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

Authors: Zhemin Zhou[1]*, Inge Lundstrøm[2], Alicia Tran-Dien[3], Sebastián Duchêne[4], Nabil-Fareed Alikhan[1], Martin J. Sergeant[1], Gemma Langridge[5]†, Anna K. Fotakis[2], Satheesh Nair[5]†, Hans K. Stenøien[6], Stian S. Hamre[7], Sherwood Casjens[8], Axel Christophersen[6], Christopher Quince[1], Nicholas R. Thomson[5], Francois-Xavier Weill[3], Simon Y. W. Ho[9], M. Thomas P. Gilbert[2,6]*, Mark Achtman[1]*

Affiliations:

[1]Warwick Medical School, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, UK.

[2]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen, Denmark.

[3]Unité des Bactéries Pathogènes Entériques, Institut Pasteur, Paris, France.

[4]Centre for Systems Genomics, University of Melbourne, Parkville, Victoria 3010, Australia.

[5]Wellcome Trust Sanger Institute, Cambridge, UK

[6]NTNU University Museum, N-7491 Trondheim, Norway

[7]Department of Archaeology, History, Cultural Studies and Religion, University of Bergen, Post Box 7805, 5020 Bergen, Norway

[8]Pathology Department, University of Utah School of Medicine, Salt Lake City, Utah 84112, U.S.A.

[9]School of Life and Environmental Sciences; University of Sydney; Sydney, NSW 2006, Australia

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

†Current addresses: GL: Molecular Microbiology, Norwich Medical School, University of East Anglia, Norwich, UK; SN: Gastrointestinal Bacteria Reference Unit, Public Health England, London, UK

Abstract: *Salmonella enterica* serovar Paratyphi C is the causative agent of enteric (paratyphoid) fever. While today a potentially lethal infection of humans that occurs in Africa and Asia, early 20th century observations in Eastern Europe suggest it may once have had a wider-ranging impact on human societies. We recovered a draft Paratyphi C genome from the 800-year-old skeleton of a young woman in Trondheim, Norway, who likely died of enteric fever. Analysis of this genome against a new, significantly expanded database of related modern genomes demonstrated that Paratyphi C is descended from the ancestors of swine pathogens, serovars Choleraesuis and Typhisuis, together forming the Para C Lineage. Our results indicate that Paratyphi C has been a pathogen of humans for at least 1,000 years, and may have evolved after zoonotic transfer from swine during the Neolithic period.

One Sentence Summary: The combination of an 800-year-old *Salmonella enterica* Paratyphi C genome with genomes from extant bacteria reshapes our understanding of this pathogen's origins and evolution.

Main Text:, According to historical records (*1*), humans have long been afflicted by bacterial infections. Yet comparative genomics of extant bacterial pathogens routinely estimate a time to the Most Recent Common Ancestor (tMRCA) of no more than a few centuries (*2*). In contrast, tMRCAs of millennia were estimated by other analyses that included ancient DNA (aDNA) (*3, 4*). One possibility to account for these apparent discrepancies is that the tMRCA of extant isolates only includes the currently existing crown lineage, and ignores older sub-lineages that are now extinct. To test the generality of this inference, we scanned metagenomic sequences from teeth and long bones of 33 skeletons who were buried in Trondheim, Norway between 1100 -1670 CE (*5*) (Fig. 1A,C). 266 *Salmonella enterica* reads were identified in one tooth from skeleton SK152. We reconstructed a genome (designated Ragna) from that *Salmonella* after additional metagenomic sequencing of nine libraries from SK152 derived from one bone, two teeth and dental calculus (table S1). We then provided an evolutionary framework for invasive salmonellosis, and its host specificity, by analyzing the ancient Ragna genome (fig. S1) within the context of ~50,000 modern genomes from *S. enterica* which we had assembled with the EnteroBase online resource (*6*).

SK152 is the skeleton of a 19-24 year old woman of 154±3 cm height who was buried in 1200±50 CE according to archaeological investigations (*5*). Calibrated radiocarbon (^{14}C) dating of two teeth estimated her burial as 100 or 200 years before this (fig. S2), a discrepancy that can be explained by the reservoir effect on radiocarbon dating when fish products were the predominant diet (*7*). Based on $\delta^{18}O_{carbon}$ isotopic measurements from her first and third molars, this woman likely migrated from the northernmost inland areas of Scandinavia or Northwest Russia during her childhood, and arrived in Trondheim by her early teens (table S2). The skeleton rested on a wooden plank (a symbolic coffin) in a grave filled and covered with anoxic, acidic and waterlogged wood chips and soil (Fig. 1B).

From the metagenomic reads, we also succeeded in reconstituting 11 clusters of contigs that correspond to nearly complete microbial genomes (Fig. 1D; fig. S3; tables S3-S4). Nine of these genomes likely reflect recent (*8*) soil contamination because their 5'-single-stranded deamination rates were low (DNA damage <0.22) *versus* values of 0.47 for human DNA and 0.9 for Ragna (table S5). Two of the nine genomes are from bacterial sulfate-reducers and a third consists of archaeal methanogens, typical of the microbes that would be common in the soil surrounding SK152. In contrast, the two other assembled genomes are likely to have been endogenous to this corpse since burial because they exhibited high levels of DNA damage. C72, a novel species of *Eubacterium*, was found almost exclusively in dental calculus (table S6), and may have been a major component of a biofilm associated with periodontal disease, as are other taxa of *Eubacterium* (*9*). C18 belongs to *Acidovorax*, which is associated with plant pathogens (*10*), and may have been introduced with the wood chips. None of the few assembled contigs related to *S. enterica* were greater than 1 kb in length, and we therefore reconstructed the Ragna genome by read mapping against its close relatives.

In order to identify close relatives of the Ragna genome, we first created phylogenetic trees of 2,964 genomes from EnteroBase that represent the genetic diversity of *S. enterica* subspecies *enterica*. We used the 711,009 SNPs (single-nucleotide polymorphisms) from 3,002 concatenated core genes (2.8 Mb) from these genomes (table S7) to infer a maximum-likelihood

tree (Fig. 2A). We also inferred an independent species tree from gene-by-gene alignments of each core gene. The topologies of the two trees were identical near the root and towards the tips, and at 75% of the intermediate branches. These trees were used for phylogenetic placement (MGPLACER) of the metagenomic reads from SK152; all reads specific to *S. enterica* were assigned to a monophyletic clade of serovar Paratyphi C genomes. That clade was in turn contained within the "Para C Lineage"; which includes related monophyletic clades of serovars Choleraesuis, and Typhisuis (*11*), plus one genome of the extremely rare serovar Lomita. We used two genomes of serovar Birkenhead as an outgroup for rooting trees of the Para C Lineage because they were the closest genetic relatives (Fig. 2).

Paratyphi C was first recognized in 1916 when it was isolated from Eastern European soldiers suffering from enteric fever (also called paratyphoid fever) (*12*). Enteric fever is a severe, occasionally fatal, illness that was also common at that time in northern Africa, South America and North America. It has since disappeared from Europe, except for occasional travelers returning from East Asia or Africa (*11*). The Ragna genome demonstrates that Paratyphi C bacteria were present in northern Norway 800 years ago, and the presence of Ragna sequences within both teeth and bones in a young adult suggests that SK152 died of septicemia associated with enteric fever. Paratyphi C only infects humans but other members of the Para C Lineage show different host specificities. Serovar Choleraesuis is associated with septicemia in swine (and occasionally humans) and Typhisuis with epidemic, chronic paratyphoid/caseous lymphadenitis in domestic pigs (*11, 13*). Although both serovars continue to cause disease in southern and eastern Asia, Choleraesuis is rare in Europe today, except in wild boar (*14*), and Typhisuis has been eradicated from European pigs through culling of infected herds.

The inference of tMRCA and phylogeographic history is best performed with bacterial strains that were collected from global sources over a broad temporal range. However, only 108 of >50,000 assembled *Salmonella* genomes in EnteroBase belonged to the Para C Lineage, and their geographical and temporal sources were very limited. We therefore combed the strain collection at the Institut Pasteur, Paris, for historical Para C Lineage strains from diverse sources, and sequenced 111 additional genomes, which resulted in a total of 219 modern Para C Lineage genomes from multiple continents and dated from 1914-2015 (table S8). These genomes defined sub-lineages within each of the serovars (Paratyphi C: PC-1, PC-2; Choleraesuis: CS Kunzendorf, CS *sensu stricto*, CS-3; Typhisuis: TS-1, TS-2) (Fig. 2B). The metagenomic reads from SK152 were then mapped (after initial processing and de-duplication) against all 219 genomes in order to identify all reads mapping to the pan-genome of the Para C Lineage. We found 1,030,108 unique Para C Lineage-specific reads in teeth (0.05-0.18% of all reads) or the femur (0.01%), but not in dental calculus (table S1). The Ragna reads covered 98.4% of a reference Paratyphi C genome (RKS4594) with a mean read depth of 7.3-fold. In order to avoid spurious SNP calls associated with DNA damage, we only called SNPs in the Ragna genome against the 95% of RKS4594 that was covered by at least two reads.

The evolutionary dynamics and selective pressures associated with local ecological interactions are often thought to result in variation of gene-content in microbes (*15*). We therefore anticipated that 800 years of evolution would have resulted in dramatically different gene content between the Ragna and modern Paratyphi C genomes, and expected even greater differences from the other serovars within the Para C Lineage. Surprisingly, this was not the case - of the 4,388±99 genes (total length 4.8±0.08 Mb) in a Para C Lineage genome, 78% were intact core genes (table S9, Fig. 2B), and only 604 core SNPs distinguished Ragna from the MRCA of modern Paratyphi

C (Fig. 3). Some of the core genes are universally present in the Para C Lineage plus Birkenhead, even though they belong to mobile genetic elements that are variably present in other *Salmonella*, e.g. the pathogenicity islands SPI-1 to SPI-4, SPI-6, SPI-9 and SPI-11 to SPI-14. Similarly, the virulence plasmid was present throughout the Para C Lineage except for Typhisuis sub-lineage TS-2. A further constant feature of the Para C Lineage was the absence of genes encoding typhoid toxin, which is thought to trigger enteric fever by serovars Typhi and Paratyphi A (*16*).

Other studies have indicated that microbial host adaptation is accompanied by the accumulation of pseudogenes (*17*). Pseudogene accumulation is thought to be due to the streamlining of genes that are no longer necessary for the infection of multiple hosts (*18*), but has also been proposed to provide an efficient mechanism for rewiring of transcriptional regulation (*19*). A median number of 40-60 pseudogenes was observed within the 2,964 representative *Salmonella* genomes. Twenty-five pseudogenes were inferred for the MRCA of the Para C Lineage and 69 for the MRCA of Paratyphi C, Choleraesuis plus Typhisuis (fig. S4), suggesting that the MRCA had not yet adapted to a particular host, or had only recently begun to do so. However, the splits into the individual serovars may well mark the beginnings of host adaptation because higher numbers of pseudogenes were inferred for the MRCAs of each of the serovars (Choleraesuis: 95; Paratyphi C: 116; Typhisuis: 181), and the number of pseudogenes continued to increase to modern times (fig. S4). It is possible that some of these pseudogenes contribute to host specificity, rather than simply representing functions that are not required for infection of those particular hosts.

We also attempted to identify mobile genetic elements that could account for the different host specificities of the individual serovars among the 3,901 genes in the accessory genome of the Para C Lineage. These accessory genes clustered together within 227 GIs (mobile elements and genomic islands), including 37 plasmids, 32 prophages, 16 IMEs (Integrative and Mobilizable Elements), SPI-5 to SPI-7, and two ICEs (Integrative and Conjugative Elements) (table S10). The GIs were acquired or lost on 311 independent occasions, of which at least 60% are unlikely to be important for host specificity because they were restricted to a single genome (table S10). Most of the other gains or losses are also unlikely to represent successful evolutionary changes in virulence or host specificity, because, as in other *Salmonella (20-22)*, they were restricted to individual sub-lineages within a serovar, and sister sub-lineages differing in the possession of those genes are also prevalent in invasive disease (Fig.2B, fig. S5, table S11). For example, the integration of a P22-like prophage designated SCP-P1 (GI076), upstream of *pgtE* in Paratyphi C has been reported to prevent opsonization via the PgtE omptin (*23*) and drastically reduce the minimal infectious dose for laboratory infections of mice (*24*). However, GI076 is absent from half of Paratyphi C genomes, including Ragna (Fig. 2B). Similarly, the *fimH102* allele of a type 1 fimbrial adhesin was reported to endow specific adhesion to porcine cells on serovar Choleraesuis (*25*). However, the entire *fimH* gene is absent in CS *sensu stricto*. Indeed, none of the virulence factors and GIs seemed likely to be consistently related to differential virulence or host specificity (table S11), with the notable exceptions of SPI-7 and SPI-6.

SPI-7 (GI107) is a pathogenicity island in serovars Typhi, Paratyphi C and Dublin which encodes the Vi capsular polysaccharide. Vi prevents the opsonization and clearance that is triggered by binding of the C3 component of complement to lipopolysaccharide (*26*), and might thereby promote enteric fever in humans. Our data showed that SPI-7 is present in all genomes of Paratyphi C, including Ragna, but absent from the other serovars in the Para C Lineage,

suggesting that SPI-7 was acquired prior to the expansion of serovar Paratyphi C, and might be associated with host specificity. A 5-gene deletion within SPI-7 was present throughout sub-lineage PC-1, but this does not affect the production of the Vi polysaccharide (*11*).

SPI-6 (GI008) is present throughout the entire Para C Lineage. It encodes a Type Six Secretion System (T6SS) as well as *Salmonella* atypical fimbriae (*saf*) and Typhi colonization factor (*tcf*) (figure S6). T6SS systems encode intracellular, inverted bacteriophage-tail-like structures that can inject lethal effectors (TaeX) into proximal eukaryotic and bacterial cells (*27*). For *S. enterica* serovar Typhimurium, SPI-6 contributes to gastrointestinal colonization and pathogenesis in mice (*28*) and chickens (*29*). Expression of *tcf* in *Escherichia coli* resulted in specific adhesion to human epithelial cells (*30*) and *saf* mutations within serovar Choleraesuis reduced gastrointestinal colonization of pigs (*31*). All T6SS genes within SPI-6 were present throughout the Para C Lineage but *tssL*, *tssM* and *vgrG* were absent in the outgroup, serovar Birkenhead, suggesting that the acquisition of an intact T6SS by horizontal gene transfer (HGT) during the evolution of the ancestor of the Para C Lineage may have facilitated its proficiency at gastrointestinal colonization and causing invasive disease. Interestingly, SPI-6 varied between the different serovars within the Para C Lineage (figure S6), and the various SPI-6 variants might also be associated with the differential host specificity of the serovar. Precise independent deletions between pairs of 43 bp-long, identical direct repeats within SPI-6 have cleanly excised the *tcf* genes within serovars Lomita and Paratyphi C, including Ragna, and the *saf* genes in serovar Typhisuis. SPI-6 in Lomita may represent an independent acquisition by HGT: The termini of the deletion of *saf* genes in Lomita differed from the termini in Typhisuis, and a Tae4/Tai4 toxin/antitoxin T6SS effector is found exclusively in Lomita's SPI-6. Furthermore, unlike other members of the Para C Lineage, the SNP density against a reference genome of Choleraesuis was fourfold higher (0.015) in SPI-6 than in the core genome. Thus, a parsimonious progression of these events would be an initial acquisition of a large SPI-6 by HGT after Lomita branched off, followed by successive deletions prior to the MRCAs of Paratyphi C and Typhisuis. However, alternative evolutionary paths in which each SPI-6 variant represents an independent HGT event cannot be excluded.

These observations immediately raised the question of evolutionary time-scales. How old are the individual serovars? And do these genomes allow an estimate of the minimal age for the infection of warm-blooded animals by *S. enterica*?

We dated key stages in the evolution of the Para C Lineage using a Bayesian phylogenetic approach (Fig. 3, tables S12-S13), and confirmed by randomization of the dates of isolation (figure S7) that there was a strong temporal signal in two individual sub-samples of each dataset. The mean molecular clock rate was the same (9.4×10^{-8} substitutions per site per year; CI: 6.9×10^{-8} - 1.2×10^{-7}) for Paratyphi C whether Ragna was included or excluded for these calculations (table S12). However, even though the dates of isolation of the extant bacteria spanned almost 100 years, and they had been obtained from multiple continents, analyses that excluded Ragna estimated the mean tMRCA as 583 y (CI: 383-927) whereas those that included the Ragna genome estimated 1,379 y (CI: 1,030-1,975) (table S13). Using the somewhat different age estimates for the Ragna genome provided by archaeological reconstructions (1200 CE) or C^{14} dating (1073 CE) had little influence on the estimated tMRCA (table S12). Thus, adding a single aDNA genome from an extinct sub-lineage to 219 genomes from a broad sample of extant bacterial isolates increased the tMRCA by twofold.

The maximum age of Paratyphi C is 3,169 y, the date when Paratyphi C split from its closest relatives, serovar Typhisuis. In turn, the estimated tMRCA for Paratyphi C, Typhisuis plus Choleraesuis was 3,943 y, and that ancestor evolved in Europe according to three independent Bayesian and maximum-likelihood methods (figure S8, table S14). We note that a European ancestry is consistent with the existence of Ragna in northern Norway 800 years ago, and of other Paratyphi C genomes in Mexico in 1,545 CE that may have been imported to the New World by Europeans (*32*). We also note that serovars Choleraesuis and Typhisuis infect swine, and that the tMRCA dates overlap with the Neolithic domestication of pigs from wild boar in Europe (*33*) (Fig. 3). It is therefore possible that this invasive sub-lineage evolved in swine or in boar, and that Paratyphi C represents a subsequent host adaptation to humans by a zoonotic pathogen of domesticated animals. This proposal is eminently testable through examining aDNA from boar and swine in Europe for genomic traces of DNA related to the Para C Lineage.

We calculated increasingly more speculative estimates of the dates of earlier branch points (table S13) by extrapolating the inferred mutation rates from the Bayesian analyses to the species tree estimated by maximum likelihood. The estimated tMRCA of the entire Para C Lineage including Lomita was ~15.500 y and the split from Birkenhead was ~32,000 y. However, because these estimates are extrapolations over long time periods, they are best considered to represent ballpark figures until substantiation has been provided by additional aDNA analyses. It is not currently feasible to distinguish mutational SNPs from recombinational SNPs at the species level in organisms which recombine as much as *S. enterica.* We therefore simply note that the estimated age of *S. enterica* subspecies *enterica* was ~70,000 y on the basis of all core SNPs, in order to provide an initial, flawed estimate of its age that will need correction.

Our analyses indicate that numerous recent estimates of tMRCAs based on comparative genomics of extant bacteria will need revision as additional genomes are derived from aDNA studies. Our results also indicate that both the core and accessory genome of bacterial pathogens are unexpectedly stable over millennia, and that much of the dramatic variation between extant genomes represents transient genetic fluctuation whose evolutionary relevance to ecological fitness is uncertain. Finally, these analyses provide a paradigm for how combining genomes from both extant and ancient bacteria can facilitate the reconstruction of their evolutionary history.

References and Notes:

1. W. H. McNeill, *Plagues and peoples* (Anchor, New York, ed. 1, 1976).

2. M. Achtman, How old are bacterial pathogens? *Proc. Biol. Sci.* **283**, 1836 (2016).

3. V. J. Schuenemann *et al.*, Genome-wide comparison of medieval and modern *Mycobacterium leprae*. *Science* **341**, 179 (2013).

4. S. Rasmussen *et al.*, Early divergent strains of *Yersinia pestis* in Eurasia 5,000 years ago. *Cell* **163**, 571 (2015).

5. A.-L. Forsåker, H. Göthberg, *Stratigrafisk analyse, delfelt FJ, FN og FW* (Fortiden i Trondheim bygrunn: Folkebibliotektomten, Riksantikvarens utgravningskontor for Trondheim, Trondheim, 1986).

6. Materials and methods are available as supplementary materials at the Science website.

7. M. Stuiver, T. F. Braziunas, Modeling atmospheric ^{14}C influences and ^{14}C ages of marine samples to 10,000 BC. *Radiocarbon* **35**, 137 (1993).

8. L. Kistler *et al.*, Thermal age, cytosine deamination and the veracity of 8,000 year old wheat DNA from sediments. http://biorxiv.org/content/early/2015/11/18/032060 (2015).

9. J. Downes *et al.*, Characterisation of *Eubacterium*-like strains isolated from oral infections. *J. Med. Microbiol.* **50**, 947 (2001).

10. M. Fegan, "Plant pathogenic members of the genera *Acidovorax* and *Herbaspirillum*" in *Plant-Associated Bacteria*, Springer Netherlands, 2006), pp. 671-702.

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

12. L. Hirschfeld, A new germ of paratyphoid. *Lancet* **193**, 296 (1919).

13. S. Uzzau *et al.*, Host adapted serotypes of *Salmonella enterica*. *Epidemiol Infect* **125**, 229 (2000).

14. U. Methner, M. Heller, H. Bocklisch, *Salmonella enterica* subspecies *enterica* serovar Choleraesuis in a wild boar population in Germany. *Eur J Wildl Res* **56**, 493 (2009).

15. O. X. Cordero, M. F. Polz, Explaining microbial genomic diversity in light of evolutionary ecology. *Nat. Rev. Microbiol.* **12**, 263 (2014).

16. J. E. Galan, Typhoid toxin provides a window into typhoid fever and the biology of *Salmonella* Typhi. *Proc. Natl. Acad. Sci. U. S. A* **113**, 6338 (2016).

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

18. A. Bäumler, F. C. Fang, Host specificity of bacterial pathogens. *Cold Spring Harb Perspect Med* **3**, a010041 (2013).

19. A. K. Hottes *et al.*, Bacterial adaptation through loss of function. *PLoS. Genet.* **9**, e1003617 (2013).

20. Z. Zhou *et al.*, Transient Darwinian selection in *Salmonella enterica* serovar Paratyphi A during 450 years of global spread of enteric fever. *Proc. Natl. Acad. Sci. U. S. A* **111**, 12199 (2014).

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

22. G. C. Langridge *et al.*, Patterns of genome evolution that have accompanied host adaptation in *Salmonella*. *Proc. Natl. Acad. Sci. U. S. A* **112**, 863 (2015).

23. R. Riva, T. K. Korhonen, S. Meri, The outer membrane protease PgtE of *Salmonella enterica* interferes with the alternative complement pathway by cleaving factors B and H. *Front Microbiol* **6**, 63 (2015).

24. Q. H. Zou *et al.*, SPC-P1: a pathogenicity-associated prophage of *Salmonella* paratyphi C. *BMC. Genomics* **11**, 729 (2010).

25. M. Yue *et al.*, Allelic variation contributes to bacterial host specificity. *Nat. Commun.* **6**, 8754 (2015).

26. R. P. Wilson *et al.*, The Vi capsular polysaccharide prevents complement receptor 3-mediated clearance of *Salmonella enterica* serotype Typhi. *Infect. Immun.* **79**, 830 (2011).

27. A. Hachani, T. E. Wood, A. Filloux, Type VI secretion and anti-host effectors. *Curr. Opin. Microbiol* **29**, 81 (2016).

28. T. G. Sana *et al.*, *Salmonella* Typhimurium utilizes a T6SS-mediated antibacterial weapon to establish in the host gut. *Proc. Natl. Acad. Sci. U. S. A* **113**, E5044 (2016).

29. D. Pezoa *et al.*, The type VI secretion system encoded in SPI-6 plays a role in gastrointestinal colonization and systemic spread of *Salmonella enterica* serovar Typhimurium in the chicken. *PLoS ONE* **8**, e63917 (2013).

30. J. M. Leclerc *et al.*, Regulation and production of Tcf, a cable-like fimbriae from *Salmonella enterica* serovar Typhi. *Microbiology* **162**, 777 (2016).

31. S. C. Carnell *et al.*, Role in virulence and protective efficacy in pigs of *Salmonella enterica* serovar Typhimurium secreted components identified by signature-tagged mutagenesis. *Microbiology* **153**, 1940 (2007).

32. Å. J. Vågene *et al.*, *Salmonella enterica* genomes recovered from victims of a major 16th century epidemic in Mexico. http://biorxiv.org/content/early/2017/02/08/106740 (2017).

33. B. Krause-Kyora *et al.*, Use of domesticated pigs by Mesolithic hunter-gatherers in northwestern Europe. *Nat Commun* **4**, 2348 (2013).

34. J. Alneberg *et al.*, Binning metagenomic contigs by coverage and composition. *Nat Methods* **11**, 1144 (2014).

35. A. Christophersen, E. Jondell, O. Marstein, S. W. Nordeide, I. Reed, *Utgravning, kronologi og bebyggelsesutvikling* (Fortiden i Trondheim bygrunn: FolkebibliotektomtenNordeide, S.W, Riksantikvarens utgravningskontor for Trondheim, Trondheim, 1988).

36. A. Christophersen, S. W. Nordeide, *Kaupangen ved Nidelva : 1000 a°rs byhistorie belyst gjennom de arkeologiske undersøkelsene pa° Folkebibliotekstomten i Trondheim 1973-1985* (Riksantikvarens skrifter, Riksantikvaren, Oslo, 1994).

37. J. I. McKinley, "Compiling a skeletal inventory: disarticulated and co-mingled remains" in *Guidelines to the standards for recording human remains*, M. Brickley, J. I. McKinley, Eds., CIfA Papers, Southampton, 2004), vol. 7, pp. 14-17.

38. M. Cox, "Ageing adults from the skeleton" in *Human osteology: in archaeology and forensic science*, M. Cox, S. Mays, Eds., Cambridge University Press, London, 2000), pp. 61-82.

39. S. Mays, M. Cox, "Sex determination in skeletal remains" in *Human osteology: in archaeology and forensic science*, M. Cox, S. Mays, Eds., Cambridge University Press, London, 2000), pp. 117-130.

40. P. Skoglund, J. Storå, A. Götherström, M. Jakobsson, Accurate sex identification of ancient human remains using DNA shotgun sequencing. *Journal of Archaeological Science* **40**, 4477 (2013).

41. C. Ramsey, T. Higham, P. Leach, Towards high-precision AMS: progress and limitations. *Radiocarbon* **46**, 17 (2016).

42. C. B. Ramsey, T. F. Higham, D. C. Owen, A. W. G. Pike, R. E. M. Hedges, Radiocarbon dates from the Oxford AMS system: archaeometry datelist 31. *Archaeometry* 1 (2016).

43. V. Daux *et al.* Oxygen isotope fractionation between human phosphate and water revisited. Journal of Human Evolution 55, 1138-1147. 2008.

44. C. Chenery, V. Pashley, A. Lamb, H. Sloane, J. Evans. The oxygen isotope relationship between the phosphate and structural carbonate fractions of human bioapatite. Rapid Communications in Mass Spectrometry 26, 309-319. 2012. .

45. S. S. Hamre, V. Daux. Stable oxygen isotope evidence for mobility in medieval and post-medieval Trondheim, Norway. Journal of Archaeological Science: Reports 8, 416-425. 2016.

46. N. Rohland, M. Hofreiter, Ancient DNA extraction from bones and teeth. *Nat Protoc.* **2**, 1756 (2007).

47. M. E. Allentoft *et al.*, Population genomics of Bronze Age Eurasia. *Nature* **522**, 167 (2015).

48. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics.* **26**, 589 (2010).

49. D. Li, C. M. Liu, R. Luo, K. Sadakane, T. W. Lam, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics.* **31**, 1674 (2015).

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

51. F. D. Ciccarelli *et al.*, Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283 (2006).

52. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772 (2013).

53. S. Capella-Gutierrez, J. M. Silla-Martinez, T. Gabaldon, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics.* **25**, 1972 (2009).

54. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).

55. K. A. Jolley *et al.*, Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology* **158**, 1005 (2012).

56. H. Jonsson, A. Ginolhac, M. Schubert, P. L. Johnson, L. Orlando, mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics.* **29**, 1682 (2013).

57. H. Cadillo-Quiroz, J. B. Yavitt, S. H. Zinder, *Methanosphaerula palustris* gen. nov., sp. nov., a hydrogenotrophic methanogen isolated from a minerotrophic fen peatland. *Int. J. Syst Evol Microbiol* **59**, 928 (2009).

58. F. E. Rey *et al.*, Metabolic niche of a prominent sulfate-reducing human gut bacterium. *Proc. Natl. Acad. Sci. U. S. A* **110**, 13582 (2013).

59. F. Nakazawa *et al.*, Taxonomic characterization of *Mogibacterium diversum* sp. nov. and *Mogibacterium neglectum* sp. nov., isolated from human oral cavities. *Int. J. Syst Evol Microbiol* **52**, 115 (2002).

60. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics.* **30**, 1312 (2014).

61. P. Vachaspati, T. Warnow, ASTRID: Accurate Species TRees from Internode Distances. *BMC. Genomics* **16 Suppl 10**, S3 (2015).

62. V. Lefort, R. Desper, O. Gascuel, FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. *Mol. Biol. Evol.* **32**, 2798 (2015).

63. S. Mirarab, T. Warnow, ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics.* **31**, i44 (2015).

64. N. Joshi, J. Fass, Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33). https://github.com/najoshi/sickle (2016).

65. F. C. Del, S. Scalabrin, M. Morgante, F. M. Giorgi, An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS. One.* **8**, e85024 (2013).

66. A. Bankevich *et al.*, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455 (2012).

67. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics.* **27**, 2987 (2011).

68. D. E. Wood, S. L. Salzberg, Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

69. C. Luo *et al.*, ConStrains identifies microbial strains in metagenomic datasets. *Nat. Biotechnol.* **33**, 1045 (2015).

70. G. L. Kay *et al.*, Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nat Commun* **6**, 6717 (2015).

71. B. Bushnell, BBMap short read aligner. http://sourceforge.net/projects/bbmap (2016).

72. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357 (2012).

73. H. Li *et al.*, The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078 (2009).

74. M. Hamada, E. Wijaya, M. C. Frith, K. Asai, Probabilistic alignments with quality scores: an application to short-read mapping toward accurate SNP/indel detection. *Bioinformatics.* **27**, 3085 (2011).

75. G. Benson, Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573 (1999).

76. R. C. Edgar, PILER-CR: fast and accurate identification of CRISPR repeats. *BMC. Bioinformatics.* **8**, 18 (2007).

77. A. D. Leache, B. L. Banbury, J. Felsenstein, A. N. de Oca, A. Stamatakis, Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for Inferring SNP phylogenies. *Syst. Biol.* **64**, 1032 (2015).

78. M. Pagel, Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proc. R. Soc. B.* **255**, 37 (1994).

79. A. J. Drummond, M. A. Suchard, D. Xie, A. Rambaut, Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* **29**, 1969 (2012).

80. V. N. Minin, E. W. Bloomquist, M. A. Suchard, Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol.* **25**, 1459 (2008).

81. T. Stadler, Sampling-through-time in birth-death trees. *J. Theor. Biol.* **267**, 396 (2010).

82. W. Xie, P. O. Lewis, Y. Fan, L. Kuo, M. H. Chen, Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.* **60**, 150 (2011).

83. C. Ramsden, F. L. Melo, L. M. Figueiredo, E. C. Holmes, P. M. Zanotto, High rates of molecular evolution in hantaviruses. *Mol. Biol. Evol.* **25**, 1488 (2008).

84. S. Duchene, D. Duchene, E. C. Holmes, S. Y. Ho, The performance of the date-randomization test in phylogenetic analyses of time-structured virus data. *Mol. Biol. Evol.* **32**, 1895 (2015).

85. A. Rambaut, T. T. Lam, C. L. Max, O. G. Pybus, Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016).

86. R Development Core Team. R: a language and environment for statistical computing. (2.10.1). 2004.

87. T. H. To, M. Jung, S. Lycett, O. Gascuel, Fast dating using least-squares criteria and algorithms. *Syst. Biol.* **65**, 82 (2016).

88. C. H. Langley, W. M. Fitch, An examination of the constancy of the rate of molecular evolution. *J. Mol. Evol.* **3**, 161 (1974).

89. J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635 (2016).

90. M. Pagel, The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Syst Biol* **48**, 612 (1999).

91. M. Pagel, A. Meade, D. Barker, Bayesian estimation of ancestral character states on phylogenies. *Syst. Biol.* **53**, 673 (2004).

92. Y. Yu, A. J. Harris, C. Blair, X. He, RASP (Reconstruct Ancestral State in Phylogenies): a tool for historical biogeography. *Mol. Phylogenet. Evol.* **87**, 46 (2015).

93. Q. Li *et al.*, Complete genome sequence of *Salmonella enterica* serovar Choleraesuis vaccine strain C500 Attenuated by chemical mutation. *Genome Announc.* **2**, (2014).

94. W. Q. Liu *et al.*, *Salmonella paratyphi* C: genetic divergence from *Salmonella choleraesuis* and pathogenic convergence with *Salmonella typhi*. *PLoS. One.* **4**, e4510 (2009).

95. C. H. Chiu *et al.*, The genome sequence of *Salmonella enterica* serovar Choleraesuis, a highly invasive and resistant zoonotic pathogen. *Nucleic Acids Res.* **33**, 1690 (2005).

96. T. Seemann, Prokka: rapid prokaryotic genome annotation. *Bioinformatics.* **30**, 2068 (2014).

97. R. C. Edgar, Search and clustering orders of magnitude faster than BLAST. *Bioinformatics.* **26**, 2460 (2010).

98. D. L. Applegate, R. E. Bixby, V. Chvátal, W. J. Cook, *The Traveling Salesman Problem - A Computational Study* (Princeton Series in Applied Mathematics, Princeton University Press, New Jersey, 2006).

99. A. M. Eren *et al.*, Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ.* **3**, e1319 (2015).

100. D. Arndt *et al.*, PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* **44**, W16 (2016).

101. S. R. Casjens, J. H. Grose, Contributions of P2- and P22-like prophages to understanding the enormous diversity and abundance of tailed bacteriophages. *Virology* **496**, 255 (2016).

102. S. H. Yoon, Y. K. Park, J. F. Kim, PAIDB v2.0: exploration and analysis of pathogenicity and resistance islands. *Nucleic Acids Res.* **43**, D624 (2015).

103. C. J. Blondel, J. C. Jimenez, I. Contreras, C. A. Santiviago, Comparative genomic analysis uncovers 3 novel loci encoding type six secretion systems differentially distributed in *Salmonella* serotypes. *BMC Genomics* **10**, 354 (2009).

104. G. S. Vernikos, J. Parkhill, Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics.* **22**, 2196 (2006).

105. J. R. Elder *et al.*, The *Salmonella* pathogenicity island 13 contributes to pathogenesis in streptomycin pre-treated mice but not in day-old chickens. *Gut Pathog.* **8**, 16 (2016).

106. D. H. Shah *et al.*, Identification of *Salmonella* Gallinarum virulence genes in a chicken infection model using PCR-based signature-tagged mutagenesis. *Microbiology* **151**, 3957 (2005).

107. J. A. Fuentes, N. Villagra, M. Castillo-Ruiz, G. C. Mora, The *Salmonella* Typhi *hlyE* gene plays a role in invasion of cultured epithelial cells and its functional transfer to *S.* Typhimurium promotes deep organ infection in mice. *Res. Microbiol.* **159**, 279 (2008).

108. P. Siguier, J. Perochon, L. Lestrade, J. Mahillon, M. Chandler, ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* **34**, D32 (2006).

109. A. Carattoli *et al.*, *In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* **58**, 3895 (2014).

110. J. Guglielmini *et al.*, Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. *Nucleic Acids Res.* **42**, 5715 (2014).

111. K. S. Baker *et al.*, The Murray collection of pre-antibiotic era *Enterobacteriacae*: a unique research resource. *Genome Med.* **7**, 97 (2015).

112. A. Folkesson *et al.*, Multiple insertions of fimbrial operons correlate with the evolution of *Salmonella* serovars responsible for human disease. *Mol. Microbiol.* **33**, 612 (1999).

113. T. Wangdi *et al.*, The Vi capsular polysaccharide enables *Salmonella enterica* serovar typhi to evade microbe-guided neutrophil chemotaxis. *PLoS. Pathog.* **10**, e1004306 (2014).

114. Y. D. Lee, H. I. Chang, J. H. Park, Complete genomic sequence of virulent *Cronobacter sakazakii* phage ESSI-2 isolated from swine feces. *Arch. Virol.* **156**, 721 (2011).

115. S. Spano, X. Gao, S. Hannemann, M. Lara-Tejero, J. E. Galan, A bacterial pathogen targets a host Rab-Family GTPase defense pathway with a GAP. *Cell Host. Microbe* **19**, 216 (2016).

116. R. W. Crawford *et al.*, Loss of very-long O-antigen chains optimizes capsule-mediated immune evasion by *Salmonella enterica* serovar Typhi. *MBio.* **4**, (2013).

117. I. Cota *et al.*, Epigenetic Control of *Salmonella enterica* O-Antigen Chain Length: A Tradeoff between Virulence and Bacteriophage Resistance. *PLoS. Genet.* **11**, e1005667 (2015).

118. P. Ramu *et al.*, The surface protease PgtE of *Salmonella enterica* affects complement activity by proteolytically cleaving C3b, C4b and C5. *FEBS Lett.* **581**, 1716 (2007).

119. D. K. Ho *et al.*, Functional recruitment of human complement inhibitor C4B-binding protein to outer membrane protein Rck of *Salmonella*. *PLoS. One.* **6**, e27546 (2011).

120. M. Nishio, N. Okada, T. Miki, T. Haneda, H. Danbara, Identification of the outer-membrane protein PagC required for the serum resistance phenotype in *Salmonella enterica* serovar Choleraesuis. *Microbiology* **151**, 863 (2005).

121. S. P. Nuccio, A. J. Baumler, Comparative analysis of *Salmonella* genomes identifies a metabolic network for escalating growth in the inflamed gut. *MBio.* **5**, e00929 (2014).

122. L. A. Singletary *et al.*, Loss of multicellular behavior in epidemic African nontyphoidal *Salmonella enterica* Serovar Typhimurium ST313 strain D23580. *MBio.* **7**, e02265 (2016).

Acknowledgments: All genomes referred to here are available from the publicly available workspaces entitled "rST representatives" (2964 genomes) and "Para C Lineage" in the

Salmonella database in EnteroBase (http://EnteroBase.warwick.ac.uk). Interactive versions of Fig. 2 can be found at https://enterobase.warwick.ac.uk/anvi_public/ParaC_pangenome. Data supplements are permanently stored at the University of Warwick (http://wrap.warwick.ac.uk/85593) and non-human metagenomic data at https://sid.erda.dk/wsgi-bin/ls.py?share_id=E56xgi8CEl. EnteroBase (BBSRC BB/L020319/1) was developed by N-F.A., M.J.S. and Z.Z. (equal contributions) under guidance by M.A. Additional grant support was from the Wellcome Trust (202792/Z/16/Z), Danish National Research Foundation (DNRF 94) and Lundbeckfonden (R52-5062). For other attributions see Supplementary Material.

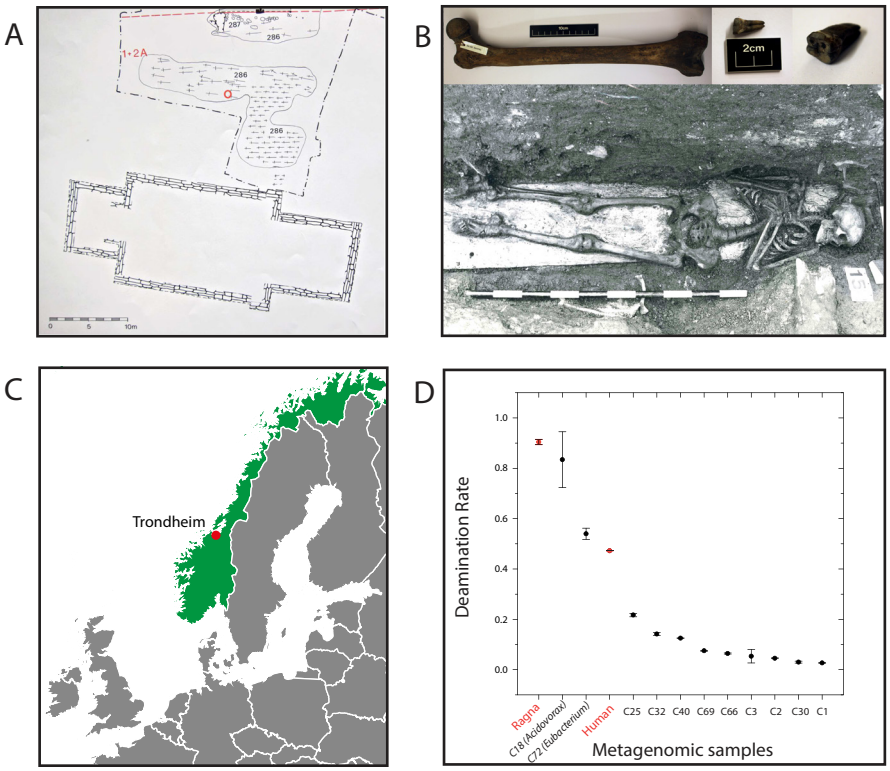

Fig. 1. Geographic, archaeological and metagenomic features of skeleton SK152. A) Excavation site (Folkebibilotekstomten, 1973-1985) of the church cemetery of St. Olav in Trondheim, Norway. The burial location of SK152 (red circle) belongs to a building phase which has been dated archaeologically (5) to 1200 CE (range 1150-1250). B) Femoral long bone plus two teeth from which *Salmonella* DNA was extracted (Top) and entire skeleton (bottom). C) Map of Europe surrounding Norway (green). D) Deamination rate for metagenomic reads in the *Salmonella* Paratyphi C Ragna genome, human DNA and 11 sets of assemblies identified by CONCOCT (34) (C1 through C72). The high deamination of C18 (*Acidovorax*) and C72 (*Eubacterium*) shows that they are as old as reads from humans or Ragna while the other assemblies likely represent modern environmental bacteria.

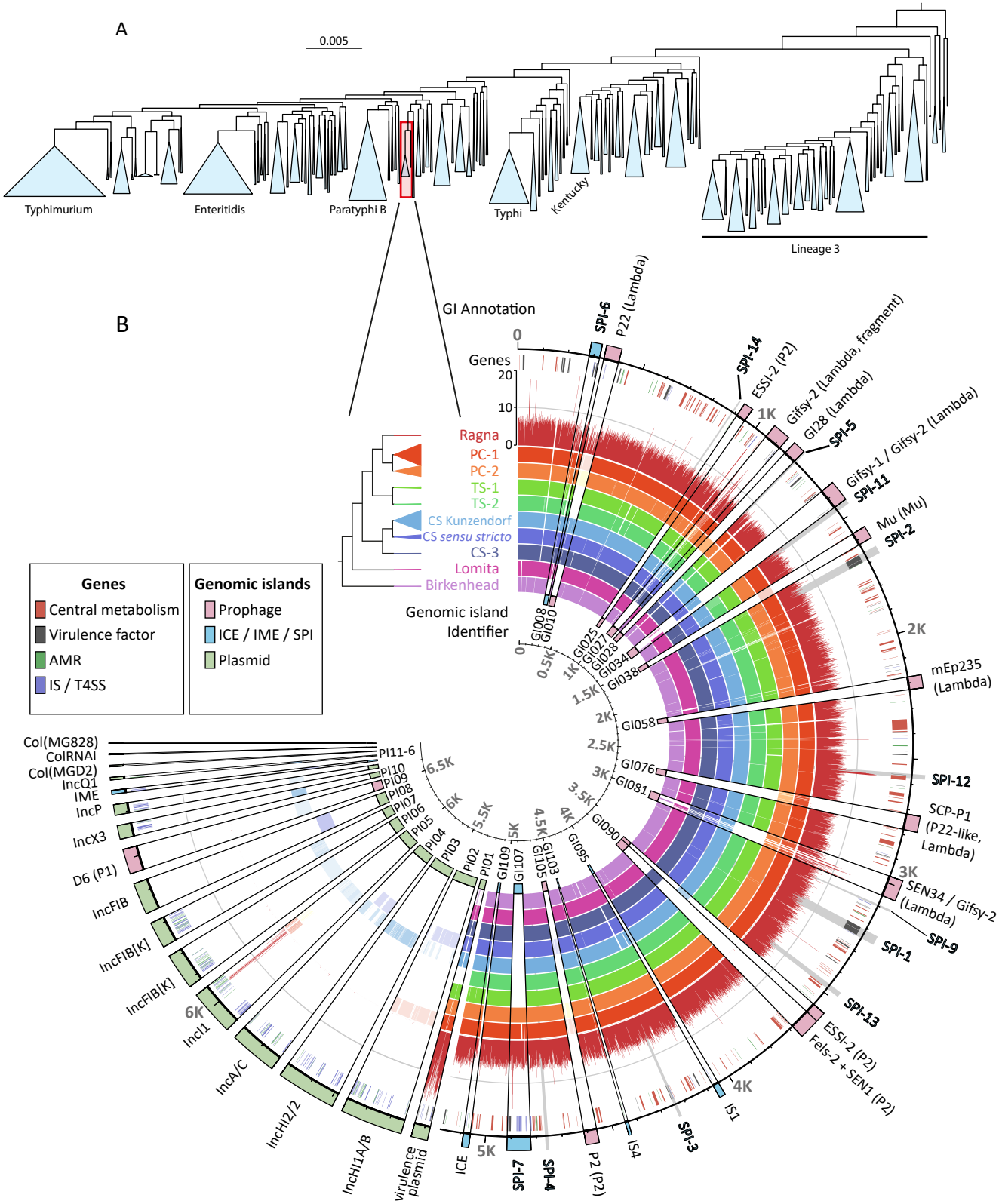

Fig. 2. Genomic phylogenies of *S. enterica* and the Para C Lineage. A) Maximum-likelihood phylogeny of representative genomes of *S. enterica* subspecies *enterica*. Each genome is a random representative of one of the 2,964 sequence types that had been defined by ribosomal multi-locus sequence typing (rMLST). The Para C Lineage and other lineages containing numerous genomes are collapsed into blue triangles, and common serovars that are associated

with several of those lineages are indicated under the tree. Red rectangle: Para C Lineage plus Birkenhead. B) Pan-genomic contents for 220 genomes of the Para C Lineage, including Ragna. Genes are ordered as in Database S1 (primary ticks every 500 genes). Circles (inner to outer): Circle 1. Sixteen major, variably present chromosomal genomic islands (GI008-GI109; Database S2) followed by sixteen cytoplasmic elements (PI01-PI16), color-coded as in the right key. Circles 2-10: The frequency of presence or absence of each gene per sub-lineage within the phylogram is indicated by color opacity. Circle 11: Coverage of aDNA reads per gene within Ragna (Scale 0-20 at 12:00). Circle 12: Genes color-coded as in the left key. Circle 13: Traditional annotations of GIs, PIs and other invariably present genomic elements, such as SPIs (gray wedges).

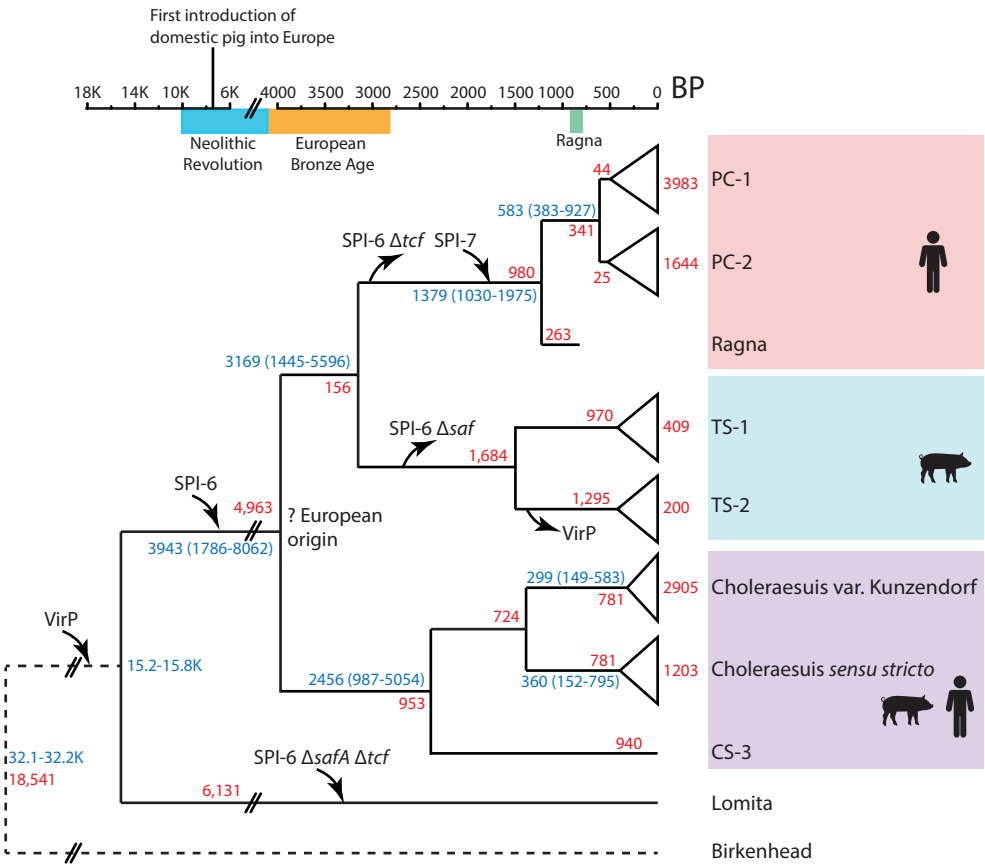

Fig. 3. Cartoon of the evolutionary history of the Para C Lineage on the historical timeframe BP for human history in Europe (top). The phylogram indicates the acquisition of the virulence plasmid (VirP), SPI-6 and SPI-7 by inward arrows and deletions of parts of SPI-6 by outwards arrows (see fig. S6). Blue numbers on branches indicate the average of the median date from two independent subsamples plus the 95% extreme credible intervals of both samples (tables S11-S13). Red numbers indicate numbers of non-recombinant core single nucleotide variants. The host specificity of the individual serovars is indicated at the right.

Supplementary Materials:

Materials and Methods

Figures S1-S9

Tables S1-S14

External Databases S1-S12

References (35 - 120)