# Ancient *Clostridium* DNA and variants of tetanus neurotoxins associated with human archaeological remains

Harold P. Hodgins[1], Pengsheng Chen[2], Briallen Lobb[1], Benjamin JM Tremblay[1], Michael J. Mansfield[3], Victoria CY Lee[1], Pyung-Gang Lee[2], Jeffrey Coffin[4], Xin Wei[1], Ana T. Duggan[5], Alexis E. Dolphin[4], Gabriel Renaud[6]*, Min Dong[2]*, Andrew C. Doxey[1]*

**Affiliations:**

[1]Department of Biology, University of Waterloo, Waterloo, Ontario, Canada

[2] Department of Urology, Boston Children's Hospital, Department of Surgery and Department of Microbiology, Harvard Medical School, Boston, MA, USA

[3]Genomics and Regulatory Systems Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa, Japan

[4]Department of Anthropology, University of Waterloo, Waterloo, Ontario, Canada

[5]McMaster Ancient DNA Centre, Department of Anthropology, McMaster University, Hamilton, Ontario, Canada

[6]Department of Health Technology, Section of Bioinformatics, Technical University of Denmark, Kongens Lyngby, Denmark.

*Corresponding authors. Email: acdoxey@uwaterloo.ca, Min.Dong@childrens.harvard.edu, gabre@dtu.dk

**Abstract:**

The analysis of microbial genomes from human archaeological samples offers a historic snapshot of ancient pathogens and provides insights into the origins of modern infectious diseases. Here, through large-scale metagenomic analysis of archeological samples, we discovered bacterial species related to modern-day *Clostridium tetani*, which produces tetanus neurotoxin (TeNT) and causes the disease tetanus. We were able to assemble 38 draft genomes from distinct human archeological samples spanning five continents from as early as ~4000 BCE. They display hallmarks of ancient DNA damage to variable degrees. While 20 fall into known *C. tetani* clades, phylogenetic analysis revealed novel *C. tetani* lineages, as well as a novel *Clostridium* species ("clade X") closely related to *C. tetani.* Within these genomes, we found 15 novel TeNT variants, including a unique lineage of TeNT found exclusively in ancient samples from South America. We experimentally produced and tested a TeNT variant selected from a ~6000-year-old Chilean mummy sample and found that it induced tetanus muscle paralysis in mice with extreme potency comparable to modern TeNT. Our work provides the first identification of neurotoxigenic *C. tetani* in ancient DNA, discovery of a new *Clostridium* species unique to ancient human samples, and a novel variant of tetanus neurotoxin that retains functional activity and ability to cause disease in mammals.

## INTRODUCTION

*Clostridium tetani*, the causative agent of the neuroparalytic disease tetanus, is an important bacterial pathogen of humans and animals. Its spores contaminate wounds, germinate and diffuse in oxygen-depleted and necrotic tissue, and produce a highly potent neurotoxin (tetanus neurotoxin, TeNT) that paralyzes hosts (*1*, *2*) leading to spastic paralysis. This may present as a local or a systemic effect, which can result in death due to paralysis of respiratory muscles and subsequent respiratory failure. As a wound-associated infectious disease, tetanus is estimated to have plagued *Homo sapiens* throughout history, from ancient wars to the first World War where it is estimated to have affected 32 in every 1000 wounded soldiers (*3*). Accounts of tetanus-like diagnostic features ("lockjaw") date back as far as the ancient Greeks through written descriptions by Hippocrates (c. 380 BCE) and the ancient Egyptians as seen in reports from the Edwin Smith papyrus (c. 1600 BCE) (*4*). These accounts highlight the ancient history of tetanus disease in humans.

*C. tetani* was first isolated and cultivated in 1889 (*5*) and an early isolate of *C. tetani* (the 1920 Harvard E88 strain) is still widely used as a reference today. Genome sequencing of the E88 reference strain revealed that it possesses a genome consisting of a single ~2.8 Mb chromosome and a ~74 kb plasmid (*6*). This genomic organization is largely maintained among all known strains of *C. tetani*, with different strains also varying in plasmid size (*7–9*). The plasmid is critical to pathogenicity as it encodes the key virulence genes including *tent* which encodes the neurotoxin and *colT* which encodes a collagenase enzyme involved in tissue degradation. Based on comparative genomic analysis, modern *C. tetani* strains cluster into two phylogenetically distinct clades (*9*), but are closely related and exhibit low genetic variation with average nucleotide identities of 96-99%. Similarly, the *tent* gene is extremely conserved and exhibits 99% to 100% amino acid identity across all strains. Currently available *C. tetani* genomes from modern times therefore offer a limited perspective on the true diversity of *C. tetani* and its evolutionary history as a human disease-causing bacterium.

The sequencing and analysis of ancient DNA (aDNA) from archeological samples provides unprecedented access to ancestral genomic information, and insights into the origins and evolution of modern species. In addition to human DNA, a significant proportion of genetic material preserved within ancient specimens is of microbial origin (*10*, *11*). Ancient microbial

DNA, including that from ancient pathogens that once impacted humans, can be found within mummified remains, fossilized feces, bones, teeth, and dental calculus (*12*). They can be distinguished from potential contaminants of modern microbial DNA based on signatures of ancient DNA damage (*13*, *14*). Particularly, damaged ancient DNA is known to accumulate an increased rate of C→T changes at the 5' ends and G→A changes at the 3' ends of sequence fragments, due to cytosine deamination in 5' overhanging ends. Damage rates exceeding 10% are indicative of genuinely ancient DNA (*15*). Once authenticated, reconstructed ancient microbial genomes can be compared with modern strains to investigate the genomic ancestry and adaptations underlying the emergence of historical epidemic strains. Groundbreaking aDNA studies on the evolutionary origins and emergence of major infectious diseases have been carried out in recent years including studies of *Mycobacterium tuberculosis* (*16*), the plague bacterium *Yersinia pestis* (*17*), *Mycobacterium leprae* (*18*), *Helicobacter pylori* (*19*), hepatitis B (*20*), and variola virus (*21*).

Here, through large-scale metagenomic data mining of millions of sequencing datasets, we report the discovery of novel *C. tetani* related genomes including neurotoxin genes in ancient human DNA samples. Some strains and neurotoxins are phylogenetically distinct from modern forms, and some strains show strong hallmarks of ancient DNA damage indicative of an ancient origin. We reconstructed a novel neurotoxin variant from a ~6,000 year old Chinchorro mummy sample and demonstrate that it possesses extreme potency comparable to modern neurotoxins. Our findings uncover a widespread occurrence of *C. tetani* and related species associated with aDNA samples, expanding our understanding of the evolution and diversity of this important human pathogen.

## RESULTS

### *Identification and assembly of C. tetani genomes from aDNA samples*

To explore the evolution and diversity of *C. tetani*, we performed a large-scale search of the entire NCBI Sequence Read Archive (SRA; 10,432,849 datasets from 291,458 studies totaling ~18 petabytes June 8, 2021) for datasets containing potential *C. tetani* DNA signatures. Since typical search methods (e.g., BLAST) could not be applied at such a large scale, we used the recently developed Sequence Taxonomic Analysis Tool (*22*) to search the SRA and identified 136 sequencing datasets possessing the highest total *C. tetani* DNA content [*k*-mer abundance >20,000 reads, k=32 base pair fragments uniquely matching the *C. tetani* genome] (Fig. 1A, table S1). Our search identified 28 previously sequenced *C. tetani* genomes (which serve as positive controls), as well as 108 uncharacterized sequencing runs (79 of human origin) with high levels of *C. tetani* DNA content. Unexpectedly, 76 (96.2%) of these are aDNA datasets collected from archeological human bone and tissue specimens (Fig. 1A), with the remaining three from modern human gut microbiome samples.

These 76 ancient DNA datasets are derived from 38 distinct archeological samples, spanning a timeframe of ~6000 years (Fig. 1B), including tooth samples from aboriginal inhabitants of the Canary Islands from the 7$^{th}$ to 11$^{th}$ centuries CE (*23*), tooth samples from the Sanganji Shell Mound of the Jomon in Japan (~1044 BCE) (*24*), Egyptian mummy remains from ~1879 BCE to 53 CE (*25*), and ancient Chilean Chinchorro mummy remains from ~3889 BCE (PRJEB9733) (*26*) (table S2). Although these archeological samples are of human origin, *k*-mer based analysis of the 38 DNA samples predicted a predominantly microbial composition (fig. S1). *C. tetani*-related DNA was consistently abundant among predicted microbial communities, detected at 13.82% average relative abundance (fig. S1, table S3). Interestingly, we also detected *M. tuberculosis* and *Y. pestis* in several of the *C. tetani* containing samples (fig. S1). These samples are associated with aDNA studies of *M. tuberculosis* and *Y. pestis* (*27–29*), but *C. tetani* was not previously identified.

To further verify the presence of *C. tetani* in aDNA samples, we mapped reads from each sample to the modern *C. tetani* reference (E88 strain) chromosome and plasmid. Coverage was evenly distributed across the chromosome for most samples (Fig. 1C), whereas coverage across the plasmid was more variable, with some samples lacking coverage for specific plasmid regions

and genes (Fig. 1D). Sequencing reads mapping to the *tent* gene were detected in 34/38 (89%) samples, whereas reads mapping to a second plasmid-encoded virulence gene, *colT,* were detected in all samples (Fig. 1E, table S4). Thus, all detected *C. tetani*-like genomes from ancient samples possess a chromosome and plasmid, and most appear to be toxigenic strains.

We next performed metagenome assembly for each individual sample and then taxonomically classified assembled contigs to identify those mapping unambiguously to *C. tetani* and not other bacterial species (fig. S2, tables S5, S6). The total length of assembled *C. tetani* contigs correlated well ($r = 0.76$, $p = 3.96\text{x}10^{-8}$, two-sided Pearson) with mapped read coverage, ranging from ~0.01 Mbp to a maximum of 2.84 Mbp, consistent with the expected genome size for *C. tetani* (fig. S3). Ultimately, this procedure resulted in 38 putative, ancient DNA associated clostridial metagenome-assembled genomes or "acMAGs", which were further assessed using CheckM (*30*) for percentage completion and contamination from divergent taxa or strain heterogeneity (Fig. 1E, table S7). Thirty-five acMAGs had low (<10%) contamination. Among them, 22 had a completeness exceeding 50% and 11 were complete at a 70% or greater level (table S7).

***A subset of C. tetani genomes from archaeological samples are of ancient origin***

Using MapDamage, we examined the 38 acMAGs for characteristic patterns of ancient DNA damage: notably, elevated patterns of deaminated cytosine residues concentrated toward the ends of molecules, and shorter fragment lengths (*13*, *14*). Seven acMAGs possessed a damage rate (5' C→T mutation rate) exceeding 10%, which is indicative of ancient DNA (*15*) (top 5 shown in Fig. 2A). The highest damage rate (19%) occurred in the acMAG from the "SLC-France-Tooth" sample (~1348 CE) (Fig. 2A). As controls, evidence of ancient DNA damage was also observed for corresponding human mitochondrial DNA (mtDNA) from the same ancient samples (Fig. 2A, fig. S4), but not for modern *C. tetani* samples (Fig. 2B). In general, we also observed a significant correlation between damage rates of acMAG DNA and human mtDNA ($R^2 = 0.46$, $p = 3.4\text{e-}06$, two-sided Pearson) (Fig. 2C), although human mtDNA rates were generally of higher magnitude (fig. S4). Finally, as expected, acMAGs had significantly shorter fragment lengths ($p = 3 \text{ x } 10^{-9}$, Mann-Whitney test) than those obtained from 21 sequencing datasets of modern *C. tetani* genomes (fig. S5, table S8). Together, these data suggest that a subset of *C. tetani* draft

genomes recovered from ancient DNA display evidence of ancient DNA damage and are plausibly of ancient origin.

### *Identification of novel* **C. tetani** *lineages and* **Clostridium** *species from ancient samples*

To explore the phylogenetic relationships involving the acMAGs and modern *C. tetani* strains, we constructed a whole genome phylogeny including 33 acMAGs and 37 known *C. tetani* genomes (*9*) (Fig. 3A). Five acMAGs were omitted due to extremely low (<1%) genome coverage, which could result in phylogenetic artifacts. We also included *C. cochlearium* as a phylogenetic outgroup, as it is the closest known relative of *C. tetani*. The genome-based phylogeny of the acMAGs and modern *C. tetani* strains (Fig. 3A) is highly consistent with expected phylogenetic structure, and contains all previously established *C. tetani* lineages (*9*). The tree topology was also supported by two independent methods of phylogenetic reconstruction (see Methods). Twenty acMAGs are clearly assigned to existing *C. tetani* lineages (Fig. 3A), including new members of clades 1B (N = 1), 1F (N = 1), 1H (N = 9), and 2 (N = 9), greatly expanding the known genomic diversity of clade 1H which previously contained a single strain and clade 2 which previously contained five strains (Fig. 3A). Interestingly, all nine newly identified lineage 1H acMAGs originate from ancient samples collected in the Americas (Fig. 3B). Four acMAGs clustered within clade 1 but fell outside of established sublineages (Fig. 3A).

The remaining nine acMAGs could not be assigned to any existing clade, and clustered as novel lineages (Fig. 3A). One sample ("GranCanaria-008-Tooth" from the Canary Islands dated to ~935 CE) forms a divergent lineage (labeled "Y") clustering outside all other *C. tetani* genomes. Based on CheckM analysis, this acMAG is of moderate quality with 74% completeness, and 0.47% contamination (table S7). Comparison of the GranCanaria-008-Tooth acMAG to the NCBI genome database revealed no close relatives (table S9). It exhibits an average nucleotide identity (ANI) of 87.4% to *C. tetani* E88, and 85.1% to *C. cochlearium*, below the 95% threshold typically used for species assignment (table S9). To further investigate the phylogenetic position of this species, we built gene-based phylogenies with ribosomal marker genes *rpsL, rpsG* and *recA* (see figs. S7-S9). Each of these three genes support the GranCanaria-008-Tooth lineage as a distinct and early-branching lineage outside of *C. tetani*. The *C. tetani*

damage level for this sample is relatively low (~4.0%), whereas its human mtDNA damage level is ~11.6% (fig. S4).

Even more surprising, however, is a novel clade of eight acMAGs (labeled lineage "X"), which clustered outside of the entire *C. tetani* tree. These samples span a large timeframe of ~2290 BCE to 1787 CE, are predominantly (7 of 8) of European origin (Fig. 3C, fig. S6), and come from variable burial contexts including single cave burials, cemeteries, mass graves and burial pits (*28*, *31–36*). Two of the samples from sites in Latvia and France are from plague (*Y. pestis*) victims (*28*, *36*), and another is from an individual with tuberculosis (*29*). The highest quality acMAG for this clade is from sample "Augsburg-Tooth" (~2253 BCE), with 59.8% estimated completeness and 5.56% contamination (table S7). Comparison of clade X acMAGs to other *Clostridium* species revealed no close relatives in the existing database. Clade X genomes are closest to *C. tetani* (86.33 +- 1.78 average nucleotide identity to E88 strain) followed by *C. cochlearium* (ANI = 85.16 +- 1.61) (Fig. 3D, table S10). As in the genome-wide tree, individual marker genes (*rpsL, rpsG* and *recA*) from clade X acMAGs also clustered as divergent lineages distinct from *C. tetani* and *C. cochlearium* (figs. S7-S9). Finally, we re-examined the damage patterns according to phylogenetic clade, and found that clade X genomes possess the highest mean damage; 6/8 clade X genomes have a damage level exceeding 5% and 3/8 exceed 10% (Fig. 3E). These analyses suggest that clade X represents a previously unidentified species of *Clostridium*, including members of ancient origin.

***Identification and experimental testing of a novel tetanus neurotoxin variants***

Given the considerable scientific and biomedical importance of clostridial neurotoxins, we next focused on *tent* and assembled a total of 20 *tent* gene sequences from aDNA: six with complete coverage, and fourteen with 75-99.9% coverage (table S11). Four are identical to modern *tent* sequences, while 16 (including two identical sequences) are novel *tent* variants with 99.1-99.9% nucleotide identity to modern *tent*, comparable to the variation seen among modern *tent* genes (98.6-100%). We then built a phylogeny including the 20 *tent* genes from aDNA and all 12 modern *tent* sequences (Fig. 4A). The *tent* genes clustered into four subgroups (Fig. 4A) with modern *tent* genes found in subgroups 1 and 3, and aDNA-associated *tent* genes found in subgroups '1' and '3', and also forming novel subgroups '2' and '4'. The *tent* sequence from

Cueva_de_los_Lagos-tooth (*C. tetani* clade "X") is the exclusive member of *tent* subgroup '4', and three *tent* sequences from clade 1H aDNA strains form the novel *tent* subgroup '2'.

We then visualized the uniqueness of aDNA-associated *tent* genes by mapping nucleotide substitutions onto the phylogeny (Fig. 4B, fig. S10), and focusing on "unique" *tent* substitutions found only in ancient samples and not in modern *tent* sequences. We identified a total of 54 such substitutions that are completely unique to one or more aDNA-associated *tent* genes (Fig. 4B, fig. S11, table S12), which were statistically supported by a stringent variant calling pipeline (table S13). Interestingly, the largest number of unique substitutions occurred in *tent* subgroup '2'. *tent*/Chinchorro, from the oldest sample in our dataset ("Chinchorro mummy bone", ~3889 BCE), possesses 18 unique substitutions not found in modern *tent*, and 12 of these are shared with *tent*/El-Yaral and 10 with *tent*/Chiribaya (Fig. 4B). The three associated acMAGs also cluster as neighbors in the phylogenomic tree (Fig. 3A), and the three associated archaeological samples originate from a similar geographic region in Peru and Chile (fig. S12). These shared patterns strongly suggest a common evolutionary origin for these *C. tetani* strains and their unique neurotoxin genes, and highlight *tent* subgroup 2 as a distinct group of *tent* variants exclusive to ancient samples (Fig. 4A).

We then focused on *tent/*Chinchorro as a representative sequence of this group as its full-length gene sequence could be completely assembled. The 18 unique substitutions present in the *tent/*Chinchorro gene result in 12 unique amino acid substitutions, absent from modern TeNT protein sequences (L140S, E141K, P144T, S145N, A147T, T148P, T149I, P445T, P531Q, V653I, V806I, H924R) (table S14). Seven of these substitutions are spatially clustered within a surface loop on the TeNT structure and represent a potential mutation "hot spot" (Fig. 4C). Interestingly, 7/12 amino acid substitutions found in TeNT/Chinchorro are also shared with TeNT/El-Yaral and 5/12 are shared with TeNT/Chiribaya (table S14). As highlighted in Fig. 4C, TeNT/Chinchorro and TeNT/El-Yaral share a divergent 9-aa segment (amino acids 141-149 in TeNT, P04958) that is distinct from all other TeNT sequences. Reads mapping to the *tent/*Chinchorro gene show a low damage level similar to that seen in *C. tetani* contigs, and it is weaker than the corresponding damage pattern from the associated human mitochondrial DNA (Fig. 4D).

Next, given the phylogenetic novelty and unique pattern of substitutions observed for the *tent*/Chinchorro gene, we sought to determine whether it encodes an active tetanus neurotoxin. For biosafety reasons, we avoided the production of a *tent*/Chinchorro gene construct and instead used sortase-mediated ligation to produce limited quantities of full-length protein toxin (fig. S13), as done previously for other neurotoxins (*37*, *38*). This involved producing two recombinant proteins in *E. coli*, one constituting the N-terminal fragment and another containing the C-terminal fragment of TeNT/Chinchorro, and then ligating these together using sortase. The resulting full-length TeNT/Chinchorro protein cleaved the canonical TeNT substrate, VAMP2, in cultured rat cortical neurons, and can be neutralized with anti-TeNT anti-sera (Fig. 4E, fig. S13). TeNT/Chinchorro induced spastic paralysis *in vivo* in mice when injected to the hind leg muscle, which displayed a classic tetanus-like phenotype identical to that seen for wild-type TeNT (Fig. 4F). Quantification of muscle rigidity following TeNT and TeNT/Chinchorro exposure demonstrated that TeNT/Chinchorro exhibits an extreme level of potency that is indistinguishable from TeNT (Fig. 4G). Together, these data demonstrate that the reconstructed *tent*/Chinchorro gene encodes an active and highly potent TeNT variant.

## DISCUSSION

In this work, large-scale data mining of millions of existing genomic datasets revealed widespread occurrence of neurotoxigenic *C. tetani* and related lineages of *Clostridium* in aDNA samples from human archaeological remains. Our study provides three main findings: 1) the first identification of neurotoxigenic *C. tetani* from archaeological samples including several *C. tetani* strains of plausibly ancient origin; 2) the discovery of novel lineages of *C. tetani* as well as an entirely new species of *Clostridium* (clade X); and 3) the identification of novel variants of TeNT including TeNT/Chinchorro which we demonstrate to be an active neurotoxin with extreme potency comparable to modern TeNT variants.

Our work is unique from previous studies of aDNA in several respects. First, using recent advances in petabase-scale genomic data-mining (*22*), we were able to perform a large-scale survey of sequencing datasets including all available DNA samples in the NCBI SRA, greatly enhancing our ability to discover patterns across spatially and temporally diverse datasets. Because of the massive size of the NCBI SRA database, traditional sequence-alignment based

methods such as BLAST are not computationally feasible. By *k*-mer based indexing of genomes, the STAT method (*22*) provided a heuristic approach to identify potential datasets from the SRA that could then be targeted for deeper analysis including metagenomic assembly and phylogenomics. Importantly, we did not specifically look for *C. tetani* in ancient DNA, but rather this came as an unexpected finding from the results of our large-scale screen. Also unexpected was the considerable diversity of ancient samples in which we identified neurotoxigenic *C. tetani*, which revealed a strong association between this organism (and related species) and human archaeological samples. Despite the abundance of environmental (e.g., soil metagenomic) samples in the SRA, these samples did not come to the surface of our genomic screen for *C. tetani*. This is consistent with the idea that, although *C. tetani* spores may be ubiquitous in terrestrial environments such as soil, these spores may be rare and so *C. tetani* DNA may not regularly appear at appreciable levels in shotgun metagenomes.

However, it is important to point out that, unlike other examples of ancient pathogens such as TB or plague, the identification of neurotoxigenic *C. tetani* in aDNA samples alone is not sufficient to implicate tetanus as a cause of death or even suggest that the corresponding *C. tetani* strains are contemporaneous with the archaeological samples. A variety of environmental factors and mechanisms may account for the presence of toxigenic clostridia in aDNA samples, including the possibility of post-mortem colonization by environmental clostridia. The majority (31/38) of *C. tetani* aDNA samples in our study originated from teeth, as teeth are commonly used in aDNA studies due to the survival and concentration of endogenous aDNA content (*39*). Interestingly, bone and tooth infections by *C. tetani* in patients have been previously but infrequently reported (*40*, *41*). Thus, it is possible that some of the identified *C. tetani* strains are the result of post-mortem bone or tooth infection by environmental microbes or even human contamination of archaeological samples after death. This explanation may account for the observation of low *C. tetani* damage rates but high human mtDNA rates in some samples. For other samples, the *C. tetani* damage levels (>10%) are indicative of an ancient origin, and it remains unknown whether these strains are the result of ancient sample colonization or contamination, or whether they are as old as the archaeological samples themselves.

Regardless of whether the identified *C. tetani* genomes are contemporaneous with the archaeological samples, an important finding of this work is the substantial expansion of genomic knowledge surrounding *C. tetani* and its relatives, such as the expansion of clade 2 and

clade 1H, as well as the discovery of lineages X and Y. Lineage 1H in particular has undergone the greatest expansion through the newly identified aDNA-associated *C. tetani* genomes, from one known sample derived from a patient in France in 2016 (*9*), to 9 additional draft genomes assembled from ancient DNA. This may indicate that a broader diversity of 1H strains exists in undersampled environments, human tissues, or other animal hosts. Interestingly, these newly identified lineage 1H strains share a common pattern of originating from the Americas, suggesting that sample handling or perhaps a common region-specific (or regionally abundant) environmental *C. tetani* strain has colonized these samples at some point in the past.

In addition to the expansion of existing lineages, our genomic analysis revealed two highly unique lineages of *Clostridium* that are closely related to, but distinct from, *C. tetani*. One of these novel lineages ("Y") was assembled from an aDNA sample (GranCanaria-008-Tooth) taken from an archeological specimen dated to 936CE. Clustering outside of the entire *C. tetani* tree based on three phylogenetic analyses, this lineage may be derived from an ancient lineage of *C. tetani* that predates the emergence of clade 1 and 2 genomes. Lineage Y also appears to be toxigenic, possessing a *tent* variant that has a unique substitution profile including substitutions not observed in any other *tent* sequences (modern or aDNA-associated).

Perhaps even more intriguing is clade "X", a group of closely related *Clostridium* strains that also formed a sister lineage to *C. tetani* and yet resemble no other species that has been sequenced to date. This clade is unlikely to have arisen by errors in genome sequencing or assembly as it is supported by the co-clustering of multiple genomes as well as the consistently divergent placement of clade X species in individual marker gene phylogenies. These organisms possess a *C. tetani*-like plasmid, and some strains (e.g., "Cueva_de_los_Lagos-Tooth") may be toxin-encoding. However, it is important to note that *tent* was only recovered from this single clade X-associated sample, and with lower coverage relative to other plasmid-associated genes (*colT*). It is therefore possible that the apparent presence of *tent* is due to contamination by other *C. tetani* strains in this sample. Indeed, CheckM estimated 2.51% contamination, 12.5% of which was estimated to be due to strain variation. Further understanding of clade X may be addressed through future efforts to sequence the microbiomes associated with archaeological samples, as well as environmental *Clostridium* isolates.

Beyond expanding *C. tetani* and *Clostridium* genomic diversity, our work also expands the known diversity of clostridial neurotoxins - the most potent family of toxins known to science. Analysis of ancient DNA revealed novel variants and lineages of *TeNT*, including the newly identified "subgroup 2" toxins: TeNT/Chinchorro, TeNT/El-Yaral toxins, and TeNT/Chiribaya-Alta. Not only do these toxins share a similar mutational profile, but they are derived from a similar geographic area (regions of Peru and Chile in South America) and their associated *C. tetani* genomes also cluster phylogenetically as the closest neighbors. Of the three subgroup 2 *tent* sequences identified, one of them (*tent*/Chinchorro) had sufficient coverage to be fully assembled, and the *tent*/Chinchorro gene also happened to be most divergent from modern *tent* sequences by possessing the greatest number of unique substitutions. Despite being the most divergent *tent*, reads mapping to the *tent*/Chinchorro gene as well as the associated *C. tetani* MAG did not show strong patterns of DNA damage, and the damage level was weaker than that for human mtDNA. This indicates that, despite originating from the oldest sample in our dataset and possessing a unique *tent* variant, it is possible that the Chinchorro mummy associated *C. tetani* strain may be a relatively "newer" strain that colonized or contaminated the sample post-mortem.

Due to the uniqueness of TeNT/Chinchorro, and its collection of amino acid substitutions that have not been observed in any modern TeNT variants, we sought to determine whether this TeNT variant is a functional neurotoxin. Lack of toxicity, for instance, might indicate a sequence artifact or even a TeNT variant that targets other non-mammalian species. We therefore utilized a previous approach based on sortase-mediated ligation to produce small quantities of the full-length protein toxin (*37*, *38*). TeNT/Chinchorro produced a classic tetanus phenotype in mouse assays, and exhibits extreme potency at a level comparable to modern TeNT. This validated our predicted neurotoxic activity for this gene sequence, and suggests that TeNT/Chinchorro's multiple unique amino acid substitutions have limited impact on potency and neurotoxicity. However, their non-random spatial clustering on a specific surface-exposed loop of the neurotoxin structure suggests that they may be a result of positive selection, a pattern that has been commonly observed in protein evolutionary studies (*42*, *43*). Such substitutions may alter yet-to-be identified TeNT protein-protein interactions. In addition to validating the predicted activity of TeNT/Chinchorro, this experimental method allowed us to directly test the disease-causing properties of a novel virulence factor variant reconstructed from aDNA, without the

necessity for growing the organism itself, which would be potentially dangerous and possibly required for other model pathogens.

In summary, using large-scale data mining, we identified ancient neurotoxigenic clostridia in archeological samples. This resulted in a substantial expansion of the known genomic diversity and occurrence of *C. tetani*, and led to the discovery of novel *C. tetani* lineages, *Clostridium* species, and tetanus neurotoxin variants that retain functional activity. The discovery of neurotoxigenic clostridial genomes in such a wide diversity of ancient samples, both geographically and temporally, is unexpected, but perhaps not inconsistent with prior hypotheses about the role of these organisms in the natural decomposition process (*44*, *45*). Although the precise origin of these strains in ancient samples remains difficult to determine at present, we anticipate that future exploration of these and additional ancient archaeological samples will shed further light on the genomic and functional diversity of these fascinating organisms, as well as the ecology and evolutionary origins of their remarkably potent neurotoxins.

## METHODS

### *Identification of C. tetani containing samples from the NCBI sequence read archive*

To identify datasets within the NCBI sequence read archive containing *C. tetani*, we performed a query of the NCBI-stat database, which stores pre-computed analyses of taxonomy using the NCBI-stat tool(*22*)). Google's cloud sdk was used to perform a Google Big Query search of the SRA STAT meta-data for matches to tax_id=1513 on March 15, 2021. All source code is available on github at: https://github.com/harohodg/aDNA-tetanus-analysis.

The query processed 86.16 GB and returned 43,620 sample hits with a *k*-mer self count ranging from 1 to 17,152,980. A threshold of 20,000 was applied which returned 136 hits including 28 *C. tetani* sequencing projects (positive controls). The SRA taxonomic profile from NCBI-stat, retrieved using a separate Google Big Query search, was used to assess the microbial community in each sample. Total counts of each mapped bacterial and archaeal taxon at the species level were extracted from the profile. Taxa counts were converted to proportional values and subsequently visualized in R v4.0.4.

FASTQ files of identified sequencing runs were downloaded using the sra-toolkit v2.9.6, and quality encodings of all runs were assessed. Eight runs were Phred+64 encoded and were

converted to Phred+33 using seqtk v1.3. Twenty-three runs had an unknown encoding and were assumed to be Phred+33 encoded based on the range of the quality scores.

***Measurement and visualization of genome coverage***

Bowtie2 v2.4.2 was used to map reads from individual runs to the E88 chromosome (accession NC_004557.1) and plasmid (accession NC_004565.1). Bowtie2 was run with the following parameters (--local -D 20 -R 3 -N 1 -L 20 -i S,1,0.50 -bSF4). Using samtools v1.12, the resultant BAM files were then sorted, indexed, and merged based on their BioSample ID. Total (average # of reads per base) and percent (number of bases with 1 or more reads divided by total number of bases) coverage was calculated for the entire chromosome and plasmid as well as the *tent* (68640 - 72587) and *colT* (39438 - 42413) regions. Coverage was visualized using Python v3.8.5 and matplotlib v3.3.2. Circular plots were created using R and a custom script.

Circular coverage plots were generated by loading the BAM files into R v4.1.0 with the Rsamtools library v2.8.0 and plotted as area plots using functions from the circlize library v0.4.12. Coverage was calculated by averaging the number of reads per base in 300bp bins for the plasmid sequences, and 11,250bp bins for the chromosome sequences. Values were capped to the 90th percentile to prevent high coverage regions from obscuring other regions. Genes were plotted as black bars using RefSeq annotations. For the plasmid plots, the *tent* (68,640-72,587) and *colT* (39,438-42,413) genes were also coloured red and blue, respectively.

***Genome reconstruction***

Reads were pre-processed using fastp v0.20.1 with default settings to perform quality filtering and remove potential adapters. FASTQ pre-processing statistics are included in table S15. Metagenome co-assembly, using all reads with the same BioSample ID, was performed using megahit v1.2.9 with default parameters(*46*). Contigs were then taxonomically classified using Kaiju v1.7.4(*47*) against the Kaiju database nr 2021-02-24 with default settings. Any contigs mapped to *C. tetani* (NCBI taxonomy ID 1513) or any of its strains (NCBI taxonomy IDs 1231072, 212717, 1172202, and 1172203) were selected for further analyses. The length of total *C. tetani* contigs was compared to the mapped read coverage with cor.test() in R v4.0.4. ANI was calculated using fastANI v1.33. CheckM v1.0.18(*30*) was used on the contigs identified as *C. tetani* with the pre-built set of *Clostridium* markers supplied with the tool to calculate

completeness, contamination, and strain heterogeneity. Contigs are available in the public github repository.

***Analysis of ancient DNA damage***

Fastq files were pre-processed using leeHom v1.2.15(*48*) to remove adapters and to perform Bayesian reconstruction of aDNA. The --ancientdna flag was applied only to paired end datasets. The leeHom output was then merged by bioSample ID (concatenated sequentially into one file). Individual and merged results were then processed using seqtk v1.3 with the "seqtk seq -L30" command to remove short sequences < 30 bp in length. For each BioSample, trimmed reads were then mapped using bwa v0.7.17 to the contigs that classified as *C. tetani* using Kaiju, and separately to the human mitochondrial reference genome (accession NC_012920.1). Read alignment was performed using "bwa aln" with the "-n 0.01 -o 2 -l 16500" options. BAM files were sorted using samtools v1.12. Misincorporation rates were measured for all samples using mapDamage v2.2.1 with the --merge-reference-sequences and --no-stats parameters.

***Whole genome based phylogenetic reconstruction***

Single base substitutions within assembled *C. tetani* contigs were identified using snippy-multi from the Snippy package v4.6.0 (https://github.com/tseemann/snippy) with *C. tetani* str. E88 as the reference genome (GCA_000007625.1_ASM762v1_genomic.gbff). A genome-wide core alignment was constructed using snippy-core. Five aDNA samples (SAMEA103957995, SAMEA103971604, SAMEA3486793, SAMEA104402285, SAMEA3937653) were removed due to very poor alignment coverage (<1%). Using the resulting alignment, we built a phylogeny using FastTree (*49*) v2.1.10 with the GTR model and aLRT metric for assessment of clade support. A maximum-likelihood phylogeny was also constructed using RAxML (*50*) with a GTR+GAMMA substitution model and 1000 rapid bootstrap inferences. The alignment and trees can be found in the public github repository.

***Sequence, structural, and phylogenetic analysis of ancient tetanus neurotoxins***

*Variant calling and construction of the MSA:* Scripts used for variant calling and generation of a *tent* multiple alignment are located in the public github repository. For the plasmid read alignments used earlier, we extracted aligned reads, and re-aligned them using BWA mem

v0.7.17-r1188 using default parameters. Read alignments were manipulated with samtools v1.12 and htslib v1.12. The read alignment was restricted to the *tent* gene locus for variant calling (using the reverse complement of NC_004565.1, bases 1496-5443). Variants were called on each individual sample using the Octopus variant caller v0.7.4 (*51*) with stringent parameters (--mask-low-quality-tails 5 --min-mapping-quality 10 --min-variant posterior 0.95 --min-pileup-base-quality 35 --min-good-base-fraction 0.75). This combination of parameters reports only variants with very high confidence and read mapping quality, minimizing identification of false positive variant calls. We then built consensus sequences of *tent* genes from each sample using the bcftools consensus tool v1.12, and htslib v1.12, replacing positions with 0 coverage with a gap character. MAFFT v7.4.80 (*52*) was used to realign fragments against the reference sequence using the --keeplength option, which notably keeps the length of the reference unchanged and therefore ignores the possibility of unique insertions. The final *tent* alignments are available in the public github repository.

*Structural modeling:* A structural model of TeNT/Chinchorro was generated by automated homology modeling using the SWISSMODEL server (*53*). Modeling was performed using two top-scoring homologous template structures of tetanus neurotoxins: PDB IDs 7by5.1.A (97.18% identity), 5N0C.1.A (97.34% identity). 7BY5.1.A was selected as the best template based on the QMEAN quality estimate (*54*). The model was visualized using PyMOL v2.4.1 and unique substitutions (present in TeNT/Chinchorro but absent in modern TeNT sequences) were highlighted.

*Phylogenetic analysis:* The *tent* consensus alignment generated as described earlier was processed to keep only sequences (N = 20) with alignment coverage exceeding 80%. The following BioSamples were removed: SAMEA104402285, SAMEA104281225, SAMEA104281219, SAMEA5054093, SAMN02799091, SAMEA103971604, SAMN02799089, SAMN12394113, SAMN06046901, SAMEA104233049, SAMEA6502100, SAMEA3486793, SAMEA3713711. We then aligned the 20 ancient *tent* gene sequences with 30 *tent* sequences from modern *C. tetani* strains, which reduced to 12 representative modern *tent* sequences after duplicates were removed using Jalview v2.9.0b2. *tent*/E88 was identical with *tent* from 11 strains (1586-U1, CN655, 641.84, C2, Strain_3, 75.97, 89.12, 46.1.08, A, 4784A, Harvard), *tent*/132CV with 1 other (Mfbjulcb2), *tent*/63.05 with 2 others (3483, 184.08), *tent*/1337 with 2 others (B4, 1240), *tent*/ATCC_453 with 1 other (3582), and *tent*/202.15 with 1

other (358.99). A phylogeny was constructed using PhyML v3.1 (*55*) with the GTR model, empirical nucleotide equilibrium frequencies, no invariable sites, across site rate variation optimized, NNI tree search, and BioNJ as the starting tree. PhyML analysis identified 362 patterns, and aLRT (SH-like) branch supports were calculated. The final newick tree is available in the public github repository.

***Experimental testing of TeNT/Chinochorro (chTeNT)***

*Antibodies and constructs:* Antibodies for Syntaxin-1 (HPC-1), SNAP25 (C171.2), VAMP1/2/3 (104102) were purchased from Synaptic Systems. Antibody against actin (AC-15) was purchased from Sigma. Rabbit antiserum of TeNT (ab53829) was purchased from Abcam, rabbit non-immunized serum (AB110) was purchased from Boston Molecules. The cDNAs encoding chTeNT-LC-$H_N$ (the N-terminal fragment, residues 1-870) and chTeNT-$H_C$ (the C-terminal fragment, residues 875-1315) were synthesized by Twist Bioscience (South San Francisco, CA). The cDNA encoding TeNT-LC-$H_N$ (residues 1-870) and TeNT-$H_C$ were synthesized by GenScript (Piscataway, NJ). A thrombin protease cleavage site was inserted between I448 and A457 in both TeNT-LC-$H_N$ and chTeNT-LC-$H_N$. LC-$H_N$ fragments were cloned into pET28a vector, with peptide sequence LPETGG fused to their C-termini, followed by a His6-tag. $H_C$ fragments were cloned into pET28a vectors with a His6-tag and thrombin recognition site on their N-termini.

*Protein purification: E. coli* BL21 (DE3) was utilized for protein expression. In general, transformed bacteria were cultured in LB medium using an orbital shaker at 37 °C until OD600 reached 0.6. Induction of protein expression was carried out with 0.1 mM IPTG at 18 °C overnight. Bacterial pellets were collected by centrifugation at 4,000 g for 10 min and disrupted by sonication in lysis buffer (50 mM Tris pH 7.5, 250 mM NaCl, 1 mM PMSF, 0.4 mM lysozyme), and supernatants were collected after centrifugation at 20,000 g for 30 min at 4 °C. Protein purification was carried out using a gravity nickel column, then purified proteins were desalted with PD-10 columns (GE, 17-0851-01) and concentrated using Centrifugal Filter Units (EMD Millipore, UFC803008).

*Sortase ligation:* $H_C$ protein fragments were cleaved by thrombin (40 mU/μL) (EMD Millipore, 605157-1KU) overnight at 4 °C. Ligation reaction was set up in 100 μL TBS buffer with LC-$H_N$ (8 μM), $H_C$ (5 μM), Ca2+ (10mM) and sortase (1.5 μM), for 1 hour at room

temperature. Then full-length proteins were activated by thrombin (40 mU/μL) at room temperature for 1 hour. Sortase ligation reaction mixtures were analyzed by Coomassie blue staining and quantified by BSA reference standards.

*Neuron culture and immunoblot analysis:* Primary rat cortical neurons were prepared from E18-19 embryos using a papain dissociation kit (Worthington Biochemical) following the manufacturer's instruction. Neurons were exposed to sortase ligation mixtures with or without antiserum in culture medium for 12 hrs. Cells were then lysed with RIPA buffer with protease inhibitor cocktail (Sigma-Aldrich). Lysates were centrifuged at 12000 g at 4 °C for 10 min. Supernatants were subjected to SDS–PAGE and immunoblot analysis.

*Animal study:* All animal studies were approved by the Boston Children's Hospital Institutional Animal Care and Use Committee (Protocol Number: 18-10-3794R). Toxins were diluted using phosphate buffer (pH 6.3) containing 0.2% gelatin. Mice (CD-1 strain, female, purchased from Envigo, 6-7 weeks old, 25–28 g, n=3) were anesthetized with isoflurane (3–4%) and injected with toxin (10 μL) using a 30-gauge needle attached to a sterile Hamilton syringe, into the gastrocnemius muscles of the right hind limb, and the left leg served as negative control. Muscle paralysis was observed for 4 days. The severity of spastic paralysis was scored with a numerical scale modified from a previous report (0, no symptoms; 4, injected limb and toes are fully rigid)(*56*).

**References:**


1. M. Popoff, Tetanus in animals. *J Vet Diagn Invest*. **32**, 184–191 (2020).
2. A. Megighian, M. Pirazzini, F. Fabris, R. Rossetto, C. Montecucco, Tetanus and tetanus neurotoxin: From peripheral uptake to central nervous tissue targets. *J Neurochem*. **158**, 1244–1253 (2021).
3. D. Bruce, Note on the incidence of tetanus among wounded soldiers. *Br Med J*. **1**, 118 (1917).
4. S. G. F. Wassilak, K. Kretsinger, in *Bacterial Infections of Humans* (Springer US, Boston, MA, 2009), pp. 813–832.
5. S. Kitasato, Ueber den Tetanusbacillus. *Z Hyg*. **7**, 225–34 (1889).
6. H. Bruggemann, S. Baumer, W. Fricke, A. Wiezer, H. Liesegang, I. Decker, C. Herzberg, R. Martinez-Arias, R. Merkl, A. Henne, G. Gottschalk, The genome sequence of Clostridium tetani, the causative agent of tetanus disease. *Proc Natl Acad Sci U S A*. **100**, 1316–1321 (2003).
7. H. Brüggemann, E. Brzuszkiewicz, D. Chapeton-Montes, L. Plourde, D. Speck, M. R. Popoff, Genomics of Clostridium tetani. *Research in Microbiology*. **166**, 326–331 (2015).
8. J. E. Cohen, R. Wang, R. F. Shen, W. W. Wu, J. E. Keller, Comparative pathogenomics of Clostridium tetani. *PLoS ONE*. **12** (2017), doi:10.1371/journal.pone.0182909.
9. D. Chapeton-Montes, L. Plourde, C. Bouchier, L. Ma, L. Diancourt, A. Criscuolo, M. Popoff, H. Brüggemann, The population structure of Clostridium tetani deduced from its pan-genome. *Sci Rep*. **9** (2019), doi:10.1038/S41598-019-47551-4.

10. K. Bos, D. Kühnert, A. Herbig, L. Esquivel-Gomez, A. Andrades Valtueña, R. Barquera, K. Giffin, A. Kumar Lankapalli, E. Nelson, S. Sabin, M. Spyrou, J. Krause, Paleomicrobiology: Diagnosis and Evolution of Ancient Pathogens. *Annu Rev Microbiol*. **73**, 639–666 (2019).
11. M. Drancourt, D. Raoult, Palaeomicrobiology: current issues and perspectives. *Nat Rev Microbiol*. **3**, 23–35 (2005).
12. C. Warinner, C. Speller, M. Collins, A new era in palaeomicrobiology: prospects for ancient dental calculus as a long-term record of the human oral microbiome. *Philos Trans R Soc Lond B Biol Sci*. **370** (2015), doi:10.1098/RSTB.2013.0376.
13. A. Briggs, U. Stenzel, P. Johnson, R. Green, J. Kelso, K. Prüfer, M. Meyer, J. Krause, M. Ronan, M. Lachmann, S. Pääbo, Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci U S A*. **104**, 14616–14621 (2007).
14. J. Dabney, M. Meyer, S. Pääbo, Ancient DNA damage. *Cold Spring Harb Perspect Biol*. **5** (2013), doi:10.1101/CSHPERSPECT.A012567.
15. S. Sawyer, J. Krause, K. Guschanski, V. Savolainen, S. Pääbo, Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS ONE*. **7** (2012), doi:10.1371/JOURNAL.PONE.0034131.
16. K. Bos, K. Harkins, A. Herbig, M. Coscolla, N. Weber, I. Comas, S. Forrest, J. Bryant, S. Harris, V. Schuenemann, T. Campbell, K. Majander, A. Wilbur, R. Guichon, D. Wolfe Steadman, D. Cook, S. Niemann, M. Behr, M. Zumarraga, R. Bastida, D. Huson, K. Nieselt, D. Young, J. Parkhill, J. Buikstra, S. Gagneux, A. Stone, J. Krause, Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature*. **514**, 494–497 (2014).
17. K. Bos, V. Schuenemann, G. Golding, H. Burbano, N. Waglechner, B. Coombes, J. McPhee, S. DeWitte, M. Meyer, S. Schmedes, J. Wood, D. Earn, D. Herring, P. Bauer, H. Poinar, J. Krause, A draft genome of Yersinia pestis from victims of the Black Death. *Nature*. **478**, 506–510 (2011).
18. V. Schuenemann, P. Singh, T. Mendum, B. Krause-Kyora, G. Jäger, K. Bos, A. Herbig, C. Economou, A. Benjak, P. Busso, A. Nebel, J. Boldsen, A. Kjellström, H. Wu, G. Stewart, G. Taylor, P. Bauer, O. Lee, H. Wu, D. Minnikin, G. Besra, K. Tucker, S. Roffey, S. Sow, S. Cole, K. Nieselt, J. Krause, Genome-wide comparison of medieval and modern Mycobacterium leprae. *Science*. **341**, 179–183 (2013).
19. F. Maixner, B. Krause-Kyora, D. Turaev, A. Herbig, M. R. Hoopmann, J. L. Hallows, U. Kusebauch, E. E. Vigl, P. Malfertheiner, F. Megraud, N. O'Sullivan, G. Cipollini, V. Coia, M. Samadelli, L. Engstrand, B. Linz, R. L. Moritz, R. Grimm, J. Krause, A. Nebel, Y. Moodley, T. Rattei, A. Zink, The 5300-year-old Helicobacter pylori genome of the Iceman. *Science*. **351**, 162–165 (2016).
20. B. Mühlemann, T. Jones, P. Damgaard, M. Allentoft, I. Shevnina, A. Logvin, E. Usmanova, I. Panyushkina, B. Boldgiv, T. Bazartseren, K. Tashbaeva, V. Merz, N. Lau, V. Smrčka, D. Voyakin, E. Kitov, A. Epimakhov, D. Pokutta, M. Vicze, T. Price, V. Moiseyev, A. Hansen, L. Orlando, S. Rasmussen, M. Sikora, L. Vinner, A. Osterhaus, D. Smith, D. Glebe, R. Fouchier, C. Drosten, K. Sjögren, K. Kristiansen, E. Willerslev, Ancient hepatitis B viruses from the Bronze Age to the Medieval period. *Nature*. **557**, 418–423 (2018).
21. A. T. Duggan, M. F. Perdomo, D. Piombino-Mascali, S. Marciniak, D. Poinar, M. v. Emery, J. P. Buchmann, S. Duchêne, R. Jankauskas, M. Humphreys, G. B. Golding, J. Southon, A. Devault, J. M. Rouillard, J. W. Sahl, O. Dutour, K. Hedman, A. Sajantila, G. L. Smith, E. C. Holmes, H. N. Poinar, 17 th Century Variola Virus Reveals the Recent History of Smallpox. *Curr Biol*. **26**, 3407–3412 (2016).
22. K. Katz, O. Shutov, R. Lapoint, M. Kimelman, J. Brister, C. O'Sullivan, STAT: a fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next-generation sequence submissions. *Genome Biol*. **22** (2021), doi:10.1186/S13059-021-02490-0.
23. R. Rodríguez-Varela, T. Günther, M. Krzewińska, J. Storå, T. Gillingwater, M. MacCallum, J. Arsuaga, K. Dobney, C. Valdiosera, M. Jakobsson, A. Götherström, L. Girdland-Flink, Genomic Analyses of Pre-European Conquest Human Remains from the Canary Islands Reveal Close Affinity to Modern North Africans. *Curr Biol*. **27**, 3396-3402.e5 (2017).
24. H. Kanzawa-Kiriyama, K. Kryukov, T. Jinam, K. Hosomichi, A. Saso, G. Suwa, S. Ueda, M. Yoneda, A. Tajima, K. Shinoda, I. Inoue, N. Saitou, A partial nuclear genome of the Jomons who lived 3000 years ago in Fukushima, Japan. *J Hum Genet*. **62**, 213–221 (2017).
25. J. Neukamm, S. Pfrengle, M. Molak, A. Seitz, M. Francken, P. Eppenberger, C. Avanzi, E. Reiter, C. Urban, B. Welte, P. Stockhammer, B. Teßmann, A. Herbig, K. Harvati, K. Nieselt, J. Krause, V. Schuenemann, 2000-year-old pathogen genomes reconstructed from metagenomic analysis of Egyptian mummified individuals. *BMC Biol*. **18** (2020), doi:10.1186/S12915-020-00839-8.

26. M. Raghavan, M. Steinrücken, K. Harris, S. Schiffels, S. Rasmussen, M. DeGiorgio, A. Albrechtsen, C. Valdiosera, M. Ávila-Arcos, A. Malaspinas, A. Eriksson, I. Moltke, M. Metspalu, J. Homburger, J. Wall, O. Cornejo, J. Moreno-Mayar, T. Korneliussen, T. Pierre, M. Rasmussen, P. Campos, P. de Barros Damgaard, M. Allentoft, J. Lindo, E. Metspalu, R. Rodríguez-Varela, J. Mansilla, C. Henrickson, A. Seguin-Orlando, H. Malmström, T. Stafford, S. Shringarpure, A. Moreno-Estrada, M. Karmin, K. Tambets, A. Bergström, Y. Xue, V. Warmuth, A. Friend, J. Singarayer, P. Valdes, F. Balloux, I. Leboreiro, J. Vera, H. Rangel-Villalobos, D. Pettener, D. Luiselli, L. Davis, E. Heyer, C. Zollikofer, M. Ponce de León, C. Smith, V. Grimes, K. Pike, M. Deal, B. Fuller, B. Arriaza, V. Standen, M. Luz, F. Ricaut, N. Guidon, L. Osipova, M. Voevoda, O. Posukh, O. Balanovsky, M. Lavryashina, Y. Bogunov, E. Khusnutdinova, M. Gubina, E. Balanovska, S. Fedorova, S. Litvinov, B. Malyarchuk, M. Derenko, M. Mosher, D. Archer, J. Cybulski, B. Petzelt, J. Mitchell, R. Worl, P. Norman, P. Parham, B. Kemp, T. Kivisild, C. Tyler-Smith, M. Sandhu, M. Crawford, R. Villems, D. Smith, M. Waters, T. Goebel, J. Johnson, R. Malhi, M. Jakobsson, D. Meltzer, A. Manica, R. Durbin, C. Bustamante, Y. Song, R. Nielsen, E. Willerslev, Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science*. **349** (2015), doi:10.1126/SCIENCE.AAB3884.

27. A. Namouchi, M. Guellil, O. Kersten, S. Hänsch, C. Ottoni, B. v. Schmid, E. Pacciani, L. Quaglia, M. Vermunt, E. L. Bauer, M. Derrick, A. Jensen, S. Kacki, S. K. Cohn, N. C. Stenseth, B. Bramanti, Integrative approach using Yersinia pestis genomes to revisit the historical landscape of plague during the Medieval Period. *Proc Natl Acad Sci U S A*. **115**, E11790–E11797 (2018).

28. K. I. Bos, A. Herbig, J. Sahl, N. Waglechner, M. Fourment, S. A. Forrest, J. Klunk, V. J. Schuenemann, D. Poinar, M. Kuch, G. B. Golding, O. Dutour, P. Keim, D. M. Wagner, E. C. Holmes, J. Krause, H. N. Poinar, Eighteenth century Yersinia pestis genomes reveal the long-term persistence of an historical plague focus. *Elife*. **5** (2016), doi:10.7554/ELIFE.12994.

29. S. Sabin, A. Herbig, Å. J. Vågene, T. Ahlström, G. Bozovic, C. Arcini, D. Kühnert, K. I. Bos, A seventeenth-century Mycobacterium tuberculosis genome supports a Neolithic emergence of the Mycobacterium tuberculosis complex. *Genome Biol*. **21** (2020), doi:10.1186/S13059-020-02112-1.

30. D. Parks, M. Imelfort, C. Skennerton, P. Hugenholtz, G. Tyson, CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. **25**, 1043–1055 (2015).

31. A. Andrades Valtueña, A. Mittnik, F. M. Key, W. Haak, R. Allmäe, A. Belinskij, M. Daubaras, M. Feldman, R. Jankauskas, I. Janković, K. Massy, M. Novak, S. Pfrengle, S. Reinhold, M. Šlaus, M. A. Spyrou, A. Szécsényi-Nagy, M. Tõrv, S. Hansen, K. I. Bos, P. W. Stockhammer, A. Herbig, J. Krause, The Stone Age Plague and Its Persistence in Eurasia. *Curr Biol*. **27**, 3683-3691.e8 (2017).

32. I. Stolarek, A. Juras, L. Handschuh, M. Marcinkowska-Swojak, A. Philips, M. Zenczak, A. Dębski, H. Kóčka-Krenz, J. Piontek, P. Kozlowski, M. Figlerowicz, A mosaic genetic structure of the human population living in the South Baltic region during the Iron Age. *Sci Rep*. **8** (2018), doi:10.1038/S41598-018-20705-6.

33. C. Valdiosera, T. Günther, J. C. Vera-Rodríguez, I. Ureña, E. Iriarte, R. Rodríguez-Varela, L. G. Simões, R. M. Martínez-Sánchez, E. M. Svensson, H. Malmström, L. Rodríguez, J. M. B. de Castro, E. Carbonell, A. Alday, J. A. H. Vera, A. Götherström, J. M. Carretero, J. L. Arsuaga, C. I. Smith, M. Jakobsson, Four millennia of Iberian biomolecular prehistory illustrate the impact of prehistoric migrations at the far end of Eurasia. *Proc Natl Acad Sci U S A*. **115**, 3428–3433 (2018).

34. H. Malmström, T. Günther, E. M. Svensson, A. Juras, M. Fraser, A. R. Munters, Ł. Pospieszny, M. Tõrv, J. Lindström, A. Götherström, J. Storå, M. Jakobsson, The genomic ancestry of the Scandinavian Battle Axe Culture people and their relation to the broader Corded Ware horizon. *Proc Biol Sci*. **286** (2019), doi:10.1098/RSPB.2019.1528.

35. C. de la Fuente, M. C. Ávila-Arcos, J. Galimany, M. L. Carpenter, J. R. Homburger, A. Blanco, P. Contreras, D. C. Dávalos, O. Reyes, M. S. Roman, A. Moreno-Estrada, P. F. Campos, C. Eng, S. Huntsman, E. G. Burchard, A. S. Malaspinas, C. D. Bustamante, E. Willerslev, E. Llop, R. A. Verdugo, M. Moraga, Genomic insights into the origin and diversification of late maritime hunter-gatherers from the Chilean Patagonia. *Proc Natl Acad Sci U S A*. **115**, E4006–E4012 (2018).

36. J. Susat, J. H. Bonczarowska, E. Pētersone-Gordina, A. Immel, A. Nebel, G. Gerhards, B. Krause-Kyora, Yersinia pestis strains from Latvia show depletion of the pla virulence gene at the end of the second plague pandemic. *Sci Rep*. **10** (2020), doi:10.1038/S41598-020-71530-9.

37. S. Zhang, F. Lebreton, M. J. Mansfield, S.-I. Miyashita, J. Zhang, J. A. Schwartzman, L. Tao, G. Masuyer, M. Martínez-Carranza, P. Stenmark, M. S. Gilmore, A. C. Doxey, M. Dong, Identification of a Botulinum

Neurotoxin-like Toxin in a Commensal Strain of Enterococcus faecium. *Cell Host & Microbe*. **23**, 169-176.e6 (2018).
38. S. Zhang, G. Masuyer, J. Zhang, Y. Shen, D. Lundin, L. Henriksson, S.-I. Miyashita, M. Martínez-Carranza, M. Dong, P. Stenmark, Identification and characterization of a novel botulinum neurotoxin. *Nature Communications*. **8**, 14130 (2017).
39. C. J. Adler, W. Haak, D. Donlon, A. Cooper, Survival and recovery of DNA from ancient teeth and bones. *Journal of Archaeological Science*. **38**, 956–964 (2011).
40. P. Levy, P. Fournier, L. Lotte, M. Million, P. Brouqui, D. Raoult, Clostridium tetani osteitis without tetanus. *Emerg Infect Dis*. **20**, 1571–1573 (2014).
41. M. Darraj, J. Stone, Y. Keynan, K. Thompson, C. Snider, A case of tetanus secondary to an odontogenic infection. *CJEM*. **19**, 497–499 (2017).
42. A. Wagner, Rapid detection of positive selection in genes and genomes through variation clusters. *Genetics*. **176**, 2451–63 (2007).
43. J. Adams, M. J. Mansfield, D. J. Richard, A. C. Doxey, Lineage-specific mutational clustering in protein structures predicts evolutionary shifts in function. *Bioinformatics*. **33**, 1338–1345 (2017).
44. C. Montecucco, M. B. Rasotto, On botulinum neurotoxin variability. *mBio*. **6**, e02131-14 (2015).
45. M. J. Mansfield, A. C. Doxey, Genomic insights into the evolution and ecology of botulinum neurotoxins. *Pathog Dis*. **76** (2018), doi:10.1093/femspd/fty040.
46. 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**, 1674–1676 (2014).
47. P. Menzel, K. Ng, A. Krogh, Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun*. **7** (2016), doi:10.1038/NCOMMS11257.
48. G. Renaud, U. Stenzel, J. Kelso, leeHom: adaptor trimming and merging for Illumina sequencing reads. *Nucleic Acids Res*. **42**, e141 (2014).
49. M. Price, P. Dehal, A. Arkin, FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One*. **5** (2010), doi:10.1371/JOURNAL.PONE.0009490.
50. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. **30**, 1312–1313 (2014).
51. D. Cooke, D. Wedge, G. Lunter, A unified haplotype-based method for accurate and comprehensive variant calling. *Nat Biotechnol*. **39**, 885–892 (2021).
52. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. **30**, 772–80 (2013).
53. A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. Heer, T. de Beer, C. Rempfer, L. Bordoli, R. Lepore, T. Schwede, SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*. **46**, W296–W303 (2018).
54. P. Benkert, M. Biasini, T. Schwede, Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*. **27**, 343–350 (2011).
55. S. Guindon, O. Gascuel, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. **52**, 696–704 (2003).
56. J. Mellanby, H. Mellanby, D. Pope, W. van Heyningen, Ganglioside as a prophylactic agent in experimental tetanus in mice. *J Gen Microbiol*. **54**, 161–168 (1968).

**Acknowledgments:**

**Funding:** This study was supported by the Natural Sciences and Engineering Research Council (NSERC) through a Discovery Grant (RGPIN-2019-04266) and Discovery Accelerator Supplement (RGPAS-2019-00004) awarded to A.C.D., by the Government of Ontario through an Early Researcher Award to A.C.D, and by a University of Waterloo Interdisciplinary Trailblazer grant awarded to A.C.D. and A.D. A.C.D. also holds a University Research Chair from the University of Waterloo. M.J.M. gratefully acknowledges funding from the Japan Society for the Promotion of Science as a JSPS International Research Fellow (Luscombe Unit, Okinawa Institute of Science and Technology Graduate University). H.H. also acknowledges funding from a NSERC Canada Graduate Scholarship. This study was also partially supported by grants from National Institute of Health (NIH) (R01NS080833 and R01NS117626 to M.D). M.D. holds the Investigator in the Pathogenesis of Infectious Disease award from the Burroughs Wellcome Fund.

**Author contributions:** A.C.D. conceived and supervised the project. H.H., B.T., B.L., M.J.M, V.L., X.W., and A.C.D. performed bioinformatic data analysis. A.C.D. and G.R. supervised aDNA analysis. P.C. and P.L. performed all experimental work, which was supervised by M.D. A.D. and J.C. performed context analysis of archaeological samples. All authors contributed to manuscript writing and preparation of figures.

**Competing interests:** Authors declare that they have no competing interests.

**Data and materials availability:** Genomic data reported in this study is available from the NCBI sequence read archive. Accession numbers for all BioSamples and sequencing runs used are listed in the Supplementary Information. Source code and additional data for this project are available at https://github.com/harohodg/aDNA-tetanus-analysis

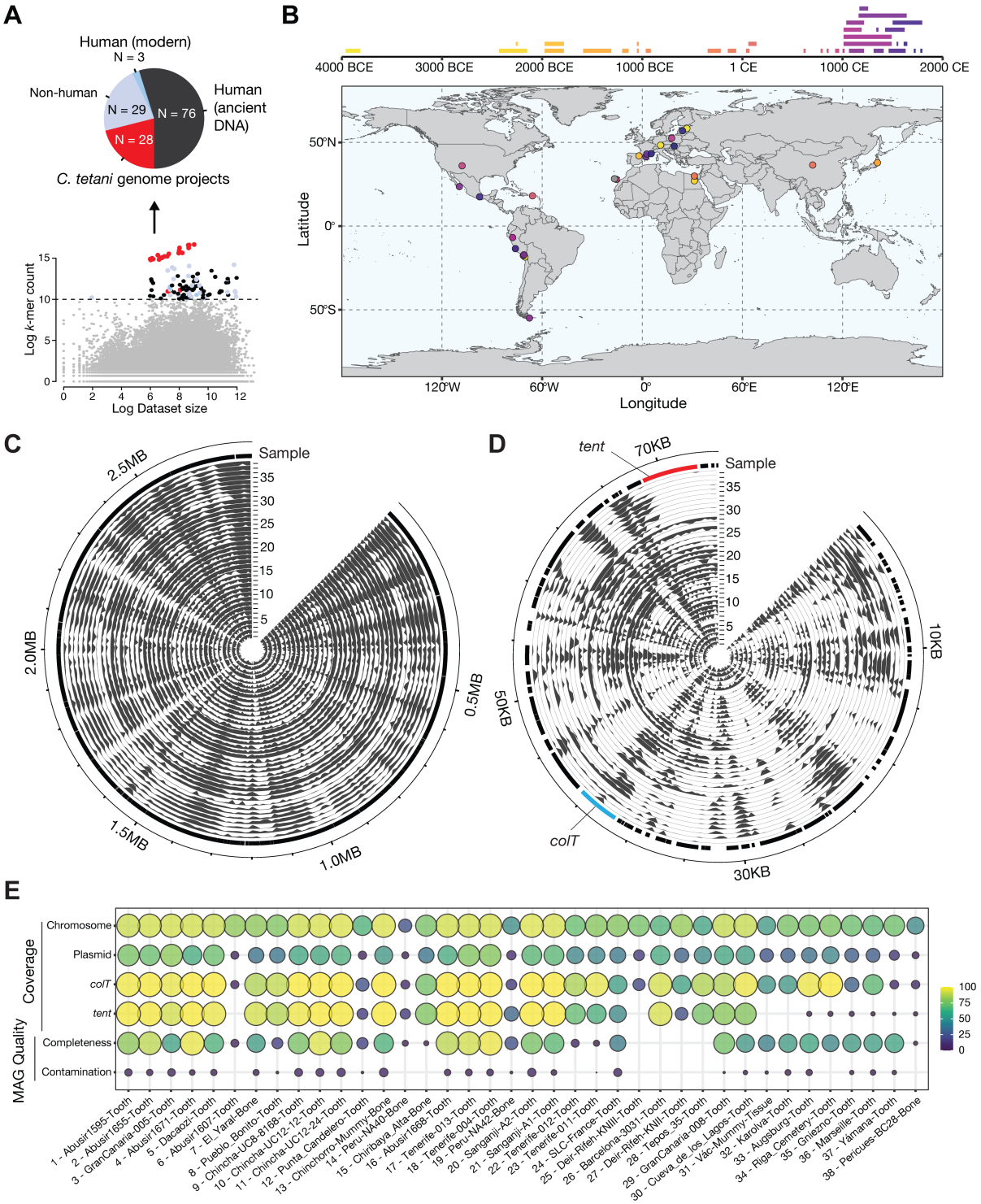


**Fig. 1. Petabase-scale screen of the NCBI sequence read archive reveals *C. tetani* related genomes in ancient human archeological samples.** (**A**) Analysis of 43,620 samples from the NCBI sequence read archive. Each sample is depicted according to its *C. tetani k*-mer abundance (y-axis) versus the overall dataset size (x-axis). A threshold was used to distinguish samples with high detected *C. tetani* DNA content, and these data points are colored by sample origin: modern

*C. tetani* genomes (red), non-human (light blue), modern human (blue), ancient human (black). The pie chart displays a breakdown of identified SRA samples with a high abundance of *C. tetani* DNA signatures. (**B**) Geographical locations and timeline of ancient DNA samples. The 76 ancient DNA datasets are associated with 38 distinct samples (BioSample IDs), which are represented as individual data points. Four samples lack date information and are absent from (B). (**C**) *C. tetani* chromosomal coverage and; (**D**) plasmid coverage detected for 38 ancient DNA associated clostridial metagenome-assembled genomes (acMAGs) using the *C. tetani* E88 genome as a reference. Samples are numbered in order of their average percent sequence identity to the reference (E88) strain, from '1' (closest to reference) to '38' (most dissimilar from reference). See table S2 for information on associated BioSample IDs and sample names. The *tent* and *colT* genes are indicated on the plasmid in red and blue, respectively. (**E**) Average per-sample coverage of *C. tetani* chromosome, plasmid, and key virulence genes, *tent* and *colT*. Also shown is the estimated completeness and contamination of *C. tetani* related MAGs assembled from aDNA samples. MAG quality estimates were performed using CheckM.

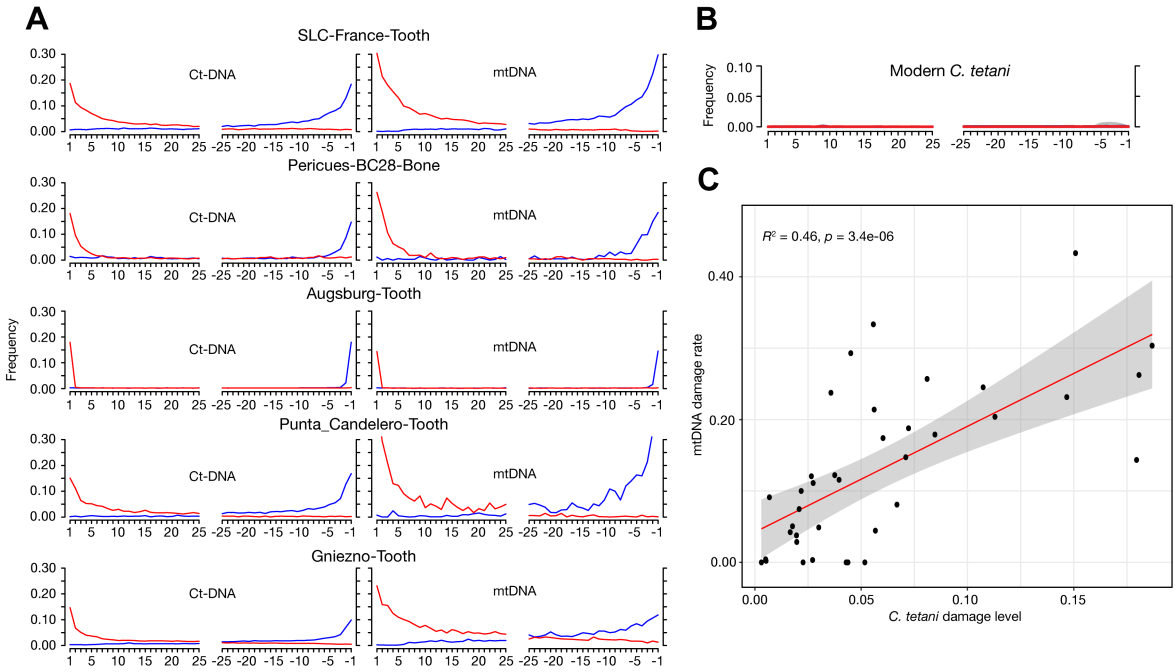


**Fig. 2. *C. tetani* DNA from a subset of ancient samples show hallmarks of ancient DNA.** (**A**) MapDamage misincorporation plots for five acMAGs displaying the highest damage levels. The plot shows the frequency of C→T (red) and G→A (blue) misincorporations at the first and last 25 bases of sequence fragments. Increased misincorporation frequency at the edges of reads is characteristic of ancient DNA, and this pattern is not observed in a representative modern *C. tetani* genomic dataset (**B**). (**C**) Correlation between damage levels of acMAGs and corresponding human mtDNA from the same sample.

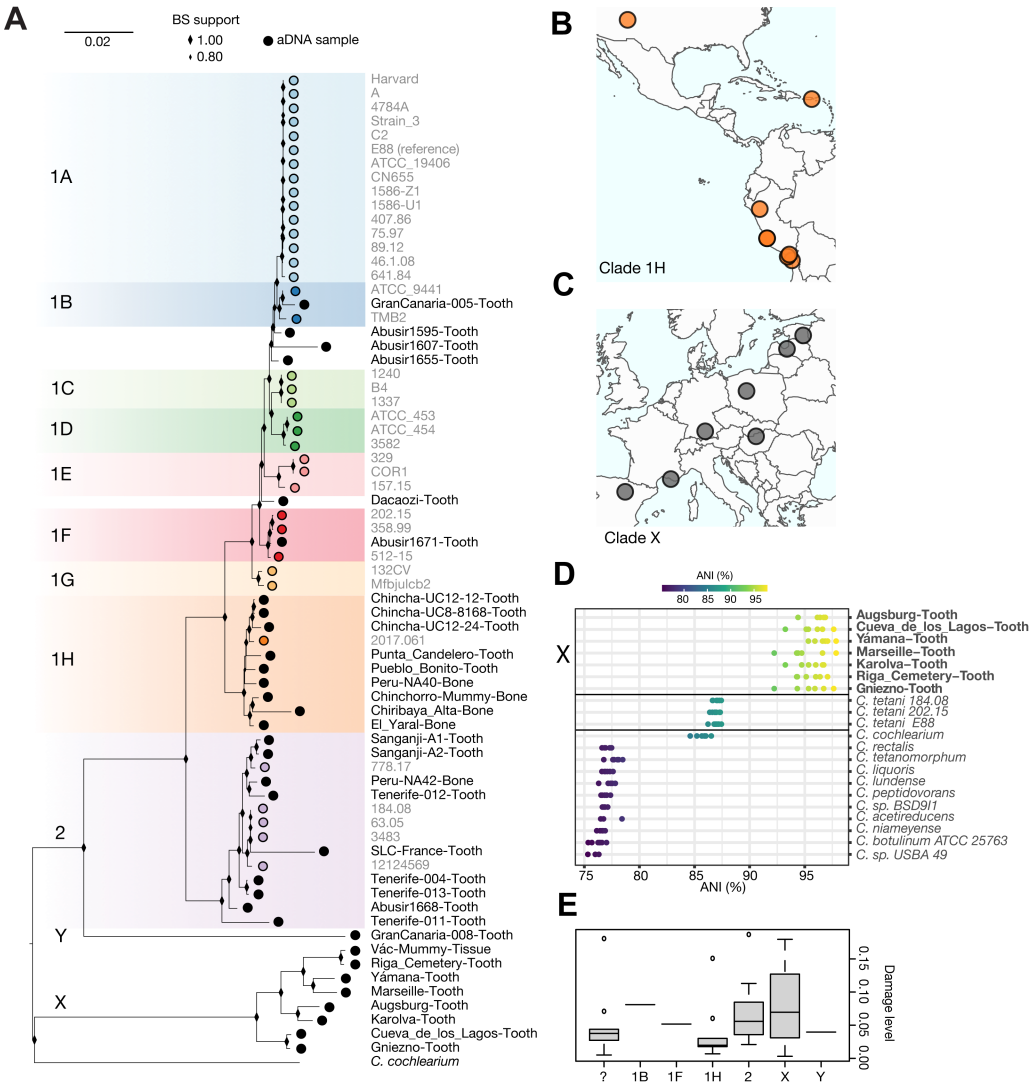


**Fig. 3. Phylogenetic analysis reveals known and novel lineages of *C. tetani* in ancient DNA samples, as well as a previously unidentified *Clostridium* species ("X").** (**A**) Whole genome phylogenetic tree of acMAGs from ancient samples and modern *C. tetani* genomes along with previously labeled phylogenetic lineages. Novel lineages are labeled "X" and "Y", which are phylogenetically distinct from existing *C. tetani* genomes. (**B**) Geographic clustering of newly identified lineage 1H acMAGs in ancient samples from the Americas, and (**C**) of newly identified clade X species in ancient samples from Europe. (**D**) Average nucleotide identity (ANI) between clade X MAGs recovered from ancient DNA and genomes of modern *Clostridium* species. Clade X MAGs show the highest ANI to *C. tetani* and *C. cochlearium* at a level that is sufficient to classify them as a novel *Clostridium* species. Note that one sample (Vac-Mummy-Tissue) was removed due to insufficient data required for fastANI. See table S10 for ANI values and genome IDs. (**E**) Distributions of damage levels for acMAGs from each phylogenetic group.

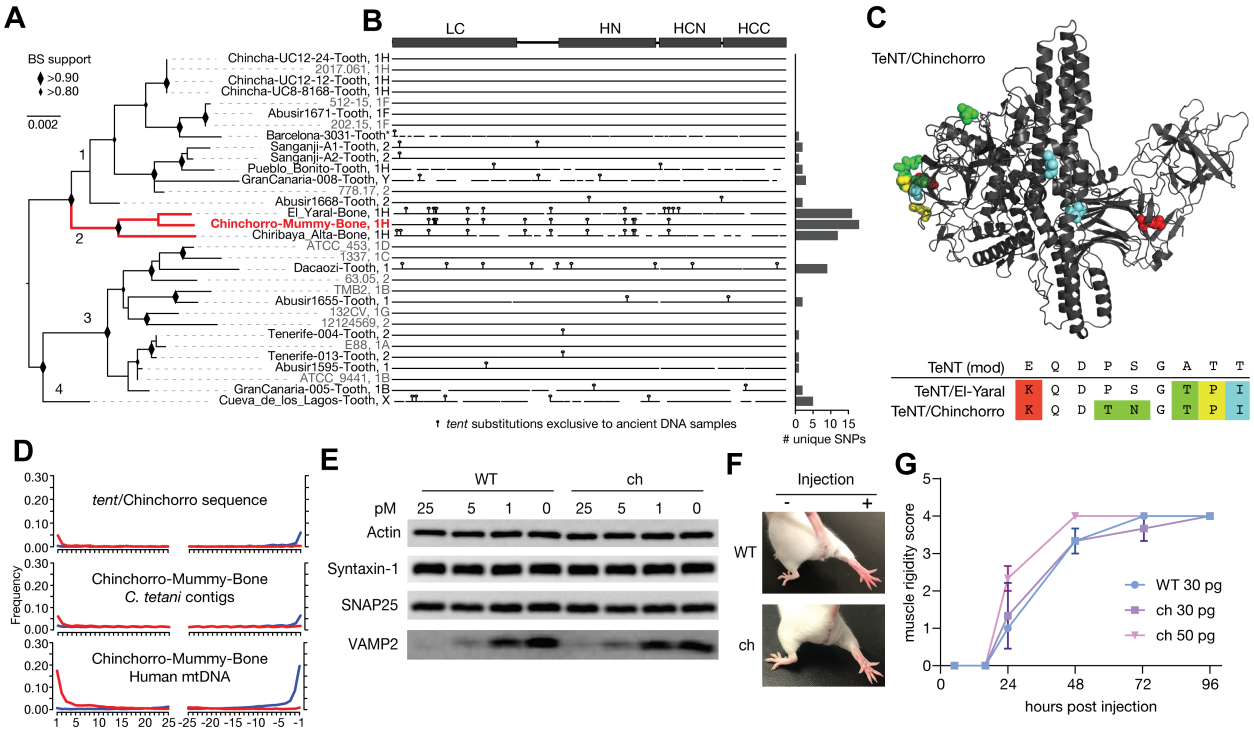


**Fig. 4. Analysis and experimental testing of a novel TeNT lineage identified from ancient DNA.** (**A**) Maximum-likelihood phylogenetic tree of *tent* genes including novel *tent* sequences assembled from ancient DNA samples and a non-redundant set of *tent* sequences from existing strains in which duplicates have been removed (see Methods for details). The phylogeny has been subdivided into four subgroups. Sequences are labeled according to sample followed by their associated clade in the genome-based tree (Fig. 3A), except for the Barcelona-3031-Tooth sequence (*) as it fell below the coverage threshold. (**B**) Visualization of *tent* sequence variation, with vertical bars representing nucleotide substitutions found uniquely in *tent* sequences from ancient DNA samples. On the right, a barplot is shown that indicates the number of unique substitutions found in each sequence, highlighting the uniqueness of subgroup 2. (**C**) Structural model of TeNT/Chinchorro indicating all of its unique amino acid substitutions, which are not observed in modern TeNT sequences. Also shown is a segment of the translated alignment for a specific N-terminal region of the TeNT protein (residues 141-149, uniprot ID P04958). This sub-alignment illustrates a segment containing a high density of unique amino acid substitutions, four of which are shared in TeNT/El-Yaral and TeNT/Chinchorro. (**D**) MapDamage analysis of the *tent*/Chinchorro gene, and associated *C. tetani* contigs and mtDNA from the Chinchorro-Mummy-Bone sample. (**E**) Cultured rat cortical neurons were exposed to full-length toxins in culture medium at indicated concentration for 12 hrs. Cell lysates were analyzed by immunoblot. WT TeNT and TeNT/Chinchorro ("ch") showed similar levels of activity in cleaving VAMP2 in neurons. (**F-G**)**,** Full-length toxins ligated by sortase reaction were injected into the gastrocnemius muscles of the right hind limb of mice. Extent of muscle rigidity was monitored

and scored for 4 days (means ± se; n=3). TeNT/Chinchorro ("ch") induced typical spastic paralysis and showed a potency similar to WT TeNT.