31.2	35.2	0.7
Controls	0	0	0	0

Table S2 – Description of reference-guided contigs showing architectural differences between the East Smithfield genome and the CO92 reference.

Table S2a – Reference-guided contigs that differ from CO92

Contig	Length (bp)	Rearrangement	Found in *Y. pestis*	Found in *Y.pseudotuberculosis*
1530	1149	20 bp indel	7	3
208	2062	insertion of IS100 in CO92	10	4
3036	1540	30kb rearrangement	7	4
3293	2519	9 bp indel	7	0
4326	963	insertion of IS100 in CO92	10	0
4746	4923	32 bp indel	9	1
6149*	9096	231 kbp rearragement	3	4
6289	659	insertion of IS285 in CO92	10	4
7016	10311	insertion of IS100 in CO92	9	4
7127	2526	insertion of IS100 in CO92	5	3

*Intact contig found in Pestoides, Microtus, Angola, B42003004 and *Y. pseudotuberculosis*

Table S2b – Incorrectly reconstructed reference-guided contigs that differ from CO92

Contig	Length (bp)	Description	Found in *Y. pestis*	Found in *Y.pseudotuberculosis*
2935	1477	in repetitive region in CO92	0	0
3181	1704	near phage protein in CO92	0	0
5765	3992	near repeated tRNA val in CO92	0	0
6107	4825	misassembly indel	0	0
6798	603	near pseudogene in C092	0	0
7052	5267	misassembly indel	0	0
7405	8587	misassembly indel	0	0
7488	4398	misassembly indel	0	0
8057	8059	misassembly indel	0	0
43500	6912	misassembly indel	0	0

Table S3 – nucleotide differences identified between the ancient genome and the CO92 reference genome used for bait design for a) the chromosome, and b) the pCD1 and pMT1 plasmids.

[Included as a separate file]

Table S4 – Known polymorphic positions[11] used for phylogenetic reconstruction.

[Included as a separate file]

Table S5 – Variant positions recorded in the East Smithfield sequences.

[Included as a separate file]

Table S6 – Substitutions of note in the two outgroups most closely related to the ES sequence. a) Differences between Microtus 91001 and the other 18 available *Yersinia* genomes. b) Differences between Microtus 91001 and ES with genetic descriptions. c) Differences between B42003004 and the other 18 available *Yersina* genomes. d) Differences between B42003004 and ES with genetic descriptions.

[Each included as separate files]

References Cited:

1. Cohn, S.K. The Black Death transformed: disease and culture in early Renaissance Europe. London: Arnoldohn (2003).
2. Schuenemann, V.J. *et al*. Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of *Yersinia pestis* from victims of the Black Death. *Proc Natl Acad Sci USA* doi: 10.1073/pnas.1105107108.
3. Cowal, L.., *et al*. The Black Death Cemetery, East Smithfield. London. Great Britain: Museum of London Archaeological Svc (2008).
4. Hawkins, D. The Black Death and new London cemeteries of 1348. *Antiquity* **64**, 637 – 642 (1990).
5. Pääbo S., *et al*. Genetic analyses from ancient DNA. *Annu Rev Genet* **38**,645-679 (2004).
6. Rohland, N., and Hofreiter M. Ancient DNA extraction from Bones and Teeth. *Nat Protoc* **2**, 1756 – 1762 (2007).
7. Meyer, M. and Kircher M. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harb Protoc* **6**; doi:10.1101/pdb.prot5448 (2010).
8. Kircher, M., Sawyer S., and Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Submitted.
9. Briggs, A.W., *et al*. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nuc. Acids Res.***38**, e87 (2009).
10. Hofreiter, M *et al*. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nuc Acids Res* **29**, 4793-4799 (2001).
11. Morelli, G., *et al*. Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity. *Nat Genet* **42**, 1140 – 1143 (2010).
12. Hodges, E., *et al*. Genome-wide *in situ* exon capture for selective resequencing. *Nat Genet* **39**, 1522 – 1527 (2007).
13. Burbano H.A., *et al*. Targeted Investigation of the Neandertal Genome by Array-Based Sequence Capture. *Science* **328**, 723 – 725 (2010).
14. Hodges, E., *et al*. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nature Protoc* **4**, 960 – 974 (2009).
15. Kircher, M., Stenzel U, and Kelso J. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol* **10**: R83 (2009).
16. Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**,1754-60 (2009).
17. Li, R., *et al*. SOAP2: and improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966 – 1967 (2009).
18. Goecks, J., *et al*. Galaxy: a comprehensive approack for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11**, R86+ (2010).

19. Blankenberg, D., *et al*. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol*. Chapter **19**, 1-21 (2010).
20. Zerbino, D. *R.,* Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821–829 (2008).
21. Altschul, S.F., *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).
22. Parkhill J., *et al*. Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **413**, 523 – 527 (2001).
23. Darling A.C., *et al*. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14**, 1394-1403 (2004).
24. Huson, D.H. and Bryant D. Application of Phylogentic Networks in Evolutionary Studies. *Mol. Biol. Evol.* **23**, 254 – 267 (2006).
25. Tamura, K., *et al*. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**, 1596 – 1599 (2007).
26. Drummond, A.J., *et al*. Relaxed Phylogenetics and Dating with Confidence. *PLoS Biol*. **4**, e88 (2006).

.