ARTICLES
https://doi.org/10.1038/s41559-017-0446-6
nature ecology & evolution

Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico

Åshild J. Vågene[1,2], Alexander Herbig[1,2*], Michael G. Campana[3,4,7], Nelly M. Robles García[5], Christina Warinner[1], Susanna Sabin[1], Maria A. Spyrou[1,2], Aida Andrades Valtueña[1], Daniel Huson[6], Noreen Tuross[3*], Kirsten I. Bos[1,2*] and Johannes Krause[1,2*]

Indigenous populations of the Americas experienced high mortality rates during the early contact period as a result of infectious diseases, many of which were introduced by Europeans. Most of the pathogenic agents that caused these outbreaks remain unknown. Through the introduction of a new metagenomic analysis tool called MALT, applied here to search for traces of ancient pathogen DNA, we were able to identify *Salmonella enterica* in individuals buried in an early contact era epidemic cemetery at Teposcolula-Yucundaa, Oaxaca in southern Mexico. This cemetery is linked, based on historical and archaeological evidence, to the 1545–1550 CE epidemic that affected large parts of Mexico. Locally, this epidemic was known as '*cocoliztli*', the pathogenic cause of which has been debated for more than a century. Here, we present genome-wide data from ten individuals for *Salmonella enterica* subsp. *enterica* serovar Paratyphi C, a bacterial cause of enteric fever. We propose that *S.* Paratyphi C be considered a strong candidate for the epidemic population decline during the 1545 *cocoliztli* outbreak at Teposcolula-Yucundaa.

Infectious diseases introduced to the New World following European contact led to successive outbreaks in many regions of the Americas that continued well into the nineteenth century. These often caused high mortality and, therefore, contributed a central, and often underappreciated, influence on the demographic collapse of many indigenous populations[1–4]. Population declines linked to regionally specific epidemics are estimated to have reached as high as 95%[3], and their genetic impact based on recent population-based studies of ancient and modern human exome and mitochondrial data attests to their scale[5,6]. One hypothesis posits that the increased susceptibility of New World populations to Old World diseases facilitated European conquest, whereby rapidly disseminating diseases severely weakened indigenous populations[2], in some cases even before European presence in the region[2,7]. Well-characterized infections, such as smallpox, measles, mumps and influenza, are known causes of later contact era outbreaks; however, the diseases that are responsible for many early contact period New World epidemics remain unknown and have been the subject of scientific debate for more than a century[2–4,7,8].

Morphological changes in skeletal remains[9] and ethnohistorical accounts[10] are often explored as sources for understanding population health in the past, although both provide only limited resolution and have generated speculative and, at times, conflicting hypotheses about the diseases introduced to New World populations[2,3,7,11,12]. Most infectious diseases do not leave characteristic markers on the skeleton due to their short periods of infectivity, the death of the victim in the acute phase before skeletal changes formed, or a lack of osteological involvement[9]. Although historical descriptions of infectious disease symptoms can be detailed, they are subject to cultural biases, are affected by translational inaccuracies, lack a foundation in germ theory and describe historical forms of a condition that may differ from modern manifestations[8,11]. In addition, differential diagnosis based on symptoms alone can be unreliable even in modern contexts, as many infectious diseases have similar clinical presentations.

Genome-wide studies of ancient pathogens have proven instrumental in both identifying and characterizing past human infectious diseases. These studies have largely been restricted to skeletal collections where individuals display physical changes consistent with particular infections[13–15], a historical context that links a specific pathogen to a known epidemiological event[16] or an organism that was identified via targeted molecular screening without prior indication of its presence[17]. Recent attempts to circumvent these limitations have concentrated on broad-spectrum molecular approaches focused on pathogen detection via fluorescence-hybridization-based microarray technology[18], identification via DNA enrichment of certain microbial regions[19] or computational screening of non-enriched sequence data against human microbiome data sets[20]. These approaches offer improvements, but remain biased in the bacterial taxa used for species-level assignments. As archaeological material is expected to harbour an abundance of bacteria that stem from the depositional context, omission of environmental taxa in species assignments can lead to false-positive identifications. Additional techniques for authenticating ancient DNA have been developed[21,22], including the identification of characteristic damage patterns caused by the deamination of cytosines[23], methods that evaluate evenness of coverage of aligned reads across a reference genome, or length distributions that consider the degree of fragmentation, where ancient molecules are expected to be shorter than those from modern contaminants[24].

[1]Max Planck Institute for the Science of Human History, Jena, Germany. [2]Institute for Archaeological Sciences, University of Tübingen, Tübingen, Germany. [3]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, USA. [4]Institute of Evolutionary Medicine, University of Zurich, Zurich, Switzerland. [5]National Institute of Anthropology and History (INAH), Mexico, Teposcolula-Yucundaa Archaeological Project, Mexico City, Mexico. [6]Center for Bioinformatics Tübingen (ZBIT), University of Tübingen, Tübingen, Germany. Present address: [7]Smithsonian Conservation Biology Institute, Center for Conservation Genomics, Washington DC, USA. Åshild J. Vågene and Alexander Herbig contributed equally to this work. *e-mail: herbig@shh.mpg.de; tuross@fas.harvard.edu; bos@shh.mpg.de; krause@shh.mpg.de

A typical next-generation sequencing data set from an ancient sample comprises millions of DNA-sequencing reads, which make taxonomic assignment and screening based on sequence alignments computationally challenging. The gold-standard tool for alignment-based analyses is the Basic Local Alignment Search Tool (BLAST)[25], owing to its sensitivity and statistical model. However, the computational time and power that BLAST requires to analyse a typical metagenomic data set is often prohibitive.

Here, we introduce the MEGAN alignment tool (MALT), a program for the fast alignment and analysis of metagenomic DNA-sequencing data. MALT contains the same taxonomic binning algorithm, that is, the naive lowest common ancestor (LCA) algorithm (for reviews, see[26,27]), implemented in the interactive metagenomics analysis software MEGAN[28]. Like BLAST, MALT computes 'local' alignments between highly conserved segments of reads and references. MALT can also calculate 'semi-global' alignments where reads are aligned end-to-end. In comparison to protein alignments or local DNA alignments, semi-global DNA alignments are more suitable for assessing various quality and authenticity criteria that are commonly applied in the field of paleogenetics.

We applied our MALT screening pipeline (Supplementary Figs. 1 and 2) using a database of all complete bacterial genomes available in National Center for Biotechnology Information (NCBI) RefSeq to non-enriched DNA sequence data from the pulp chamber of teeth collected from indigenous individuals excavated at the site of Teposcolula-Yucundaa, located in the highland Mixteca Alta region of Oaxaca, Mexico[29,30]. The site contains both pre-contact and contact era burials, including the earliest identified contact era epidemic burial ground in Mexico[30,31] (Fig. 1 and Supplementary Methods 1). This is the only known cemetery historically linked to the *cocoliztli* epidemic of 1545–1550 CE[30], described as one of the principal epidemiological events responsible for the cataclysmic population decline of sixteenth century Mesoamerica[7,32]. This outbreak affected large areas of central Mexico and Guatemala, spreading perhaps as far south as Peru[7,30]. Through the MALT screening approach, we were able to identify ancient *Salmonella enterica* DNA in the sequence data generated from this archaeological material, to the exclusion of DNA stemming from the complex environmental background. Although the pathogenic cause of the *cocoliztli* epidemic is ambiguous based on ethnohistorical evidence[7,8,30], we report the molecular evidence of microbial infection with genome-wide data from *S. enterica* subsp. *enterica* serovar Paratyphi C (a known cause of enteric fever in humans) isolated from ten epidemic-associated contact era burials.

Results

The individuals included in this investigation were excavated from the contact era epidemic cemetery located in the Grand Plaza (administrative square) ($n=24$) and the pre-contact churchyard cemetery ($n=5$) at Teposcolula-Yucundaa between 2004 and 2010[30] (Fig. 1, Supplementary Table 1 and Supplementary Methods 1). Previous work demonstrated ancient DNA preservation at the site through the identification of New World mitochondrial haplogroups in 48 individuals, 28 of which overlap with this study[30]. In addition, oxygen isotope analysis identified them as local inhabitants[30]. Thirteen individuals included in this study were previously radiocarbon dated[31], yielding dates that support archaeological evidence that the Grand Plaza ($n=10$) and the churchyard ($n=3$) contain contact and pre-contact era burials, respectively (Supplementary Table 1). The Grand Plaza is estimated to contain >800 individuals, most interred in graves that contain multiple individuals. The excavated individuals contribute to a demographic profile consistent with an epidemic event[29,30].

Tooth samples were processed according to protocols designed for ancient DNA work (Supplementary Methods 2). An aggregate soil sample from the two burial grounds was analysed in parallel to gain an overview of environmental bacteria that may have infiltrated our samples. Pre-processed sequencing data of ~1 million paired-end reads per tooth were analysed with MALT using a curated reference database of 6,247 complete bacterial genomes,

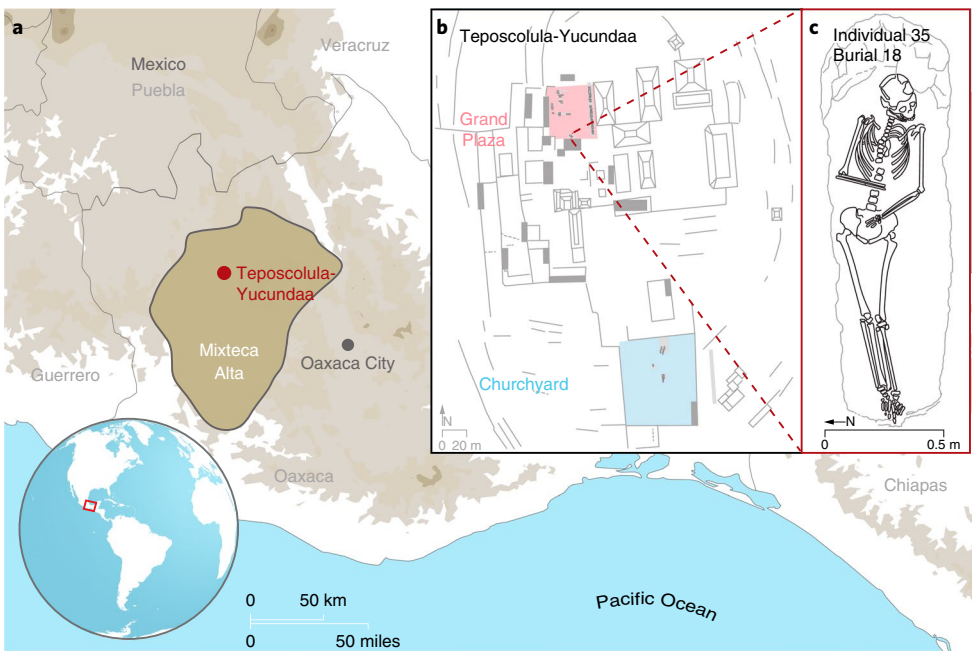

Fig. 1 | Overview of Teposcolula-Yucundaa. a, The location of the Teposcolula-Yucundaa (17.502500° N, 97.467493° W) site in the Mixteca Alta region of Oaxaca, Mexico. **b**, The central administrative area of Teposcolula-Yucundaa showing the relative positioning of the Grand Plaza and the churchyard cemetery sites. Burials within each cemetery are indicated with dark grey outlines, and the excavation area is shaded in grey. **c**, Drawing of individual 35 from which the Tepos_35 *S.* Paratyphi C genome was isolated. Panels **b** and **c** are adapted with permission from drawings provided by the Teposcolula-Yucundaa archaeological project archives-INAH and Christina Warinner

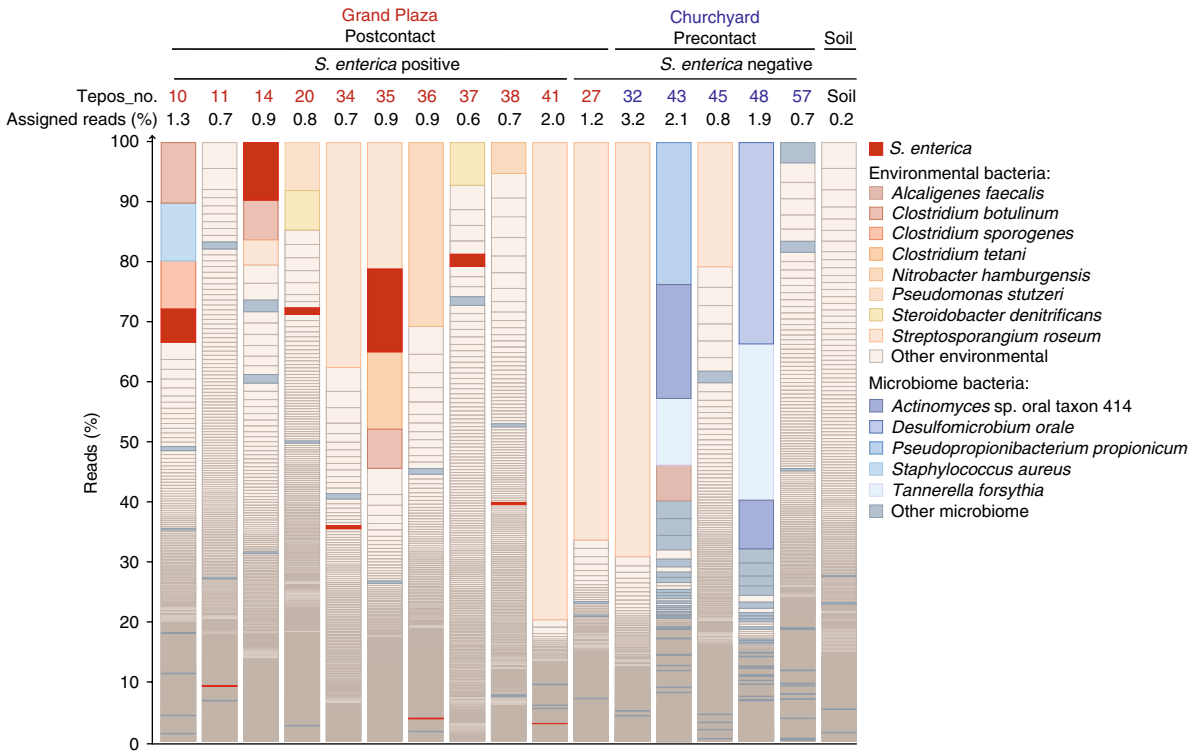

Fig. 2 | MALT analysis and pathogen screening of shotgun data. Shotgun data were analysed with MALT using a database constructed from all bacterial genomes available through NCBI RefSeq (December 2016). MALT results were visualized using MEGAN6 (ref. [28]). The bar chart was constructed from the MEGAN6 output and is based on the per cent reads assigned to bacterial species when using a 95% identity filter. Reads assigned to *S. enterica* are coloured red regardless. Other taxa to which ≥3% reads, per sample, were assigned are colour-coded depending on whether they are 'environmental' or 'human oral microbiome' bacteria. Remaining taxa are sorted into two categories: 'other environmental' or 'other microbiome' (Supplementary Methods 3). Samples from the post-contact Grand Plaza epidemic cemetery containing *S. enterica* reads, pre-contact era samples from the churchyard cemetery and the soil sample are illustrated. In addition, a sample negative for *S. enterica* from the Grand Plaza cemetery (Tepos_27) is shown. Samples whose names are coloured in red are from the Grand Plaza and those in blue are from the churchyard cemetery. The percentage of reads in the shotgun data assigned by MALT per sample is indicated at the top of each column. Only taxa with ≥4 reads assigned are visualized.

comprising all those available in NCBI RefSeq (December 2016). Our approach limits ascertainment biases and false-positive assignments that could result from databases deficient in environmental taxa (Supplementary Methods 3). A runtime analysis revealed a 200-fold improvement in computation time for MALT compared to BLASTn (see Methods). Results were visualized in MEGAN[28], and taxonomic assignments were evaluated with attention to known pathogenic species. Reads that were taxonomically assigned by MALT ranged from 4,842 to 44,315 for the samples. Assigned reads belonging to bacterial constituents of human oral and soil microbiota are present in varying proportions among the samples (Fig. 2 and Supplementary Table 2). Of note, three teeth (Tepos_10, Tepos_14 and Tepos_35) had between 365 and 659 reads assigned to *S. enterica*. Of the *S. enterica* strains present in the database, *S.* Paratyphi C had the highest number of assigned reads (Supplementary Methods 3). Mapping these three metagenomic data sets to the *S.* Paratyphi C RKS4594 genome (NC_012125.1) revealed the characteristic pattern of damage expected of ancient DNA (Supplementary Fig. 3, Supplementary Methods 3 and Supplementary Table 4). Subsequently, the sequencing data for all samples were mapped to the human genome (hg19), revealing a similar level of damage in the human reads for Tepos_10, Tepos_14 and Tepos_35, thus providing further support for the ancient origin of the *S. enterica* reads (Supplementary Methods 4 and Supplementary Table 5). An additional seven individuals from the Grand Plaza cemetery and one negative control harboured low numbers of assigned *S. enterica* reads ranging from 4 to 51 (Supplementary Table 2). These were considered as potential weak-positive samples. One negative control was found to contain 15 reads assigned to *S. enterica*, and a further four contained one or two reads, as did nine sample libraries, seven of which were not included in downstream analyses. The soil library and the remaining sample libraries were void of *S. enterica* reads (Fig. 2, Supplementary Methods 3 and Supplementary Table 2). An additional MALT screen for traces of viral DNA revealed one notable taxonomic hit to *Salmonella* phage Vi II-E1, which is a phage associated with *Salmonella* serovars, including *S.* Paratyphi C, that produce the Vi capsule antigen[33] (Supplementary Methods 3 and Supplementary Table 3).

To further authenticate and elucidate our findings, we performed whole-genome targeted array and in-solution hybridization capture[34,35], using probes designed to encompass modern *S. enterica* genome diversity (Supplementary Methods 5–7 and Supplementary Table 6). All five pre-contact samples, the soil sample, one post-contact sample putatively negative for *S. enterica* based on our MALT screening, all negative controls, and both uracil DNA glycosylase (UDG)-treated (DNA damage removed) and non-UDG-treated libraries from the ten putatively positive samples (Tepos_10, Tepos_11, Tepos_20, Tepos_14, Tepos_34, Tepos_35, Tepos_36, Tepos_37, Tepos_38 and Tepos_41) were included in the capture (Supplementary Methods 6 and 7).

Mapping and genotyping of the captured Illumina-sequenced reads were performed using the *S.* Paratyphi C reference genome (NC_012125.1) (Supplementary Methods 6–8 and Supplementary Table 7). Capture of *S. enterica* DNA was successful for the ten

Table 1 | Overview of mapping statistics of captured sample libraries from the Grand Plaza (contact era) and the churchyard (pre-contact)

Sample ID	Cemetery site	Library treatment	Processed reads before mapping (*n*)	Unique mapped reads (*n*)	Endogenous DNA (%): quality-filtered reads	Mean fold coverage	Percentage of genome covered at least 3-fold
Tepos_10	Grand Plaza	Non-UDG	16,945,834	399,561	20.56	4.35	52.17
		UDG	68,628,270	2,903,258	16.30	32.84	95.49
Tepos_14	Grand Plaza	Non-UDG	20,559,478	1,222,402	23.51	14.41	95.77
		UDG	73,204,225	3,410,610	18.62	36.44	97.67
Tepos_35	Grand Plaza	Non-UDG	27,248,720	1,803,043	31.37	25.50	97.67
		UDG	90,815,050	7,025,774	30.00	96.43	98.06
Tepos_11	Grand Plaza	Non-UDG	21,941,119	19,576	0.87	0.21	0.93
		UDG	48,959,732	103,492	0.75	1.21	14.56
Tepos_20	Grand Plaza	Non-UDG	771,431	15,236	6.94	0.15	0.26
		UDG	20,123,713	427,781	4.75	4.59	67.53
Tepos_34	Grand Plaza	Non-UDG	18,934,710	123,307	2.55	1.35	14.65
		UDG	26,284,766	157,930	2.05	1.74	21.67
Tepos_36	Grand Plaza	Non-UDG	23,147,904	36,224	0.75	0.40	1.76
		UDG	21,910,196	33,327	0.42	0.36	1.4
Tepos_37	Grand Plaza	Non-UDG	5,223,138	218,874	9.28	2.55	42.12
		UDG	9,603,890	416,449	7.71	5.49	74.48
Tepos_38	Grand Plaza	Non-UDG	8,280,412	18,308	0.91	0.19	0.97
		UDG	47,835,731	65,812	0.54	0.67	4.22
Tepos_41	Grand Plaza	Non-UDG	17,608,445	33,664	0.72	0.37	1.47
		UDG	19,966,958	36,208	0.48	0.40	1.34
Tepos_27	Grand Plaza	Non-UDG	17,931,300	4,778	0.07	0.04	0.27
Tepos_32	Churchyard	Non-UDG	25,721,427	6,665	0.08	0.06	0.47
Tepos_43	Churchyard	Non-UDG	31,129,662	3,426	0.05	0.03	0.25
Tepos_45	Churchyard	Non-UDG	18,027,289	6,879	0.12	0.06	0.34
Tepos_48	Churchyard	Non-UDG	17,915,341	4,312	0.06	0.04	0.25
Tepos_57	Churchyard	Non-UDG	24,478,844	5,527	0.07	0.05	0.28
Soil	Grand Plaza and churchyard	Non-UDG	10,875,300	796	0.02	0.01	0.07

positive samples, yielding a minimum of 33,327 unique reads per UDG-treated library. The remaining bone samples, soil sample and negative controls were determined to be negative for ancient *S. enterica* DNA, with the exception of one negative control (EB2-091013) that had probably become cross-contaminated during processing (see Supplementary Methods 8 and Supplementary Table 7). Five complete genomes were constructed for Tepos_10, Tepos_14, Tepos_20, Tepos_35 and Tepos_37, covering 95%, 97%, 67%, 98% and 74% of the reference at a minimum of 3-fold coverage and yielding an average genomic coverage of 33-, 36-, 4.6-, 96- and 5.5-fold, respectively (Table 1). Artificial reads generated in silico for 23 complete genomes included in the probe design were also mapped to the *S.* Paratyphi C RKS4594 reference (Supplementary Methods 8 and Supplementary Table 6), and phylogenetic comparison revealed that the five ancient genomes clustered with *S.* Paratyphi C (Fig. 3, Supplementary Figs. 4 and 6 and Supplementary Methods 8). The phylogenetic positioning was retained when the whole data set was mapped to and genotyped against the *S. enterica* subsp. *enterica* serovar Typhi CT18 reference genome (NC_003198.1) (Supplementary Fig. 5 and Supplementary Table 8), the most common bacterial cause of enteric fever in humans today. This result excludes the possibility of a reference bias. Despite all five ancient genomes being contemporaneous, the Tepos_10 genome was observed to contain many more derived positions. An investigation of heterozygous variant calls showed that Tepos_10 has a much higher number of heterozygous sites. We believe that this is best explained by the presence of genetically similar non-target DNA that co-enriched in the capture for this sample alone. Based on the pattern of allele frequencies, this genome was excluded from downstream analyses (Supplementary Methods 9 and Supplementary Fig. 7). Tepos_20 and Tepos_37 were also excluded owing to their genomic coverage of <6-fold, which allowed more reliable single-nucleotide polymorphism (SNP) calling at a minimum of 5-fold coverage for Tepos_14 and Tepos_35. Subsequent phylogenetic tree construction with 1,000 bootstrap replicates revealed 100% support and branch shortening for the Tepos_14 and Tepos_35 genomes in all phylogenies, supporting their ancient origin (Fig. 3 and Supplementary Fig. 8).

SNP analysis for the ancient genomes together with the reference data set yielded a total of 203,256 variant positions among all 25 genomes. Our analyses identified 681 positions present in one or both of the ancient genomes, where 133 are unique to the ancient lineage (Supplementary Methods 10 and Supplementary Table 9). Of these, 130 unique SNPs are shared between Tepos_14 and Tepos_35, supporting their close relationship and shared ancestry. The *ydiD* gene, which is involved in the breakdown of fatty acids[36], and the *tsr* gene, which is related to the chemotaxic response system[37], were found to contain multiple non-synonymous SNPs unique to the ancient genomes (Supplementary Methods 10). Seven homoplastic and four tri-allelic variant positions were detected in the ancient genomes (Supplementary Methods 10 and Supplementary Tables 10a and 10b).

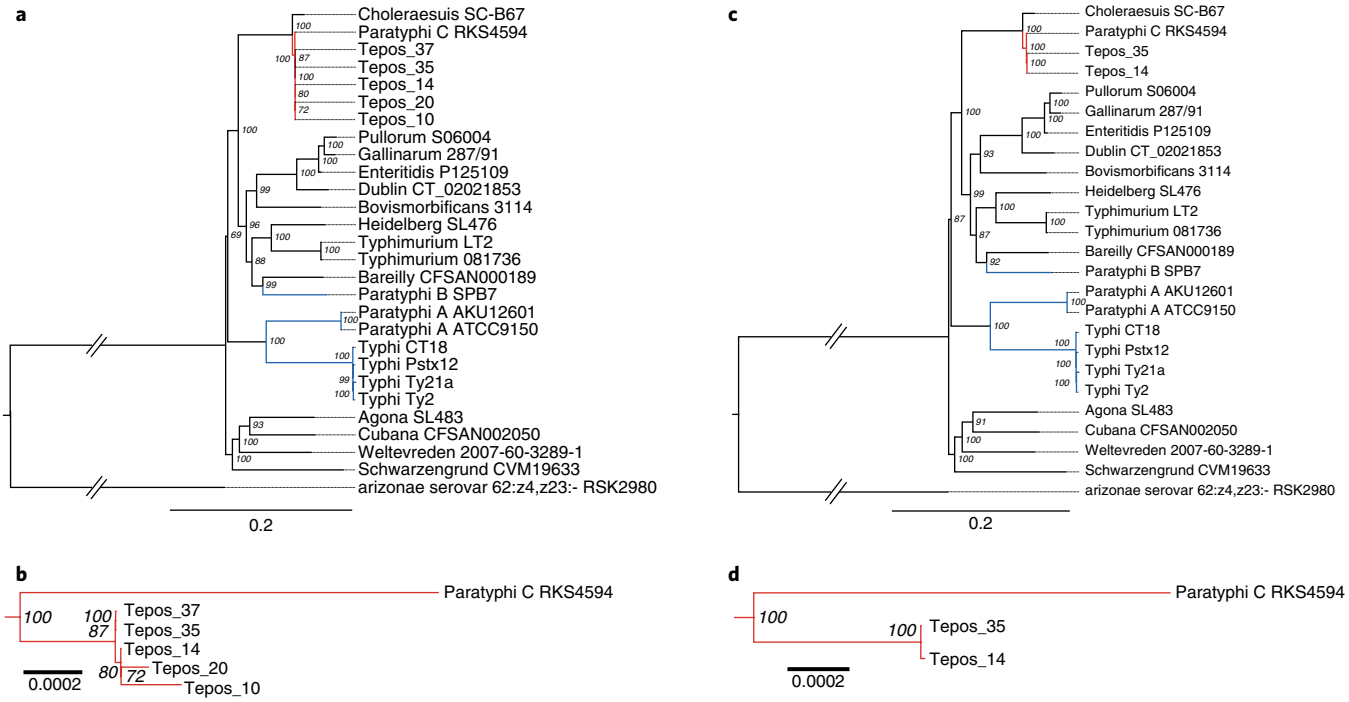

Fig. 3 | Maximum likelihood trees for *S. enterica* subsp. *enterica* phylogeny. Two maximum likelihood trees were produced. Positions with missing data were excluded in both cases. **a**, The tree includes all five ancient genomes, and is based on 3-fold SNP calls and 51,456 variant positions. **b**, A zoomed in view of the *S.* Paratyphi C genomes. **c**, The tree includes two high-coverage ancient genomes, and is based on 5-fold SNP calls and 81,474 variant positions. **d**, A zoomed in view of the *S.* Paratyphi C genomes, illustrating the branch shortening of the two ancient genomes (Tepos_14 and Tepos_35). Both trees were built with RAxML[72]. Branches that are coloured red indicate *S.* Paratyphi C genomes, and branches in blue indicate other genomes that are human-specific and cause enteric (typhoid/paratyphoid) fever.

A region of the *pil* operon consisting of five genes, *pilS*, *pilU*, *pilT*, *pilV* and *rci*, was found in our ancient genomes and was absent in the *S.* Paratyphi C RKS4594 genome[38] (Supplementary Methods 12 and Supplementary Table 12). This region is located in *Salmonella* pathogenicity island 7 (SPI-7), and encodes a type IVB pili[39,40]. The version of *pilV* in our ancient genomes is thought to facilitate bacterial self-aggregation, a phenomenon that potentially aids in invasion of host tissues[39,40] (for details, see Supplementary Methods 12). A further presence/absence analysis was performed to evaluate additional virulence factors. These results are summarized in Supplementary Fig. 9 and Supplementary Methods 13 (See also Supplementary Table 13).

The *S.* Paratyphi C RKS4594 strain harbours a virulence plasmid, pSPCV, which was included in our capture design. It is present at 10-fold to 224-fold average coverage for the five genomes (Supplementary Methods 14 and Supplementary Tables 14 and 15).

Non-UDG capture reads mapped to the *S.* Paratyphi C genome (NC_012125.1) for Tepos_11, Tepos_34, Tepos_36, Tepos_38 and Tepos_41, that is, those that did not yield full genomes, had damage patterns that are characteristic of ancient DNA (Supplementary Fig. 3). To further verify these reads as true ancient *S.* Paratyphi C reads, we investigated 45 SNPs unique to Tepos_14 and Tepos_35, which are included in our phylogenetic analysis (Supplementary Methods 11 and Supplementary Table 11). Of the 45 positions, between 6 and 29 were identified at minimum 1-fold in these lower coverage genomes. All of these were in agreement with the unique SNPs present in the high-coverage ancient genomes, thus confirming their shared ancestry.

Discussion

Interpretations of ethnohistorical documents have suggested some form of typhus or enteric (typhoid/paratyphoid) fever (from the Spanish 'tabardillo', 'tabardete' and '*tifus mortal*'), viral haemorrhagic fever, measles or pneumonic plague as potential causes of the *cocoliztli* epidemic of 1545 CE (for refs, see Supplementary Discussion 1). These diseases present symptoms similar to those that were recorded in the *cocoliztli* outbreak, such as red spots on the skin, bleeding from various body orifices and vomiting (Supplementary Discussion 1 and Supplementary Fig. 10). Given the non-specific nature of these symptoms, additional sources of data are needed to clarify which disease (or diseases) was circulating. Previous investigation of sequencing data generated from the Teposcolula-Yucundaa material did not identify DNA traces of ancient pathogens; however, *S. enterica* was not considered as a candidate[41]. Here, we have isolated genome-wide data of ancient *S.* Paratyphi C from ten Mixtec individuals buried in the Grand Plaza epidemic cemetery at Teposcolula-Yucundaa, indicating that enteric fever was circulating in the indigenous population during the *cocoliztli* epidemic of 1545–1550 CE. As demonstrated here, MALT offers a sensitive approach for screening non-enriched sequence data in search for unknown candidate bacterial pathogens that were involved in past disease outbreaks, even to the exclusion of a dominant environmental microbial background. Most importantly, it offers the advantage of extensive genome-level screening without the need to specify a target organism, thus avoiding ascertainment biases that are common to other screening approaches. Fast metagenomic profiling tools that are based on k-mer matching, such as Kraken[42], or specific diagnostic marker regions, such as MetaPhlan2 (ref.[43]), have limitations in ancient DNA applications. Complete alignments are needed to authenticate candidate taxonomic assignments, and a small number of marker regions might not provide sufficient resolution for identification, as target DNA is often present in low amounts. Our focus on only bacterial and DNA viral taxa limits our resolution in identifying other infectious agents that may have been present in the population during the Teposcolula-Yucundaa epidemic.

Although our discussions here have focused on a single pathogenic organism, the potential of its having acted synergistically with other pathogens circulating during the epidemic must be considered. The concept of syndemics and the complex biosocial factors that influence infectious disease transmission and severity are well documented in both modern and historical contexts[44,45]. We are currently limited to the detection of bacterial pathogens and DNA viruses included in the NCBI genomic database, although the resolution offered by MALT analyses will increase as this database grows. We have not investigated the presence of RNA viruses, as methods for RNA retrieval from archaeological tissues are underdeveloped and not supported by our current protocols[46].

We confidently exclude an environmental organism as the source for our ancient genomes on the basis that: (1) *S.* Paratyphi C is restricted to humans, (2) it is not known to freely inhabit soil (our soil sample was negative for *Salmonella* during screening and after capture), (3) the deamination patterns observed for the ancient human and *S.* Paratyphi C reads are characteristic of authentic ancient DNA, and (4) the ancient *S.* Paratyphi C genomes display expected branch shortening in all constructed phylogenies. Moreover, we recovered all ancient genomic data from the pulp chambers of teeth collected in situ, increasing the likelihood of having identified a bacterium that was present in the individual's blood at the time of death. *S. enterica* introduction via post-burial disturbance is unlikely because the graves in the Grand Plaza were dug directly into the thickly paved floor at the site, and historical records indicate that Teposcolula-Yucundaa was abandoned shortly after the epidemic ended in 1552 CE[29,30].

S. Paratyphi C is one of >2,600 identified *S. enterica* serovars distinguished by their antigenic formula[47]. Only four serovars (*S.* Typhi and *S.* Paratyphi A, B, C), all of which cause enteric fever, are restricted to the human host[47]. Today, *S.* Typhi and *S.* Paratyphi A cause the majority of reported cases[48]. *S.* Paratyphi C is rarely reported[38,48]. Infected individuals shed bacteria long after the termination of symptoms[47], and in the case of *S.* Typhi infection, 1–6% of individuals become asymptomatic carriers[49]. Following the hypothesis that this disease was introduced through European contact, it is conceivable that asymptomatic European carriers who withstood the cross-Atlantic voyage could have introduced *S.* Paratyphi C to Mesoamerican populations in the sixteenth century. First-hand descriptions of the 1545 *cocoliztli* epidemic suggest that both European and Mixtec individuals were susceptible to the disease[7,50], with one estimate of a 60–90% population decline in New Spain during this period[7].

The additional SPI-7 genes detected through insertion and deletion (indel) analysis are reported to vary in presence/absence among modern *S.* Paratyphi C strains[38,40], and are suspected to cause increased virulence when the inverted repeats in *pilV* allow the Rci recombinase to shuffle between its two protein states (Supplementary Methods 12). This may support an increased capacity for our ancient strains to cause an epidemic outbreak. However, the overall mechanisms through which *S.* Paratyphi causes enteric fever remain unclear. The non-synonymous SNPs in the *ydiD* and *tsr* genes may signify adaptive processes, and comparison with a greater number of *S.* Paratyphi C genomes may clarify this[51].

Today, *S.* Paratyphi C is rare in Europe and the Americas, with more cases identified across Africa and Asia[52,53]. Based on multilocus sequence typing (MLST) data from modern *S.* Paratyphi C strains, no clear phylogeographical pattern has been observed[52]. However, the presence of a 1200 CE *S.* Paratyphi C genome in Norway indicates its presence in Europe in the pre-contact era[51], which would be necessary for it to be considered an Old World disease. However, based on the small number of pre-contact individuals that we have screened, we cannot exclude the presence of *S.* Paratyphi C at Teposcolula-Yucundaa prior to European arrival. A local origin for the *cocoliztli* disease has been proposed elsewhere[54]. Historical accounts offer little perspective on its origin as neither the indigenous population nor the European colonizers had a pre-existing name for the disease[7,8,30]. Spanish colonial documents refer to it as "*pujamiento de sangre*" ('full bloodiness'), whereas the indigenous Aztec population of Central Mexico called it *cocoliztli*, a generic term meaning 'pestilence' in Nahuatl[7,8] (see Supplementary Discussion 1).

Little is known about the past severity and worldwide incidence of enteric fever, which was first determined to be distinct from typhus in the mid-nineteenth century[55]. Enteric fevers are regarded as major health threats worldwide[48], causing an estimated ~27 million illnesses in 2000, the majority of which were attributed to *S.* Typhi[56]. Owing to the rarity of *S.* Paratyphi C diagnoses, mortality rates are not established for this particular serovar. Today, outbreaks predominantly occur in developing countries. *S.* Typhi and *S.* Paratyphi are commonly transmitted through the faecal–oral route via ingestion of contaminated food or water[57]. Changes imposed under Spanish rule, such as forced relocations under the policy of '*congregación*', altered living arrangements, and new subsistence farming practices[29,30] compounded by drought conditions[32] could have disrupted existing hygiene measures, facilitating *S.* Paratyphi C transmission.

Our study represents a first step towards a molecular understanding of disease exchange in contact era Mexico. The 1545 *cocoliztli* epidemic is regarded as one of the most devastating epidemics in New World history[7,32]. Our findings contribute to the debate concerning the causative agent of this epidemic at Teposcolula-Yucundaa, where we propose that *S.* Paratyphi C be considered. We introduced MALT, a novel fast alignment and taxonomic assignment method. Its application to the identification of ancient *S. enterica* DNA within a complex background of environmental microbial contaminants speaks to the suitability of this approach, and its resolution will improve as the number of available reference genomes increases. This method may be eminently useful for studies wishing to identify pathogenic agents involved in ancient and modern disease, particularly in cases for which candidate organisms are not known a priori.

Methods

The MALT algorithm. MALT is based on the seed-and-extend paradigm and consists of two programs: malt-build and malt-run.

First, malt-build is used to construct an index for the given database of reference sequences. To do so, malt-build determines all occurrences of spaced seeds[58–60] in the reference sequences and places them into a hash table[61].

Following this, malt-run is used to align a set of query sequences against the reference database. To this end, the program generates a list of spaced seeds for each query and then looks them up in the reference hash table, which is kept in the main memory. Using the x-drop extension heuristic[25], a high-scoring ungapped alignment that is anchored at the seed is computed and is used to decide whether a full alignment should be constructed. Local or semi-global alignments are computed using a banded implementation[62] of the Smith–Waterman[63] or Needleman–Wunsch[64] algorithms, respectively. The program then computes the bit-score and the expected value (E-value) of the alignment and decides whether to keep or discard the alignment depending on user-specified thresholds for the bit-score, the E-value or the per cent identity. The application of malt-run is illustrated in Supplementary Fig. 1.

The MALT screening pipeline. To use MALT in ancient DNA contexts to screen for bacterial DNA and to assess the taxonomic composition of ancient bacterial communities, we applied the following workflow (Supplementary Fig. 2). First, we used malt-build to construct a MALT index on all complete bacterial genomes in GenBank[65]. This was done only once, and is rebuilt only when the target database requires updating. We align reads to the reference database using malt-run in semi-global mode. MALT generates output in RMA format and in SAM format. The RMA format can be used for interactive analysis of taxonomic composition in MEGAN[28], and the SAM format can be used for alignment-based assessment of damage patterns and other authenticity criteria.

Sample provenience. The site of Teposcolula-Yucundaa is situated on a mountain ridge in the Mixteca Alta region of Oaxaca, Mexico. Prior archaeological excavation at this site revealed a large epidemic cemetery located in the Grand

Plaza—the town's administrative centre—and an additional cemetery in the churchyard (Fig. 1 and Supplementary Methods 1). Twenty-four teeth were collected from individuals buried in the Grand Plaza cemetery and five from individuals buried in the churchyard cemetery (Supplementary Methods 2 and Supplementary Table 1). Soil samples were also collected from both cemetery sites (Supplementary Methods 2).

DNA extraction and library preparation. DNA extracts and double-stranded indexed libraries that are compatible with Illumina sequencing were generated using methods tailor-made for ancient DNA[66–68]. This work was carried out in dedicated ancient DNA cleanroom facilities at the University of Tübingen and Harvard University (Supplementary Methods 2).

Screening with MALT. Amplified libraries were shotgun sequenced. The reads were adapter clipped and merged before being analysed with MALT, and the results were visualized in MEGAN6 (ref. [28]) (Supplementary Methods 2 and 3). Two MALT runs were executed: the first using all complete bacterial genomes that were available through NCBI RefSeq (December 2016), and the second using the full NCBI Nucleotide database (ftp://ftp-trace.ncbi.nih.gov/blast/db/FASTA/) as reference to screen for viral DNA (Fig. 2, Supplementary Methods 3 and Supplementary Tables 2 and 3). Both runs used 'semi-global' alignment and a minimum per cent identity of 95 (Supplementary Methods 3). The shotgun data were also mapped to the *S.* Paratyphi C RKS4594 reference (NC_012125.1) and the human genome (hg19), and damage plots were generated (Supplementary Methods 3 and 4, Supplementary Fig. 3 and Supplementary Tables 4 and 5).

Runtime comparison. The programs MALT (version 0.3.8) and BLAST[25] (version 2.6.0+) were applied to the shotgun screening data of Tepos_35, consisting of 952,511 reads. For both programs, the DNA alignment mode (BLASTn) was chosen. The maximal E-value was set to 1.0. The maximal number of alignments for each query was set to 100. The minimal per cent identity was set to 95. The number of threads was set to 16. The alignment type of MALT was set to 'local' to be comparable to BLAST. The total amount of random acess memory (RAM) required by MALT during this run was 252.7 GB.

For MALT, the runtime was measured excluding the initial loading of our reference database, which happens only once when screening multiple samples. The loading of the database takes 27.27 min. Including taxonomic binning, the application of MALT to our complete shotgun screening data took 123.36 min. As a comparison, processing only the screening data of a single sample (Tepos_35) with BLASTn took 1,420.58 min without any taxonomic analysis. Processing of this sample alone with MALT, including taxonomic binning, took 6.48 min, which constitutes a 200-fold improvement in terms of computation time.

The computations were performed on a Dell PowerEdge R820 with four Intel Xeon E5-4620 2.2 GHz central processing units (CPUs) and 768 GB of RAM.

Probe design and whole-genome capture. Array probes were designed based on 67 publicly available *S. enterica* chromosomes/assemblies and 45 associated plasmid sequences (Supplementary Methods 5 and Supplementary Table 6). Extracts from samples that were deemed to be positive for *S.* Paratyphi C were converted into additional rich UDG-treated libraries[69] for whole-genome capture (Supplementary Methods 6 and 7). Pre-contact and post-contact samples were serially captured using our custom probe design, according to two established methods[35,70]. The eluate from both array and in-solution capture was sequenced to a sufficient depth to allow high-coverage genome reconstruction (Supplementary Methods 6 and 7).

Sequence data processing, initial phylogenetic assessment and authenticity. The sequence data were adapter clipped and quality filtered before being mapped to the *S.* Paratyphi C reference (NC_012125.1) (Supplementary Methods 8 and Supplementary Table 7). Deamination patterns for the DNA were generated to assess the authenticity of the ancient *S.* Paratyphi C DNA using mapDamage[23] (Supplementary Methods 8, Supplementary Table 7 and Supplementary Fig. 3). Artificial read data were generated for a data set of 23 genomes that were selected for comparative phylogenetic analysis; this data were also mapped to the *S.* Paratyphi C reference (Supplementary Methods 8). SNP calling was carried out with Genome Analysis Toolkit (GATK) using a quality score of ≥30 for the five *S.* Paratyphi C genomes and the artificial read data set. A neighbour-joining tree was constructed using MEGA6 (ref. [71]), based on homozygous SNPs called at a minimum of 3-fold coverage where at least 90% of reads are in agreement (Supplementary Methods 8 and Supplementary Fig. 4). To exclude a reference bias in the ascertainment of the phylogenetic positioning of the five ancient genomes (Tepos_10, Tepos_14 Tepos_20, Tepos_35 and Tepos_37), mapping, SNP calling and tree construction were repeated for the *S.* Typhi CT18 reference (NC_003198.1) (Supplementary Methods 8, Supplementary Table 8 and Supplementary Fig. 5).

SNP typing and phylogenetic analysis. Homozygous SNPs were called from the complete data set (5 ancient and 23 modern) based on our criteria using a tool called MultiVCFAnalyzer (Supplementary Methods 9). Repetitive and highly conserved regions of the *S.* Paratyphi C genome (NC_012125.1) were excluded from SNP calling to avoid spurious mapping reads. Maximum parsimony[71] and maximum likelihood[72] trees were made that included the five genomes (Fig. 3 and Supplementary Fig. 6). Heterozygous positions were also called, and their allele frequency distributions plotted using R[73] (Supplementary Methods 9 and Supplementary Fig. 7). SNP calling and phylogenetic tree construction were repeated, excluding the Tepos_10, Tepos_20 and Tepos_37 genomes (Fig. 3, Supplementary Methods 10 and Supplementary Fig. 8).

The five weak-positive samples—Tepos_11, Tepos_34, Tepos_36, Tepos_38 and Tepos_41—that did not yield enough data for genome reconstruction were investigated for 46 SNPs unique to the ancient genomes to verify that the captured reads for these samples are true ancient *S.* Paratyphi C reads (Supplementary Methods 11 and Supplementary Table 11).

SNP effect and indel analyses. SNP effect analysis was carried out for the two ancient genomes (Tepos_14 and Tepos_35) alongside the modern data set (see Supplementary Methods 10). SNPs unique to the ancient genomes, pseudogenes and homoplastic positions were investigated (Supplementary Methods 10 and Supplementary Tables 9 and 10). Indels, ≥700 base pairs, in the two ancient genomes were identified through two approaches. Deletions were visually detected by mapping the ancient data to the *S.* Paratyphi C reference using a mapping quality threshold of 0 and manually viewing the genome alignment in the Integrative Genomics Viewer (IGV) (Supplementary Methods 12). To detect insertions, or regions present in the ancient genomes that are missing the modern reference, the ancient data were mapped to concatenated reference pairs. One reference was in all cases, the *S.* Paratyphi C RKS4594 reference (NC_012125.1), and the other was one of four *S. enterica* genomes of interest. A mapping quality threshold of 37 was used, thus allowing only regions unique to one or the other genome in the pair to map (Supplementary Methods 12 and Supplementary Table 12).

Virulence factor analysis. Forty-three effector genes identified within *S. enterica* subsp. *enterica*[74] were investigated using the BEDTools suite[75]. The percentage of each gene that was covered at least onefold in the ancient and modern genomes in our data set was plotted using the ggplot2package[76] in R[73] (Supplementary Methods 13 and Supplementary Fig. 9).

Plasmid analysis. The ancient data were mapped to the *S.* Paratyphi C virulence plasmid, pSPCV. SNP effect analysis was carried out in comparison to three other similar plasmid references (Supplementary Methods 14 and Supplementary Tables 14 and 15).

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Data availability. Sequence data that support the findings of this study have been submitted to the European Nucleotide Archive under accession number PRJEB23438 (https://www.ebi.ac.uk/ena/data/view/PRJEB23438). MALT is open source and freely available from: http://ab.inf.uni-tuebingen.de/software/malt. The program MultiVCFAnalyzer is available on GitHub: https://github.com/alexherbig/MultiVCFAnalyzer. Source data for figures are available upon request.

Received: 7 June 2017; Accepted: 7 December 2017;
Published online: 15 January 2018

References

1. Ubelaker, D. H. Prehistoric New World population size: historical review and current appraisal of North American estimates. *Am. J. Phys. Anthropol.* **45**, 661–665 (1976).
2. Crosby, A. W. Virgin soil epidemics as a factor in the aboriginal depopulation in America. *William Mary Q.* **33**, 289–299 (1976).
3. Dobyns, H. F. Disease transfer at contact. *Annu. Rev. Anthropol.* **22**, 273–291 (1993).
4. Acuna-Soto, R., Stahle, D. W., Therrell, M. D., Griffin, R. D. & Cleaveland, M. K. When half of the population died: the epidemic of hemorrhagic fevers of 1576 in Mexico. *FEMS Microbiol. Lett.* **240**, 1–5 (2004).
5. Llamas, B. et al. Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Sci. Adv.* **2**, e1501385 (2016).
6. Lindo, J. et al. A time transect of exomes from a Native American population before and after European contact. *Nat. Commun.* **7**, 13175 (2016).
7. Cook, N. D. & Lovell, W. G. *Secret Judgments of God: Old World Disease in Colonial Spanish America* (Univ. Oklahoma Press, Norman, 2001).
8. Fields, S. L. *Pestilence and Headcolds: Encountering Illness in Colonial Mexico* (Columbia Univ. Press, New York, 2008).
9. Ortner, D. J. *Identification of Pathological Conditions in Human Skeletal Remains* 2nd edn (Academic Press, Cambridge, 2003).
10. Walker, R. S., Sattenspiel, L. & Hill, K. R. Mortality from contact-related epidemics among indigenous populations in Greater Amazonia. *Sci. Rep.* **5**, 14032 (2015).

11. Joralemon, D. New World depopulation and the case of disease. *J. Anthropol. Res.* **38**, 108–127 (1982).
12. Larsen, C. S. In the wake of Columbus: native population biology in the postcontact Americas. *Am. J. Phys. Anthropol.* **37**, 109–154 (1994).
13. Bos, K. I. et al. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* **514**, 494–497 (2014).
14. Schuenemann, V. J. et al. Genome-wide comparison of medieval and modern *Mycobacterium leprae*. *Science* **341**, 179–183 (2013).
15. Warinner, C. et al. Pathogens and host immunity in the ancient human oral cavity. *Nat. Genet.* **46**, 336–344 (2014).
16. Bos, K. I. et al. A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature* **478**, 506–510 (2011).
17. Maixner, F. et al. The 5300-year-old *Helicobacter pylori* genome of the Iceman. *Science* **351**, 162–165 (2016).
18. Devault, A. M. et al. Ancient pathogen DNA in archaeological samples detected with a microbial detection array. *Sci. Rep.* **4**, 4245 (2014).
19. Bos, K. I. et al. Parallel detection of ancient pathogens via array-based DNA capture. *Phil. Trans. R. Soc. B* **370**, 20130375 (2015).
20. Devault, A. M. et al. A molecular portrait of maternal sepsis from Byzantine Troy. *eLife* **6**, e20983 (2017).
21. Warinner, C. et al. A robust framework for microbial archaeology. *Annu. Rev. Genomics Hum. Genet.* **18**, 321–356 (2017).
22. Key, F. M., Posth, C., Krause, J., Herbig, A. & Bos, K. I. Mining metagenomic data sets for ancient DNA: recommended protocols for authentication. *Trends Genet.* **33**, 508–520 (2017).
23. Jonsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. & Orlando, L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682–1684 (2013).
24. Prufer, K. et al. Computational challenges in the analysis of ancient DNA. *Genome Biol.* **11**, R47 (2010).
25. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
26. Peabody, M. A., Van Rossum, T., Lo, R. & Brinkman, F. S. Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. *BMC Bioinforma.* **16**, 363 (2015).
27. Lindgreen, S., Adair, K. L. & Gardner, P. P. An evaluation of the accuracy and speed of metagenome analysis tools. *Sci. Rep.* **6**, 19233 (2016).
28. Huson, D. H. et al. MEGAN community edition—interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS. Comput. Biol.* **12**, e1004957 (2016).
29. Spores, R. & Robles García, N. A prehispanic (postclassic) capital center in colonial transition: excavations at Yucundaa Pueblo Viejo de Teposcolula, Oaxaca, Mexico. *Lat. Am. Antiq.* **18**, 333–353 (2007).
30. Warinner, C., Robles García, N., Spores, R. & Tuross, N. Disease, demography, and diet in early colonial New Spain: investigation of a sixteenth-century Mixtec cemetery at Teposcolula Yucundaa. *Lat. Am. Antiq.* **23**, 467–489 (2012).
31. Tuross, N., Warinner, C. & Robles García, N. in *Yucundaa: La Cuidad Mixteca Yucundaa-Pueblo Viejo de Teposcolula y su Transformación Prehispánica-Colonial* Vol. 2 (eds Spores, R. & Robles García, N.) 541–546 (Instituto Nacional de Antropología e Historia, Mexico City, 2014).
32. Acuna-Soto, R., Stahle, D. W., Cleaveland, M. K. & Therrell, M. D. Megadrought and megadeath in 16th century Mexico. *Emerg. Infect. Dis.* **8**, 360–362 (2002).
33. Pickard, D. et al. Molecular characterization of the *Salmonella enterica* serovar Typhi Vi-typing bacteriophage E1. *J. Bacteriol.* **190**, 2580–2587 (2008).
34. Burbano, H. A. et al. Targeted investigation of the Neandertal genome by array-based sequence capture. *Science* **328**, 723–725 (2010).
35. Fu, Q. et al. DNA analysis of an early modern human from Tianyuan Cave, China. *Proc. Natl. Acad. Sci. USA* **110**, 2223–2227 (2013).
36. Campbell, J. W., Morgan-Kiss, R. M. & Cronan, J. E. Jr A new *Escherichia coli* metabolic competency: growth on fatty acids by a novel anaerobic beta-oxidation pathway. *Mol. Microbiol.* **47**, 793–805 (2003).
37. Rivera-Chavez, F. et al. *Salmonella* uses energy taxis to benefit from intestinal inflammation. *PLoS. Pathog.* **9**, e1003267 (2013).
38. Liu, W. Q. et al. *Salmonella* Paratyphi C: genetic divergence from *Salmonella choleraesuis* and pathogenic convergence with *Salmonella typhi*. *PLoS ONE* **4**, e4510 (2009).
39. Tam, C. K., Morris, C. & Hackett, J. The *Salmonella enterica* serovar Typhi type IVB self-association pili are detached from the bacterial cell by the PilV minor pilus proteins. *Infect. Immun.* **74**, 5414–5418 (2006).
40. Tam, C. K., Hackett, J. & Morris, C. *Salmonella enterica* serovar Paratyphi C carries an inactive shufflon. *Infect. Immun.* **72**, 22–28 (2004).
41. Campana, M. G., Robles Garcia, N., Ruhli, F. J. & Tuross, N. False positives complicate ancient pathogen identifications using high-throughput shotgun sequencing. *BMC Res. Notes* **7**, 111 (2014).
42. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
43. Segata, N. et al. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814 (2012).
44. Singer, M. & Clair, S. Syndemics and public health: reconceptualizing disease in bio-social context. *Med. Anthropol. Q.* **17**, 423–441 (2003).
45. Herring, D. A. & Sattenspiel, L. Social contexts, syndemics, and infectious disease in northern Aboriginal populations. *Am. J. Hum. Biol.* **19**, 190–202 (2007).
46. Guy, P. L. Prospects for analyzing ancient RNA in preserved materials. *Wiley Interdiscip. Rev. RNA* **5**, 87–94 (2014).
47. Gal-Mor, O., Boyle, E. C. & Grassl, G. A. Same species, different diseases: how and why typhoidal and non-typhoidal *Salmonella enterica* serovars differ. *Front. Microbiol.* **5**, 391 (2014).
48. Wain, J., Hendriksen, R. S., Mikoleit, M. L., Keddy, K. H. & Ochiai, R. L. Typhoid fever. *Lancet* **385**, 1136–1145 (2015).
49. Monack, D. M., Mueller, A. & Falkow, S. Persistent bacterial infections: the interface of the pathogen and the host immune system. *Nat. Rev. Microbiol.* **2**, 747–765 (2004).
50. de Sahagún, B. *General History of the Things of New Spain: Florentine Codex* (School of American Research, Santa Fe, 1950–1982).
51. Zhou, Z. et al. Millennia of genomic stability within the invasive Para C lineage of *Salmonella enterica*. Preprint at https://www.biorxiv.org/content/early/2017/02/14/105759 (2017).
52. Achtman, M. et al. Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. *PLoS. Pathog.* **8**, e1002776 (2012).
53. *National Typhoid and Paratyphoid Fever Surveillance Annual Summary, 2014* (CDC, 2016).
54. Acuna-Soto, R., Romero, L. C. & Maguire, J. H. Large epidemics of hemorrhagic fevers in Mexico 1545–1815. *Am. J. Trop. Med. Hyg.* **62**, 733–739 (2000).
55. Smith, D. C. Gerhard's distinction between typhoid and typhus and its reception in America, 1833–1860. *Bull. Hist. Med.* **54**, 368–385 (1980).
56. Crump, J. A., Luby, S. P. & Mintz, E. D. The global burden of typhoid fever. *Bull. World Health Organ.* **82**, 346–353 (2004).
57. *Typhoid Fever—Uganda*(WHO, 2015); http://who.int/csr/don/17-march-2015-uganda/en/
58. Burkhardt, S. & Kärkkäinen, J. in *Combinatorial Pattern Matching: 12th Annual Symposium, CPM 2001 Jerusalem, Israel, July 1–4, 2001 Proceedings* (eds Amihood, A. & Landau, G. M.) 73–85 (Springer, Berlin, 2001).
59. Ma, B., Tromp, J. & Li, M. PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18**, 440–445 (2002).
60. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
61. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).
62. Chao, K. M., Pearson, W. R. & Miller, W. Aligning two sequences within a specified diagonal band. *Comput. Appl. Biosci.* **8**, 481–487 (1992).
63. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
64. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
65. Benson, D. A. et al. GenBank. *Nucleic Acids Res.* **41**, D36–D42 (2013).
66. Dabney, J. et al. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. USA* **110**, 15758–15763 (2013).
67. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010**, prot5448 (2010).
68. Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* **40**, e3 (2012).
69. Briggs, A. W. et al. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res.* **38**, e87 (2010).
70. Hodges, E. et al. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat. Protoc.* **4**, 960–974 (2009).
71. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729 (2013).
72. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
73. R Development Core Team *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2013).
74. Connor, T. R. et al. What's in a name? Species-wide whole-genome sequencing resolves invasive and noninvasive lineages of *Salmonella enterica* serotype Paratyphi B. *mBio* **7**, e00527-16 (2016).
75. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

76. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, New York, 2009).

Acknowledgements

This work was supported by the Max Planck Society (J.K.), the European Research Council (ERC) starting grant APGREID (to J.K.), Social Sciences and Humanities Research Council of Canada postdoctoral fellowship grant 756-2011-501 (to K.I.B.) and the Mäxi Foundation (M.G.C.). We thank the Archaeology Council at Mexico's INAH and the Teposcolula-Yucundaa Archaeological Project for sampling permissions. We are grateful to A. Wissgott, G. Brandt and V. Schuenemann for assistance with laboratory work, A. Günzel for providing graphical support for Fig. 1 and Supplementary Fig. 10, and R. Barquera, J. Hackett and M. Pi for thoughts and discussion on the manuscript. Part of the data storage and analysis was performed on the computational resource bwGRiD Cluster Tübingen funded by the Ministry of Science, Research and the Arts Baden-Württemberg, and the Universities of the State of Baden-Württemberg, Germany, within the framework programme bwHPC. We thank the MALT user community for helpful comments and bug reports.

Author contributions

K.I.B., M.G.C., A.H., N.T. and J.K. conceived the investigation. K.I.B., A.H. Å.J.V., M.G.C. and J.K. designed the experiments. N.M.R.G. provided archaeological information and drawings, submitted INAH permits and assisted in the sampling processes. Å.J.V., M.G.C., S.S., M.A.S. and K.I.B. performed the laboratory work. Å.J.V., A.H., K.I.B., C.W. and A.A.V. performed the analyses. D.H. implemented the MALT algorithm. A.H., J.K. and D.H. designed and set up the MALT ancient DNA analysis pipeline. C.W. performed the ethnohistorical analyses. Å.J.V. and K.I.B. wrote the manuscript with contributions from all authors.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-017-0446-6.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to A.H. or N.T. or K.I.B. or J.K.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding Author: Johannes Krause

Date: 11/15/2017

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

   Describe how sample size was determined.

   Our sample sizes were determined by the number of individuals excavated from the cemetery sites and the number of individuals to which we were permitted access to by the The Archaeology Council at Mexico's National Institute of Anthropology and History (INAH) and the Teposcolula-Yucundaa Archaeological Project.

2. Data exclusions

   Describe any data exclusions.

   No data were excluded from the analyses. Three genomes (Tepos_10, Tepos_20, Tepos_37) were excluded partway through analyses due to the lower quality or lower coverage of these genomes, the reasons for which are explained in the main text and supplementary information.

3. Replication

   Describe whether the experimental findings were reliably reproduced.

   Experimental findings were reproduced in that both non-UDG and UDG libraries from the same DNA extracts, for each of the ten positive individuals from the epidemic cemetery, yielded Salmonella enterica Paratyphi C genomic DNA after capture.

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   Samples were allocated to experimental groups based on the cemetery site from which the individuals were excavated from.

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   Blinding was not necessary to this study as neither human participants or animals were used.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☒	☐ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly.
☒	☐ A statement indicating how many times each experiment was replicated
☒	☐ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☒	☐ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☒	☐ The test results (e.g. *p* values) given as exact values whenever possible and with confidence intervals noted
☒	☐ A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☒	☐ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

All software used to analyze the data in this study is publicly available online. MALT is open source and freely available from http://ab.inf.uni-tuebingen.de/software/malt. Custom scripts were used in some instances to parse the data generated by publicly available software.

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The *Nature Methods* guidance for providing algorithms and software for publication may be useful for any submission.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No unique materials were used

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No eukaryotic cell lines were used

b. Describe the method of cell line authentication used.

No eukaryotic cell lines were used

c. Report whether the cell lines were tested for mycoplasma contamination.

No eukaryotic cell lines were used

d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No commonly misidentified cell lines were used

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

No animals were used

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	This study did not involve human research participants