Supplementary Materials for

Title: *Salmonella enterica* genomes recovered from victims of a major 16th century epidemic in Mexico

Authors: Åshild J. Vågene, Michael G. Campana, Nelly M. Robles García, Christina Warinner, Maria A. Spyrou, Aida Andrades Valtueña, Daniel Huson, Noreen Tuross, Alexander Herbig, Kirsten I. Bos and Johannes Krause

Correspondence to: N.T.: tuross@fas.harvard.edu; A.H.: herbig@shh.mpg.de; K.I.B.: bos@shh.mpg.de; J.K.: krause@shh.mpg.de

This PDF file includes:

Materials and Methods
Figs. S1 to S7

Other Supplementary Materials for this manuscript includes the following:

Additional Data Tables S1-S13

Materials and Methods

S1. Archaeological context

The site of Teposcolula-Yucundaa sits on a mountain ridge in the Mixteca Alta, northwest of the city of Oaxaca, Mexico. Prior to the arrival of the Spanish, the *señorío* of Teposcolula-Yucundaa controlled a large Mixtec territory, and was a tribute subject of the Aztec Triple Alliance (*28*). Following the conquest of Mexico the site became of the focus of Dominican evangelism in the 1530s (*53, 54*) and a large stone church was built on the site by the indigenous peoples. The toll taken by disease is evidenced by a large plaster covered cemetery located in the Grand Plaza (administrative square), which is estimated to contain as many as 800 individuals (*28, 55, 56*). In 1552 the Yucundaa site was abandoned and its people were relocated to a new location 2 km south that exists today as the modern town of San Pedro y San Pablo Teposcolula (*27, 28*).

During 2004-2010, archaeologists led by Dr. Ronald Spores and Dr. Nelly Robles García (INAH) excavated the ancient town of Teposcolula-Yucundaa, including burials encountered in the churchyard and the Grand Plaza. A detailed description of the excavation is found in Spores and Robles García, 2007. Multiple occupation periods were noted, including a Postclassic phase that predates the arrival of the Spanish. In addition, an early Convento phase was associated with the Dominican presence (*57*). Direct radiocarbon dating confirmed the presence of pre-contact individuals in the churchyard, while the majority of the radiocarbon dates from individuals excavated from the Grand Plaza overlap with the arrival of the Spanish (*29*). Multiple simultaneous burials found in the Grand Plaza, as well as disproportional mortality of young adults in the absence of skeletal trauma, suggests a high rate of disease causing death among the indigenous Mixtec peoples (*28*).

S2. DNA extraction and library preparation

Teeth were collected from 29 individuals (one from each individual) excavated at the Grand Plaza (n=24) and churchyard (n=5) cemeteries at Teposcolula-Yucundaa (table S1). Each tooth was sectioned at the cementoenamel junction and a sample was drilled from the crown pulp chamber. Samples were processed according to an established protocol tailored for extracting DNA from archaeological bone (*58*), samples were rotated during the lysis step for at least 16 hours. One extraction blank was added for every ten samples processed per batch and a positive control was included in every batch. DNA extracts were eluted in 100µl of TET (10 mM Tris-Cl, pH 8.0; 1 mM EDTA, pH 8.0; 0.05% Tween-20).

Libraries for screening were generated using 10µl of extract (*59*). One library blank was added for every ten samples processed. All libraries were double-indexed (*60*) using a combination of custom 8nt P5 and P7 Illumina indexes in a 10-cycle PCR reaction. Indexed libraries were further amplified using AccuPrime (Thermo Scientific) enzyme with IS5/IS6 primers to reach a concentration of 1x10^13 copies per reaction. All steps from sampling to

setting up the indexing reactions were carried out in facilities dedicated to ancient DNA work at the University of Tübingen. All libraries were diluted and pooled into an equimolar solution of 10nmol/l and shotgun sequenced on a HiSeq 2500. Sequencing yielded between 1,708,868-4,457,666 paired-end reads for the samples and 208,188-887,974 for the blanks (table S3).

Additionally, an aggregate soil sample consisting of soil taken near three skeletons – two in the Grand Plaza (individuals 2 (not investigated in this study) and 26) and one in the churchyard (individual 32) – was collected for DNA screening. The aggregate soil sample was extracted at Harvard University using the PowerMax Soil Maxi Kit, concentrated via ultracentrifugation (30 kDA) and sheared using a Covaris (430 sec, PeakPower 175.0, Duty Factor 10.0, cycles/burst 200). A library was generated using 10µl of extract and indexed in the modern laboratory facilities at the University of Tübingen, using the same method as above. Shotgun sequencing was carried out on a HiSeq 2500 yielding 12,100,244 reads for the soil sample and 748,026 reads for its associated library blank (table S3).

S3. Screening with MALT

The shotgun data generated for all tooth pulp chamber samples, the soil sample and negative controls were screened for ancient bacterial pathogen DNA using the bioinformatics tool MALT (Megan ALignment Tool) (*26*), a tool specifically designed for the analysis of metagenomic data. MALT is a rapid sequence alignment tool that uses a reference database – in this case one consisting of all available bacterial genomes in NCBI RefSeq (December 2015) allowing for the identification of both pathogenic and non-pathogenic bacteria (*26*). MALT uses a taxonomic binning approach (LCA) to assign reads to their taxonomic node of best fit based on the alignment of each read against every reference genome in the database.

The data from all samples was de-indexed and the EAGER pipeline (*61*) was used to perform adapter clipping and paired-end read merging. Only merged reads were used as input for MALT (version 0.3.6). The MALT run was performed using 95 as the "minimum percent identity" parameter (--minPercentIdentity). The minimum support parameter (--minSupport) was set to 5, i.e. only nodes with a minimum support of 5 reads are kept. BlastN mode and SemiGlobal alignment were applied and a top percent value (--topPercent) of 1 was set. All other parameters were set to default. MALT results were viewed in MEGAN6 (*31*).

Using this approach we identified samples from eight individuals (Tepos_10, Tepos_11, Tepos_14, Tepos_20, Tepos_34, Tepos_35, Tepos_37 and Tepos_38) and one extraction blank (EB2-091013) that contained reads ranging from 5 to 653 assigned to *Salmonella enterica* (table S2). The 14 reads assigned to *S. enterica* in EB2-091013 likely arose from contaminants stemming from the lab or elsewhere. In order to consider a sample positive for *S. enterica* and a candidate for whole-genome capture we required 100 or more reads to be assigned to *S. enterica* by MALT. Three samples met this requirement: Tepos_10, Tepos_14 and Tepos_35, respectively harboring 374, 387 and 653 assigned reads. The majority of reads cluster at the downstream *S. enterica* subsp. *enterica* node, where the reference sequence for all three samples with the highest number of aligned reads was *Salmonella enterica subsp. enterica* serovar

Paratyphi C (NC_012125.1). For samples Tepos_10, Tepos_14 and Tepos_35, 12, 5 and 22 reads were specifically assigned to *S.* Paratyphi C, respectively. These three individuals were excavated from the Grand Plaza epidemic cemetery. All other samples investigated from the Grand Plaza and churchyard cemeteries and the soil sample were negative for *S. enterica* DNA in the MALT screening.

A taxon table of the MALT results for shotgun data from all samples and blanks is shown in table S2. The metagenomic profiles of selected samples are visualized in Figure 2. For Figure 2, The Human Oral Microbiome Database (HOMD) (*62*) was consulted in order to classify the microbial species as 'human oral microbiome' or 'environmental'. However, bacteria that are included in the HOMD, but predominantly occur in the environment were counted as exceptions and were classified as 'environmental'. These species were: *Achromobacter xylosoxidans*, *Acinetobacter baumannii*, *Agrobacterium tumefaciens*, *Burkholderia cepacia*, *Comamonas testosterone*, *Kytococcus sedentarius*, *Mesorhizobium loti*, *Proteus mirabilis*, *Pseudomonas aeruginosa*, *Pseudomonas fluorescens*, *Pseudomonas stutzeri*, *Ralstonia pickettii*, *Rhodobacter capsulatus*, *Sanguibacter keddieii* and *Variovorax paradoxus*.

Visual inspection of the reads assigned within *S. enterica* by MALT for samples Tepos_10, Tepos_14 and Tepos_35 revealed mismatches consistent with an ancient DNA damage pattern. In order to confirm the visually observed deamination pattern the merged reads used as MALT input were mapped against the *S.* Paratyphi C RKS4594 reference (NC_012125.1) using the Burrows-Wheeler-Aligner (BWA) (*63*) with parameters adjusted to accommodate deaminated bases (-l 16, -n 0.01, -q 37). Mapping statistics are shown in table S3. mapDamage (*64*) plots were subsequently generated from the mapping reads, where the first base on the 5-prime end was deaminated in 22.86%, 21.62% and 17.21% of reads respectively for Tepos_10, Tepos_14 and Tepos_35 (fig. S1; table S3).

S4. Array design

To further verify the finding of *S. enterica* DNA in samples Tepos_10, Tepos_14 and Tepos_35 via MALT based screening and to attempt whole-genome reconstruction, we designed a set of probes for whole-genome enrichment of *S. enterica* for array-based hybridization capture.

Probes were designed based on 112 publicly available reference sequences (67 chromosomes/assemblies and 45 plasmids; see table S4 for details). These reference genomes were selected based on modern strain diversity within the species *Salmonella enterica*. The probes were designed to be 60bp long and have a tiling density of 7bp across the template. This was achieved by generating two different sets of probes with 15bp tiling density each, which differ from each other by a coordinate offset of 7bp (versions A and B). Low complexity repetitive and duplicate probes were excluded from the final probe set. This produced 928,395 and 928,078 unique probes for versions A and B respectively. By randomly sampling probes each probe set was enlarged to 968,000 probes, as this is the maximum number of probes that can be included on an Agilent One-million feature array.

S5. Array capture

Concentrated libraries were made using 30-40µl extract from samples Tepos_10, Tepos_14 and Tepos_35. Prior to library preparation the DNA extracts were pre-treated with USER enzyme (New England BioLabs), which contains Uracil DNA glycosylase (UDG) and endonuclease VIII (endoVIII). UDG removes uracil residues located in 5-prime and 3-prime overhangs in ancient DNA, creating an abasic site that is cleaved and removed by endoVIII. This is done to avoid the incorporation of incorrect bases during amplification (*65*). Allowing for more stringent mapping parameters to be used in the reconstruction of full genomes and excludes erroneous nucleotide substitutions from interfering with downstream data analyses. Indexing and further amplification was done as described above using the enzyme Herculase II Fusion DNA Polymerase (Agilent).

UDG and non-UDG libraries for samples Tepos_10, Tepos_14 and Tepos_35 were amplified to make two pools of 20µg, where 75% of each pool was dedicated to equimolar quantities of the UDG treated libraries and the remaining 25% to equimolar quantities of the non-UDG libraries used for screening. Each pool was serially captured using versions A and B of the array; together we term these the 'MALT-positives' array.

The five non-UDG screening libraries made from the pre-contact churchyard samples (see table S1), one sample (Tepos_27) from the Grand Plaza cemetery negative for *S. enterica* in the MALT screening and the soil sample were amplified and pooled in equimolar amounts to make a 20µg pool. This pool was serially captured on a version A array; we term this the 'MALT-negatives' array. A fourth 10µg equimolar pool was made consisting of negative controls carried along during extraction and library preparation. The 'negative controls' pool was captured in a single round on a version A array.

Array capture was performed according to an established method (*32*). The eluate from the first round of capture performed for all arrays was quantified on the qPCR using IS5/IS6 primers and amplified using Herculase II Fusion DNA Polymerase. The 'MALT-positives' and 'MALT-negatives' array eluate was further amplified up to 17µg and serially captured on identical arrays to those used in the first round. The eluted product was quantified as above and re-amplified using IS5/IS6 primers. The product from the 'MALT-positives' and 'MALT-negatives' arrays were diluted and pooled in equimolar amounts to create a 10nmol/l sequencing pool. The pool was paired-end sequenced (2x75bp cycles) on a NextSeq 500. The capture product for the 'negative controls' array was sequenced separately on part of a HiSeq 4000 paired-end run (2x75bp cycles).

S6. Read processing, mapping and ascertainment of phylogenetic positioning

The sequenced paired-end data were de-indexed using bcl2fastq (Illumina; http://support.illumina.com/downloads/bcl2fastq-conversion-software-v217.html) and further processed using the EAGER pipeline (*61*) to clip adapters, merge paired-end reads, map the data using

BWA (*63*), remove duplicates, execute mapDamage (*64*) and carry out SNP calling with the GATK UnifiedGenotyper (*66*). Only merged reads were used in all mapping based analyses.

All data from the 'MALT-positives', 'MALT-negatives' and 'negative controls' arrays were mapped against the *S.* Paratyphi C RKS4594 reference (NC_012125.1). BWA mapping parameters were adjusted depending on whether the library was pre-treated with UDG or not. UDG treated libraries were mapped with more stringent parameters (BWA parameters: -l 32; -n 0.1; -q 37) than non-UDG libraries (BWA parameters: -l 16; -n 0.01; -q 37). Mapping results show all libraries captured on the 'MALT-negatives' array contained 1,864 or fewer unique mapping reads (table S5). Whereas non-UDG libraries captured on the 'MALT-positives' array contain between 289,468-1,430,852 and UDG treated libraries between 2,056,326-5,852,171 unique mapping reads (table S5). Data from the non-UDG libraries yielded average coverages of 3-, 8- and 19-fold and the UDG treated data yielded average coverages of 22-, 27- and 77-fold for Tepos_10, Tepos_14 and Tepos_35 respectively.

All 'MALT-negatives' and 'negative controls' array capture libraries were negative for *S. enterica* DNA (table S5). The EB2-091013 extraction blank that had 14 reads assigned to *S. enterica* in the MALT screening had 244 unique mapping reads after capture with a read duplication factor of 6.7. Deamination patterns generated with mapDamage (*64*) for the non-UDG capture data for Tepos_10, Tepos_14 and Tepos_35 yielded 20.52%, 29.06% and 20.38% deamination on the first base on the 5-prime ends of the reads. These numbers differ from the deamination values yielded by the shotgun data, likely because they are based on a higher number of reads than previously estimated using the shotgun reads (fig. S1; tables S3, S5).

Artificial read data (100bp reads with 1bp tiling density) was generated from a subset of the genomes used in the array design consisting of 23 fully scaffolded or assembled *S. enterica* genomes using an in-house script. The artificial read data was mapped against the *S.* Paratyphi C RKS4594 reference (NC_012125.1) using stringent UDG mapping parameters.

The dataset used for downstream analyses consisted of the genomes reconstructed from the UDG treated capture data for samples Tepos_10, Tepos_14 and Tepos_35, in addition to the 23 genomes reconstructed from the artificial read data mapped against *S.* Paratyphi C RKS4594. SNP calling was carried out for all dataset genomes using the 'EMIT_ALL_SITES' function, providing a call for all variant or non-variant bases in the *vcf* file output.

We used an in-house tool (MultiVCFanalyzer) to collate homozygous SNPs (90% of reads covering a position must be in agreement) called at a minimum of 5X coverage against the *S.* Paratyphi C RKS4594 reference for the three captured UDG treated positive samples and the artificial reference genome dataset. The collated SNP alignment was used to generate a Neighbor-joining tree in MEGA6 (*67*) (fig. S2). This tree shows that the three captured genomes cluster with *S.* Paratyphi C with a high bootstrap support of 100, confirming the initial taxonomic indication provided by MALT. Oddly, the Tepos_10 genome has a much longer branch length compared to the other ancient genomes (Supplementary materials S8).

In order to exclude the possibility of a reference bias in the ascertainment of the phylogenetic positioning, we mapped the sample libraries from the 'MALT-positives' array

using the approach outlined above, but this time against the *S.* Typhi CT18 reference (NC_003198.1). All sample libraries captured on the 'MALT-positives' array show decrease in percentage of the reference genome covered and a decrease in average coverage from *S.* Paratyphi C (table S5) to *S.* Typhi (table S6). SNP calling was also repeated for the genome dataset with the parameters described above. A neighbor-joining tree was constructed, confirming the positioning of the ancient genomes with *S.* Paratyphi C (fig. S3).

Array capture efficiency was calculated using the non-UDG shotgun and capture data. 4bp were trimmed from each end of all the non-UDG treated reads in order to remove the majority of deaminated DNA bases. All data were subsequently mapped to the *S.* Paratyphi C RKS4594 reference genome (NC_012125.1) using stringent UDG mapping parameters. Based on the quality filtered mapping reads before duplicate removal, array capture efficiency was estimated to 564-, 585- and 457-fold increase for Tepos_10, Tepos_14 and Tepos_35 respectively.

S7. Human DNA analysis

The non-UDG shotgun data for samples and negative controls were mapped to the human genome (hg19) using non-UDG parameters. Endogenous human DNA ranged from 0.005 to 27% for the samples and 0.1 to 35% for the blanks (table S7). Damage patterns were estimated using mapDamage (*64*). The human DNA in the blanks does not have any damage pattern and is likely stemming from human DNA contamination introduced from reagents and plastic ware during extraction and/or library preparation. The characteristic ancient DNA damage pattern is present in 25 of the 29 human tooth pulp chamber samples (table S7). The remaining four samples (Tepos_9, Tepos_12, Tepos_13 and Tepos_19) do not exhibit a damage pattern, likely due to the low amount of preserved human DNA. It may also be that no human DNA was preserved in these four samples and the DNA that is present stems from modern contaminant sources.

S8. SNP typing and phylogenetic analysis

SNP calls generated for all 26 genomes in the dataset were compared in parallel using an in-house Java tool (MultiVCFanalyzer). MultiVCFanalyzer outputs a multi-genome SNP-alignment with an entry for all genomes where at least one variant position is called within the dataset. Homozygous positions were called where 90% or more of the reads covering a position were in agreement, a minimum of 5 reads were covering the position and the position's GATK quality score was a minimum of 30. Homozygous calls were also made in cases where GATK had called a heterozygous position, but the above requirements were still met. Non-variant positions meeting the above criteria were called as the reference base and positions with missing data or those that did not meet the above requirements were inserted as an 'N' in the SNP-alignment. Positions called in repetitive regions, phage-related regions, recombination-related regions or regions prone to cross-mapping from other organisms were excluded. These regions were identified based on two genome annotation (*gff*) files for the *S.* Paratyphi C RKS4595 reference genome (NC_012125.1). The two *gff* files can be found respectively through the

current version (ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/Salmonella_enterica/) and the archived older version (ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/Salmonella_enterica_serovar_Paratyphi_C_RKS4594_uid59063/) of the FTP archive of publicly available bacterial genomes. Regions excluded were identified in the *gff* files as: mobile elements, phages or phage-related, transposase, resolvase, rRNA, tRNA, repeat-region, insertion sequence or as a recombination protein.

203,287 variant positions were called from the total dataset. The three ancient genomes had a respective number of SNP calls of 609, 655 and 679 for Tepos_10, Tepos_14 and Tepos_35. The multi-genome SNP alignment consisting of homozygous calls for the dataset was used to construct a Neighbor-joining tree in MEGA6 (*67*) (fig. S3). Complete deletion was used, restricting phylogenetic analysis to the core genome.

Heterozygous positions in the three ancient genomes were investigated due to the extensively long branch observed for the Tepos_10 genome compared to the two other ancient genomes (figs. S2-S3). Heterozygous positions were called in MultiVCFanalyzer using the parameters described above, where additionally all positions with a SNP allele frequency between 10-90% were typed as heterozygous. 2,838, 326 and 375 heterozygous positions were called respectively for Tepos_10, Tepos_14 and Tepos_35. The distribution of SNP allele frequencies for each of the three ancient samples is shown in figure S4. The tail of the distribution of heterozygous positions in the Tepos_10 genome infringes upon the threshold of 90% set for calling homozygous positions. This indicates that the homozygous calls for this genome cannot be considered to be reliable. In light of this, Tepos_10 was excluded from further analyses.

A new SNP alignment was generated for the dataset excluding the Tepos_10 genome, comprising 203,257 variant positions. Neighbour Joining, Maximum Parsimony and Maximum Likelihood trees were generated, using complete deletion, based on homozygous positions from the dataset, excluding the Tepos_10 genome (see Figs. 2, S5, S6). All trees show a bootstrap support of 100 for the phylogenetic positioning of the ancient genomes with *S.* Paratyphi C. The two ancient genomes, Tepos_14 and Tepos_35, exhibit branch shortening in comparison to the modern strain.

S9. SNP analysis of protein coding genes

A SNP table of variant positions occurring in at least one strain within the dataset (excluding the Tepos_10 genome) was generated using MultiVCFanalyzer applying the same parameters as described above for generating the SNP alignment. The SNP table was annotated using the *gff* file for the *S.* Paratyphi C RKS4594 genome (NC_012125.1) located in the archived version of the FTP archive.

The annotated SNP table of variant positions was used as input for the bioinformatics tool snpEff (*68*), which predicts the amino acid changes and effects of the variant SNP positions on genes. MultiVCFanalyzer was subsequently used to collate the output from snpEff with the SNP

table dataset. In total 207,521 homozygous SNPs are present in at least one genome in the dataset. 681 SNP positions are present in one or both of the two ancient genomes, Tepos_14 and Tepos_35. 339 of these are non-synonymous SNPs (nsSNPs): with 326 nsSNPs leading to an amino-acid codon change, 9 cause the loss of a stop-codon (STOP_LOST) and 4 are stop mutations (STOP_GAINED). 210 are synonymous changes and 13 are considered as non-coding because they occur in RNA related genes.

131 variant positions are unique to one or both of the ancient strains. Tepos_14 has two unique nsSNPs occurring in the *yfbU* and *yhck* genes. The *yfbU* gene is annotated as encoding a 'hypothetical protein'. Whilst the *yhck* gene or *nanR* is involved in sialic acid utilization and regulates the *nan* operon by repressing it in the absence of sialic acid (*69*). Only one SNP is unique to Tepos_35 and it is intergenic. A table comprising all SNPs occurring in one or both of the ancient strains is shown in table S8.

44 genes contain two or more variant SNPs (table S8). Two of these genes are of particular note with regard to the ancient strains. In the *ydiD* gene two nsSNPS and one sSNP occur, the sSNP and one of the nsSNPs are unique to the ancient strains. The *ydiD* gene encodes for a putative acyl-CoA synthetase involved in the breaking down of fatty acids (*34*). In the *tsr* gene there are two SNPs, one non-synonymous and one synonymous, which are uniquely shared by both the ancient strains within the dataset. *Tsr* codes for a methyl-accepting chemotaxis protein involved in serine sensing, steering the bacterium towards host-sources of nitrate (*35*). In *S.* Typhimurium, the *tsr* gene is associated with enhanced rates of infection in the mouse intestine (*35, 70*).

Nine 'STOP_LOST' mutations were identified in the ancient genomes in comparison to the *S.* Paratyphi C reference (table S8). However, these nine pseudogenes present in the *S.* Paratyphi C reference are active genes or missing in the other 22 modern genomes. This suggests that the modern *S.* Paratyphi C RKS4594 strain is the odd one out based on the diversity of *Salmonella enterica* strains included in this analysis. Three of the four 'STOP_GAINED' mutations are unique to the two ancient strains within the dataset. Two affect genes coding for hypothetical proteins (SPC_0289 and SPC_4426), the third affects the *rbsA* gene, coding for part of an ATP-binding cassette, which is involved in nutrient uptake (*71*).

Six homoplastic SNP positions were detected (table S9A). Only one is non-synonymous and is located in the *phsC* gene. Within the dataset this homoplasy is shared with strains *S.* Typhimurium 08-1736 and *S.* Dublin CT-02021853. The *phsC* gene is part of the *phsABC* operon in *S. enterica* that encodes a thiosulfate reductase, which metabolizes, or reduces, thiosulfate into hydrogen sulfide (*72*) providing an alternate energy source in anaerobic environments. Thiosulfate is readily available in the mammalian-gut, and is suggested to support bacterial growth of *S. enterica* during gut colonization and potentially contributing to pathogenesis (*73*). Additionally, the ancient genomes share two nsSNPs, one each in the *phsA* and *phsB* genes which make up the other two genes in the *phsABC* operon with all other strains included in the analysis (excluding *S. arizonae* where it is absent). This may be some indication that the ancient *S.* Paratyphi C strains metabolized thiosulfate differently than the modern *S.*

Paratyphi C RKS4594 reference strain. Three SNP positions were also found to be tri-allelic within the analyzed dataset (table S9B). Only one tri-allelic SNP position is non-synonymous and occurs in the *bah* gene, which putatively codes for acetyl esterase (http://www.uniprot.org/uniprot/Q57HI6).

S10. Indel analysis

Insertions and deletions (indels) in the ancient genomes were identified through two different approaches. Deletions in the ancient genomes larger than 700bp in comparison to the *S.* Paratyphi C RKS4594 reference were identified based on visual inspection using the IGV browser (*74*). The ancient data was mapped to the reference using UDG parameters with a mapping quality (-q) of 0. Thus, reads that map equally well at more than one point in the genome are kept in the alignment. Only one region absent in both of the ancient genomes (Tepos_14 and Tepos_35) was identified. This region is ~1,784bp in length, spanning positions 1,355,468 to 1,357,252 in the *S.* Paratyphi C RKS4594 genome (NC_012125.1). This region contains the prophage related gene SPC_1297, which encodes for the terminase large subunit (TerL). TerL is part of the machinery that prophages use to translocate DNA and package it into empty capsid heads (*75*). This absent region was a part of the array probe design and its absence is not due to a capture bias. An additional mapping was performed where the less stringent (non-UDG) parameters (with –q 0) were used to verify the absence of the region in the case that it should be present, but with a high number of mismatches to the reference. However, it was still not detected as part of the ancient genomes.

In order to investigate regions present in the ancient genomes that are absent in the *S.* Paratyphi C RKS4594 genome, the ancient UDG treated genome data was mapped to four concatenated reference pairs, where one genome in every pair was the *S.* Paratyphi C RKS4594 (NC_012125.1) and the other was one of the following: *S.* Choleraesuis SC-B67 (NC_006905.1), *S.* Paratyphi A ATCC-9150 (NC_006511.1), *S.* Paratyphi B SGSC-4150 (NC_010102.1) or *S.* Typhi CT18 (NC_003198.1). Mapping was done with standard UDG mapping parameters, and because the parameter for mapping quality (-q) was set to 37, reads that map equally well at more than one position across the two concatenated references were discarded, meaning only reads unique to either the *S.* Paratyphi C or its paired genomes were mapped. Through this method several genes were determined to be present in the ancient genome capture data that were not present in the *S.* Paratyphi C RKS4594 strain. Mapping to the concatenated pair with *S.* Typhi CT18 yielded a number of regions, including a portion of the Salmonella Pathogenicity Island 7 (SPI-7) containing five genes, which are absent or degenerate in the *S.* Paratyphi C RKS4594 reference (NC_012125.1). Notably, SPI-7 also contains the *viaB*-locus, an important *Salmonella* virulence factor (*37*). An overview of the regions identified is shown in table S10. These extra regions were captured due to the *S. enterica* genome diversity included in the array probe design.

The five additional SPI-7 genes (*pilS*, *pilT*, *pilU*, *pilV*, *rci*) present in our ancient genomes, which are absent or degenerate (*rci*) in the Paratyphi C RKS4594 strain, form part of the *pil* operon that encodes the type IVB pili (*37*) (table S10). The *pil* operon is concluded with a

shufflon that encodes the Rci recombinase. Rci recombinase switches rapidly between the two different 19bp C termini of the PilV protein (PilV1 and PilV2), causing the IVB pili to be malformed or not synthesized (*36, 37*). This activity causes the bacteria to self-adhere (*36*). Modern *S.* Paratyphi C strains have been found to lack these five genes, such as in Paratyphi C RKS4594, or to carry an intact version where the Rci protein cannot act upon its target sites due to the insertion of an additional base (A) in each of the, now 20bp, C termini of *pilV* (*37, 38*). Thus, causing *pilV* to be locked by mutation. The *pilV* gene in the two ancient genomes carries the 19bp inverted repeats present in the *S.* Typhi CT18 genome and all other *S.* Typhi strains included in our modern dataset. When the ancient genomes are mapped to the intact *pil* operon sequence of *S.* Paratyphi C SGSC 2712 (AY249242.1) we observe a SNP deletion in both the 20bp inverted repeats of the *pilV* gene. This indicates that the *rci* shufflon in our ancient strains may have had full functionality. It is thought that the ability for the bacteria to self-adhere is an early step in *S.* Typhi pathogenesis and that it aids in the invasion of epithelial cells in the human gut (*36, 37*). Whilst the absence/degenerative state or inactive versions of the *pilV* and *rci* genes carried by modern *S.* Paratyphi C strains has been suggested to account for their inability to cause epidemic scale outbreaks (*37, 38*). Finding the active versions of these genes in our ancient genomes may indicate an increased capacity for these strains to cause an epidemic scale outbreak, potentially supported by the fact that they were isolated from the Grand Plaza epidemic cemetery. However, *S.* Paratyphi A lacks SPI-7 completely and is one of the major causes of human enteric fever today (*37, 76*).

S11. Presence/absence analysis of virulence factors

A set of 43 effector genes identified within the *Salmonella enterica* subsp. *enterica* serovars presented in a recent paper by Connor *et al.* (*77*) were used to make a concatenated reference file (table S11). All genomes (including plasmids) in our dataset were mapped to this reference using BWA with the following parameters –l 32, -n 0.1 and –q 0. When the mapping quality is reduced to zero, reads that map equally well to two or more genes will be kept and randomly mapped to one of the positions. The bedtools bioinformatics suite (*78*) was used to generate the percentage of each gene covered at least 1-fold in each genome in the dataset. This information was plotted using the ggplot2 package (*79*) in R (*80*) and is shown in figure S7. Notably, in comparison to the modern *S.* Paratyphi C RKS4594 reference, the *pipB2* gene has a higher percentage of the gene covered in the two ancient genomes with ~92 and 94% covered, respectively, in Tepos_14 and Tepos_35 and only ~71% covered in the *S.* Paratyphi C RKS4594 genome. *pipB2* is an effector protein secreted via the type III secretion system and it is involved in regulating the recruitment of kinesin-1 (*81*).

S12. Plasmid analysis

S. Paratyphi C strains harbor a virulence plasmid called pSPCV, which was included in the design of our capture probes. To investigate the presence of this plasmid in relation to our captured ancient strains we mapped our UDG treated data with the corresponding parameters to

the pSPCV reference sequence (NC_012124.1). pSPCV is present in the two ancient strains with a coverage of 56X and 178X, respectively for Tepos_14 and Tepos_35 (table S12). The pSPCV plasmid is also present at 44X coverage in the Tepos_10 genome indicating that this sample is indeed positive for *S*. Paratyphi C (table S12), despite the chromosome containing too many heterozygous positions to allow SNP analysis. This plasmid is estimated to be present in 1-2 copy numbers per bacterial cell (*82*), which may account for the near double coverage of the pSPCV in comparison to the rest of the genome for all three samples.

SNP analysis of the pSPCV plasmids present in our two ancient strains (Tepos_14 and Tepos_35) was carried out in comparison to three virulence plasmids present in other *S. enterica* subsp. *enterica* strains that share a high degree of sequence similarity with pSPCV (*39, 83*). Artificial sequencing data (100bp reads with 1bp tiling density) was generated for these three plasmids and mapped to the pSPCV reference. The additional plasmids comprise pSCV50 (NC_006855.1) present in *S.* Choleraesuis SC-B67, pKDSC50 (NC_002638.1) present in *S.* Choleraesuis RF-1 and pSLT (NC_003277.1) present in *S.* Typhimurium LT2.

411 homozygous SNP positions were called from our dataset using the same parameters as outlined previously in supplementary materials S8. Ten SNPs were shared by or unique to the ancient pSPCV plasmids, they are listed in table S13. Two non-synonymous SNPs were identified to be specific to one or both of the ancient pSPCV plasmids. One nsSNP specific to Tepos_14 occurs in the *pefD* gene that is part of the fimbrial *pef* operon involved in bacterial adherence to the intestinal epithelium, where *pefD* specifically encodes for the periplasmic chaperone. The second nsSNP occurs in the replication related *parA* gene carried by both Tepos_14 and Tepos_35 (*82, 83*).

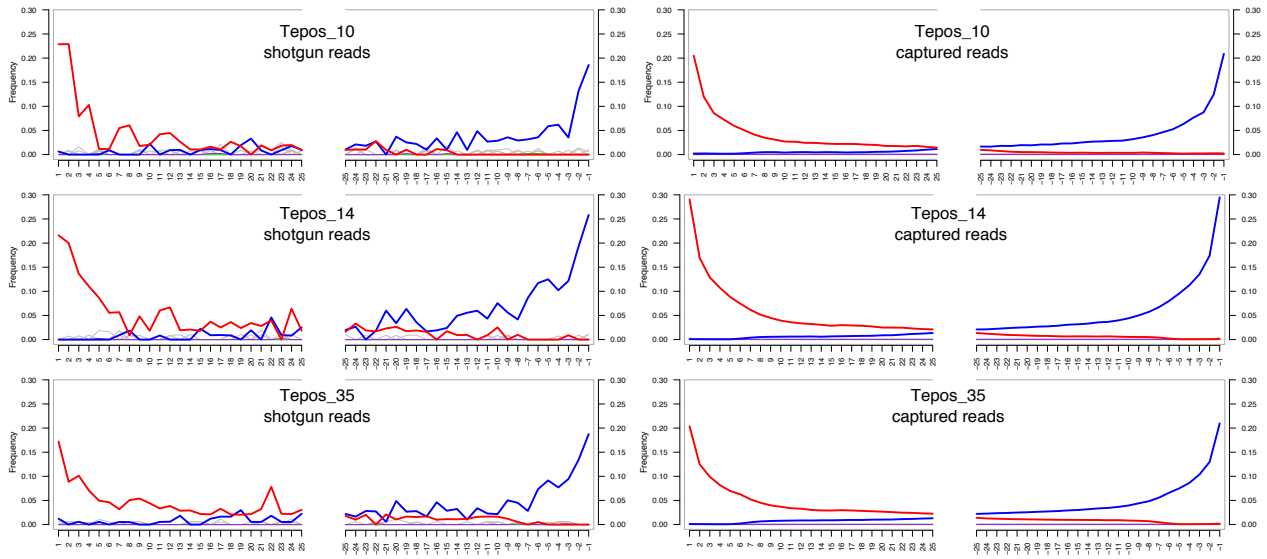

Fig. S1. Comparison of damage plots generated from shotgun data versus capture data for the three ancient *S.* Paratyphi C genomes.

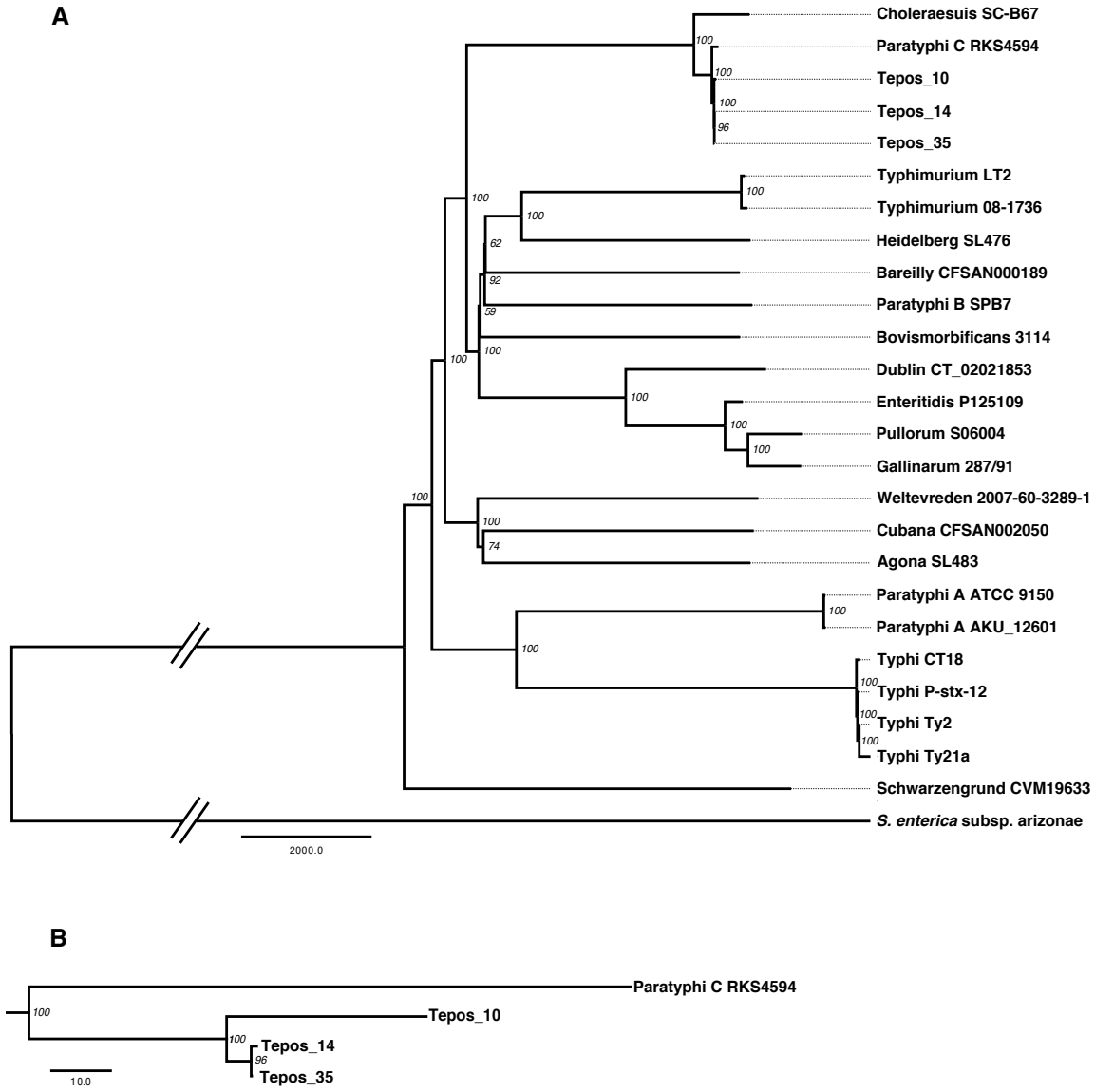

Fig. S2. *S.* Paratyphi C reference based Neighbour-joining *S. enterica* phylogeny including the Tepos_10 genome. A) Neighbour-joining tree constructed using the dataset, including the Tepos_10 genome, when mapped to the *S.* Paratyphi C RKS4594 reference (NC_012125.1). The tree was constructed by excluding all missing and ambiguous data, using 1000 bootstrap replicates and is based on 67835 positions. The three ancient genomes cluster with *S.* Paratyphi C, with high bootstrap support. The tree was constructed using MEGA6 *(67)*. B) An enlarged view of the *S.* Paratyphi C clade. The long branch length of the Tepos_10 genome is clearly shown in comparison to the other two ancient genomes.

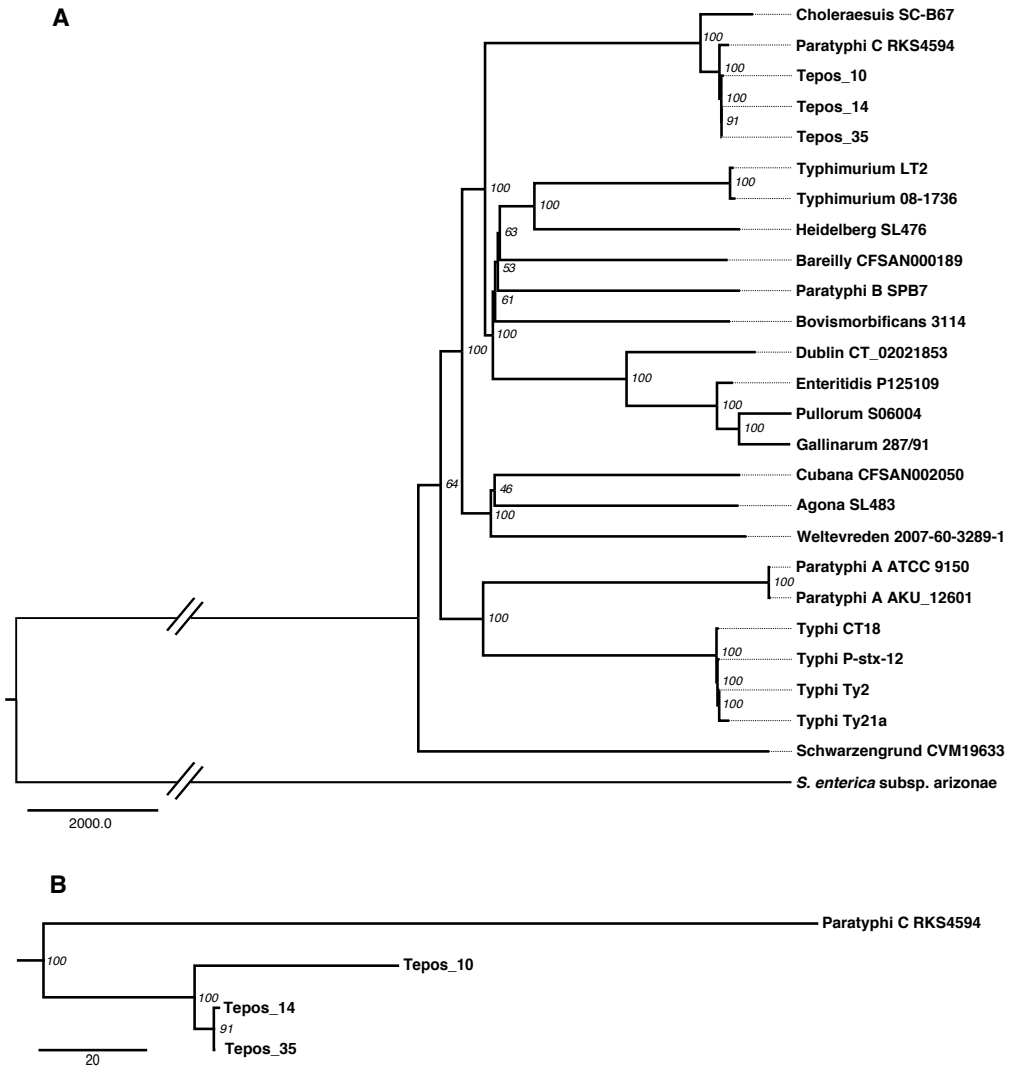

Fig. S3. *S.* Typhi reference based Neighbour-joining *S.* enterica phylogeny including the Tepos_10 genome. A) Neighbour-joining tree constructed using the dataset, including the Tepos_10 genome, mapped to the *S.* Typhi CT18 genome (NC_003198.1). The tree was constructed by excluding all missing and ambiguous data, using 1000 bootstrap replicates and is based on 62656 positions. The three ancient genomes cluster with *S.* Paratyphi C, with high bootstrap support. The tree was constructed using MEGA6 *(67)*. B) An enlarged view of the *S.* Paratyphi C clade. The long branch length of the Tepos_10 genome is clearly shown in comparison to the other two ancient genomes.

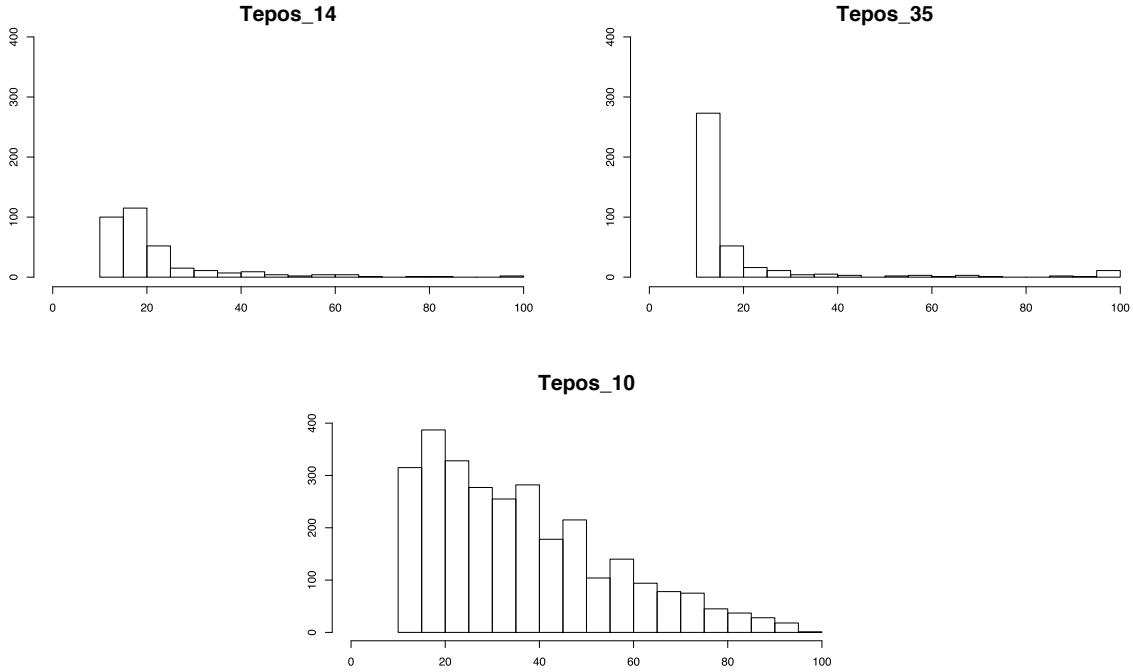

Fig. S4. Histograms of SNP allele frequency distributions for the three ancient *S.* Paratyphi C genomes. The *x*-axis shows the SNP allele frequencies as a percentage. All variants where the SNP allele frequency is higher than 10% and lower than 100% are shown.

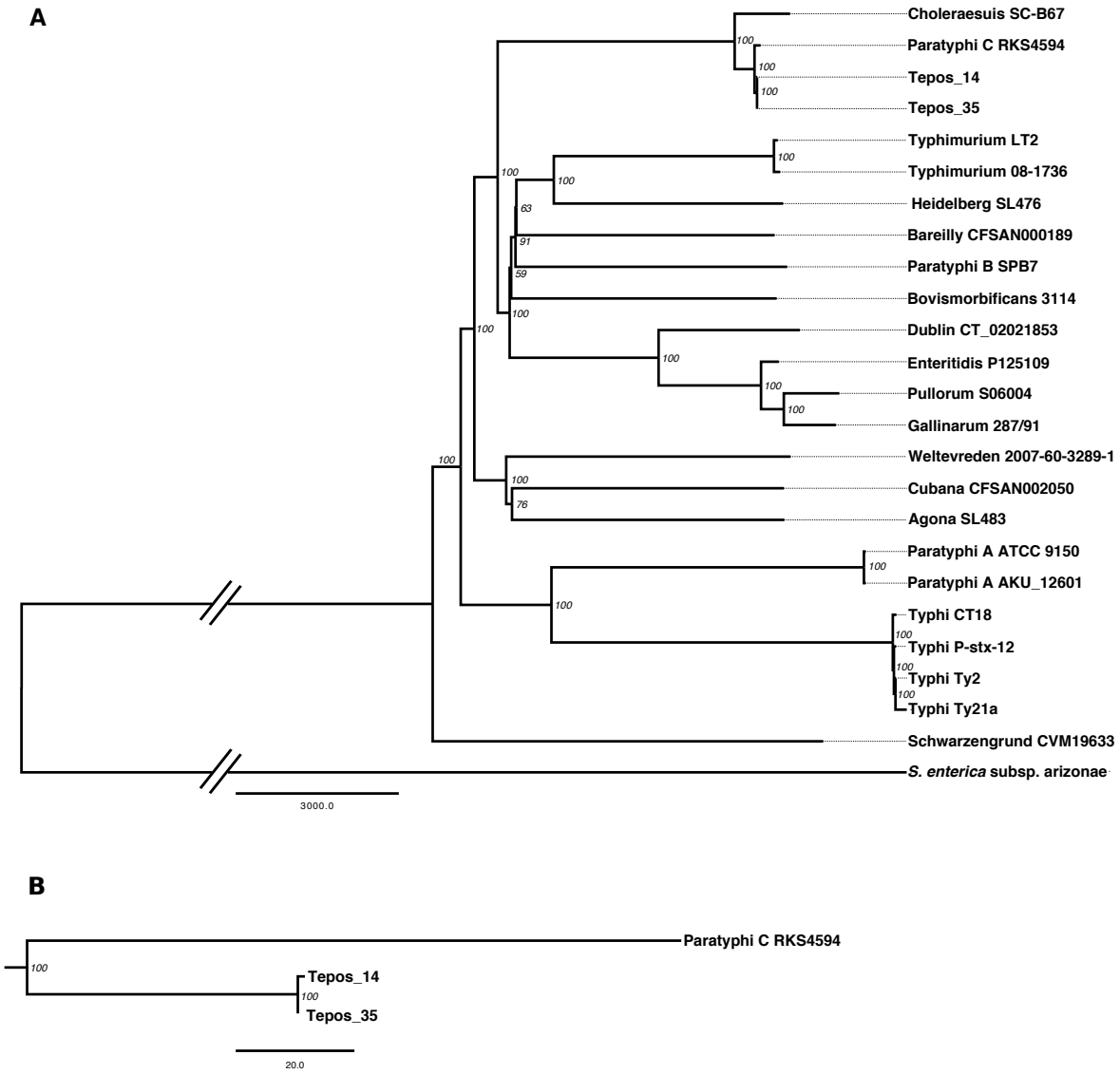

Fig. S5. Neighbour-joining *S. enterica* phylogeny. A) Neighbour-joining tree constructed using the dataset when mapped to the *S.* Paratyphi C RKS4594 reference (NC_012125.1). The tree was constructed by excluding all missing and ambiguous data, using 1000 bootstrap replicates and is based on 81110 positions. The two ancient genomes cluster with *S.* Paratyphi C, with a bootstrap support of 100%. The tree was constructed using MEGA6 *(67)*. B) An enlarged view of the *S.* Paratyphi C clade.

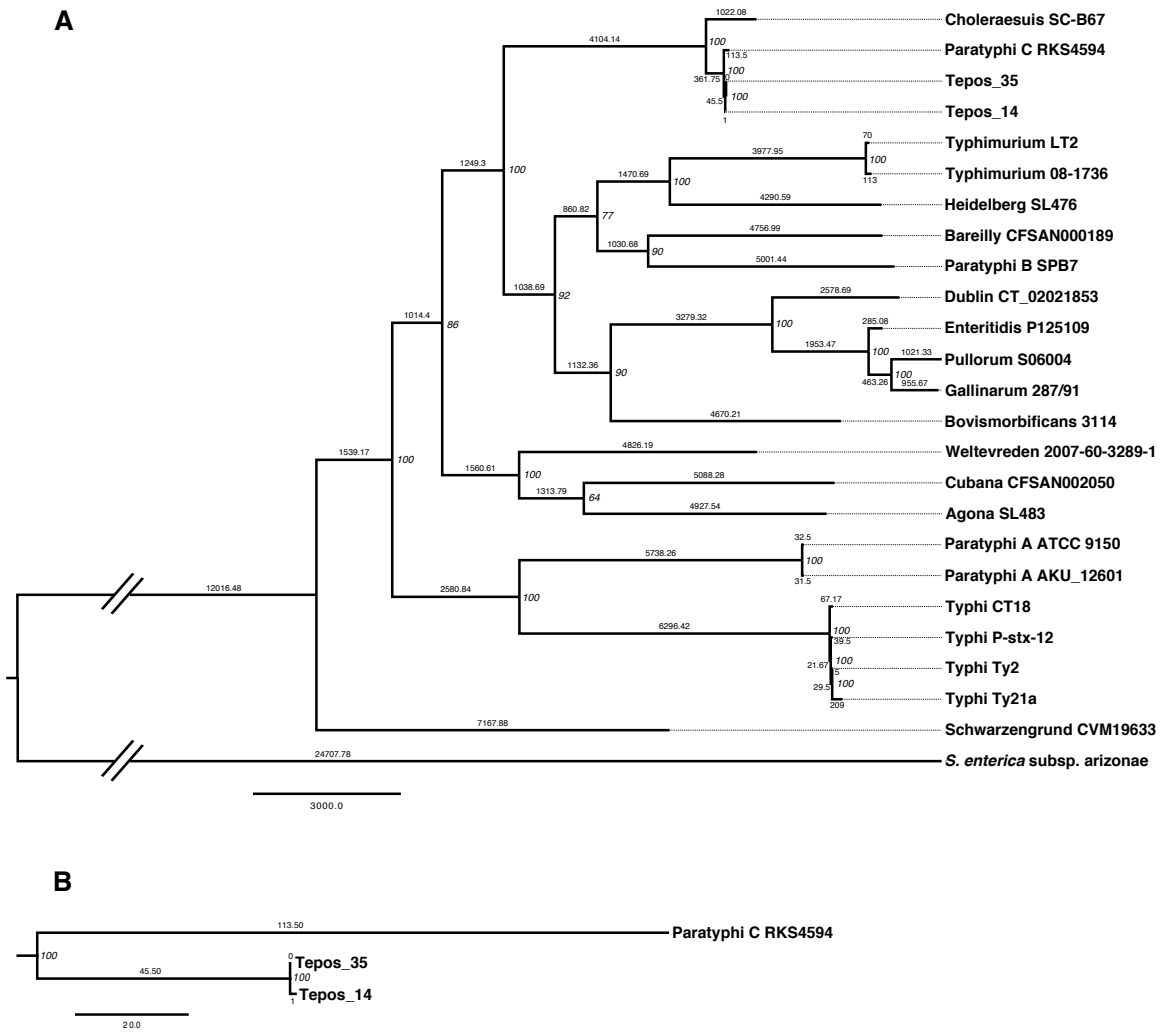

Fig. S6. Maximum Parsimony *S. enterica* phylogeny. A) Maximum Parsimony tree constructed using the dataset when mapped to the *S.* Paratyphi C RKS4594 reference (NC_012125.1). The tree was constructed by excluding all missing and ambiguous data, using 1000 bootstrap replicates and is based on 81110 positions. The two ancient genomes cluster with *S.* Paratyphi C, with a bootstrap support of 100%. The tree was constructed using MEGA6 *(67)*. Branch lengths are displayed. B) An enlarged view of the *S.* Paratyphi C clade.

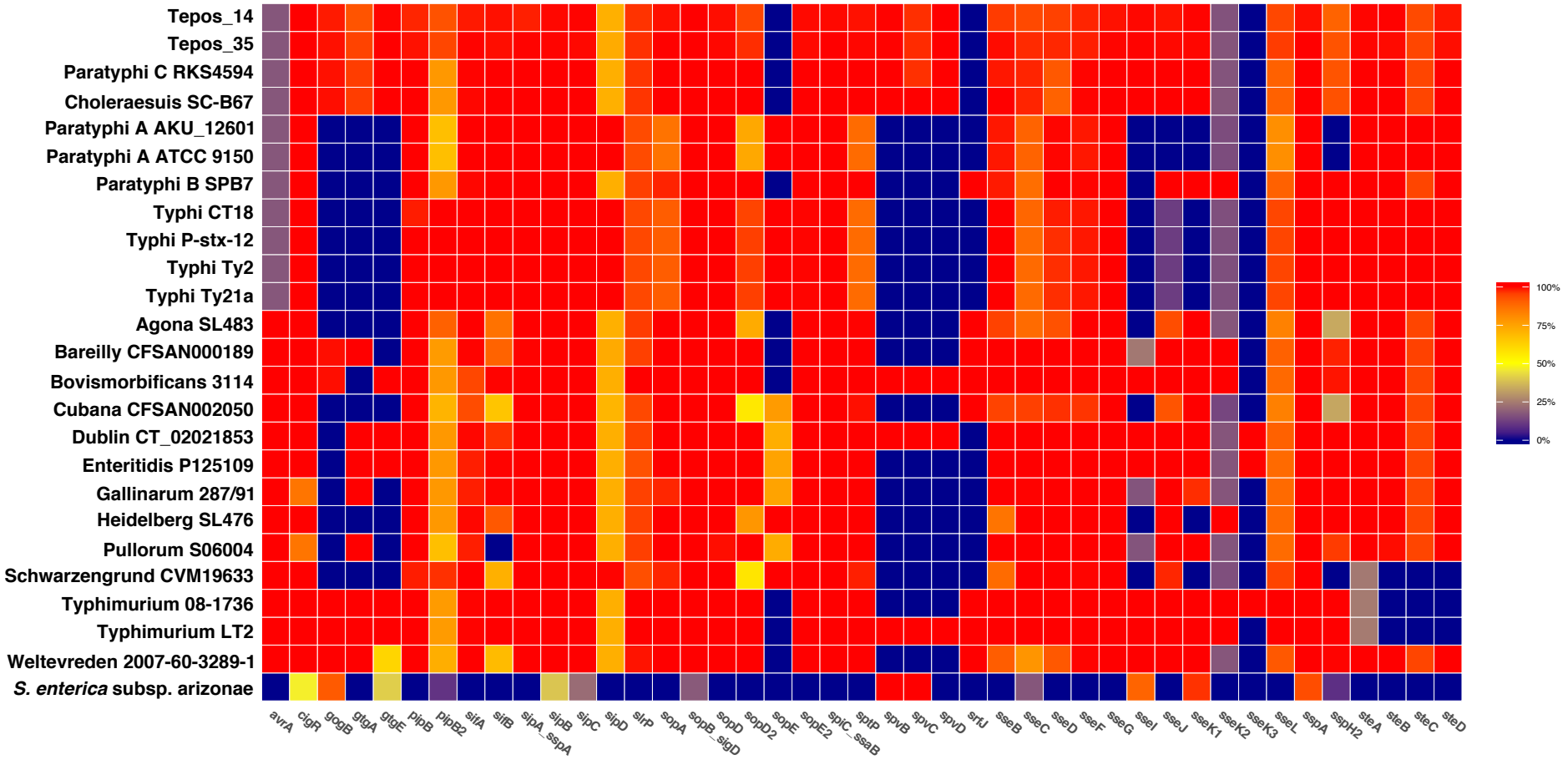

Fig. S7. Heatmap visualizing the presence/absence of effector protein coding genes. The color scale signifies the percent of each gene covered at least 1-fold.