Supplementary Materials for

Millennia of genomic stability within the invasive Para C Lineage of *Salmonella enterica*

Authors: Zhemin Zhou, Inge Lundstrøm, Alicia Tran-Dien, Sebastián Duchêne, Nabil-Fareed Alikhan, Martin J. Sergeant, Gemma Langridge, Anna K. Fotakis, Satheesh Nair, Hans K. Stenøien, Stian S. Hamre, Sherwood Casjens, Axel Christophersen, Christopher Quince, Nicholas R. Thomson, Francois-Xavier Weill, Simon Y. W. Ho, M. Thomas P. Gilbert, Mark Achtman

Correspondence to: M.A.: m.achtman@warwick.ac.uk; M.T.P.G.: tgilbert@snm.ku.dk; Z.Z.: zhemin.zhou@warwick.ac.uk

Additional Acknowledgements.: We thank A. Murat Eren for assistance with implementation of Anvi'o in EnteroBase, and Siavash Mirarab and Tandy Warnow for advice on ASTRID and running ASTRAL-II. We thank Birgitte Skar for permission to sample the skeletal material and the Danish National High-Throughput DNA Sequencing Centre for technical assistance with data generation. We thank Monica H. Green, Timothy P. Newfield and Francois Balloux for helpful comments on the manuscript.

This PDF file includes:

Materials and Methods
Figs. S1 to S9
Tables S1 to S14

Other Supplementary Materials for this manuscript includes the following:

Databases S1 to S12 as zipped archives at http://wrap.warwick.ac.uk/85593 :

- Database S1. The pan-genome of the Para C Lineage based on 6,665 single copy genes from 220 genomes.
- Database S2. Genomic islands and plasmids.
- Database S3. SNPs in the core genome of the Para C Lineage plus Birkenhead.
- Database S4. A species tree from 2,964 representative genomes from Salmonella enterica subsp. I estimated with ASTRID on the basis of 3,002 core gene trees.
- Database S5. A RAxML maximum likelihood phylogeny (Fig. 2A) for 2,964 representative genomes from Salmonella enterica subsp. I based on 3,002 core genes concatenated into a 2.8 Mbp sequence.

- Database S6. Pseudogenes associated with sub-lineages and clades of the Para C Lineage.
- Database S7. Maximum likelihood phylogeny and inferred geographical origins based on 49,610 SNPs in the non-recombinant, non-repetitive core genome of the Para C Lineage plus Birkenhead.
- Database S8. Contigs of 11 MAGs that likely represent unique genomes. Additional details on MAGs are in tables S4 - S6.
- Database S9. Metadata for 2,964 representative genomes in S. enterica subsp. I.
- Database S10. Detailed metadata for 220 strains in the Para C Lineage plus 2 Birkenhead strains.
- Database S11. Raw Maximum Clade Credibility trees from BEAST inferences.
- Database S12. Sequences, original gene names and annotations of the reference sequences for 6,665 pan genes in the Para C Lineage.

Materials and Methods

Properties of SK152 (summarized in table S2)

SK152 was excavated from grave 523E in a cemetery on the north side of a church in Trondheim, Norway (Fig. 1A) in June, 1985. The skeleton rested on a wooden plank (symbolic coffin) (Fig. 1B). According to the daybook written in the field, the grave was filled with “grey, sandy loam, wood chips” which likely represents a mixture of minerogenic soil and waste from wood-building activities in the nearby, densely built-up area. The grave itself was covered with a wooden chip layer derived from later building activity in the area. The stratigraphical information indicates that the environment surrounding SK152 was anoxic, acidic and waterlogged, which explains why the skeleton was very well preserved.

Grave 523E was part of sub-phase FN7/level II within the extension of the cemetery, corresponding to a building phase when the cemetery had recently been extended northwards, and covered older/earlier buildings from sub-phase FN6. The construction of FN6 began ca. 1150. Sub-phase FNT was initially dated to “mid 1200s” with the aid of ceramics, coins, ^{14}C-dating and archaeological lead artefacts (Reed in (*35*) p. 192), but this estimate was later revised to 1175-1225 CE on the basis of additional dendrographic dating (Fig. 23 in (*36*)).

Preservation, age and sex of SK152. The overall surface preservation of bone elements was very good, grade 1 according to the seven-category grading system defined by McKinley (*37*)). The skeleton was over 90% complete, with almost all bone elements present. Only minimal localized areas of bone damage due to post mortem disturbance were noted, and a small degree of fragmentation. Discoloration due to burial taphonomy was also noted, but this was characteristic of the whole assemblage and did not impact on the visual inspection of bone surface.

Skeleton 152 was assigned to the young adult category (19-24 years) (*38*) on the basis of dental eruption and dental wear pattern, degeneration of the pelvis and fusion of epiphyses. Standard techniques for osteological sex determination (*39*) indicated that SK152 was female, and this conclusion was confirmed by the ratio of sequences aligning to the Y chromosome relative to the total number of sequences aligning to both sex chromosomes (*40*).

Accelerator Mass Spectrometry (AMS) dating of SK152. AMS dating was performed in two independent institutes, DirectAMS and Oxford Radiocarbon Accelerator Unit, yielding dates in radiocarbon years BP (Before Present - AD1950) using the ^{14}C half-life of 5,568 years. Isotopic fractionation was corrected using the δ^{13}C values measured on the AMS. The quoted δ^{13}C values were measured independently on a stable isotope mass spectrometer (to ±0:3 per mil relative to VPDB). For details of the chemical pre-treatment, target preparation and AMS measurement see (*41, 42*). Fig. S2 shows calculated calendar age ranges generated by the OxCal computer program (v4.2) of C. Bronk Ramsey, using the `IntCal13' dataset. These calculations support dates of 994-1052 with 53% likelihood and 1081-1152 with 43%.

Isotope analysis δ^{18}O of SK152. The oxygen composition in enamel apatite carbonates from the first (M1) and third (M3) molars yielded M1 and M3 $\delta^{18}O_{C/SMOW}$ values of 21.13‰ and 24.72‰ on the V-SMOW scale, respectively. The $\delta^{18}O_{C/SMOW}$ data was converted to $\delta^{18}O_{water}$ (oxygen composition in water/precipitation) by using equation 6 from (*43*) as modified by (*44*), to yield $\delta^{18}O_{water}$ values of -15‰ to -16.5‰ for M1 and -9.3‰ ±1‰ for M3 (*45*). The values from the first molar suggest that SK152 was born inland in northern Scandinavia or in the northwestern regions of Russia, whereas the values from the third molar indicate that she arrived in Trondheim in her childhood years (*45*) (table S2).

Metagenomic sequencing of samples from SK152

All molecular work including pre-library amplification was conducted in dedicated aDNA clean laboratory facilities at the Centre for GeoGenetics, Natural History Museum, University of Copenhagen. All samples were collected and processed using strict aDNA guidelines. Nine libraries were prepared from the upper 3rd left molar root including dental pulp, the upper 2nd right molar dentine/cementum (200 mg) and pulp (interior root canal, 200 mg), femoral long bone (300 mg), and a mixture of mineralized dental plaque (calculus, 30 mg) taken from multiple teeth, and sequenced on an Illumina HiSeq in 15 runs (table S1).

These samples were processed in a specialized drill room within the dedicated aDNA facilities. In the initial investigation, the entire root of the upper 3rd left molar was crushed, and DNA extracted according to Rohland and Hofreiter (*46*), with a predigestion step as described by Allentoft *et al. (47)*. The same protocol was used for the interior dental pulp.

For the upper 2nd right molar, the entire tooth was removed from the mandible, and the tooth crown separated horizontally from the tooth root with a diamond-dust-coated cutting disk in a mechanical drill. The tooth root surface was then cleaned with a new cutting disk before using a small pointed drill bit to remove the interior dental pulp. DNA was extracted from this pulp as above. The remaining dentine and cementum fractions were crushed with a hammer, and also extracted. (We extracted dentine as well as cementum because this has previously maximized the yield of endogenous DNA.) The extraction protocol for this material and for dental calculus followed the protocols of Allentoft et al. (*47*) which are based on silica powder-based extraction, except that silica powder was only incubated for 1 h in the supernatant rather than the full 3 h.

DNA libraries for sequencing were prepared using NEBNext DNA sample preparation reagents (E6070) and Illumina specific adapters following established protocols. The libraries were shotgun sequenced in pools using Illumina HiSeq 2500 and 4000 platforms, and with a mix of 80 and 100-bp single read and 150-bp paired end chemistry. All pools submitted for sequencing contained between 5-15nM of DNA.

After initial screening, the highest proportion of Salmonella-specific reads were found in the upper 2nd right molar dentine/cementum. Additional libraries were constructed from the same extracts with the same protocol, and also shotgun sequenced to increase the depth of coverage.

Direct metagenomic assembly from SK152

Assembly and binning. Reads in the nine metagenomic libraries (table S1) that mapped to the human genome according to BWA (*48*) were excluded from the following analyses. The non-human reads were co-assembled into contiguous sequences (contigs) with MEGAHIT (*49*), using default parameters, and producing the assembly statistics in table S3 (All contigs). Contigs which exceeded 20 kb were split into 10 kb fragments as described (*34*), and the 39,016 contigs greater than 1 kb in length were clustered by CONCOCT into 76 bins using both sequence composition and coverage across all samples (*34*). All of the contigs in each bin were potentially derived from a single bin-specific genomic assembly. ORFs on these contigs were called using PRODIGAL (*50*), and then annotated to Clusters of Orthologous Groups of proteins (COGs) (*51*). We have previously identified 36 single copy core COGs (SCGs) that are found in a single copy in all bacterial genomes (*34*). We therefore tested each of the bins for the number of copies of these 36 SCGs, and found that 11 bins likely represented a unique genome because they contained at least 27 SCGs (75% of 36) in a single copy (fig. S3). These are henceforth referred to as metagenome assembled genomes (MAGs) and were used in the following downstream analyses. The contigs that belong to the 11 MAGs were stored in Database S8.

Taxonomic assignments. We constructed a phylogenetic tree of the 36 SCGs from 1,755 reference genomes plus the 11 MAGs. The reference genomes consisted of one representative of each bacterial genus and each archaeal species with complete genome sequences on NCBI. Where exceptional genomes, or MAGs, contained multiple SCGs we chose one of the sequence variants at random. Each SCG was then aligned separately using MAFFT (*52*), variable bases were trimmed with TRIMAL (*53*), and all 36 genes were concatenated for each genome, with gaps added if that gene was absent. The concatenated gene alignments were used to construct a single tree with FASTTREE (*54*).

For taxonomic assignments, we used the most frequent taxon among the reference genomes that clustered most tightly with each of the MAGs in the resulting tree. To this end we identified the ancestral nodes containing each MAG plus one or more reference genomes, and the most frequent taxonomic designation among the neighbouring reference genomes was assigned to each MAG (table S4).

Taxonomy assignment for Cluster C72. Cluster C72 was designated as *Mogibacterium timidum* according to the analysis of 36 SCGs. However, only few related reference genomes existed in NCBI, and we were uncertain about the accuracy of this taxonomic assignment. We therefore further the ribosomal genes from C72 against the rMLST database (*55*) , which contains >200K genomes representing over 6,000 bacterial species. To this end, the translated amino acids of representative alleles in the rMLST database were mapped onto C72 using TBLASTN. This identified the positions of 46 of the 53 ribosomal genes. Those 46 ribosomal genes were then compared against all genomes in the rMLST database. Analyses of both concatenated sequences and gene-by-gene analyses showed that C72 was more similar to *Eubacterium sulci* (74% identity) than *Mogibacterium timidum* (70%). We therefore refer to C72 as a *Eubacterium.*

Calculation of deamination rates. The mapping files were mapped with BWA onto each contigs from each MAG, and equivalent mappings with BWA were performed onto the human and Ragna reference genomes. We then used MAPDAMAGE 2.0 (*56*), to separately characterise the DNA damage associated with each of the 13 organisms. lists The posterior mean estimates and standard deviations for δS, the single-stranded deamination rate for all genomes and MAGs, are listed in table S5 and depicted graphically in Fig. 1D.

Environmental sources of MAGs. Several MAGs were clearly of environmental origin. C69 was unambiguously classified as the methanogen *Methanosphaerula palustris*. These organisms are associated with acidic peat bogs (*57*), and belong to a group of hydrogenotrophic methanogens with no human-associated relatives. Similarly, C40 and C66 were assigned to the environmental sulphate-reducing species *Desulfatiglans anilini* and *Desulfomonile tiedjei,* respectively, whereas human-associated sulphate reducers usually belong to *Desulfovibrio* species (*58*). C72 is a *Eubacterium,* which are found in periodontal biofilms, and are associated with dental plaque (*59*). In support of this role, C72 was recovered almost exclusively from dental calculus whereas reads from the putative environmental taxa were present in multiple sources (table S6).

Multilocus Sequence Typing

EnteroBase automatically assembles genomes using the ‘Genomic assemblies’ pipeline described below from all publicly available Illumina short reads as well as from short reads that are uploaded by users. Genomes that pass internal criteria for purity and manifest a mean coverage of ≥20fold are automatically genotyped by a second, common pipeline into Sequence Types of unique allele numbers (STs) for multiple multi-locus sequence typing (MLST) schemes. These MLST schemes currently include a 7 housekeeping gene legacy scheme (*11*), rMLST based on ribosomal proteins (*55*), wgMLST based on a pan-genome and core genome MLST (cgMLST). Details of these schemes are available on the Help pages at EnteroBase, which also described the detailed stages involved in defining the *Salmonella* schemes. In brief, the 7-gene scheme is co-ordinated with the MLST scheme defined by Achtman *et al*. (*11*), and co-ordinated with the legacy MLST database at University of Warwick. The rMLST scheme is based on the 51 ribosomal gene scheme defined for *S. enterica* by Jolley *et al. (55)*, and is fully co-ordinated with the primary rMLST database at Oxford University. The wgMLST scheme encompasses the 21,065 unique sets of homologs with ≥70% pairwise amino acid similarity over 50% of their length which were identified in 537 diverse, high quality *Salmonella* genomes. A few of those homolog sets contain duplicated genes for which more than one copy is present per genome. In order to define cgMLST V2, we examined the presence of all wgMLST genes in 3,144 genomes, one for each rMLST ST that had been defined by May, 2016. cgMLST V2 consists of that subset of 3,002 loci from the wgMLST scheme which were present in ≥98% of those genomes, intact in ≥94%, and were not exceptionally diverse in comparison to other genes. When rare genomes contained two or more copies of any cgMLST locus, that locus was scored as duplicated for that genome, and hidden to normal public access for that genome.

Species trees of *S. enterica* subspecies I

We assembled a representative collection of genomes from *S. enterica* subsp. I by choosing a random representative of each of the 2,964 rMLST STs in EnteroBase Workspace of Representative Strains (May, 2016) (Database S9). The sequences of the 3,002 core genes defined in the cgMLST V2 scheme in EnteroBase were aligned with MAFFT (*52*) with the parameter '--auto'. These alignments were used to calculated two species trees by two different algorithms. i) A Maximum Likelihood tree (Database S5) was generated using a GTRCAT model in RAxML v8.2.4 (*60*) from the concatenated, aligned sequences of all core gene sequences (2.8 Mb). ii) RAxML was used to generate a single tree per core gene alignment, and these 3,002 gene trees were fed into ASTRID (*61*), which was used with default parameters to generate a consensus tree based on the minimum-evolution criterion as implemented in FASTME (*62*). ASTRID is a recently developed coalescent-based method for species tree estimation that is statistically consistent under the multi-species coalescent and can therefore account for incomplete lineage sorting. As a result, ASTRID is competitive with the best current methods for accuracy, and can handle large amounts of data. Astrid calculates topologies but not branch lengths. We therefore estimated the branch lengths in the ASTRID consensus tree (Database S4) from the corresponding concatenated alignments using RAXML with a GTRGAMMA model. We also compared these results with those from a species tree (data not shown) calculated with the help of a GPU compilation of ASTRAL-II (*63*) by S. Mirarab and T. Warnow. The species trees were almost indistinguishable between ASTRAL and ASTRID (data not shown).

Illumina sequencing of modern genomes

Total DNA was extracted with a Maxwell 16 cell DNA purification kit (Promega, Madison WI), in accordance with the manufacturer's recommendations. Libraries were constructed using the Nextera XT kit (Illumina) with the following modifications. The initial tagmentation reaction was performed using 2 µl of 0.7 ng/µl of template DNA and 2/5 of the specified volume for other reagents, resulting in a volume of 10 µl after neutralization. For the PCR step, 25 µl of 2x

Extensor Hi-Fidelity PCR master mix (Thermo Scientific), 5 µl of each index primer (4 µM) and 5 µl of sterile distilled water was added to the tagmentation reaction. The standard PCR reaction was extended by an extra 3 cycles, and the extension step was lowered from 72^{o}C to 68^{o}C. The libraries were purified using 25 µl of Ampure XP beads (Beckman Coulter) with two 200 µl washes with 80% ethanol before elution in 30 µl of RSB from the Nextera XT kit. Libraries were quantified using the Qubit dsDNA HS Assay Kit (Thermo Scientific), and diluted to 3.2 ng/µl (approx. 8 nM). Pooled libraries (40 samples per run) were denatured following the Illumina protocol, and 600 µl (approx. 20 pM) was loaded onto a MiSeq V2 -500 cycle cartridge (Illumina), and sequenced on a MiSeq to produce FASTQ files.

Genomic assemblies

Strains sequenced as part of this study and other sequenced reads from public databases (Database S10, EnteroBase Workspace of Para C Lineage) were assembled in EnteroBase using its default assembly pipeline. In brief, the ends of sequenced reads with base quality less than 5 were removed (trimmed) using SICKLE (*64, 65*) and assembled into contigs using SPADES 3.5 (*66*) with the parameter '--only-assembler' and k-mers equal to 0.3, 0.5, 0.7 and 0.9 of the average read lengths. To validate the consensus call for each base in the assemblies, the original trimmed, sequenced reads were mapped back to the corresponding assembled contigs using BWA (*48*), and analysed with SAMTOOLS/BCFTOOLS 1.2 0.7.12-r1039 (*67*). The quality scores for consensus calling for each base were stored together with the assemblies in standard FASTQ format. Finally, assembled contigs were assigned taxonomic labels using KRAKEN (*68*) in order to exclude potential contamination from genera other than *Salmonella*.

Identification of Paratyphi C aDNA within metagenomic data

KRAKEN (*68*) was used to examine metagenomic reads from 33 human skeletons from Trondheim, Norway. These analyses scored 266 reads in the metagenome from skeleton SK152 as *Salmonella,* but reads from *Salmonella* and other potentially invasive bacterial pathogens were not identified in any of the other 32 skeletons. We used BLASTN to align the putative *Salmonella*-specific reads from SK152 against the non-redundant nucleotide database in GenBank, which confirmed that they were specific for *Salmonella*, and also indicated that they were most similar to either serovar Paratyphi C or Choleraesuis. Recovering strain-level genotypes from metagenomic data is a long-lasting problem (*69*), especially because aDNA samples can contain reads from multiple genotypes due to mixed infections (*70*). We therefore used MGPlacer (*70*) to iteratively map Salmonella-specific metagenomic reads from SK152 onto a core SNP phylogeny that was reconstructed from 20 complete genomes, representing one genome per serovar in RefSeq (May, 2016). All *Salmonella*-specific reads were placed exclusively on the branch to *S. enterica* serovar Paratyphi C strain RKS4594, suggesting that SK152 is a close relative of modern Paratyphi C.

Identification of Ragna-specific reads in metagenomic libraries.

Firstly, the metagenomic reads from nine sequencing libraries were pre-processed by two scripts from the BBmap (*71*) package: BBmerge to collapse paired end reads arising from a common DNA fragment and BBduk2 to remove adapters and low quality bases from the single ended reads.

The reads were then filtered against genomes of the modern relatives of Paratyphi C in order to identify sequences specific to the Ragna genome. To this end, the pre-processed reads were aligned onto two sets of references using BOWTIE2 (*72*): i) The "ParaC" set, consisting of all 219 modern genomes in the Para C Lineage and ii) the “outgroup” set, consisting of all representative bacterial genomes in RefSeq which were not *Salmonella*, plus the human genome hg38. Sequenced reads were scored as belonging to Ragna which showed equal or better alignment scores to the ParaC set than to the outgroup, and which differed from the most similar genome in the ParaC lineage by ≤4% sequence variation.

Duplicate reads resulting from PCR amplification during sample preparation distorts both estimates of coverage and can lead to false SNP calling. We therefore removed all potential duplicates by retaining only one representative Ragna-specific read sequence when multiple identical reads were identified. This deduplication resulted in even Ragna read coverage across the Para C Lineage pan genome (Fig. 2B).

Reference based mapping of aDNA

The Ragna-specific, de-duplicated reads were aligned against RKS4594 using BOWTIE2 (*72*) with the end-to-end option and analysed using SAMTOOLS/BCFTOOLS 1.2 (*73*). In order to exclude potential spurious SNPs due to deamination, the consensus base was only called on sites that were covered by at least two reads with a consensus base quality ≥10 and which were located at least 5 bases from both read ends. Exceptionally, SNP calls were included with single coverage matching these criteria if the same call was provided by previously excluded 5’ and 3’ ends of reads.

Reference based SNP calling in modern genomes

Assemblies within EnteroBase from modern strains within the Para C lineage + Birkenhead (Database S10) were aligned against RKS4594 using LAST (*74*), and SNPs from these alignments were filtered to remove regions with low base qualities (Q < 10) or ambiguous alignment (ambiguity of alignment ≥ 0.1). Sites were also removed if they aligned with ≥ 95% identity to disperse repetitive regions that were longer than ≥ 100 bps according to BLASTN; overlapped with tandem repeats as identified by TRF (*75*); or overlapped with CRISPR regions as identified by PILER-CR (*76*). After eliminating these repetitive regions, a total of 61,451core SNPs were identified within core genomic regions that were conserved in ≥95% of the genomes from the ParaC Lineage + Birkenhead (Database S3).

Removal of recombinant SNPs from the Para C lineage phylogeny

RECHMM (*20*) was used to identify clustered SNPs within the Para C lineage phylogeny + Birkenhead which are putatively derived from homologous recombination. Briefly, an initial phylogeny was calculated on all core SNPs using RAxML v8.2.4 (*60*) under a GTRCAT model with Stamatakis ascertainment correction for invariant sites (*77*). SNPs were then assigned onto branches in that initial phylogeny using a maximum likelihood method with a symmetric transition model (*78*). RecHMM then identifies clusters of SNPs on each branch using hidden Markov models, and these were excluded from further analysis. The remaining 49,610 SNPs were used to generate a mutational phylogeny (Figs. 2B and 3; fig. S8; Database S7) using RAxML as above.

Date estimates for the Para C Lineage

To infer the evolutionary timescale of the Para C Lineage, we performed molecular-clock analyses using three different methods, with the results summarised in table S13.

i) **Bayesian phylogenetic approach (BEAST v1.8.3)** (*79*). Sequence diversity was analysed using the GTR+G model of nucleotide substitution on mutational SNPs, with a discrete gamma distribution and six rate categories to account for rate heterogeneity across sites. To account for SNP ascertainment bias, we applied a correction that incorporated the nucleotide frequencies across all of the constant sites. For computational tractability, we analysed two subsamples each from the Para C Lineage, the Paratyphi C serovar and the modern Paratyphi C strains without Ragna. Each subsample comprised 50 sequences, including the Ragna sequence, the sole strain in the CS-3 cluster and 48 other randomly sampled sequences. All subsamples were run with either of two dates for Ragna (1200 CE from archaeological dates and 1073 CE from AMS calibrated ages). Thus, our Bayesian phylogenetic analyses involved 12 data sets (table S12 and Database S11).

Posterior distributions of parameters, including the tree, were estimated using Markov chain Monte Carlo sampling. Samples were drawn every 5000 steps over a total of 5×10^7 steps, with the first 10% of samples discarded as burn-in.

Marginal likelihoods (table S12) were used to compare three demographic models for the tree prior: constant population size, Bayesian Skyride coalescent (*80*), and birth-death process with serial sampling (*81*). Marginal likelihoods were estimated using stepping-stone sampling (*82*), with 20 path steps and a chain length of 10^6 per path step.

ii) date-randomisation test (*83*). This test involves randomly reassigning the ages of the sequences to produce a number of date-randomised replicate data sets. If the 95% credibility interval of the rate estimate from the original data does not overlap with any of the multiple rate estimates from the date-randomised replicates, the data set is considered to have sufficient temporal structure (*84*). For this test, we used 10 date randomisations (fig. S7).

iii) root-to-tip regression and least-squares dating. Each of these methods was used to analyse five data sets: the whole Para C Lineage and the Paratyphi C serovar with two possible ages of Ragna, in addition to the modern Paratyphi C strains without Ragna.

A root-to-tip regression analysis was performed in TEMPEST (*85*) using the best-scoring likelihood tree. In this method, the genetic distance between the root and each tip (given in substitutions per site) is plotted against the sampling year of each sequence. A regression analysis is then used to identify the line of best fit, of which the slope reflects the substitution rate. The significance of the slope was evaluated in R v3.0.1 (*86*).

Least-squares dating was performed in LSD (*87*) using the best-scoring likelihood tree. This method assumes a strict molecular clock and it estimates the rate and node ages using a normal approximation of the Langley-Fitch algorithm (*88*). To test for the presence of temporal structure in the data set, substitution rates were also estimated from 1,000 date-randomised replicates. We verified that the rate estimate from the original data set was not contained within the range of values obtained from the date-randomised replicates.

Extrapolation of the BEAST mutation rate to *S. enterica* subsp. I. We extrapolated the median of all estimates of the BEAST mutation rate for the Para C Lineage to several interesting nodes in the two species trees in Databases S4 and S5. Dates (table S13) were estimated for these nodes as the average genetic distances between each node to all its descending tips divided by the median mutation rate.

Inferring ancestral geographic sources

A subtree was extracted from the Para C Lineage phylogeny plus Birkenhead with the ETE3 Python package (*89*) that contained only genomes with geographic source information (thus lacking Lomita). The ancestral geographic states of internal nodes in this subtree were inferred by three independent algorithms: i) the Maximum Likelihood comparison (*90*) in BayesTraits; ii) the Markov chain Monte Carlo (MCMC) approach (*91*) in BayesTraits; and iii) Bayesian Binary

MCMC in RASP (*92*). Independent tests were run on the entire subtree and after excluding Ragna and serovar Birkenhead, to test whether there was oversampling of European strains in deep branches. All three algorithms indicated a European origin for the Para C Lineage (fig. S8, table S14).

Reconstruction of the pan genome of the Para C Lineage

A wide collection of annotated reference genes was collected from three sources: 1) all 21,065 unique sets of homologs in the wgMLST scheme in Enterobase; 2) published annotations for the complete genomes of strains C500 (Choleraesuis) (*93*), RKS4594 (Paratyphi C) (*94*) and SC-B67 (Choleraesuis) (*95*); 3) PROKKA (*96*) annotations of all draft genomes in EnteroBase. All these reference genes were grouped into 29,436 clusters using UCLUST (*97*) with parameters '-id 0.95 -query_cov 0.9 –target_cov 0.9 -centroids'. The centroid sequences of all these clusters were aligned with all the modern genomes in the Para C Lineage using BLASTN, in order to obtain sets of homolog regions that cover ≥50% of the length of the centroid sequences with ≥70% identity.

Overlapping paralogs between the homolog sets were identified through the same iterative methodology used to construct the entire *Salmonella* wgMLST scheme. After removal of those paralogs, the remaining 6,665 homolog sets were treated as the pan genes of the Para C Lineage (Database S1). The sequences, original designations and annotations of the centroid sequences for the 6,665 homolog sets were stored in Database S12.

Pseudogenes

The pan gene alignments used to reconstruct the pan genome of the Para C Lineage were screened for potential gene disruptions. Genes with stop codons or frameshifts anywhere in the coding regions were scored as pseudogenes (Database S1), and assigned to branches in the core SNP phylogeny using the Maximum Likelihood algorithm (*78*) (Database S6; fig. S4).

Reconstruction of pan genome synteny

Synteny was inferred through constructing and traversing a graph of assembled sequences from genes in the Para C Lineage pan-genome. Firstly, the graph was seeded with one node for each unique, single copy gene. The connections between nodes were weighted using the following criteria: 1) An edge had the maximal weighting if the two genes connected by that edge were co-located on a single contig; 2) edges were given an intermediate weighting when they connected two genes at the ends of distinct contigs due to read-pairs that would map across both contigs. This weighting was 2*(number of read-pairs joining the two contigs)/(total number of unpaired reads at the ends of contigs); 3) Edges were not assigned for pairs of genes which did not co-locate according to these two criteria.

CONCORDE (*98*) was used to find the shortest possible path that visited all the nodes in the graph, which equates to the most likely gene order within the Para C Lineage. Ambiguous paths were inspected manually to identify the duplicated genes and collapsed repeats, which are usually associated with prophages or plasmids. These were then manually broken and re-joined as appropriate. Finally, all repetitive genes were inserted into the gene order according to their location within the assemblies, in order to reconstruct the synteny of the pan genome of the Para C Lineage (Fig. 2B and Database S1). A total of 227 genomic islands (Database S2; figs. S5 and S9; summarised in table S12) were identified as continuous blocks of gene gain/loss in the pan genome of Para C Lineage.

The conservation of genes within the pan genome across the ParaC lineage was illustrated using Anvi'o (*99*) (Fig. 2B) and can be examined in detail in a publicly accessible, interactive local version of Anvi'o server Pan-genome. The synteny of pan genes was enforced within the Anvi'o rendering by help of an artificial guide tree based on the gene order in Database S1, where each sequential gene bifurcates from the previous gene at a constant distance. (That artificial tree was deleted from the printed version in Fig. 2B). The figures are dependent on manually generated input files based on the annotations and sub-divisions that are indicated in Database S1, which contains the frequencies of pan genes in all sub-lineages, the locations of major genomic islands, as well as additional metadata. These input files can be downloaded from the interactive version (Anvi'o server Pan-genome).

Identification of phages

Prophages in modern genomes of Para C Lineage were identified using the online API in PHASTER (*100*), and those pan genes that were found to be associated with prophages were annotated in accordance with their most similar genes in the PHASTER database. The identified prophages were taxonomically labelled according to similarities of the major capsid protein (MCP) to the clusters of bacteriophages described by Casjens *et al.* (*101*). Further manual modifications were based on comparisons of all genes in each prophage against the published phage clusters. Prophage names were the same as the previously described phages which they most closely resembled, except for GI28, which does not have any known close relatives. SPC-P1 has been previously described in Paratyphi C (*24*), and this name was retained.

SPIs, IS elements, plasmid Inc (incompatibility) groups and type IV secretion systems

The reference sequences of SPI-1 through SPI-12 were downloaded from PAIDB (*102*), as were SGI-1 and SGI-2. SPI-13 through SPI-21 were obtained from publications in which they were described (*103-107*). We defined genomic islands within the pan genome using BLASTN as tight clusters of genes in which ≥60% of the sequences aligned with the reference SPI/SGI ssequences with ≥ 80% identity. The SPI-6 island, which has undergone multiple events of gene gain/loss in the Para C Lineage, is presented in detail in a publicly accessible, interactive local version of Anvi'o server SPI-6 (*99*) and summarised in fig. S6.

Genes in the pan genome were also screened for genes associated with mobile elements against ISfinder (IS elements) (*108*), PlasmidFinder 1.3 (Inc groups of plasmids) (*109*) and CONJscan-T4SSscan (relaxases and key components of type IV secretion systems) (*110*). Strong hits were annotated as provided by those sources (Fig. 2B; Databases S1 and S2).

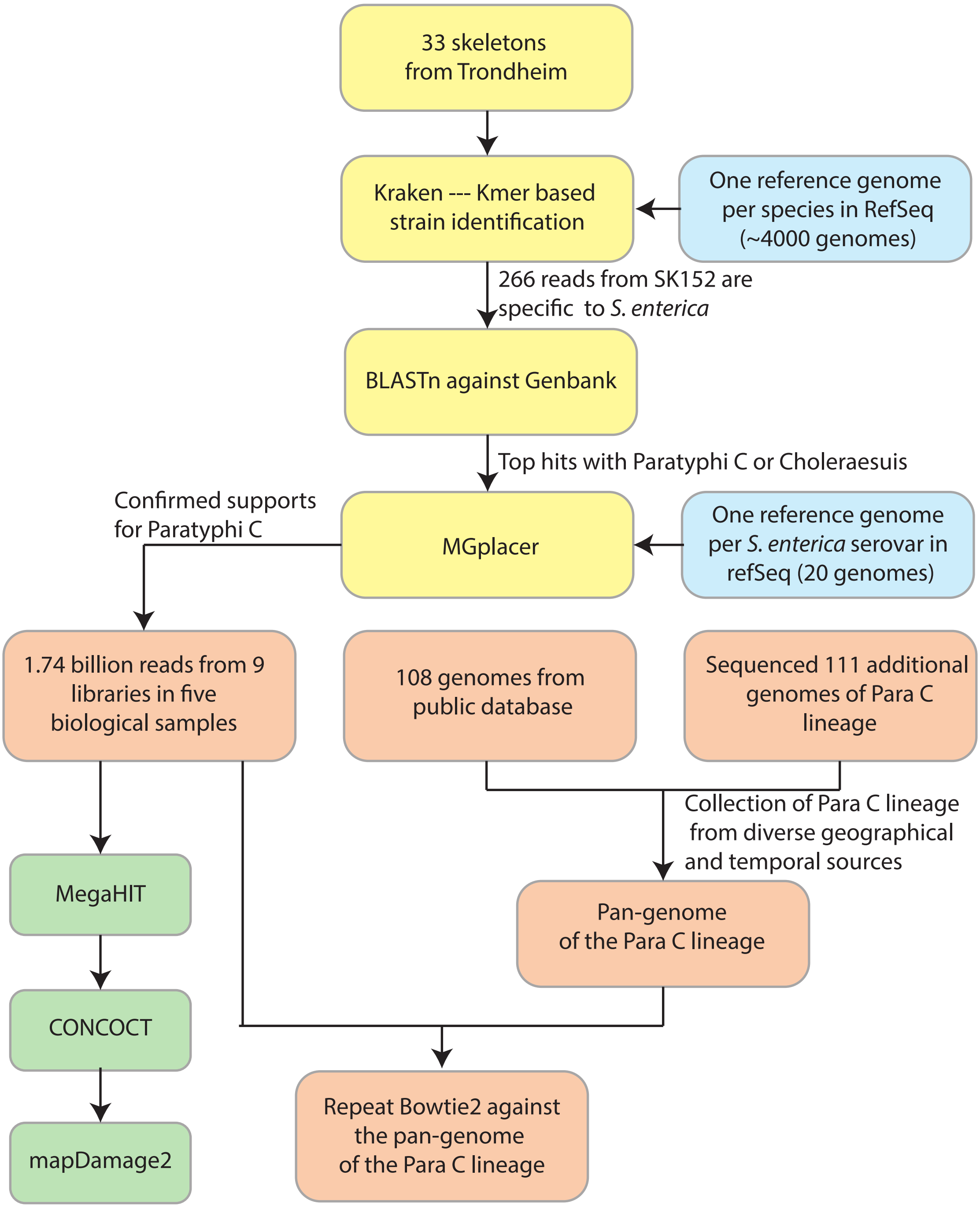
33 skeletons from Trondheim
Kraken --- Kmer based strain identification
One reference genome per species in RefSeq (~4000 genomes)
266 reads from SK152 are specific to *S. enterica*
BLASTn against Genbank
Top hits with Paratyphi C or Choleraesuis
Confirmed supports for Paratyphi C
MGplacer
One reference genome per *S. enterica* serovar in refSeq (20 genomes)
1.74 billion reads from 9 libraries in five biological samples
108 genomes from public database
Sequenced 111 additional genomes of Para C lineage
Collection of Para C lineage from diverse geographical and temporal sources
MegaHIT
CONCOCT
mapDamage2
Pan-genome of the Para C lineage
Repeat Bowtie2 against the pan-genome of the Para C lineage

Figure S1. Overview of the workflow used in this study.

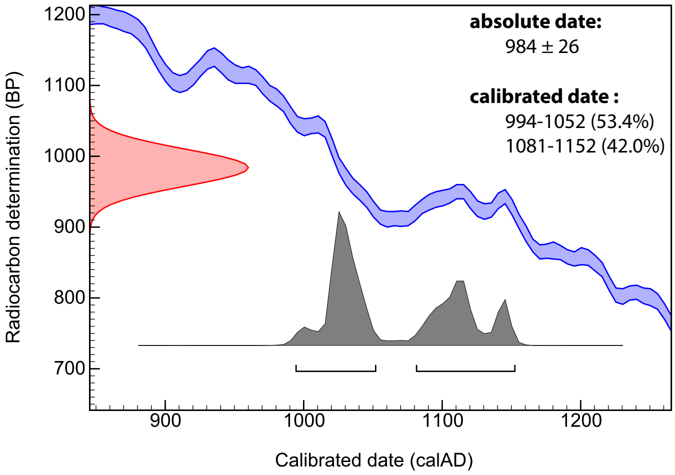

Figure S2. Radiocarbon (^{14}C) dating of tooth from SK152. The red curve on the Y axis shows the uncalibrated ^{14}C estimate of years before present. The blue line is the calibration curve, and the grey curves on the X axis are the calibrated dates according to that calibration curve. These calibrated dates fall into two time periods with almost equal likelihood.

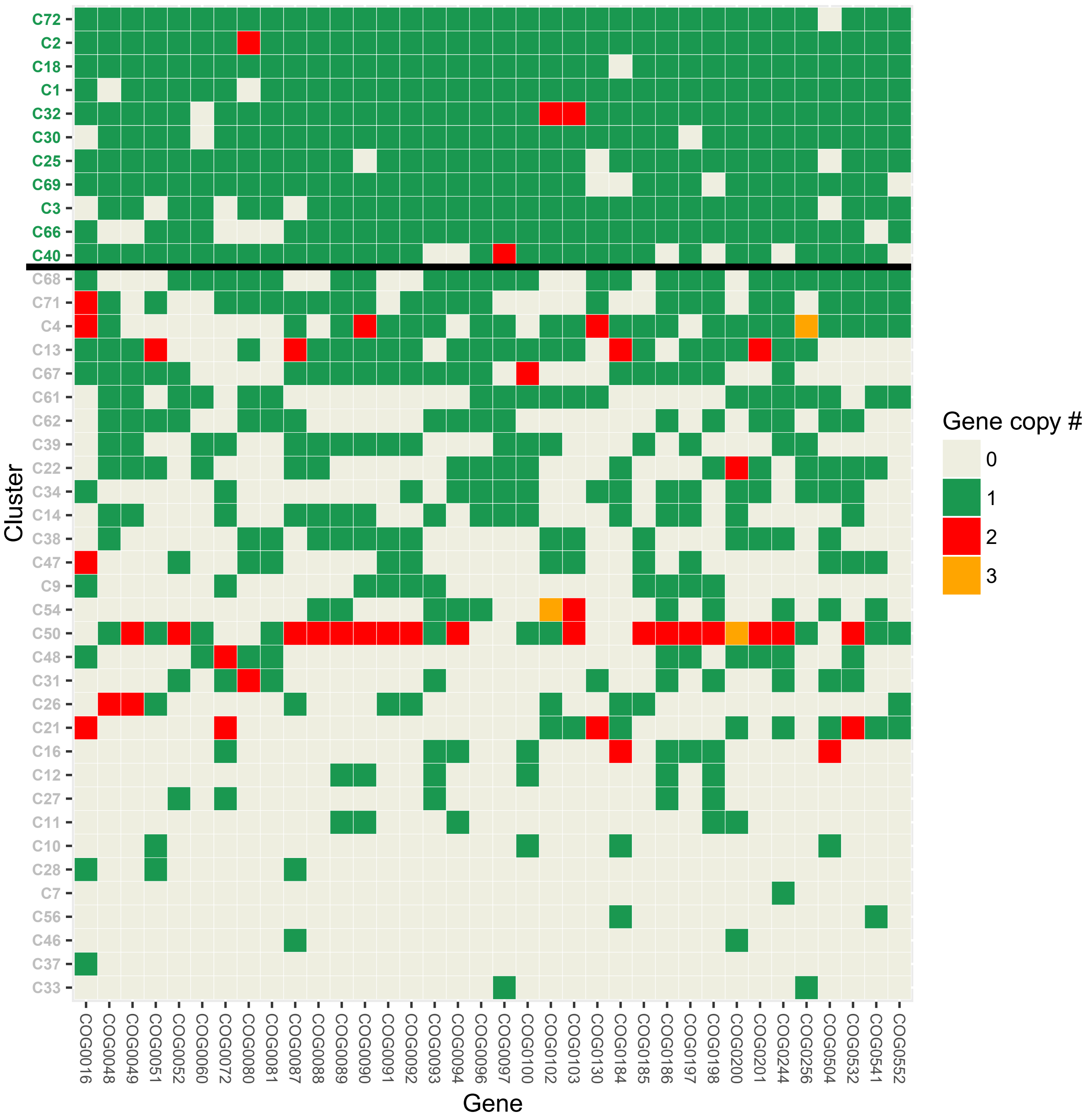

Figure S3. Single core gene (SCG) frequencies in the 76 CONCOCT clusters generated after binning contigs ≥1 kb. Each row corresponds to a cluster and columns are SCGs. The top 11 clusters, delineated by a thick black line, included ≥75% of the 36 SCGs. Gene copy number is indicated by the key.

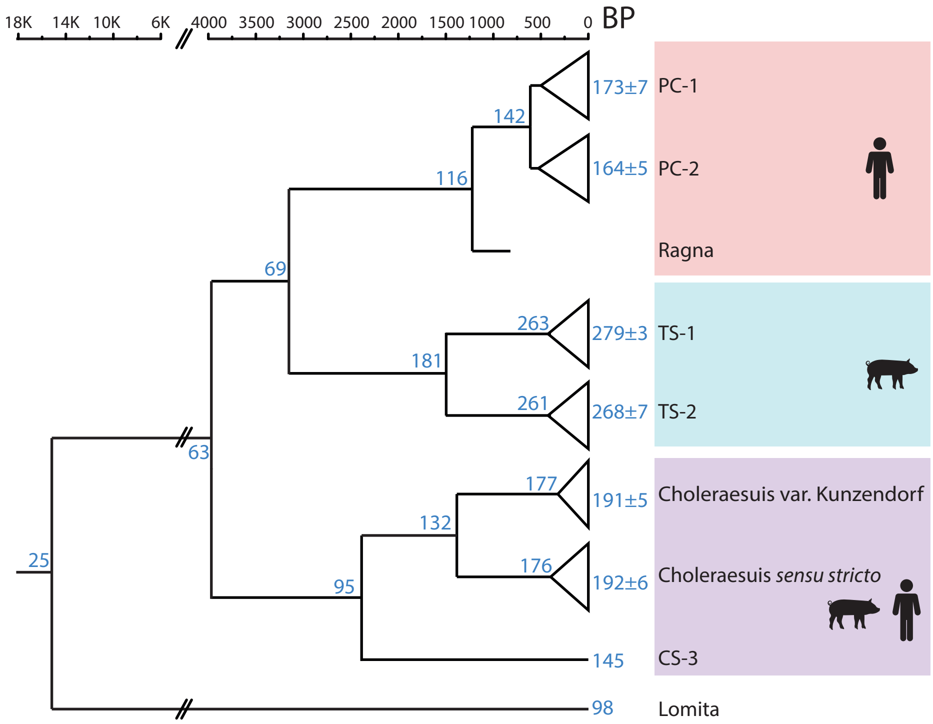

Figure S4. A cartoon tree of the Para C Lineage scaled with dates before present estimated by BEAST. The total number of pseudogenes accumulated since the root per major sub-lineage are shown along the branches.

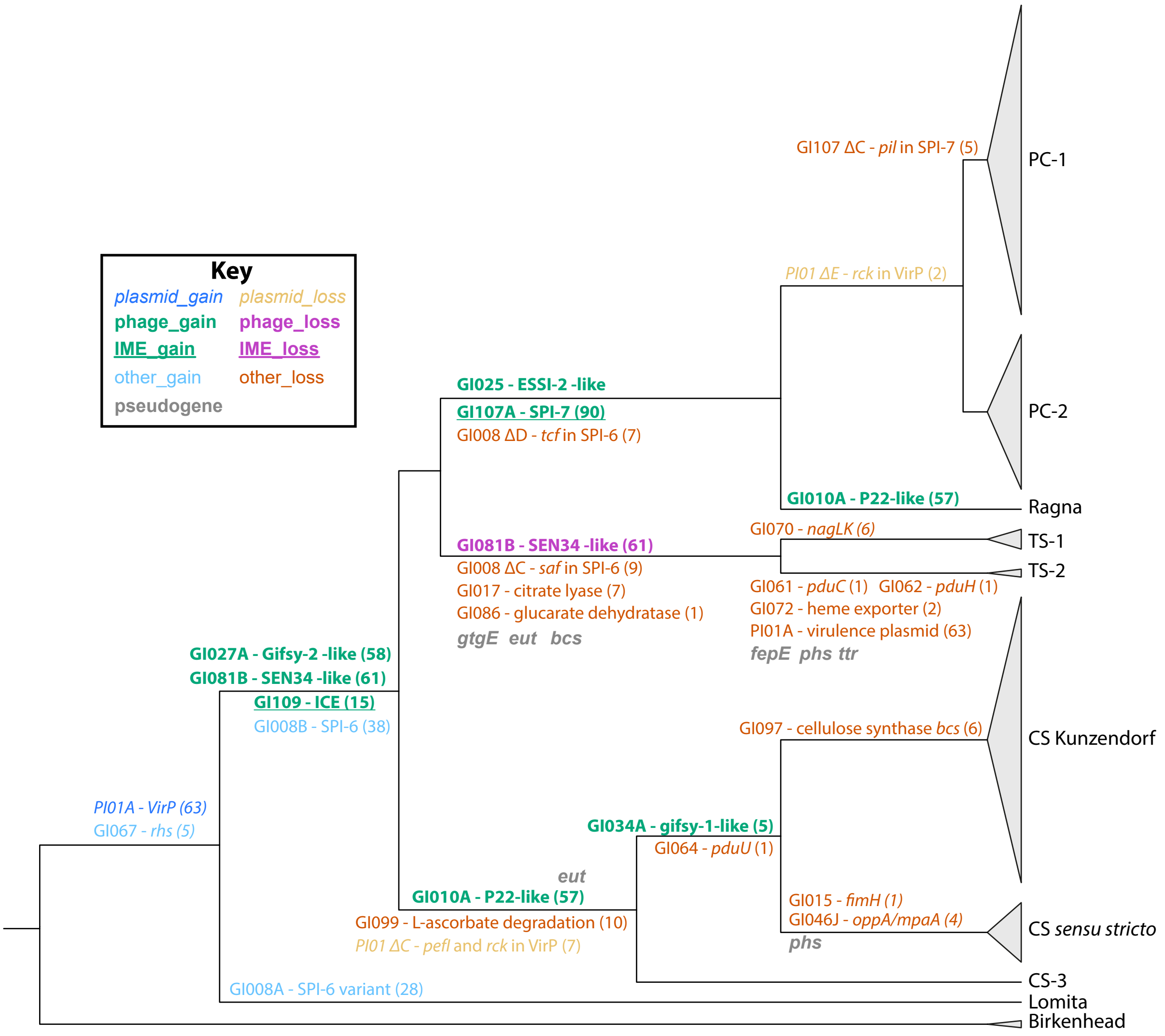
Key
plasmid_gain
plasmid_loss
phage_gain
phage_loss
IME_gain
IME_loss
other_gain
other_loss
pseudogene
GI107 ΔC - pil in SPI-7 (5)
PC-1
PI01 ΔE - rck in VirP (2)
PC-2
GI025 - ESSI-2 -like
GI107A - SPI-7 (90)
GI008 ΔD - tcf in SPI-6 (7)
GI010A - P22-like (57)
Ragna
GI070 - nagLK (6)
TS-1
GI081B - SEN34 -like (61)
GI008 ΔC - saf in SPI-6 (9)
GI017 - citrate lyase (7)
GI086 - glucarate dehydratase (1)
gtgE eut bcs
TS-2
GI061 - pduC (1) GI062 - pduH (1)
GI072 - heme exporter (2)
PI01A - virulence plasmid (63)
fepE phs ttr
GI027A - Gifsy-2 -like (58)
GI081B - SEN34 -like (61)
GI109 - ICE (15)
GI008B - SPI-6 (38)
GI097 - cellulose synthase bcs (6)
CS Kunzendorf
PI01A - VirP (63)
GI067 - rhs (5)
GI034A - gifsy-1-like (5)
GI064 - pduU (1)
eut
GI010A - P22-like (57)
GI099 - L-ascorbate degradation (10)
PI01 ΔC - pefI and rck in VirP (7)
GI015 - fimH (1)
GI046J - oppA/mpaA (4)
phs
CS sensu stricto
GI008A - SPI-6 variant (28)
CS-3
Lomita
Birkenhead

Figure S5. Summary of the genomic changes in the accessory genome of selected genes in deep branches of the Para C Lineage plus Birkenhead. The sizes of the collapsed triangles at the tips of the tree are proportional to the numbers of genomes. Numbers in parentheses indicate the numbers of genes affected by the change. Partial deletions in the virulence plasmid (VirP) and SPI-6 are indicated by "Δ". The gain or loss of individual islands is described in detail in Database S2 and the acquisition of pseudogenes in Database S6. Colours in the Key indicate the types of GIs and gain/loss. Note that the basal branches are not continuously scaled to their true branch length.

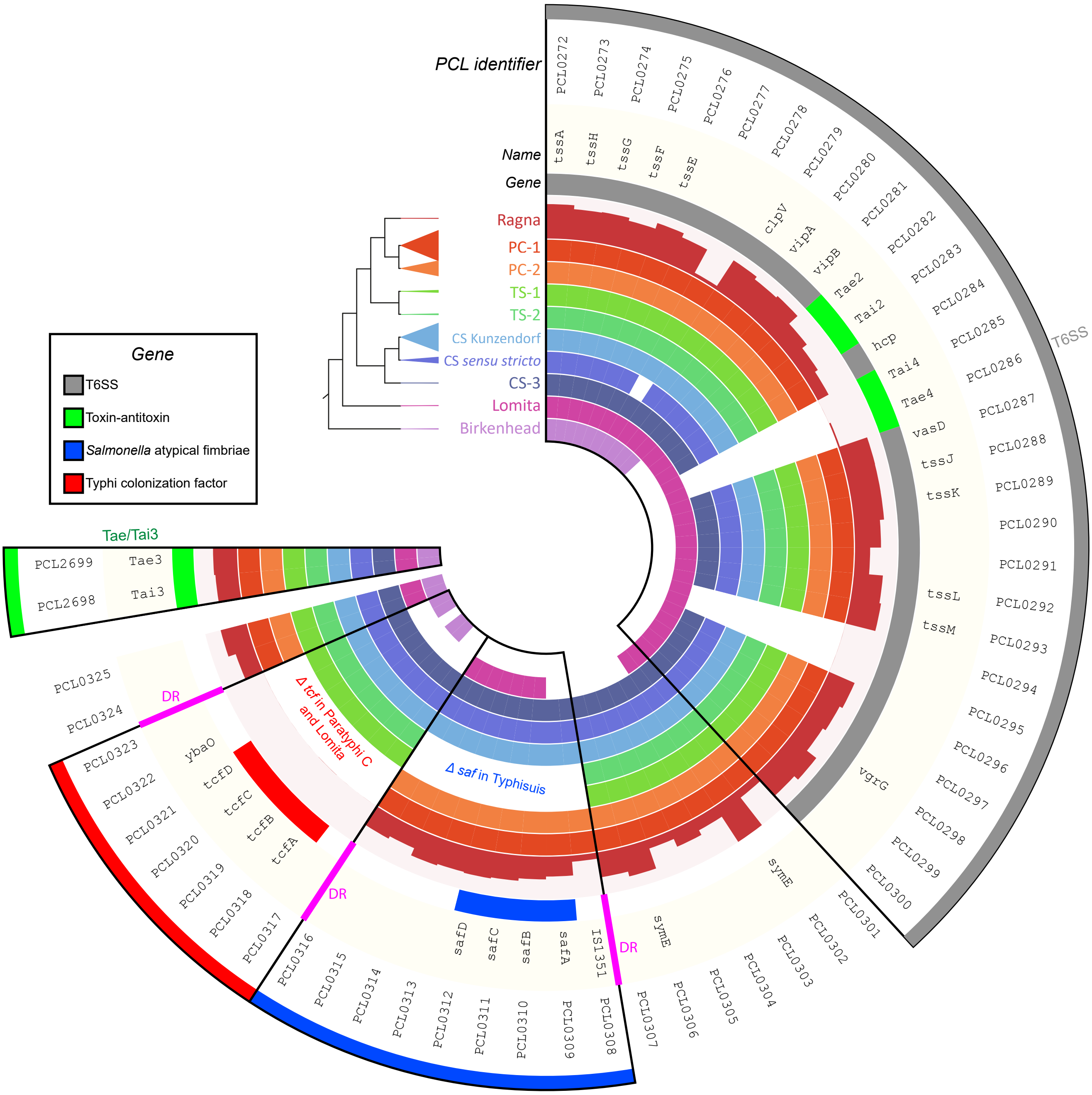

Figure S6. SPI-6 gene within Para C Lineage plus Birkenhead. The entire *saf* and *tcf* clusters have been deleted in Typhisuis and Paratyphi C plus Lomita, respectively, due to excision via bounding direct-repeats (DR, purple) that are associated with IS1351. Other mechanisms have resulted in the absence of some *saf* genes in Lomita and of both *saf* and *tcf* genes in Birkenhead, possible reflecting independent acquisitions by HGT. An interactive version is also available at https://enterobase.warwick.ac.uk/anvi_public/ParaC_SPI6

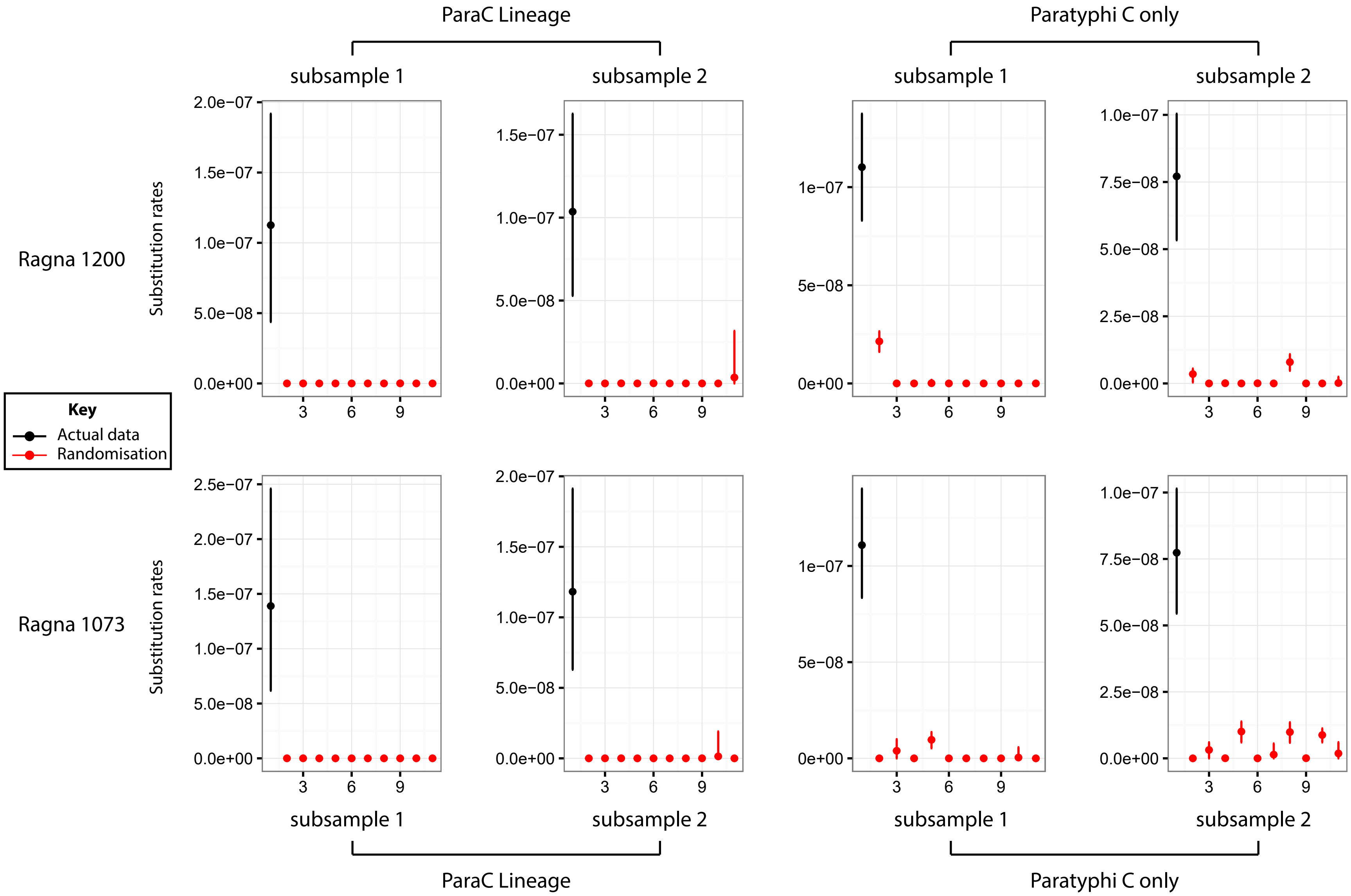

Figure S7. Estimates of substitution rates for multiple replicate data sets. Solid circles represent mean estimates, and errors bars show the 95% credible interval. Black symbols were estimated with real data and red symbols were obtained after randomizing the dates, with 10 randomizations per data set.

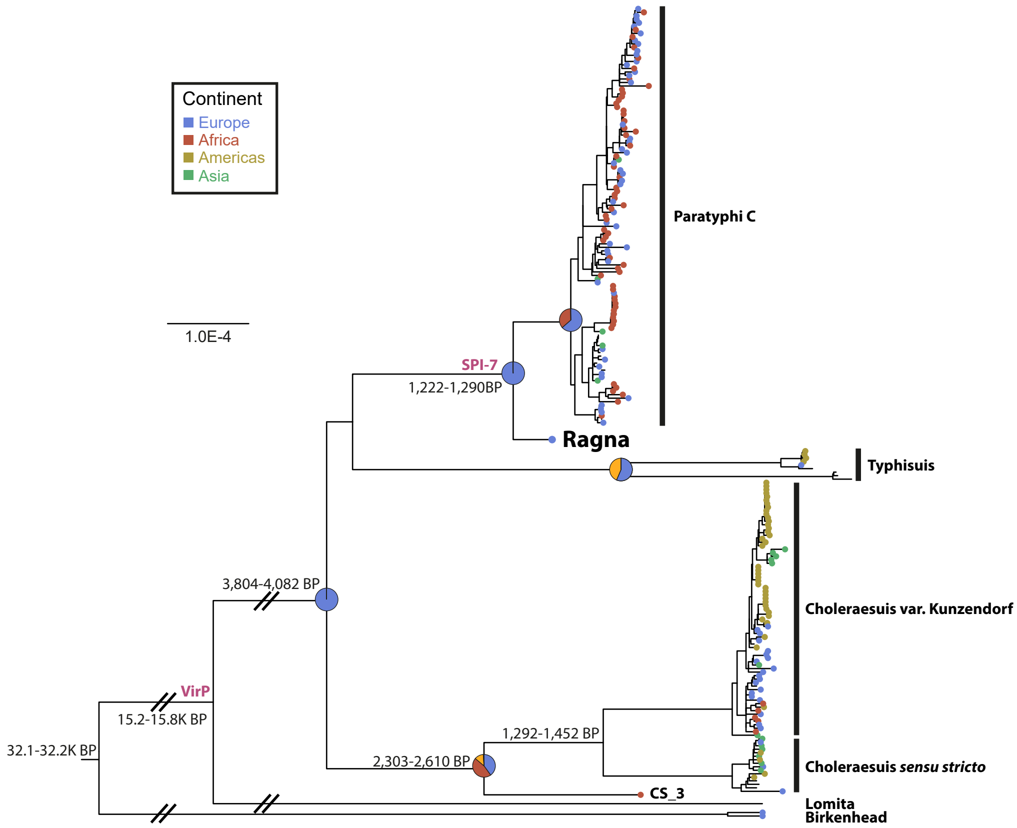

Figure S8. Maximum likelihood phylogeny and geographical origins derived from 49,610 SNPs in the non-recombinant, non-repetitive core genome of the Para C Lineage plus Birkenhead. Filled circles on the tips are color-coded according to the source continent, as indicated in the key. Pie-charts on internal nodes are geographical inference estimates of the likelihoods of continental sources according to standard MCMC in BayesTraits (table S14). The number on branches indicate the range between two median dating estimates for independent subsamples of the Para C Lineage, except that the dates for the two deepest nodes were extrapolated from deeper trees (Fig. 2A) that included representatives of all *S. enterica* subspecies I. Note that the basal branches are not continuously scaled to their true branch length.

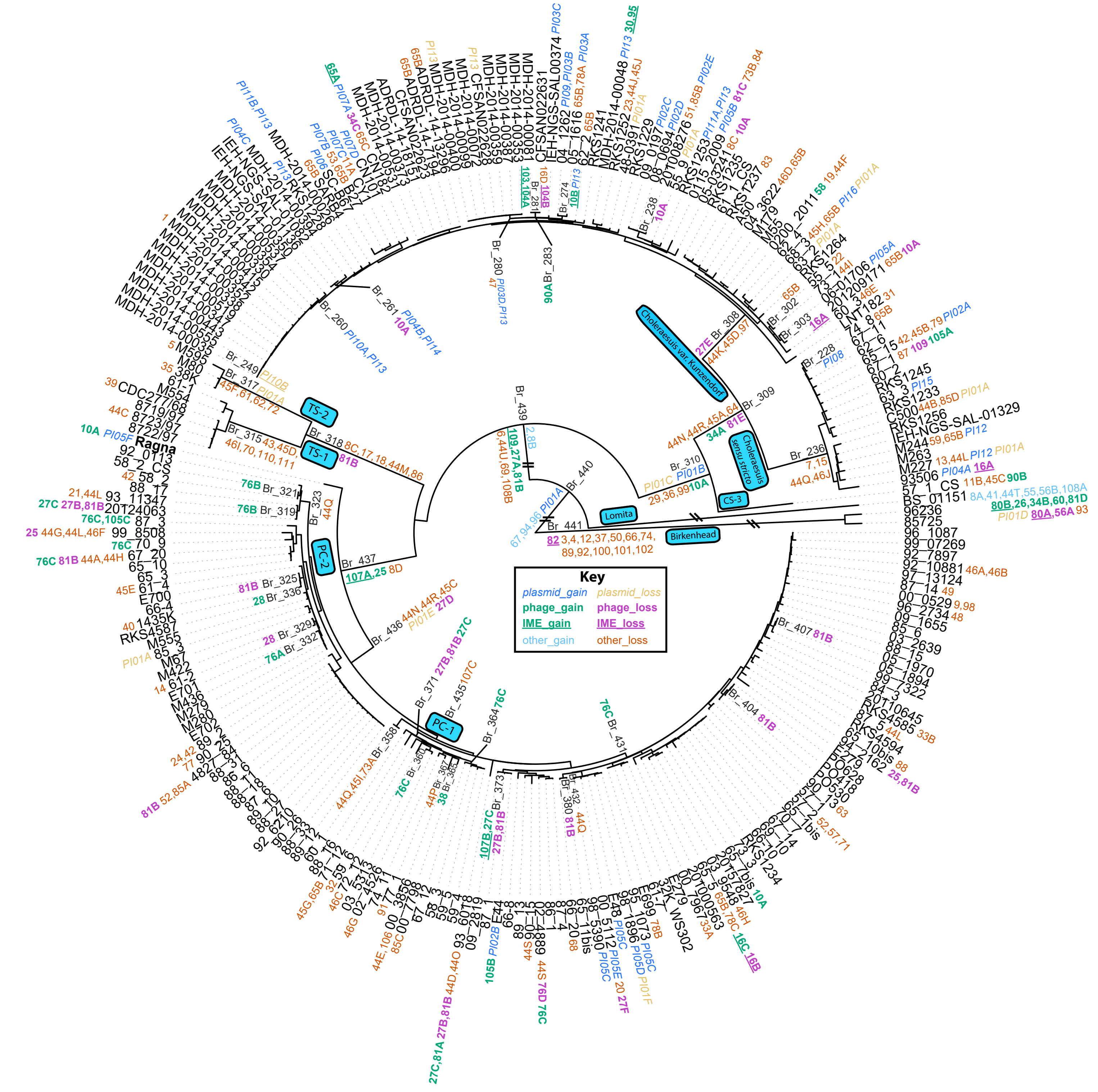

Figure S9. Maximum Likelihood phylogeny derived from 49,610 non-recombinant, non-repetitive core SNPs within the Para C Lineage plus Birkenhead. Internal branches are indicated in black while internal nodes for major sub-lineages are indicated by cyan rectangles. Other numbers refer to gain or loss of genomic islands (42 for GI042, *etc.*) or plasmids (P108) in the accessory genome coloured according to the key (Database S1).

Table S1. Details of metagenomic sequencing of samples from SK152

Source	No. of libraries	Total unique reads	Total non-human reads	% specific reads	% duplicates	No. of specific unique reads	Mean read length (bp)	Genome coverage[1]
Upper 3rd left molar root & pulp	2	237,735,419	179,666,553	0.050	77	26,853	56	0.29
Upper 2nd right molar dentine/cementum	4	1,077,156,946	949,170,574	0.183	53	920,267	35	6.39
Upper 2nd right molar pulp	1	119,308,674	92,583,145	0.088	26	77,928	43	0.60
Femoral long bone	1	73,372,819	60,607,690	0.013	49	5,060	41	0.04
dental calculus (multiple teeth)	1	235,375,745	214,638,090	0.000	0	0	0	0
Total: 15 HiSeq runs	9	1,742,949,603	1,496,666,052	0.126	53	1,030,108	36	7.32

[1] Coverage relative to RKS4594

Table S2. Archaeological findings for SK152

Skeleton number	SK152
Sex	Female
Age (y)	19-24
Height (cm)	155.4 ± 4.5
Date (Archaeology)	1200 ± 50 AD
AMS result (Oxford)	1073 ± 79 AD
AMS result (DirectAMS)	911 ± 30 AD
possible birth place ($\delta^{18}O_{Carbon}$ from first molar)	northernmost inland areas of Scandinavia / Northwest Russia
travel history ($\delta^{18}O_{Carbon}$ from third molar)	travelled south at age of 3-14 and arrived Trondheim in early teens

Table S3. Statistics after clustering all metagenomes with MEGAHIT (*49*).

Data set	No. contigs	Total length (bp)	Max. length (bp)	N50	N90
All contigs	327,964	219,615,560	144,829	723	335
Contigs > 1kb	39,016	85,978,945	19,990	2,375	1,148

Table S4. Taxonomic assignments for 11 metagenome assembled genomes (MAGs)

Cluster	**Phylum**	**Class**	**Species**	**Environment associated with taxon**
	Taxon (Proportion of clustered reference genomes)			
C66	Proteobacteria (1.0).	*Deltaproteobacteria.* (1.0)	*Desulfomonile tiedjei* (1.0)	Sulfate reducer
C72	Firmicutes (1.0)	*Clostridia* (1.0)	*Eubacterium sulci*	Plaque associated
C18	Proteobacteria (1.0)	*Betaproteobacteria* (1.0)	*Acidovorax delafieldii* (1.0)	Plant associated
C25	Firmicutes (1.0)	*Clostridia* (1.0)	*Clostridium acetobutylicum* (1.0)	
C30	Firmicutes (0.5)	*Clostridia* (0.5)	*Pseudothermotoga elfii* (0.03)	
C69	Euryarchaeota (1.0)	*Methanomicrobia* (1.0)	*Methanosphaerula palustris*(1.0)	Methanogen
C1	Bacteroidetes (1.0)	*Bacteroidia* (1.0)	*Rikenella microfusus* (0.3)	
C32	Proteobacteria (0.9)	*Deltaproteobacteria* (0.9).	*Desulfomicrobium baculatum*(0.03)	
C2	Proteobacteria (1.0)	*Deltaproteobacteria* (1.0)	*Babela massiliensis* (1.0)	
C3	Chloroflexi (0.9)	*Thermomicrobia* (0.6)	*Sphaerobacter thermophiles* (0.1)	
C40	Proteobacteria (1.0)	*Deltaproteobacteria* (1.0)	*Desulfatiglans aniline* (1.0)	Sulfate reducer

Table S5. Single-stranded deamination rates for 11 MAGs plus Ragna and human DNA.

Sample	**Taxon**	**Mean**	**Standard Deviation**
Ragna	*Salmonella enterica* Paratyphi C	0.904	0.0103
C18	*Acidovorax delafieldii*	0.834	0.111
C72	*Eubacterium sulci*	0.540	0.022
Human	*Homo sapiens*	0.472	0.0004
C25	*Clostridium acetobutilicum*	0.217	0.007
C32	*Desulfomicrobium baculatum*	0.142	0.0057
C40	*Desulfatiglans aniline*	0.125	0.001
C69	*Methanosphaerula palustris*	0.0753	0.0015
C66	*Desulfomonile tiedjei*	0.0643	0.002
C3	*Sphaerobacter thermophiles*	0.0536	0.027
C2	*Babela massiliensis*	0.0452	0.002
C30	*Pseudothermotoga elfii*	0.0300	0.004
C1	*Rikenella microfusus*	0.026	0.001

Table S6. Genomic coverage of 11 MAGs by source and sequencing library (Lib).

Cluster	Dental calculus	Femoral long bone	Upper 2nd right molar dentine				Upper 2nd right molar pulp	Upper 3rd left molar	
			Lib1	Lib2	Lib3	Lib4		Lib1	Lib2
C1	1.01	0.621	4.64	0.74	0.015	0.053	1.24	4.26	18.2
C18	0.368	0.104	0.643	0.086	0.001	0.004	160	0.126	0.508
C2	0.125	0.086	0.542	0.072	0.002	0.005	103	0.146	0.698
C25	1.16	18	12.7	0.638	0.037	0.133	4.04	5.32	25.8
C3	0.454	0.475	1.24	0.203	0.008	0.024	17.5	1.09	6.94
C30	2.45	3.09	13	2.43	0.038	0.139	6.29	13.3	51.2
C32	3.4	0.26	18.7	2.56	0.051	0.168	1.4	0.618	2.82
C40	6.49	1.45	46.9	4.8	0.141	0.445	1.8	9.95	49
C66	2.69	0.125	14.7	1.52	0.054	0.174	0.996	1.63	7.57
C69	0.819	0.645	2.56	0.357	0.013	0.043	206	2.15	14.8
C72	127	0.033	0.066	0.005	0.0003	0.001	2.11	0.004	0.018

Table S7. Number of mutational and recombinational SNPs per clade within the Para C Lineage and Birkenhead

Clade	**No. Core SNPs**[1]	
	Mutation	**Recombination**
PC-1	3,983	75
PC-2	1,644	79
Modern Paratyphi C (w/o Ragna)	5,696	154
Paratyphi C	6,300	161
Ragna	263	5
Paratyphi C & Typhisuis	11,838	273
Choleraesuis var. Kunzendorf	2,905	829
Choleraesuis *sensu stricto*	1,203	94
Choleraesuis	7,334	1,286
Paratyphi C, Typhisuis, Choleraesuis	20,281	1,572
Para C Lineage	31,480	6,493
Para C Lineage plus Birkenhead	49,610	11,841
S. enterica subsp. I	711,009	-

[1] SNPs (Database S3) were extracted from core genomic regions in the Para C Lineage with RECHMM except for *S. enterica* subsp. I, where they correspond to all SNPs within the concatenated core genes.

Table S8. Summary properties of 219 bacterial strains plus Ragna in the Para C Lineage.

Categories	Count
Lineages	
Choleraesuis var Kunzendorf	73
Choleraesuis *sensu stricto*	16
Choleraesuis CS-3	1
Paratyphi C PC-1	62
Paratyphi C PC-2	57
Ragna	1
Typhisuis TS-1	6
Typhisuis TS-2	3
Lomita	1
Collection years	
1200±50	1
1914-1954[1]	21
1955-1979	59
1980-1999	57
2000-2015	67
Continents	
Africa	65
Asia	18
Europe	73
North America	45
South America	2

[1] The exact year of isolation of strains M263 (NCTC20356) and M80 (NCTC348) from the Murray collection (*111*) is unknown except that they were isolated between 1917 and 1954.

Table S9. Summary statistics for pan-genomic contents of Ragna and 219 modern genomes of the Para C Lineage.

	Ragna	**Modern genomes of Para C Lineage**				
		Average	**Median**	**S.D.**	**Min.**	**Max.**
Average Read Length	36	171	201	67	37	263
Coverage	7.3	53	41	33	15	187
N50	-	159,880	175,531	58,783	27,234	302,177
Total length	4,674,067	4,789,681	4,795,755	77,844	4,508,504	4,985,340
Contigs (≥200 bp)	-	87	71	41	38	316
Genes	4,522	4,388	4,398	99	3,990	4,590
Pseudogenes	124	181	178	27	64	282
Core genes	3,463	3,419	3,424	23	3,327	3,501
Core pseudogenes	79	123	118	23	41	215
Accessory genes	1,059	969	968	89	646	1,178
Accessory pseudogenes	45	58	59	6	23	70
Prophages	5	5	5	1	1	6
Prophage genes	182	177	178	60	28	328
Prophage pseudogenes	7	9	10	2	2	13
Plasmids	2	1	1	1	0	3
Plasmid genes	70	72	55	53	0	273
Plasmid pseudogenes	5	5	5	2	0	16
ICEs	2	2	2	1	0	2
ICE genes	100	57	93	41	0	100
ICE pseudogenes	5	4	5	3	0	9
IMEs	4	4	4	0	3	6
IME genes	19	19	19	3	11	52
IME pseudogenes	0	0	0	0	0	2
Frameshifts	71	100	98	19	35	187
Stop codons	53	80	77	12	27	98

Table S10. Summary statistics for the pan-genome of the Para C Lineage (220 genomes).

	No. of Islands	**No. of genes**	**Total Length (bp)**	**Terminal nodes**		**Internal nodes**	
				Gain	**Loss**	**Gain**	**Loss**
Core genes	-	3,542	3,387,468	-	-	-	-
All accessory genes	227	3,123	2,436,825	63	130	37	81
Prophages							
Lambda supercluster	17	459	305,175	13	11	13	12
P2 supercluster	5	172	126,438	3	2	2	0
P1 supercluster	1	97	71,289	1	0	0	0
Mu supercluster	1	50	34,377	0	0	1	0
Others	8	21	15,364	2	3	0	3
Subtotal:	32	799	552,663	19	16	16	15
ICE	2	107	89,160	0	1	2	0
IME	16	86	65,079	7	3	4	4
Plasmids	37	1,377	1,073,457	31	11	10	3
Other islands	140	754	656,466	6	99	5	59

Table S11 Distribution across the Para C lineage of GIs and selected genes associated with host specificity, serum resistance, adhesion, central metabolism and multicellular behaviour.

Genes/GIs	Category	Description	Status	Citation
GI008	Pathogenicity island	SPI-6	Present in Para C Lineage.	Fig. S6
tcf	Host specificity / fimbriae	SPI-6 Typhi colonization factor	Deleted in Paratyphi C and Lomita	(*30*, *112*)
saf	fimbriae	SPI-6 *Salmonella* atypical fimbriae	Deleted in Typhisuis and Lomita	(68, 69)
GI107	Pathogenicity island	SPI-7	Present in all PC-1, PC-2 and Ragna. 4.5 kb internal deletion of 5 genes in PC-1.	Fig. 2
tvi	Prevents Complement Receptor 3-mediated clearance	SPI-7 Vi capsular polysaccharide	All Paratyphi C	(*26*, *113*)
GI010A	prophage	P22-like (lambda)	Present in Ragna and one other Paratyphi C genome and all but four Choleraesuis genomes	Fig. 2
GI025	prophage	ESSI-2-like (P2)	Present in Ragna and all but two Paratyphi C genomes	Fig. 2 (*114*)
GI081	prophage	SEN34 (Gifsy-2)	Absent in Typhisuis	Fig. 2
pltAB	Host specificity	Typhi toxin	Absent	(*16*)
gtgE	Host specificity	Rab32-Dependent Pathway to human restriction in Typhi	Pseudogene in Typhisuis	(*115*)
sopD2 (SPI-2)	Host specificity	Rab32-Dependent Pathway to human restriction in Typhi	Present with SNP variation	(*115*)
fimH	Host specificity / fimbriae	Host-specific adhesion	Deleted in Choleraesuis *sensu stricto*	(*25*)
fepE	Serum resistance	LPS length regulation	Pseudogene in TS-2	(*116*)
opvAB	Serum resistance	LPS length regulation	Present with SNP variation	(*117*)
pgtE	Serum resistance	Cleaves complement factors B and H	SCP-P1 (GI1076) inserts upstream in some PC-1 strains and increases virulence in a murine model	(*24*, *118*)
rck	Serum resistance	Binds to complement factor H	Deleted in both Choleraesuis and modern Paratyphi C	(*119*)
pagC	Serum resistance		Present with SNP variation	(*120*)
cbi, cbo	Central metabolism	Vitamin B12 biosynthesis	Present with SNP variation	(*22*, *121*)
pdu	Central metabolism	Propanediol utilization	Deleted in Choleraesuis	(*22*, *121*)
eut	Central metabolism	Ethanolamine utilisation	Pseudogene in Choleraesuis and Typhisuis	(*22*, *121*)
ttr, phs, asr	Central metabolism	Tetrathionate reduction	Pseudogene in Choleraesuis *sensu stricto* and TS-2	(*22*, *121*)
bcs	Multicellular Behaviour		Deleted in Choleraesuis var. Kunzendorf and disrupted in Typhisuis	(*122*)
katE	Multicellular Behaviour		Present with SNP variation	(*122*)

Table S12. Comparison of dating estimates for MRCA using ancient plus extant, and only extant strains.

Parameters			Median MRCA (CI95%)			Mutation rate		Model[1]					
								Strict			UCLD		
Ragna	Strains[2]	Subsample	Para C Lineage (w/o Lomita)	Paratyphi C	Modern Paratyphi C (w/o Ragna)	Median	CI 95%	Constant	Skyride	BDP-SS	Constant	Skyride	BDP-SS
1200	Para C Lineage (w/o Lomita)	1	3804 (1786-8062)	1222 (884-2093)	485 (241-969)	1.34E-7	4.82E-8 - 2.22E-7	-6845066	-6845084	-6845158	**-6845001**	-6845038	-6845047
		2	4081 (2040-7250)	1290 (899-2024)	595 (328-1001)	1.18E-7	5.63E-8 - 1.93E-7	-6832249	-6832264	-6832331	**-6832246**	-6832279	-6832293
	Paratyphi C (incl. Ragna)	1	-	1206 (1030-1432)	477 (383-611)	1.11E-7	8.40E-8 - 1.38E-7	**-6742493**	-6742523	-6742512	-6742654	-6742675	-6742665
		2	-	1553 (1255-1975)	689 (513-927)	7.79E-8	5.35E-8 - 1.00E-7	**-6744367**	-6744389	-6744389	-6744563	-6744571	-6744562
1073	Para C Lineage (w/o Lomita)	1	4066 (1758-8687)	1370 (1040-2307)	517 (250-1055)	1.29E-7	4.35E-8 - 2.22E-7	-6845060	-6845097	-6845151	**-6845003**	-6845046	-6845055
		2	4359 (2100-7601)	1437 (1056-2264)	624 (331-1085)	1.15E-7	5.36E-8 - 1.86E-7	-6832246	-6832277	-6832296	**-6832228**	-6832269	-6832337
	Paratyphi C (incl. Ragna)	1	-	1276 (1088-1504)	481 (382-614)	1.10E-7	8.23E-8 - 1.37E-7	**-6742498**	-6742520	-6742495	-6742670	-6742682	-6742670
		2	-	1611 (1298-2047)	688 (512-941)	7.80E-8	5.47E-8 - 1.03E-7	**-6744370**	-6744387	-6744381	-6744554	-6744577	-6744564
no Ragna	Modern Paratyphi C (w/o Ragna)	1	-	-	484 (384-617)	1.09E-7	8.16E-8 - 1.36E-7	/	/	/	/	/	/
		2	-	-	681 (507-922)	7.86E-8	5.39E-8 - 1.02E-7	/	/	/	/	/	/

NOTE: Bold face indicates the likeliest model for that subsample

[1] Strict -- strict clock model. UCLD -- uncorrelated relaxed clock model with an underlying lognormal distribution. Constant -- Constant population size model. Skyride -- Bayesian skyride model. BDP-SS -- General birth-death process model.

[2] Only genomes whose collection year was known were analysed.

Table S13. Dating estimates within Para C lineage and *Salmonella enterica* subspecies I

			BEAST (Paratyphi C only)				BEAST (Para C Lineage)					
			Subsample 1		**Subsample 2**		**Subsample 1**		**Subsample 2**		**Species Tree[2]**	
Lineages	**LSD**	**TempEst[1]**	**Median**	**CI95%**	**Median**	**CI95%**	**Median**	**CI95%**	**Median**	**CI95%**	**ASTRID**	**Concatenated**
PC-1	-	-	403	320-509	593	445-798	403	206-813	458	253-775	-	-
PC-2	-	-	422	332-529	599	448-806	318	182-627	370	202-638	-	-
Modern Paratyphi C (w/o Ragna)	1,076	141	477	382-611	689	513-927	485	241-969	595	328-1001	-	-
Paratyphi C	2,434	788	1,204	1031-1437	1,553	1255-1975	1,222	884-2093	1,290	899-2024	-	-
Paratyphi C, Typhisuis	-	-	-	-	-	-	3,198	1522-6946	3,140	1445-5596	-	-
Choleraesuis var. Kunzendorf	-	-	-	-	-	-	287	149-583	312	169-552	-	-
Choleraesuis *sensu stricto*	-	-	-	-	-	-	361	152-795	360	189-654	-	-
Choleraesuis	-	-	-	-	-	-	2,303	987-5054	2,610	1120-4971	-	-
Paratyphi C, Typhisuis, Choleraesuis	3,204	2,156	-	-	-	-	3,804	1786-8062	4,082	2040-7250	-	-
Para C Lineage	-	-	-	-	-	-	-	-	-	-	15,791	15,238
Para C Lineage + Birkenhead	-	-	-	-	-	-	-	-	-	-	32,110	32,248
S. enterica subsp. I	-	-	-	-	-	-	-	-	-	-	67,030	71,726

[1] Originally designated Path-O-Gen

[2] ASTRID – super-tree calculated by ASTRID (*61*) (Database S4). Concatenated – super-tree based on concatenated core genes (Database S5)

Table S14. Geographic inference of the sources of clades within the Para C Lineage.

Clade	Para C Lineage plus Birkenhead minus Lomita			Para C Lineage minus Ragna minus Lomita		
	BayesTraits MCMC[1]	RASP BBM	BayesTraits ML	BayesTraits MCMC	RASP BBM	BayesTraits ML
modern Paratyphi C	Africa:36.16%; **Europe:63.57%**	**Europe:98.28%**	Africa:36.16%; **Europe:63.57%**	-	-	-
Paratyphi C including Ragna	**Europe:99.73%**	**Europe:99.55%**	**Europe:96.04%**	Africa:34.25%; **Europe:65.75%**	**Europe:94.28%**	Africa:37.38%; **Europe:62.40%**
Typhisuis	Europe:56.43%; Americas:43.37%	**Europe:96.22%**	Europe:48.88%; Americas:48.63%	Europe:56.31%; Americas:43.49%	**Americas:87.20%**; Europe-Americas:7.53%	Europe:48.77%; Americas:48.77%
Paratyphi C & Typhisuis	**Europe:98.35%**	**Europe:98.94%**	Africa:12.66%; Asia:11.75%; **Europe:60.19%;** Americas:15.39%	**Europe:93.08%**	**Europe:94.51%**	Africa:26.65%; Asia:14.51%; Europe:42.97%; Americas:15.87%
Choleraesuis var. Kunzendorf	Asia:13.20%; **Europe:86.72%**	Asia:21.19%; **Europe:62.63%;** Asia-Europe:14.75%	Asia:14.98%; **Europe:83.71%**	Asia:12.96%; **Europe:86.96%**	Asia:50.54%; Europe:36.08%; Asia-Europe:10.47%	Asia:15.00%; **Europe:83.69%**
Choleraesuis *s.s.*	Asia:48.86%; Europe:16.24%; Americas:34.83%	Asia:9.87%; **Europe:76.89%;** Asia-Europe:10.76%	Asia:46.43%; Europe:10.01%; Americas:42.54%	Asia:48.25%; Europe:16.28%; Americas:35.38%	Asia:40.11%; Europe:45.58%; Asia-Europe:9.68%	Asia:46.39%; Europe:9.97%; Americas:42.65%
All Choleraesuis	Africa:46.12%; Asia:12.75%; Europe:40.06%	Africa:7.50%; **Europe:82.44%**; Africa-Europe:6.38%	Africa:46.88%; Asia:19.71%; Europe:23.83%; Americas:9.58%	Africa:46.15%; Asia:12.42%; Europe:40.32%	Africa:39.79%; Asia:6.34%; Europe:45.65%; Africa-Europe:5.58%	Africa:51.07%; Asia:19.43%; Europe:24.84%
Para C Lineage (w/o Lomita)	**Europe:99.16%**	**Europe:97.43%**	Africa:20.49%; Asia:12.38%; Europe:56.22%; Americas:10.91%	Africa:5.97%; **Europe:93.66%**	Africa:12.11%; **Europe:80.83%**	Africa:38.92%; Asia:13.70%; Europe:40.02%; Americas:7.36%

NOTE: Lomita was excluded due to the absence of geographical data. Bold font: Assignment probabilities of >60%

[1] See fig. S8.