A 16th century *Escherichia coli* draft genome associated with an opportunistic bile infection

George S. Long, Jennifer Klunk, Ana T. Duggan, Madeline Tapson, Valentina Giuffra, Lavinia Gazzè, Antonio Fornaciari, Sebastian Duchene, Gino Fornaciari, Olivier Clermont, Erick Denamur, G. Brian Golding & Hendrik Poinar

*Escherichia coli* – one of the most characterized bacteria and a major public health concern – remains invisible across the temporal landscape. Here, we present the meticulous reconstruction of the first ancient *E. coli* genome from a 16th century gallstone from an Italian mummy with chronic cholecystitis. We isolated ancient DNA and reconstructed the ancient *E. coli* genome. It consisted of one chromosome of 4446 genes and two putative plasmids with 52 genes. The *E. coli* strain belonged to the phylogroup A and an exceptionally rare sequence type 4995. The type VI secretion system component genes appears to be horizontally acquired from *Klebsiella aerogenes*, however we could not identify any pathovar specific genes nor any acquired antibiotic resistances. A sepsis mouse assay showed that a closely related contemporary *E. coli* strain was avirulent. Our reconstruction of this ancient *E. coli* helps paint a more complete picture of the burden of opportunistic infections of the past.
The recovery of ancient pathogen DNA (aDNA) from human victims has almost exclusively focused on historically significant mortality events, such as the Black Death, revealing the evolutionary history of canonical pathogens such as *Yersinia pestis*, *Mycobacterium tuberculosis*, and *Variola virus*. In contrast, much human morbidity and mortality is the result of opportunistic infections, ones that often remain invisible in the past. Opportunistic pathogens, those without historical records—such as *Escherichia coli*, *Pseudomonas aeruginosa*, and *Staphylococcus aureus*—have been understudied relative to their contemporary burdens on people today. They are defined by their ability to infect during periods of stress, imbalance, or disturbance while being otherwise commensal or saprophytic. Opportunistic infections likely played an important role in human mortality in our shared past and thus have had a broader impact on human health than can be or has been measured.

A confounding factor in the identification of historically understudied pathogens is that the resultant infections were likely primarily opportunistic. That is, they colonized their host environments asymptomatically, leaving no identifiable pathologies and as such are not easily identified in human remains. Ancient DNA studies typically focus on obligatory pathogens such as *M. leprae*, *Salmonella enterica*, and *Y. pestis* that are easily correlated with pathologically distinct, or historically relevant mortality events and as such are easily distinguished as foreign in ancient human metagenomic DNA read data. Opportunistic infections have the added burden of authentication due to their ubiquity as environmental contaminants and of modern commensal strains. Historical identification of these pathogens would allow for a careful assessment of their evolutionary history and the commensal-pathogen continuum defined by gene content gain and loss as strains modulate toward or away from host virulence.

*E. coli* is one such pathogen. It is a common commensal bacterium found in vertebrate gut microbiome that can also become an opportunistic pathogen under specific conditions. *E. coli* has such a large impact on our health care systems that it is the subject of several vaccine development efforts to mitigate the effects of the most pathogenic strains. Having been responsible for several food poisoning outbreaks and becoming a leading pathogen for deaths caused by antimicrobial resistance, *E. coli* is thus a key bacterium used in public health surveillance.

Global sampling of *E. coli* strains produces a tree with several unique phylogenetic groups and interspersed pathovars. While the phylogenetic relationships between strains are constructed based on genetic similarity, pathovars are defined by the virulence traits of their members. In many cases, members of the same pathovar do cluster together in the same clade, however, as virulence genes can be acquired horizontally, some pathovars—like enteroaggregative *E. coli*—are distributed across multiple phylogenetic groups. This striking diversity and the evolutionary transitory states among *E. coli* strains highlight their genomic plasticity and evolutionary versatility along this aforementioned continuum. Ancient *E. coli* genomes would provide useful insights into the forces that influence the emergence of commensalism and pathogenicity in bacteria. Here, we describe the reconstruction of a 16th-century *E. coli* genome characterized from the gallstone of an Italian noble—Giovani d’Avalos (1538–1586)—highlighting a genome with commensal characteristics.

## Results

The mummiﬁed remains. In 1983, the mummiﬁed remains of several Italian nobles were recovered from the Abbey of Saint Domenico Maggiore in Naples, Italy (Supplemental Fig. 1). A careful paleopathological and histological survey of one of the individuals—Giovani d’Avalos (NASD1), a Neapolitan noble who died in 1586 at the age of 48 (further information can be found in the Supplementary Notes section of the Supplemental Materials) —revealed thickened gallbladder walls, Rokitanski-Aschoff sinuses, and several intact gallstones. These features suggested that NASD1 may have suffered from chronic cholecystitis. While not the only cause of cholecystitis, chronic bacterial infections from *E. coli*, *Bacteroidales spp.*, and *Clostridium spp.* can lead to the formation of gallstones. These infections are typically indicated by a brown pigment, as can be seen in the gallstone from NASD1 (Fig. 1a, b).

Authenticating the ancient DNA. Metagenomic proﬁles generated from the DNA extracts using Kraken provided preliminary evidence for the substantial and increasing presence of *Enterobacteriales* in digest rounds 2 (outer layer), 3–4 (middle layer), and 5–6 (inner layer) from the gallstone while indicating its absence in the tissue samples of NASD1 (Supplemental Fig. 2). *Enterobacteriales* reads are also present in digest round 1 however they are comparable with those in the reagent blanks (Supplemental Fig. 2), thus likely to be a mixture of endogenous and contaminant DNA such as *Bradyrhizobium*. Species-level identiﬁcation suggested that *E. coli* made up the largest proportion of reads identiﬁed as *Enterobacteriales*. Fortunately, *E. coli*-speciﬁc reads are virtually absent (≤0.002%) from the blanks and present in the second through sixth digests. In comparison, other taxa such as *Alcaligenes* and *Bradyrhizobium* were present in both the experimental samples and the blanks and are likely to be contaminants. Human assigned reads were present in all digests but were also detected in the blanks. *Klebsiella aerogenes* was also detected, however only in digest rounds 3–4. To assess the authenticity of the Human, *E. coli*, and *K. aerogenes* reads we examined the deamination and depurination kinetics (Fig. 1c, d and Supplemental Fig. 3).

Reconstructing the ancient *E. coli* genome and putative plasmids. The reconstructed ancient *E. coli* genome had a mean read depth of 28.46[28.31, 28.62]×, mean heterozygosity of 7.66 × 10⁻⁴[6.66 × 10⁻⁴, 8.66 × 10⁻⁴], and 4446 genes when mapped to the *E. coli* pan-genome dataset and a total estimated size of 4050, 4299pb when the detected genes are concatenated. Significant differences (two-sided t-test, *P* = 7.951 × 10⁻⁰⁵) were detected when comparing read depths of the core (3122 genes, 28.73[28.61, 28.86]×) and accessory (1324 genes, 27.82[27.39, 28.26]×) genomes (Fig. 2a and Table 1); the same was true for their mean heterozygosity with the core at 2.74 × 10⁻⁴ and accessory genome with 1.92 × 10⁻³. Eighty-three genes were identiﬁed as having a substantially (at least two standard deviations) deeper gene coverage. These results were then compared to the closely related *A. phylogroup* *E. coli* genomes of FSIS11816402, an sequent type (ST) 4995 strain, and K-12 MG1655, an ST10 strain (Fig. 2b and Supplemental Fig. 5). Both genome comparisons returned a lower mean depth with a larger variance of coverage. This is especially true for FSIS11816402 as 25.24% of its contigs had a mean coverage of less than 1× (Supplemental Fig. 6).

We identiﬁed two potential plasmids within the assembled scaffolds with M0B-suite: 17; *E. coli* MDR-56 plasmid unnamed 5 (CP019906.1) and *S. flexneri* 1a strain 0228 plasmids (CP012732.1). Subsequent mappings of the *E. coli* reads back to these reference plasmids conﬁrmed their presence in the ancient strain and authenticated their origin (Fig. 2c and Supplemental Figs. 5, 7). One plasmid, CP019906.1 had a read coverage similar to that of K-12 (20.95[20.80, 21.10]×) and 34 genes, whereas, CP012732.1 had lower coverage...
and only 19 genes. These plasmids contain regions with no read coverage—including a ~6.6 Kbp region in CP012732.1—but we could not confirm if the genes were part of the chromosome.

A subset of the sequenced reads were also classified as belonging to *K. aerogenes* by Kraken 2. To confirm or refute this result, unmapped reads were aligned against the *K. aerogenes* (NC_015663.1) reference genome. The overall read depth (0.38[0.37, 0.38] ×) and genome coverage was low (3.43%), suggesting that only a subset of genes from *K. aerogenes* was present. However, a 38 Kbp section of the genome (Fig. 2d) had a read mapping depth similar to the ancient *E. coli* results (20.36[20.23, 20.50] ×).

**Phylogenetic reconstruction.** In order to place the ancient *E. coli* within the global phylogeny and help identify the strain, we produced a core SNP alignment of 451 *E. coli* and four *Shigella* genomes representing the current breadth of *E. coli* diversity. After removing redundancy, the resulting phylogeny contained 107 genomes. This pruned sample set was then used to create a new alignment of 5007 core SNPs which returned an overall topology resembling previously published results and placed the ancient strain into a strongly supported phylogeny within phylogroup A (Fig. 3a).

To more carefully refine the position of the ancient strain within group A, we generated a reduced ML tree using 94 genomes (Fig. 3b) consisting of 291 SNPs. The ancient genome clustered closely with the ST4995 strains with some statistical support (bootstrap support of 65%), signifying that it is likely part of the same sequence type (Table 2). An even further refined ML tree was then generated using the 22 available ST4995 genomes from Enterobase18 (Fig. 3c). The core SNP alignment contained 16,026 nucleotides, providing a much greater resolution than the global and A0 subgroup alignments. The ancient strain clearly clusters within the ST4995 genomes.

We scanned for temporal signals across all phylogenies using TEMpest19 and found one when the phylogeny was restricted to the ST4995 genomes (Supplemental Fig. 8), a result confirmed using a date-randomization test in LSD (**P** = 0.003). The ST4995 phylogeny (Fig. 3c) suggests an 11th century (1027[787, 1220] CE) tMCRA for ST4995 strains, nearly 500 years before the diversification of the modern strains. The estimated evolutionary rate across the phylogeny was $2.555 \times 10^{-6}[1.567 \times 10^{-6}, 3.992 \times 10^{-6}]$ subs/site/year, similar to previously published results20.

**Global gene content.** The majority (95%) of the *E. coli* strains contained 3190 core genes, 3122 of which were detected in our database, including a ~6.6 Kbp region in CP012732.1—but we could not confirm if the genes were part of the chromosome.

Phylogenetic reconstruction. In order to place the ancient *E. coli* within the global phylogeny and help identify the strain, we produced a core SNP alignment of 451 *E. coli* and four *Shigella* genomes representing the current breadth of *E. coli* diversity. After removing redundancy, the resulting phylogeny contained 107 genomes. This pruned sample set was then used to create a new alignment of 5007 core SNPs which returned an overall topology resembling previously published results and placed the ancient strain into a strongly supported phylogeny within phylogroup A (Fig. 3a).

To more carefully refine the position of the ancient strain within group A, we generated a reduced ML tree using 94 genomes (Fig. 3b) consisting of 291 SNPs. The ancient genome clustered closely with the ST4995 strains with some statistical support (bootstrap support of 65%), signifying that it is likely part of the same sequence type (Table 2). An even further refined ML tree was then generated using the 22 available ST4995 genomes from Enterobase18 (Fig. 3c). The core SNP alignment contained 16,026 nucleotides, providing a much greater resolution than the global and A0 subgroup alignments. The ancient strain clearly clusters within the ST4995 genomes.

We scanned for temporal signals across all phylogenies using TEMpest19 and found one when the phylogeny was restricted to the ST4995 genomes (Supplemental Fig. 8), a result confirmed using a date-randomization test in LSD (**P** = 0.003). The ST4995 phylogeny (Fig. 3c) suggests an 11th century (1027[787, 1220] CE) tMCRA for ST4995 strains, nearly 500 years before the diversification of the modern strains. The estimated evolutionary rate across the phylogeny was $2.555 \times 10^{-6}[1.567 \times 10^{-6}, 3.992 \times 10^{-6}]$ subs/site/year, similar to previously published results20.

Multilocus sequence and typing of our ancient *E. coli* genome confirmed its placement in the ST4995—a rare sequence type with only 22 strains in the Enterobase18 database which encompasses 182,476 entries18. It exhibited an Onovel15:H? serotype, similar to its sister taxa (see the Supplementary Discussion section in the Supplementary Materials for full results). A different fimH variant – fimH86 – was present in our *E. coli* strain.
ancient genome (Supplemental Fig. 9). Enrichment analysis of the missing core genes revealed a significant lack of flagellar assembly genes \((P = 0.0445)\), however, the protein-protein network contained the expected number of interactions \((P = 0.11)\). An analysis of regions with mapped ancient reads in CP019906.1 revealed the presence of genes involved in biofilm formation \((\text{NP}_{311382.1})\) and a multidrug efflux pump \((\text{acrD})\). For CP012732.1, in contrast, reads mapped to pseudogenes are likely related to osmoprotection and transcription regulations.

A presence/absence (P/A) analysis of the accessory genome recapitulates the results from the ML phylogeny results (Fig. 4a) and previously published results\(^2\). The accessory PCoA indicates that the genome is a member of phylogroup A and an ancestral member of ST4995. The latter is confirmed by a P/A analysis of an ST4995-only pan-genome (Supplemental Fig. 10).

A total of 91 virulence factors were identified in the ancient genome, 37 of which were not found in \(E. coli\) K-12 MG1655 (see Table 3 for gene families with more than one hit). Type VI secretion system (T6SS) components consisted of the majority of these genes and contained high copy numbers. The T6SS—formed by the \(\text{tss}\) gene family as well as \(\text{hcp}, \text{vasK}\), and \(\text{vgrG}\)—mediates antagonistic interactions between competing bacteria. In addition to its role in bacterial killing, T6SS is involved in interbacterial signaling, biofilm formation, and phage defense\(^2\). Parts of a type III secretion system were also detected at much lower levels.

We also observed incomplete virulence gene complements involved in several mechanisms. Several genes belonging to the \(\text{ecp}\) and \(\text{fim}\) fimbriae families were found in commensal strains. Fimbriae known to be associated with virulence such as \(\text{cfaB}\) and \(\text{csg}\) were also identified\(^2\) yet they do not provide enough information to pinpoint the specific pathovar of the ancient strain\(^9\). The key virulence factors for the \(E. coli\) pathovars such as \(\text{eae}\) and \(\text{stx}\) for Shiga toxin-producing \(E. coli\) (STEC), heat-labile and heat-stable genes for Enterotoxigenic \(E. coli\) (ETEC), and the type III secretion system in EPEC and EHEC were absent\(^9\).

### Table 1 Read coverage metrics for the \(E. coli\) pan-genome, FSIS11816402, K-12 MG1655, CP019906.1, and CP012732.1 references.

| Sequence       | % Positions with coverage \(\geq 1\times\) | Mean coverage |
|----------------|------------------------------------------|---------------|
| Pan-Genome     | 97.93%                                   | 28.46[28.31, 28.62] × |
| FSIS11816402   | 77.84%                                   | 18.00[16.17, 19.83] × |
| K-12 MG1655    | 78.30%                                   | 20.94[20.93, 20.95] × |
| CP019906.1     | 80.62%                                   | 20.95[20.80, 21.10] × |
| CP012732.1     | 64.07%                                   | 16.46[16.28, 16.64] × |

FSIS11816402 used the mean of means greater than 1 × as it was an assembly. The 95% confidence interval is listed in the square brackets.
In contrast, astA, a key virulence factor in Enteroaggregative E. coli (EAEC)\(^{25}\), was detected in the ancient genome. In combination with evidence that T6SS is commonly found in EAEC strains\(^9\), this may indicate that the ancient strain was a member of this pathovar. The remaining EAEC virulence factors in the pan-genome—aggA, aggR, and aap—however, were not detected\(^{25}\). The ancient genome also contains a complete enterobactin system which is responsible for iron uptake\(^{23}\). The enterobactin system and astA gene were also found in E. coli K-12 MG1655, once again indicating that they do not confer enough virulence on their own.

To evaluate the extraintestinal virulence potential of ST4995 we tested a close relative (507 core SNPs separating the two strains) of the ancient genome, the reference strain ATCC11229 (see Fig. 3b and Supplemental Fig. 11 for similarities to the modern strain) in a well-calibrated mouse sepsis assay where ten mice were inoculated with the strain and their death recorded\(^{26}\). None of the mice were killed by ATCC11229, whereas all of the mice infected by the positive control B2 phylogroup CFT073 strain were killed. The phenotype of the strain which killed the mice is linked to the presence of specific extraintestinal virulence genes with a major role in the iron capture system encoding genes as those bared by the high-pathogenicity island (HPI)\(^{27,28}\). This result is in agreement with the absence of typical extraintestinal virulence genes present in the ATCC11229 and ancient strain.

We searched our ancient genome for the presence of antimicrobial resistance (AMR) genes using the Resistance
Gene Identifier\textsuperscript{29} and found 47 distinct genes in the scaffold assembly of our ancient \textit{E. coli} genome. Of these, 35 genes were also identified in the global pan-genome P/A analysis. Eight genes contained duplicates with \textit{mdtB} being the most commonly detected. The remaining subset spanned thirteen antimicrobial drug resistance classes with nine classes represented more than once (Table 4). Resistance cassettes targeting the five most common drug classes were detected in the ancient \textit{E. coli} strain. The majority of these genes are multidrug efflux pumps, which is typical of \textit{E. coli}\textsuperscript{30}. There were no unexpected AMR genes present in our ancient genome with all 35 being found in \textit{E. coli} K-12 MG1655.

**Discussion**

The DNA isolated from the stone showed clear evidence of DNA damage. Specifically, the deamination plots for the ancient reads mapping to the \textit{E. coli} pan-genome reference contained characteristic aDNA profiles which indicated increased deamination rates at or near the terminal bases (Fig. 1c). In addition depurination kinetics plots of the \textit{E. coli} mapped reads from libraries with sufficient \textit{E. coli} read depth (digests 2, 3–4, 5–6) showed that fragments from internal sections of the gallstone (better protected from hydrolysis) were on average longer (34 bp median in digest 2 to 37 bp median in digests 5–6) (Fig. 1d). We obtained similar results for the \textit{K. aerogenes} data across digests two through six.
Interestingly, the mapped human reads lack these signature features (Supplemental Fig. 3). More importantly, however, is that they confirm that the *E. coli* and *K. aerogenes* sequences are indeed ancient and were not the result of modern contamination.

Two potential explanations exist for the lack of deamination signal in the mapped human reads. The first is that the sample likely contains a mixture of both contaminating and to a lesser degree authentic ancient human DNA. This is almost certainly the case for the first digest, as it represents the outer layer of the gallstone, which would have been subjected to contamination from handling, storage, sampling, etc. However, as we moved to deeper layers of the gallstone, presumably better protected from contamination, the lack of endogenous human DNA was surprising. This is especially striking with the edit distances for both human and *E. coli* reads from across the different digests (Supplemental Fig. 12). The more likely explanation is that there are simply not enough human DNA reads for meaningful damage analysis. Gallstones are typically formed from a combination of cholesterol, bile salts, and phosphatidylcholine. Furthermore, in the highly acidic environment of the gallbladder, only bacteria that are specialized in surviving these conditions can thrive and potentially form a stone. Since gallstones are not formed from direct human cellular components its interior is likely virtually devoid of human DNA, which is very different from the results we obtained from the DNA of ancient abscesses.

The gene depth results of the ancient *E. coli* genome confirm the lack of modern contaminants as they are part of a unimodal normal distribution. The sole exception to these coverage stats are 83 genes with substantially larger mean depths (Supplemental Data 2). These high copy number genes, however, are with the greater heterozygosity and presence of transposons (Table 1). Most plasmids are typically present in greater copy numbers than chromosomes, however, this was not the case as the ancient chromosome had a significant ($P = 0$ using Tukey’s method to compare coverage estimates between the plasmids and ancient chromosome) deeper mean coverage than CP019906.1 and CP012732.1. One possibility is that these two plasmids are actually not present, but rather the genes to which the ancient reads mapped were actually located in the ancient chromosome. This is supported by the fact that the mean gene depths for the plasmids fall firmly within the coverage distribution of the chromosome (Supplemental Table 2). This could explain large regions of the plasmids that lack any mapped reads, as well as the lack of reads mapping to intergenic regions (see Fig. 2c and Supplementary Fig. 5B).

The reads mapping to the *K. aerogenes* genome could represent a similar situation to the plasmid data. Interestingly, the 38 Kbp region in the *K. aerogenes* chromosome which had a read depth similar yet significantly different ($P > 0$ using Tukey’s method to compare coverage estimates between the *K. aerogenes* T6SS and the ancient *E. coli* chromosome) from the plasmids and *E. coli* chromosome (20.36 ± 13.56 ×) contains the Type VI Secretion System (T6SS), which is ubiquitous across Gram-negative bacteria and plays an important role in antagonistic interactions.

Given that T6SS is widespread among bacteria, its presence is not an exclusive marker for *K. aerogenes*. However, competitive read mapping between *E. coli* and *K. aerogenes* does suggest that this T6SS does indeed belong to *K. aerogenes* rather than the homologous system in *E. coli* suggesting a recent horizontal transfer, which is supported by a comparison of their relative GC contents (Supplemental Fig. 13).

The structure of our global phylogeny resembles those previously published and places the ancient genome in phylogroup A with meaningful support (100% bootstrap for a clade of three strains). Interestingly, other historical (~1800s) *E. coli* strains have also been placed in this phylogroup. Strains that have been isolated from humans living in less industrialized and more rural communities typically harbor strains falling within phylogroup A. The late Medieval date of our sample, along with its phylogenetic placement helps confirm its authenticity. In addition, a study of *E. coli* strains isolated from bile infections showed that they belong mainly to the phylogenetic A group.

We identified a temporal signal in the reduced phylogeny (Fig. 3c) which provides additional context for the evolutionary history of ST4995 strains. The majority of the sequence type can trace their lineage to a strain from approximately 1592[1479,1690] CE, however, the divergence of our ancient genome and two others – *ESC AA9618AA AS* and *ESC VA4573AA AS* – has a deeper time to most recent common ancestor than the other two strains, it is likely that ST4995 is actually composed of at least two subgroups. Evidence supporting this can be seen in an MDS of the P/A analysis of all existing ST4995 genomes (Supplemental Fig. 9).

The P/A analysis of the accessory genes recapitulates the results from the ML phylogenies (Fig. 4a). The accessory PCoA confirms that the ancient strain is a member of phylogenetic group A while a subsequent P/A analysis with an ST4995-only pan-genome supports the proposition that the genome is a member of ST4995 (Supplemental Fig. 9). Phylogenetic group A—the group from which ST4995 stems—is a well-described clade and is the most commonly isolated type from human commensals. *E. coli* strain diversity is driven primarily by socioeconomic factors as individuals living in modern industrialized countries are more likely to carry B2 strains while those in less industrialized and rural communities harbor mostly A. Given the time period NASD1 lived, our ancient *E. coli* is a member of phylogenetic group A.

No acquired AMR genes were detected in the ancient genome confirming the age of the strain. Post removal of the multidrug efflux pumps from our genome, no other AMR genes were found when compared to *E. coli* K-12 MG1655. The latter is extremely pertinent as strong resistance to a particular drug class requires...
several mechanisms. Thus, it is likely that only mild resistances to these antimicrobial compounds were needed.

The identified ancient genome likely exhibited commensal characteristics as it shares most of its virulence factors with *E. coli* K-12 MG1655. The ancient strain lacks the canonical virulence factors for STEC and ETEC while also missing the components of the type III secretion system used in EHEC. Components of EAE—*astA* and the T6SS—were detected, yet other key virulence factors were missing and *astA* was also detected in K-12 MG1655. Results from the P/A analysis of the accessory genes with pathovar labels (Fig. 4c) and the virulence genes (Fig. 4d) present a similar picture. The ancient genome is positioned in a relatively general cluster containing a variety of different pathovars which include EHEC, STEC, ExPEC, and STEC. The mouse sepsis model further reinforces that the ancient genome was an opportunistic pathogen as the proxy strain—ATCC 11229—did not infect any mice. This information, in combination with the identified virulence genes and P/A clustering of the ancient *E. coli* suggests that the ancient genome is an opportunistic pathogen that acquired a *K. aerogenes* T6SS gene cassette during expansion in the gallbladder, consistent with previous work showing *Klebsiella* sp. often associated with *E. coli* in contemporary bile infections.

### Methods

**Sample, DNA extraction, library preparation, and sequencing.** We performed DNA extractions on a single gallstone sample, DNA extraction, library preparation, and sequencing. We performed a P/A analysis using the classic *opportunistic pathogen as the proxy strain* sepsis model further reinforces that the ancient genome was an opportunistic pathogen—measured by Root-to-Tip length—and which returned 107 genomes. A new SNP alignment and phylogeny was then created using these pruned genomes. The modern strains were labeled based on known *E. coli* phylogenetic groups using ClermonTyping.

Four *E. coli* strains which were identified as part of potential *Shigella* strains. This is due to the absence of an *ipil* gene in two of the genomes and their positioning in the phylogeny.

Upon determining the phylogroup of the ancient genome, the subgroup A0 *E. coli* genomes were used to carefully position the ancient strain within the same clade. This subgroup is defined by the Clermon genotype (+ - + -) and the enterotoxin genes.

**Bioinformatics.** Adapters were identified through AdapterRemoval with trimming and merging performed by trimmomatic. Sequencing reads were then aligned against the reference genomes using a maximum alignment score of 50. We excluded reads that did not have at least 90% identity to the reference genome.

The dated phylogeny for modern strains was generated by 

The dated ST4995 phylogeny was created by dating the bootstrap trees with LDDZ (using a 1.0 CV against the RefSeq non-redundant protein database from 2019-11-29 with a maximum E-value of 10). To determine if the data contained any temporal signature, sampling dates or sequencing dates were assigned to the pruned phylogeny and a Root-to-tip (RTT) regression was performed for the global, A0 subgroup, and ST4995 phylogenies. The presence of a temporal signal was then assessed via a date-randomization test and with TEMPEST. A molecular clock was fit to the data using east-squares dating using LDDZ.

The dated ST4995 phylogeny was created by dating the bootstrap trees with LDDZ (using a 1.0 CV against the RefSeq non-redundant protein database from 2019-11-29 with a maximum E-value of 10). To determine if the data contained any temporal signature, sampling dates or sequencing dates were assigned to the pruned phylogeny and a Root-to-tip (RTT) regression was performed for the global, A0 subgroup, and ST4995 phylogenies. The presence of a temporal signal was then assessed via a date-randomization test and with TEMPEST. A molecular clock was fit to the data using east-squares dating using LDDZ.

The dated ST4995 phylogeny was created by dating the bootstrap trees with LDDZ (using a 1.0 CV against the RefSeq non-redundant protein database from 2019-11-29 with a maximum E-value of 10). To determine if the data contained any temporal signature, sampling dates or sequencing dates were assigned to the pruned phylogeny and a Root-to-tip (RTT) regression was performed for the global, A0 subgroup, and ST4995 phylogenies. The presence of a temporal signal was then assessed via a date-randomization test and with TEMPEST. A molecular clock was fit to the data using east-squares dating using LDDZ.

Four *E. coli* strains which were identified as part of potential *Shigella* strains. This is due to the absence of an *ipil* gene in two of the genomes and their positioning in the phylogeny.

Upon determining the phylogroup of the ancient genome, the subgroup A0 *E. coli* genomes were used to carefully position the ancient strain within the same clade. This subgroup is defined by the Clermon genotype (+ - + -) and the enterotoxin genes. The strains were transformed into a binary distance matrix and clustered using a maximum alignment score of 50. We excluded reads that did not have at least 90% identity to the reference genome.

The dated ST4995 phylogeny was created by dating the bootstrap trees with LDDZ (using a 1.0 CV against the RefSeq non-redundant protein database from 2019-11-29 with a maximum E-value of 10). To determine if the data contained any temporal signature, sampling dates or sequencing dates were assigned to the pruned phylogeny and a Root-to-tip (RTT) regression was performed for the global, A0 subgroup, and ST4995 phylogenies. The presence of a temporal signal was then assessed via a date-randomization test and with TEMPEST. A molecular clock was fit to the data using east-squares dating using LDDZ.

The dated ST4995 phylogeny was created by dating the bootstrap trees with LDDZ (using a 1.0 CV against the RefSeq non-redundant protein database from 2019-11-29 with a maximum E-value of 10). To determine if the data contained any temporal signature, sampling dates or sequencing dates were assigned to the pruned phylogeny and a Root-to-tip (RTT) regression was performed for the global, A0 subgroup, and ST4995 phylogenies. The presence of a temporal signal was then assessed via a date-randomization test and with TEMPEST. A molecular clock was fit to the data using east-squares dating using LDDZ.

The dated ST4995 phylogeny was created by dating the bootstrap trees with LDDZ (using a 1.0 CV against the RefSeq non-redundant protein database from 2019-11-29 with a maximum E-value of 10). To determine if the data contained any temporal signature, sampling dates or sequencing dates were assigned to the pruned phylogeny and a Root-to-tip (RTT) regression was performed for the global, A0 subgroup, and ST4995 phylogenies. The presence of a temporal signal was then assessed via a date-randomization test and with TEMPEST. A molecular clock was fit to the data using east-squares dating using LDDZ.

The dated ST4995 phylogeny was created by dating the bootstrap trees with LDDZ (using a 1.0 CV against the RefSeq non-redundant protein database from 2019-11-29 with a maximum E-value of 10). To determine if the data contained any temporal signature, sampling dates or sequencing dates were assigned to the pruned phylogeny and a Root-to-tip (RTT) regression was performed for the global, A0 subgroup, and ST4995 phylogenies. The presence of a temporal signal was then assessed via a date-randomization test and with TEMPEST. A molecular clock was fit to the data using east-squares dating using LDDZ.

The dated ST4995 phylogeny was created by dating the bootstrap trees with LDDZ (using a 1.0 CV against the RefSeq non-redundant protein database from 2019-11-29 with a maximum E-value of 10). To determine if the data contained any temporal signature, sampling dates or sequencing dates were assigned to the pruned phylogeny and a Root-to-tip (RTT) regression was performed for the global, A0 subgroup, and ST4995 phylogenies. The presence of a temporal signal was then assessed via a date-randomization test and with TEMPEST. A molecular clock was fit to the data using east-squares dating using LDDZ.
Reference
45.  Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863–864 (2011).
46.  Schneider, V. A. et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 27, 849–864 (2017).
47.  Li, H. & Durbin, R. Fast and accurate short read alignment. Bioinformatics 25, 1754–1760 (2009).
48.  Duggan, A. T. et al. 17th century variola virus reveals the recent history of smallpox. Curr. Biol. 26, 3407–3412 (2016).
49.  Wattam, A. R. et al. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. Nucleic Acids Res. 45, D535–D542 (2016).
50.  Seemann, T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2062–2064 (2014).
51.  Page, A. J. et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31, 3691–3693 (2015).
52.  Jönsson, 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).
53.  Beghain, J., Bridier-Nahmias, A., Le Nagard, H., Denamur, E. & Clermont, O. Clermontyping: an easy-to-use and accurate in silico method for Escherichia coli strain phylotyping. Curr. Biol. 26, 11 (2016).
54.  Nurk, S. et al. Assembling genomes and mini-metagenomes from highly chimeric reads. In Annual International Conference on Research in Computational Molecular Biology (Springer, 2013).
55.  Croucher, N. J., Marsh, L. S. & Maiden, M. C. J. Genomes as maps: understanding pathogenicity in microbial evolution. Nat. Rev. Microbiol. 10, 587–597 (2012).
56.  Menardo, F. et al. Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. BMC Bioinformatics 19, 1–8 (2018).
57.  Beghain, J., Bridier-Nahmias, A., Le Nagard, H., Denamur, E. & Clermont, O. Clermontyping: an easy-to-use and accurate in silico method for Escherichia coli strain phylotyping. Microb. Genom. 4, e000192 (2018).
58.  Hazen, T. H. et al. Investigating the relatedness of enteroinvasive Escherichia coli to other E. coli and Shigella isolates by using comparative genomics. Infect. Immun. 84, 2362–2371 (2016).
59.  Clermont, O., Christensen, J. K., Denamur, E. & Gordon, D. M. The Clermont Escherichia coli phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. Environ. Microbiol. Rep. 5, 58–65 (2013).
60.  van Mentzer, A. et al. Identification of enterotoxigenic Escherichia coli (ETEC) clades with long-term global distribution. Nat. Genet. 46, 1321 (2014).
61.  To, T.-H., Jung, M., Rueda, S. M., Gascuel, O. & Fast dating using least-squares criteria and algorithms. Syst. Biol. 65, 82–97 (2016).
62.  Duchene, S. et al. Inferring demographic parameters in bacterial genomic data using bayesian and hybrid phylogenetic methods. BMC Evol. Biol. 18, 1–11 (2018).
63.  Sachard, M. A. et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 4, vey016 (2018).
64.  Riehle, M. M., Bennett, A. F. & Long, A. D. Genetic architecture of thermal adaptation in Escherichia coli. Proc. Natl Acad. Sci. USA 98, 525–530 (2001).
65.  Szkarczyk, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 47, D607–D613 (2019).
66.  de Lastours, V. et al. Mortality in Escherichia coli bloodstream infections: antibiotic resistance still does not make it. J. Antimicrob. Chemother. 75, 2334–2343 (2020).
67.  Camacho, C. et al. BLAST+: architecture and applications. BMC Bioinformatics 10, 421 (2009).
68.  Touchon, M. et al. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet. 5, e1000344 (2009).
69.  Long2. Long2/AncientEcoli: published version. Zenodo https://doi.org/10.5281/zenodo.6403348 (2022).
70.  Long2. Long2/AncientEcoli: human Ecoli version. Zenodo https://doi.org/10.5281/zenodo.6403351 (2022).

Acknowledgements
We would like to thank Wael Elhenawy and Brian Coombes for their correspondence as well as Sara Dorn for her technical assistance with the mouse model. Funding for this work was supported by an SSHRC Insight Grant, research funding from CIFAR’s human and the microbiome program and the Boris Family Fund to H.P. as well as the NSERC grant RGPIN-2020-05733 awarded to G.B.G. This work was also partially supported by the "Fondation pour la Recherche Médicale" Equipe FRM 2016, grant number DEQ20161136698 to E.D.

Author contributions
G.S.L. and J.K. were responsible for the investigation, methodology, formal analysis, and original draft of this study. G.S.L. also performed the data curation. M.T., A.T.D., V.G., L.G., and A.F. aided in the investigation. A.T.D. also contributed to the methodology. G.F. provided the sample. E.D. and O.C. validated the E. coli function and typing results. E.D. also provided funding for the mouse sepsis assay. S.D. validated the temporal signal and aided in its investigation. G.B.G. supervised and provided computational resources and funding. H.P. conceived and administered the project, provided funding, and also supervised. All authors approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-022-03527-1.

Correspondence and requests for materials should be addressed to George S. Long, Erick Denamur or Hendrik Poinar.

Peer review information Communications Biology thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Luke R. Grinham. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022