


ARTICLE

DOI: 10.1038/s41467-017-01449-9

OPEN

Distinct *Campylobacter fetus* lineages adapted as livestock pathogens and human pathobionts in the intestinal microbiota

Gregorio Iraola et al.[#] 

Campylobacter fetus is a venereal pathogen of cattle and sheep, and an opportunistic human pathogen. It is often assumed that *C. fetus* infection occurs in humans as a zoonosis through food chain transmission. Here we show that mammalian *C. fetus* consists of distinct evolutionary lineages, primarily associated with either human or bovine hosts. We use whole-genome phylogenetics on 182 strains from 17 countries to provide evidence that *C. fetus* may have originated in humans around 10,500 years ago and may have “jumped” into cattle during the livestock domestication period. We detect *C. fetus* genomes in 8% of healthy human fecal metagenomes, where the human-associated lineages are the dominant type (78%). Thus, our work suggests that *C. fetus* is an unappreciated human intestinal pathobiont likely spread by human to human transmission. This genome-based evolutionary framework will facilitate *C. fetus* epidemiology research and the development of improved molecular diagnostics and prevention schemes for this neglected pathogen.

Correspondence and requests for materials should be addressed to G.I. (email: giraola@pasteur.edu.uy) or to T.D.L. (email: tl2@sanger.ac.uk)
[#]A full list of authors and their affiliations appears at the end of the paper

The species *Campylobacter fetus* is currently divided in three subspecies based on traditional biochemical and genotyping methods: *C. fetus* subsp. *venerealis* (Cfv) and *C. fetus* subsp. *fetus* (Cff) are associated with infections in mammals^{1–3}, while *C. fetus* subsp. *testudinum* (Cft) is primarily isolated from reptiles⁴. The application of multilocus sequence typing (MLST) has shown that reptile-associated Cft are genetically distant from mammal-associated Cfv and Cff, which show high genetic relatedness^{5, 6}. More recently the comparison of whole-genome sequences confirmed that Cft and Cff/Cfv represent distinct and divergent evolutionary lineages associated with reptiles and mammals, respectively⁷. Despite the importance of Cff/Cfv in livestock and human health, its genomic evolution within mammals remains poorly understood. Two recent whole genome based studies revealed incongruences between the phylogenetic structure of the Cff/Cfv population and the biochemical features used for their discrimination, questioning the clinical relevance of subtyping mammal strains^{8, 9}. Accordingly, effective infection prevention and control schemes require a robust phylogenetic framework describing the host-associated evolution of *C. fetus* in mammals. Many important questions remain unanswered regarding the evolutionary relationship between strains isolated from bovine and human hosts, the transmission patterns of *C. fetus* between mammal hosts, and the actual potential of this species as a zoonotic pathogen.

In this work, we evidence the presence of distinct *C. fetus* lineages that have primarily adapted to humans or cattle. We propose that *C. fetus* may have originated as a human pathobiont present in the intestinal microbiota of healthy individuals and then host jumped to cattle and adapted as a venereal pathogen. Our work provides the phylogenetic and evolutionary framework to guide the development of methods for differentiation and epidemiological surveillance of the bovine and human lineages.

Results

A global *C. fetus* collection. To investigate the population structure and genomic evolution of mammal associated *C. fetus* we whole-genome sequenced 177 *C. fetus* strains isolated from 13 different countries and 5 different hosts, and combined our data with 11 published genomes sampled from 4 additional countries, giving a total data set of 188 genomes coming from strains isolated from 1952 to 2015 (63 years). To confirm species membership, the average nucleotide identity (ANI) was calculated for all possible pairs of genomes¹⁰. This analysis revealed a group of 6 genomes with an ANI < 95% compared to the other genomes (two public and four sequenced as part of our study), which were assigned to the subspecies Cft due to high genomic divergence and were therefore removed from subsequent analysis. It is noteworthy that Cff/Cfv and Cft could be classified in two distinct species (ANI < 95%). The remaining 182 Cff/Cfv genomes belong to a single species (ANI > 95%; Supplementary Fig. 1) and were isolated from 4 different mammal hosts: 91 from bovines, 77 from humans, 13 from ovines and 1 from a monkey in captivity (Supplementary Data 1 and <https://microreact.org/project/Bke4QRtHx>).

Phylogenetic structure, host-association and transmission. We built a time-scaled phylogeny and applied a Bayesian-clustering method (BAPS)¹¹ over the core genome to understand the genetic structure of the *C. fetus* population (Supplementary Fig. 2). BAPS identified eight clusters consistent with the observed phylogenetic structure (Fig. 1a). The average number of core genome SNPs for isolates belonging to the same cluster was 90 (IQR = 38–142) while it increased to 558 for isolates belonging to different clusters (IQR = 231–885). Based on the phylogeny we estimated the time

of divergence from the most recent common ancestor for the eight clusters to ~10,500 years ago (95% HPD = 8000–14,000). This estimation was not affected by confounding effects caused by previously identified sampling biases^{12–14} (Supplementary Figs. 3, 4; Methods section). The average substitution rate was 2.9×10^{-5} s/s/y (95% HPD: 2.1×10^{-5} – 5.1×10^{-6}), which is comparable to estimations made for *C. jejuni* (around 3.2×10^{-5} s/s/y)¹⁵ and falls within the interval inferred for different gram-negative and gram-positive bacterial data sets with temporal structure¹⁶. Comparison of the genetic structure defined by the phylogeny and the host type of each isolate using a Bayesian Tip-association Significance (BaTS) test¹⁷ revealed a significant association of the population structure to each host type, particularly for human and bovine hosts (Supplementary Fig. 5). The uneven distribution of host type across the different *C. fetus* clusters is shown Fig. 1b. As a general trend, cluster 1 is associated to bovine hosts (with 99% of bovine strains) and is referred to as the modern bovine lineage. The origin of this lineage was traced around 2500 years ago (95% HPD = 1500–4000), suggesting the establishment of a successful genotype adapted to bovine hosts. Clusters 2 to 8 are mainly associated to non-bovine hosts, with up to 100% of human strains, hence are hereinafter referred as the human lineages. Thus, our results are consistent with a strong host-associated evolution of *C. fetus* lineages that began around 10,500 years ago (95% HPD = 8000–14,000). Interestingly, this corresponds to the time when humans began to domesticate cattle¹⁸.

To infer the host-to-host transmission patterns of *C. fetus* we reconstructed the ancestral states for the internal nodes of the phylogeny and quantified each specific type of host transmission. An asymmetric transition model was found to be more consistent with the data than a symmetric model (Bayes factor = 31), and a median of 37 host jumps were inferred along the evolutionary history of *C. fetus*. By far the most significant directional jump was from human to bovine hosts (Bayes factor = 124), however the reverse transmission from bovines to humans was also supported (Fig. 1c). For most ancestral nodes, including the *C. fetus* MRCA, the posterior probability for a human ancestral host was consistently higher (~0.5) than for the other hosts (Supplementary Figs. 6, 7; Methods section) supporting a human origin of the currently sampled mammal *C. fetus* population. This was also suggested by the identification of significant switches in the distribution of bovine hosts across the *C. fetus* phylogeny (Supplementary Fig. 8) and by the higher nucleotide sequence diversity (Π) observed for human isolates in comparison with bovine isolates (Supplementary Fig. 9). Our results support that humans were the original mammalian host of the currently sampled *C. fetus* population and provide a potential connection between cattle domestication and the modern bovine lineage that subsequently evolved within these hosts.

Adaptive selective pressures. To investigate the adaptive evolution of the bovine and human lineages of *C. fetus* we analyzed the core genes under positive selection (Fig. 2a and Supplementary Table 2). A total of 32 genes were positively selected regardless of the host. A set of 33 genes were under positive selection exclusively in human lineages and the strongest signal was observed for the flagellar hook cap protein FlgD. The presence of diversifying alleles of flagellar genes, including *flgD*, has been found as a defining feature of hyper-invasive *C. jejuni* strains¹⁹. This may suggest parallels with the invasive and systemic bacteremia observed in many human-associated *C. fetus* infections. A distinct set of 30 genes were positively selected only in the bovine lineage and the enterobactin uptake receptor *cfrA* presented the strongest signal. The expression of the CfrA protein is induced under iron-

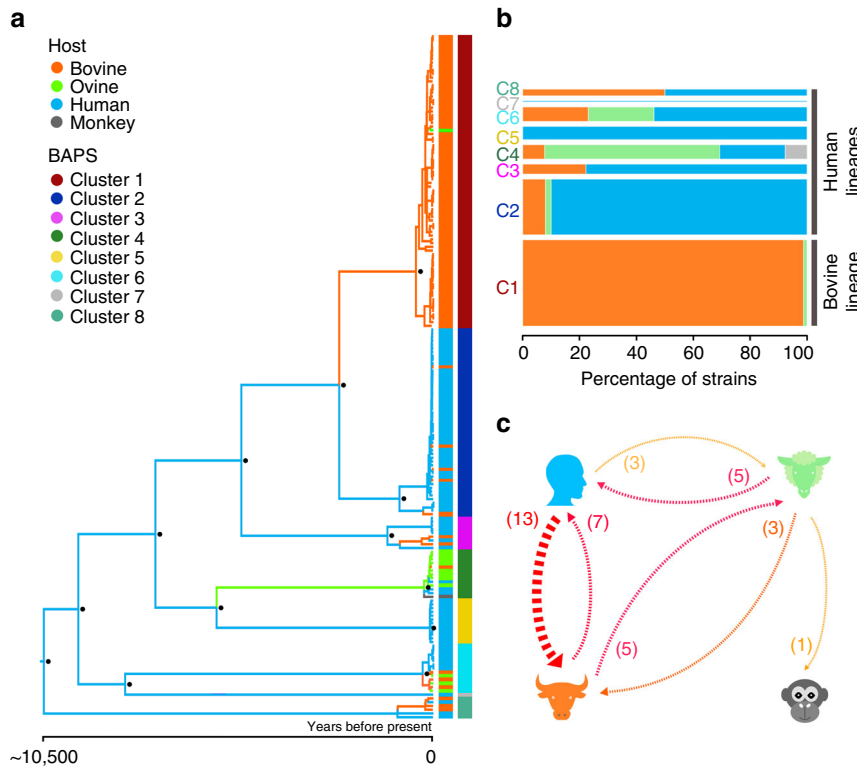


Fig. 1 Phylogenetic structure and transmission. **a** Time-scaled phylogeny of the *C. fetus* population constructed from the core genome alignment (980 genes). Branches are colored according the most probable ancestral host. Vertical color strips are colored according to host type and to BAPS clustering, respectively. **b** Bar plot showing the distribution of hosts in each BAPS cluster, bar width is proportional to the percentage of strains present at each cluster. **c** Asymmetric graph showing significant host-to-host transmissions. Arrows redness is proportional to the transmission rate and the width is proportional to the Bayes factor

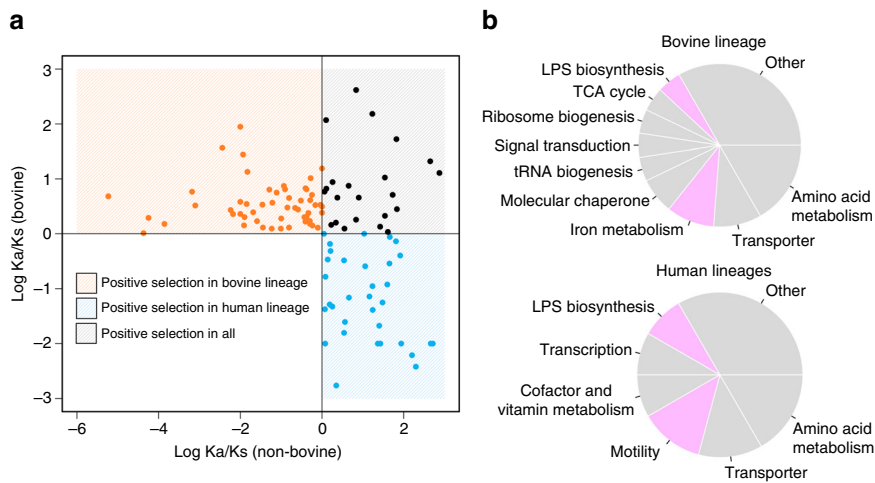


Fig. 2 Positive selection in the core genome. **a** The scatter plot shows genes with $Ka/Ks > 1$ in the bovine or human lineages. For better visualization the values are shown as the logarithm of the Ka/Ks ratio. **b** The pie charts show the distribution of biological process assignments to the genes with positive selection exclusively in bovine and human lineages. Categories involved in bacterial virulence are highlighted in violet. Categories with at least two genes were plotted

restricted conditions and plays a crucial role in iron scavenging and in vivo colonization of other species like *C. jejuni*²⁰. Thus, the core genomes of the bovine and human lineages of *C. fetus* contain signatures of diversifying evolution as they are adapted in a unique way to interact with their hosts.

A host-adapted accessory genome. To further explore the host-association of the *C. fetus* populations, we defined the accessory

genomes and observed signals of adaptive evolution. First, we evidenced the presence of recombinant blocks shared by different phylogenetic clusters that are unlikely to be explained by the clonal frame (Supplementary Fig. 10a). Second, we explored the relationship between the synonymous divergence in the core genome and the genomic fluidity of the accessory genome to demonstrate the absence of a significant correlation between these two measures (Supplementary Fig. 10b), hence supporting

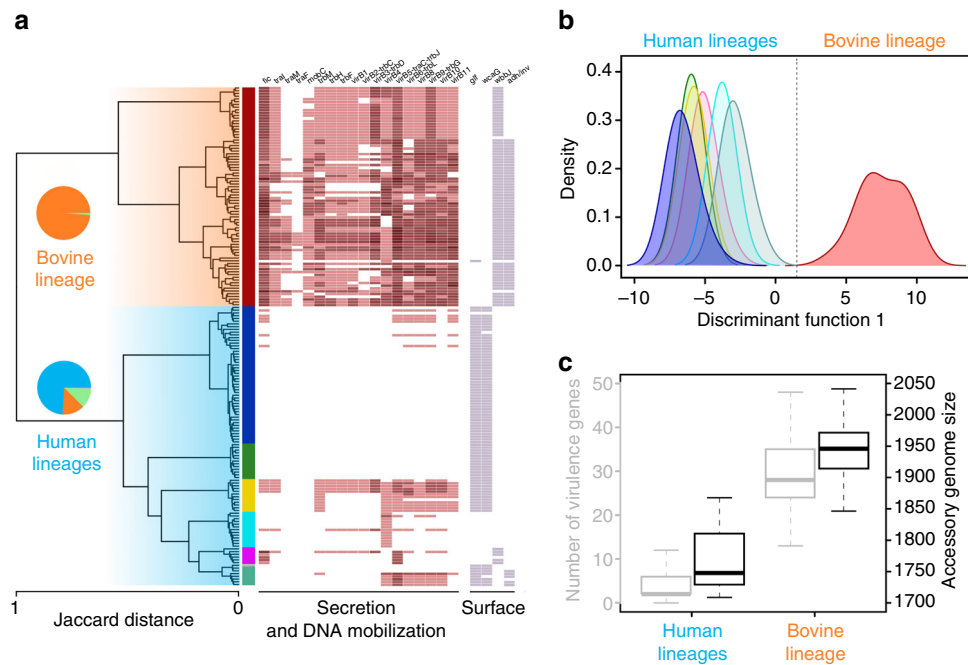


Fig. 3 Accessory genome analysis. **a** Hierarchical clustering based on pairwise Jaccard distances, the distribution of hosts in the two main groups is presented as pie charts. The vertical color strip shows the correspondence of the accessory genome groups with the BAPS clusters identified with the core genome. The presence of virulence genes is shown as heatmap blocks whose opacity is proportional to the copy number. **b** Discriminant analysis of principal components (DAPC) showing the complete separation between bovine and human lineages based on the accessory genome. **c** Boxplots showing the abundance of virulence genes (gray) and all the accessory genes (black) in bovine and human lineages

adaptive evolution. Third, we found a positive and significant correlation between the genomic fluidity and the number of positions within recombinations (Supplementary Fig. 10c). Finally, we found substantial signals of homoplasy in more than 50% of the accessory genes using a maximum-parsimony approach (Supplementary Fig. 11a, b; Methods section). These findings led us to conclude that *C. fetus* accessory gene patterns are substantially influenced by an adaptive rather than neutral evolutionary forces.

Figure 3a shows how accessory gene patterns are differentially distributed across host-adapted lineages. Also we revealed that the accessory genome of the bovine lineage is substantially larger than that of human lineages, and contains a unique set of accessory genes including those putatively involved in host–pathogen interactions (Fig. 3b, c). For example, conjugative transfer (*traItrb*) systems and type IV secretion system (T4SS) genes (*virB1–virB11*) were more frequent in the bovine lineage (Supplementary Fig. 12). T4SS genes have been found previously in *C. fetus* carried on plasmids or pathogenicity islands^{21, 22}, and here we evidence they were potentially horizontally acquired from other *Campylobacter* and *Helicobacter* species (Supplementary Table 3 and Supplementary Fig. 13). T4SS are used to transfer plasmids between bacteria or to deliver virulence effectors to host cells in a variety of pathogens such as *Helicobacter pylori*²³. We also noted differences between the lineages in the LPS biosynthesis pathway genes which has important roles in LPS structures and host-interactions. For example, the *glf* (UDP-glucopyranose mutase) and *wcaG* (GDP-fucose biosynthesis) genes were more abundant in human lineages while *wbbJ* (galactoside O-acetyltransferase) was more frequent in the bovine lineage (Supplementary Fig. 12). Thus, the accessory genomes of bovine and human lineages harbor distinct gene sets that are potentially important for genome evolution and host–pathogen interactions, and likely reflect the adaptations to different hosts.

Interestingly, other accessory genes involved in blocking the incorporation of foreign DNA, like Cas protein coding genes, were more abundant in human lineages (Supplementary Fig. 14a). The CRISPR spacer loci were also unevenly distributed in bovine and human lineages (Supplementary Fig. 14b), supporting previous evidence of their suitability for host type tracking²⁴. Restriction–modification system genes (R–M), which also play a role in maintaining genome integrity after foreign DNA incorporation²⁵, also grouped genomes according to host type (Supplementary Fig. 15a). In particular, the abundance of type I R–M genes was greater in bovine respect to human lineages (Supplementary Fig. 15b). Hence, the differences in CRISPR/Cas and R–M systems may explain the smaller accessory genomes in human lineages and the stabilization of horizontally acquired DNA in the bovine lineage, contributing to host adaptation.

A pathobiont in the intestinal microbiota. Taken together, the genomic distinctions observed in the bovine and human lineages raise the possibility of having distinct natural reservoirs. Considering that other farm animals, like poultry and pigs, are not considered reservoir hosts^{26, 27} we hypothesized that *C. fetus* could be an unrecognized member of the human gastrointestinal microbiota. To determine if healthy humans may act as natural reservoir of *C. fetus*, we compared all genomes ($n = 182$) to the Human Pan-Microbe Communities database²⁸. This allowed us to scan 7121 shotgun metagenomic data sets representing an international sampling of human feces. Interestingly, *C. fetus* was detected in ~8% of the samples from healthy humans. In contrast, the commensal *Escherichia coli* was detectable in 7% of the samples and the human pathogens *C. jejuni* and *C. coli* were detected in less than 0.5% of the samples. Of the 10 *C. fetus* genotypes detected in the gastrointestinal microbiota of healthy individuals, 22.11% belonged to strains from the modern bovine lineage and 77.89% belonged to strains from human lineages

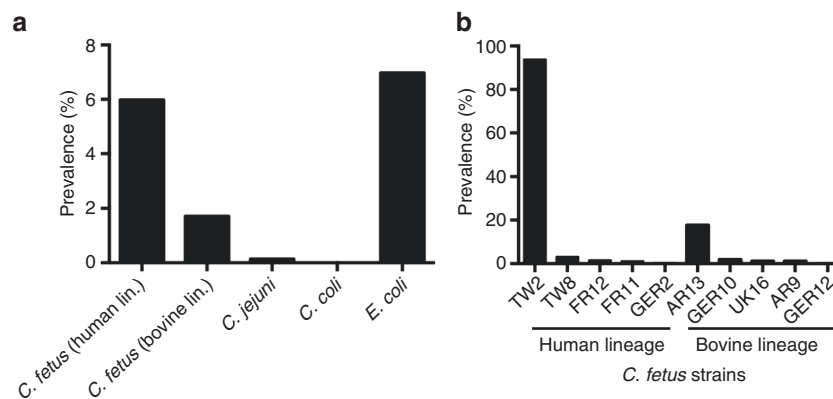


Fig. 4 Human gut microbiome analysis. **a** Prevalence of *C. fetus* strains of human and bovine origin compared to prevalence of common pathogenic *Campylobacter* species (*C. jejuni* and *C. coli*) and *E. coli* in the gut microbiota of healthy humans. **b** Prevalence of *C. fetus* genotypes found in the intestinal microbiota of healthy individuals

(Fig. 4a). No correlation was observed with gender, sex, age, ethnicity, or geographical origin of the donor or originating sample. Equally, co-abundance data showed no enrichment in other members of the microbiota. Taken together, this data suggests that microbiota community structure is not a determinant for *C. fetus* colonization success. Surprisingly, of the strains found from the human lineages, TW2 represented 94% of strains found within healthy individuals suggesting clonal expansion of this genotype in the sampled human population (Fig. 4b). The TW2 genotype belongs to cluster 5 which also includes the CA24 genotype that was responsible for a *C. fetus* outbreak among men who have sex with men, where human-to-human transmission was highly probable²⁹.

Discussion

Our results support a broad view of the natural history of *C. fetus* where the bacterium originally colonized humans, likely the gut microbiota. Then within the past 10,500 years *C. fetus* likely began to colonize and adapt to cattle, giving rise to a modern day bovine lineage that displays significant genomic distinctions from the other genomes belonging to the human lineages, linked to host-microbe interactions and genome stability. This observation is reminiscent to evolutionary patterns observed in other bacteria that moved between humans and animals^{30, 31}. Furthermore, the presence of distinctive patterns of accessory genes in the bovine and human lineages is mostly explained by adaptive evolution. This is supported both by empirical evidence of homoplasy provided here and by recent theoretical models^{32, 33} that explain the emergence of these genomic distinctions as an adaptive process linked to niche transitions, such as the host jump observed in *C. fetus*. We also provide evidence that human-adapted *C. fetus* acts as an intestinal pathobiont capable of asymptomatic carriage and likely human to human transmission. Our work supports a previously unappreciated source of infection and transmission essential to properly understand the epidemiology of symptomatic *C. fetus* infections in humans. Interestingly, the absence of the bovine associated lineage in symptomatic individuals in our analysis suggests that zoonotic transmission of the bovine *C. fetus* lineage to humans is not as common as human-to-human transmission. In contrast, the detection of the human-associated lineages in bovine samples introduces humans as a possible source of infection to cattle. Together, these observations will guide the development of phylogenetically based genetic markers capable of differentiating bovine and human lineages, and inform *C. fetus* diagnosis and epidemiology.

Methods

Sequencing, species typing, and pan-genome. Strains were retrieved from independent collections over the world (information of each sequenced strain can be found in Supplementary Data 1). Genomic DNA was prepared and sequenced using the Illumina Hi-Seq platform with library fragment sizes of 200–300 bp and read length of 100 bp at the Wellcome Trust Sanger Institute, as previously described³⁴. Each sequenced genome was de novo assembled and improved using an in-house pipeline developed at the Wellcome Trust Sanger Institute³⁵. Briefly, raw reads are first screened with Kraken³⁶ software in order to discard significant amounts of contaminating reads from other organisms. Data summary statistics are then generated using FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc). Then, reads are assembled using Velvet³⁷ to generate multiple assemblies by varying the k-mer size between 66 and 90 % of the read length using with VelvetOptimiser³⁸. From this set of assemblies the one with highest N50 is chosen. Iterative assembly improvement steps are subsequently run to scaffold the contigs using SSPACE³⁹ and to fill in sequence gaps using GapFiller⁴⁰. Finally, contigs are excluded from the assembly if they are shorter than the target fragment size (normally 300–500 bases). Remaining contigs are sorted by size and renamed in a standardized manner to include the raw sequencing data accession number. To assess the quality of the assembly and to produce summary statistics, the reads are aligned back to the final assembly using SMALT⁴¹. The assemblies are then automatically annotated using Prokka⁴² with a genus-specific database from RefSeq. The resulting annotated assemblies are in GFF3 format. A more detailed explanation of parameters used for each software at each step can be found in Page et al.³⁵ The membership to the species *C. fetus* was assessed by comparing each assembled genome against each other and the published genomes using the ANI, as previously implemented⁴³.

The annotated assemblies in GFF3 format were input to Roary, a tool that builds large-scale pan-genomes, to construct the core and accessory genomes⁴⁴. Briefly, coding sequences are extracted from annotation files and converted into protein sequences, filtered to remove partial sequences and iteratively pre-clustered with CD-HIT⁴⁵. Then all-against-all comparisons of pre-cluster representatives are performed with BLASTP using an user defined percentage sequence identity cut-off. Sequences are then clustered with MCL⁴⁶ and finally the pre-clustering results from CD-HIT are merged together with the results of MCL. Then homologous groups containing paralogs are split into groups of true orthologs using conserved gene neighborhood information. A graph of relationships between clusters is constructed based on their order of occurrence in the input sequences for providing context to each gene. Full details of the method and outputs are provided in the original paper by Page et al.⁴⁴ As the percentage sequence identity cut-off is an important parameter for defining core and accessory gene sets we ran multiple pan-genome estimations by varying this parameter from 80 to 95% by a step of 5%. Results for this analysis are shown in Supplementary Fig. 2. As no significant differences were found fundamentally in the size of the inferred core genome, all the analyses were performed with 90% of identity. Additionally, to check the core genome quality we inspected that the housekeeping genes belonging to the *C. fetus* MLST scheme clustered in the corresponding homologous groups.

Phylogenetics, population structure, and host jump analyses. The core genome alignment extracted with Roary (1,098,169 aligned positions representing 980 concatenated single copy genes) was filtered using Gubbins⁴⁷ to remove high SNP density regions which indicate putative recombination events (44,213 positions removed representing ~4% of the core genome alignment). We used the BEAST v1.7.5 package⁴⁸ to jointly estimate the substitution rate, a dated phylogeny and the reconstruction of host ancestral states using a continuous-time Markov chain (CTMC) discrete model. Briefly, we tested various combinations for the molecular clock prior (strict vs. relaxed), the demographic function prior (constant,

exponential, and logistic) and the host jump model (symmetric vs. asymmetric). We analyzed four independent runs using 100 million generations sampled every hundred steps, checking for effective sample sizes (greater than 200 for each parameter) to ensure convergence. A burn-in of 10% of states was discarded from each run. Models were compared by calculating Bayes factors (BFs) using the stepping stone algorithm⁴⁹. Since no significant differences were found for the molecular clock and demographic function priors we used a relaxed-lognormal clock with a constant population function. A significant BF was found supporting an asymmetric over a symmetric CTMC discrete model. The significance of each directional host jump was assessed by calculating BFs over the non-zero rates. We also used the approach described in TreeBreaker⁵⁰ for identifying branches in the *C. fetus* tree where the distribution of hosts has significantly changed. Sequence nucleotide diversity (Π) was calculated with the APE v4.0 package⁵¹. Phylogenetic trees were visualized using R v3.2.1⁵² and Microreact⁵³.

To evaluate the population structure observed in the core genome phylogeny, we compared it with the output of HierBAPS¹¹. This approach performs a Bayesian analysis of population structure to cluster similar samples based on their genomic relatedness. We used two clustering layers and 5, 10, 20, and 30 expected numbers of clusters (k) as input parameters. The monophyletic clades seen in the phylogeny totally agreed with the first layer of clustering. The association between the phylogenetic structure and host types was assessed with Bepi-BaTS software v0.1.1¹⁷.

Accounting for confounding effects for time-dating and sampling bias. To account for the confounding effects due to sampling biases of the *C. fetus* collection we implemented different approaches. First, we built two additional data sets subsampled from the complete data set, one just with European strains to minimize the uneven sampling at distant geographic regions, and other with the same number of strains belonging to each host type to minimize the overrepresentation of bovine and human strains at some phylogenetic clusters. Second, we built another two data sets by filtering out alignment sites likely to be subject to selection: one approach consisted in removing 106 core genes with evidence of negative selection ($dN/dS \ll 1$) and the 95 previously identified core genes with evidence of positive selection ($dN/dS > 1$); giving a filtered core genome alignment representing 779 neutrally evolving genes (839,185 aligned positions). The other data set was generated by calculating position-specific dN/dS values across the core genome alignment using the “kaksCodon” function from the CorMut package⁶⁵, this resulted in masking 9032 positions with evidence of negative ($dN/dS < 1$) or positive ($dN/dS > 1$) selection. Over all data sets we first evaluated the presence of time signal by calculating the linear regression between the root-to-tip distance and the isolation year. Null distributions of R^2 values were generated by 1000 permutations on the actual isolation year using three approaches: (i) random permutations, (ii) clustered permutations as suggested by Murray et al.⁵⁴, and (iii) clustered permutations by inputting the clusters according to BAPS results. Then, we implemented the Bayesian clustered permutation approach also described in Murray et al.⁵⁴ using the “rand.xmls.R” R script to generate the randomized inputs to BEAST analyses. To account for confounding effects due to sampling bias we applied the CTMC host state reconstruction analysis in BEAST by performing 10 independent runs with the complete data set and both subsampled data sets (European and host-balanced). Posterior estimates for the bovine and human host states at the MRCA node are displayed as boxplots in Supplementary Fig. 7.

Analysis of selective pressures. The aligned nucleotide sequences of each 980 single copy core genes were extracted from Roary’s output. The ratio between the number of non-synonymous mutations (K_a) and the number of synonymous mutations (K_s) was calculated for the whole alignment and for the respective subsets of strains belonging to the bovine and human lineages. The K_a/K_s ratio for each gene alignment was calculated with SeqinR v3.1. A $K_a/K_s > 1$ was considered as the threshold for identifying genes under positive selection. To evaluate the strength of the positive pressure acting over the same gene in bovine and non-bovine strains, we used the ranked absolute difference between the individual K_a/K_s ratios calculated over each subset of the corresponding alignments.

Accessory genome analyses. The accessory genome was defined as gene sets present in less than 100% of the analyzed genomes. The Jaccard pair-wise distance between accessory gene patterns was calculated with the APE v4.0 package⁵¹. A discriminant analysis of principal components was used to identify differences in the distribution of accessory genes across the genomes, as implemented in the AdeGenet v2.0.1 package⁵⁵. The identification of virulence genes was performed by comparing the annotated protein sequences of each genome with Blast + blastp⁵⁶ against the VFDB database⁵⁷, the Victors database (<http://www.phidias.us/victors/>) and the PATRIC database of manually curated virulence genes⁵⁸. First, genes were considered as present with query coverage or subject coverage $\geq 80\%$ and identity $\geq 80\%$. Then, a second round of blastp was performed as previously but incorporating into the databases the *C. fetus* genes recovered from the first round, in order to minimize possible false negatives due to sequence divergence. The same approach was used to confirm the presence or absence of accessory

genes present in bovine-associated or human-associated lineages. The Cas genes were identified by running hmmsearch⁵⁹ (e -value $< 1 \times 10^{-5}$) against the Cas HMM profiles retrieved from Pfam⁶⁰. The CRISPR spacers were identified with CRISPRdetect⁶¹.

Evaluation of adaptive evolution. To evaluate the presence of adaptive signals governing the evolution of accessory genes we implemented two main approaches: (i) as described by Andreani et al.⁶², the correlation between synonym diversity (a measure of effective population size) in the core genome and genomic fluidity⁶³ (a robust measure of accessory genome diversity) can be used to test for neutral evolution. We calculated genomic fluidity from accessory gene patterns as the ratio of unique gene families to the sum of gene families in pairs of genomes averaged over randomly chosen genome pairs from within a group of sampled genomes. For this we took 10,000 random samples of 50 genomes. Then we performed a linear regression analysis of genomic fluidity values against synonym diversity values calculated over the same random samples by applying natural logarithms to both measures. The same regression analysis was performed using the genomic fluidity values against the number of recombinant bases present on each sample. (ii) To test if the observed accessory gene patterns are not likely to be explained only by the clonal frame (evidence of homoplasy), for each accessory gene we calculated its consistency index (CI) using phangorn⁶⁴. CI is defined as minimum number of changes divided by the number of changes required on the tree by maximum parsimony, and is equal to one if there is no homoplasy. We then set an empirical threshold of $CI < 0.25$ to define genes with evidence of homoplasy.

Analysis of human gut metagenomes. Human gastrointestinal metagenome analysis was performed using the Human Pan-Microbe Community database (<http://www.hpmcd.org/>). Genomes were included in the normalized database generation and scanned using the standard scanning algorithm. Abundance was scaled by genome uniqueness as described previously²⁸. Abundance was scaled by read count for sample comparison.

Data availability. The assembled genomic data and raw sequences have been deposited in the European Nucleotide Archive under the accession codes provided in Supplementary Data 1. The authors declare that all other data supporting the findings of the study are available in this article and its Supplementary Information files, or from the corresponding authors upon request.

Received: 13 February 2017 Accepted: 15 September 2017

Published online: 08 November 2017

References

- Mshelia, G. D., Amin, J. D., Woldehiwet, Z., Murray, R. D. & Egwu, G. O. Epidemiology of bovine venereal campylobacteriosis: geographic distribution and recent advances in molecular diagnostic techniques. *Reprod. Domest. Anim.* **45**, e221–e230 (2010).
- Wagenaar, J. A. et al. *Campylobacter fetus* infections in humans: exposure and disease. *Clin. Infect. Dis.* **58**, 1579–1586 (2014).
- Fernández-Cruz, A. et al. *Campylobacter* bacteremia: clinical characteristics, incidence, and outcome over 23 years. *Medicine* **89**, 319–330 (2010).
- Fitzgerald, C. et al. *Campylobacter fetus* subsp. *testudinum* subsp. nov., isolated from humans and reptiles. *Int. J. Sys. Evol. Microbiol.* **64**, 2944–2948 (2014).
- Iraola, G. et al. A rural worker infected with a bovine-prevalent genotype of *Campylobacter fetus* subsp. *fetus* supports zoonotic transmission and inconsistency of MLST and whole-genome typing. *Eur. J. Clin. Microbiol. Infect. Dis.* **34**, 1593 (2015).
- Dingle, K. E. et al. Genetic relationships among reptilian and mammalian *Campylobacter fetus* strains determined by multilocus sequence typing. *J. Clin. Microbiol.* **48**, 977–980 (2010).
- Gilbert, M. J. et al. Comparative genomics of *Campylobacter fetus* from reptiles and mammals reveals divergent evolution in host-associated lineages. *Genome Biol. Evol.* **8**, 2006–2019 (2016).
- van der Graaf-van Bloois, L. et al. Inconsistency of phenotypic and genomic characteristics of *Campylobacter fetus* subspecies requires reevaluation of current diagnostics. *J. Clin. Microbiol.* **52**, 4183–4188 (2014).
- van der Graaf-van Bloois, L. et al. Whole genome sequence analysis indicates recent diversification of mammal-associated *Campylobacter fetus* and implicates a genetic factor associated with H_2S production. *BMC Genomics* **17**, 713 (2016).
- Goris, J. et al. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Sys. Evol. Microbiol.* **57**, 81–91 (2007).

11. Cheng, L., Connor, T. R., Sirén, J., Aanensen, D. M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol. Biol. Evol.* **30**, 1224–1228 (2013).
12. Duchêne, D., Duchêne, S. & Ho, S. Y. W. Tree imbalance causes a bias in phylogenetic estimation of evolutionary timescales using heterochronous sequences. *Mol. Ecol. Res.* **15**, 785–794 (2015).
13. Murray, G. G. R. et al. The effect of genetic structure on molecular dating and tests for temporal signal. *Methods Ecol. Evol.* **7**, 80–89 (2016).
14. Ho, S. Y. et al. Time-dependent rates of molecular evolution. *Mol. Ecol.* **20**, 3087–3101 (2011).
15. Dearlove, B. L. et al. Rapid host switching in generalist *Campylobacter* strains erodes the signal for tracing human infections. *ISME J.* **10**, 721–729 (2016).
16. Duchêne, S. et al. Genome-scale rates of evolutionary change in bacteria. *Microb. Genom.* **2**, 11 (2016).
17. Parker, J., Rambaut, A. & Pybus, O. G. Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infect. Genet. Evol.* **8**, 239–246 (2008).
18. Loftus, R. T. et al. Evidence for two independent domestications of cattle. *Proc. Natl Acad. Sci. USA* **99**, 2757–2761 (1994).
19. Baig, A. et al. Genetic import and phenotype specific alleles associated with hyper-invasion in *Campylobacter jejuni*. *BMC Genomics* **16**, 852 (2015).
20. Palyada, K., Threadgill, D. & Stintzi, A. Iron acquisition and regulation in *Campylobacter jejuni*. *J. Bacteriol.* **186**, 4714–4729 (2004).
21. van der Graaf-van Bloois, L. et al. *Campylobacter fetus* subspecies contain conserved type IV secretion systems on multiple genomic islands and plasmids. *PLoS ONE* **11**, e0152832 (2016).
22. Iraola, G. et al. Genomic evidence for the emergence and evolution of pathogenicity and niche preferences in the genus *Campylobacter*. *Genome Biol. Evol.* **6**, 2392–2405 (2014).
23. Brandt, S. NF- κ B activation and potentiation of proinflammatory responses by the *Helicobacter pylori* CagA protein. *Proc. Natl Acad. Sci. USA* **102**, 9300–9305 (2005).
24. Calleros, L. et al. Assessing the intra-species genetic variability in the clonal pathogen *Campylobacter fetus*: CRISPRs are highly polymorphic DNA markers. *J. Microbiol. Methods* **132**, 86–94 (2017).
25. Kommireddy, V. & Nagaraja, V. Diverse functions of restriction-modification systems in addition to cellular defense. *Mol. Biol. Rev.* **77**, 53–72 (2009).
26. Logue, C. M. et al. The incidence of *Campylobacter* spp. on processed turkey from processing plants in the midwestern United States. *J. Appl. Microbiol.* **95**, 234–241 (2003).
27. Kempf, I. et al. Broilers do not play a dominant role in the *Campylobacter fetus* contamination of humans. *J. Med. Microbiol.* **55**, 1277–1278 (2006).
28. Forster, S. et al. HPMCD: the database of human microbial communities from metagenomic datasets and microbial reference genomes. *Nucleic Acids Res.* **44**, D604–D609 (2016).
29. Marchand-Sénécal, X., Bekal, S., Pilon, P. A., Sylvestre, J.-L. & Gaudreau, C. *Campylobacter fetus* outbreak in Montreal, Quebec, Canada: a retrospective analysis. *Open Forum Infect. Dis.* **3**, S357 (2016).
30. Guinane, C. M. et al. Evolutionary genomics of *Staphylococcus aureus* reveals insights into the origin and molecular basis of ruminant host adaptation. *Genome Biol. Evol.* **2**, 454–466 (2010).
31. Carter, M. Q. Decoding the ecological function of accessory genome. *Trends Microbiol.* **25**, 6–8 (2017).
32. Niehus, R., Mitri, S., Fletcher, A. G. & Foster, K. R. Migration and horizontal gene transfer divide microbial genomes into multiple niches. *Nat. Commun.* **6**, 8924 (2015).
33. McInerney, J. O., McNally, A. & O’Connell, J. O. Why prokaryotes have pangenomes. *Nat. Microbiol.* **2**, 17040 (2016).
34. He, M. et al. Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. *Nat. Genet.* **45**, 109–113 (2013).
35. Page, A. J. et al. Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. *Microb. Genome* **2**, e000083 (2016).
36. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
37. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
38. Gladman, S. & Seemann, T. VelvetOptimiser. <https://github.com/tseemann/VelvetOptimiser> (2008).
39. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2010).
40. Nadalin, F., Vezzi, F. & Pollcriti, A. GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics* **13**, S8 (2012).
41. Ponstingl, H. & Ning, Z. SMALT. <http://www.sanger.ac.uk/science/tools/smalt-0>. (2015).
42. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
43. Piccirillo, A. et al. *Campylobacter geochelonis* sp. nov., isolated from the western Hermann’s tortoise (*Testudo hermanni hermanni*). *Int. J. Sys. Evol. Microbiol.* **66**, 3468–3476 (2016).
44. Page, A. J. et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
45. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
46. Enright, A. J., Van Dongen, S. & Ouzonis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
47. Croucher, N. J. et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).
48. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with beauti and the beast 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).
49. Xie, W., Lewis, P. O., Fan, Y., Kuo, L. & Chen, M. H. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.* **60**, 150–160 (2011).
50. Ansari, M. A. & Didelot, X. Bayesian inference of the evolution of a phenotype distribution on a phylogenetic tree. *Genetics* **204**, 89–98 (2016).
51. Popescu, A. A., Huber, K. T. & Paradis, E. ape 3.0: new tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics* **28**, 1536–1537 (2012).
52. R Development Core Team. *R: a language and environment for statistical computing*. (R Foundation for Statistical Computing, Vienna, Austria, 2008).
53. Argimón, S. et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb. Genome* **2**, 11 (2016).
54. Murray, G. G. R. et al. The effect of genetic structure on molecular dating and tests for temporal signal. *Methods Ecol. Evol.* **7**, 80–89 (2016).
55. Jombart, T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405 (2008).
56. Camacho, C. et al. Blast+: architecture and applications. *BMC Bioinformatics* **10**, 1 (2009).
57. Cheng, L. et al. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.* **33**, D325–D328 (2005).
58. Wattam, A. R. et al. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* **42**, D581–D591 (2014).
59. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
60. Bateman, A. et al. The Pfam protein families database. *Nucleic Acids Res.* **32**, D138–D141 (2004).
61. Biswas, A., Staals, R. H. J., Morales, S. E., Fineran, P. C. & Brown, C. M. CRISPRDetect: a flexible algorithm to define CRISPR arrays. *BMC Genomics* **17**, 356 (2016).
62. Andreani, N. A., Hesse, E. & Vos, M. Prokaryote genome fluidity is dependent on effective population size. *ISME J.* **11**, 1719–1721 (2017).
63. Kislyuk, A. O., Haegeman, B., Bergman, N. H. & Weitz, J. S. Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genomics* **12**, 32 (2011).
64. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2010).
65. Li, Z. CorMut: an R/Bioconductor package for computing correlated mutations based on selection pressure. *Bioinformatics* **30**, 2073–2075 (2014).

Acknowledgements

We acknowledge the Pathogen Informatics and Sequencing groups at the Wellcome Trust Sanger Institute for technical support. We also thank to Mark Stares and Hilary Browne at the Host-Microbiota Interactions Laboratory, Wellcome Trust Sanger Institute, for their technical support. G.I. is supported by the Comisión Sectorial de Investigación Científica (CSIC, Uruguay) and by the Agencia Nacional de Investigación e Innovación (ANII, Uruguay) grant FSSA_X_2014_1_105252. S.C.F. is supported by the Australian National Health and Medical Research Council (1091097) and the Victorian Government’s Operational Infrastructure Support Program. This work received financial support from Fondo de Convergencia Estructural del Mercosur (FOCEM) grant COF 03/11, the Wellcome Trust grant number 098051 and the Medical Research Council UK grant number PF451.

Author contributions

G.I. and T.D.L. conceived the idea. G.I., S.C.F., N.K. and T.D.L. designed the experiments. G.I., S.C.F. and N.K. performed the experiments and analyzed the data. P.L., S.B., F.J.G.-P., F.P., C.M., H.H., P.-R.H., A.Vi., S.L., W.Y., C.B., A.Va., A.P., B.C., J.E.H., L.B., L.Co., I.T., A.C.M., H.T.D., F.M., L.Ca., R.P. and H.N. collected and

provided samples and/or contributed to data analysis and interpretation. G.I., S.C.F., N.K. and T.D.L. wrote the paper with suggestions from all authors. All authors approved the manuscript prior to submission.

Additional information

Supplementary Information accompanies this paper at doi:10.1038/s41467-017-01449-9.

Competing interests: The authors declare no competing financial interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

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) 2017

Gregorio Iraola^{1,2,3}, Samuel C. Forster^{3,4,5}, Nitin Kumar³, Philippe Lehours^{6,7}, Sadjia Bekal^{8,9}, Francisco J. García-Peña¹⁰, Fernando Paolicchi¹¹, Claudia Morsella¹¹, Helmut Hotzel¹², Po-Ren Hsueh¹³, Ana Vidal¹⁴, Simon Lévesque⁸, Wataru Yamazaki¹⁵, Claudia Balzan¹⁶, Agueda Vargas¹⁶, Alessandra Piccirillo¹⁷, Bonnie Chaban¹⁸, Janet E. Hill¹⁹, Laura Betancor²⁰, Luis Collado²¹, Isabelle Truysers²², Anne C. Midwinter²³, Hatice T. Dagi²⁴, Francis Mégraud^{6,7}, Lucía Calleros², Ruben Pérez², Hugo Naya^{1,25} & Trevor D. Lawley³

¹Unidad de Bioinformática, Institut Pasteur Montevideo, 11400 Montevideo, Uruguay. ²Sección Genética Evolutiva, Facultad de Ciencias, Universidad de la República, 11400 Montevideo, Uruguay. ³Host-Microbiota Interactions Laboratory, Wellcome Trust Sanger Institute, CB10 1SA Hinxton, UK. ⁴Centre for Innate Immunity and Infectious Diseases, Hudson Institute of Medical Research, Clayton, VIC 3168, Australia. ⁵Department of Molecular and Translational Sciences, Monash University, Clayton, VIC 3168, Australia. ⁶Bordeaux Research in Translational Oncology, INSERM UMR1053, University of Bordeaux, 33076 Bordeaux, France. ⁷French National Reference Center for Campylobacters and Helicobacters, University of Bordeaux, 33076 Bordeaux, France. ⁸Laboratoire de Santé Publique du Québec, Institut National de Santé Publique du Québec, Sainte-Anne-de-Bellevue, QC, Canada H9X 3Y3. ⁹Département de Microbiologie, Immunologie et Infectiologie, Université de Montréal, Montréal, QC, Canada H3T 1J4. ¹⁰Departamento de Bacteriología, Laboratorio Central de Veterinaria de Algete (MAGRAMA), 28110 Algete, Spain. ¹¹Laboratorio de Bacteriología, EEA-INTA Balcarce, Balcarce 7620, Argentina. ¹²Friedrich-Loeffler-Institut, Institute of Bacterial Infections and Zoonoses, 07743 Jena, Germany. ¹³Departments of Laboratory Medicine and Internal Medicine, National Taiwan University Hospital, Taipei 10617, Taiwan. ¹⁴Animal and Plant Health Association (APHA), Addlestone KT15 3NB, UK. ¹⁵Department of Veterinary Science, Faculty of Agriculture, University of Miyazaki, Miyazaki 889-2192, Japan. ¹⁶Departamento de Medicina Veterinária Preventiva, Universidade Federal de Santa Maria, Santa Maria 97105-900, Brazil. ¹⁷Department of Comparative Biomedicine and Food Science, University of Padova, Padova 35122, Italy. ¹⁸Faculty of Science, Health, Education and Engineering, University of the Sunshine Coast, Sippy Downs, QLD 4556, Australia. ¹⁹Department of Veterinary Microbiology, University of Saskatchewan, Saskatchewan, SK, Canada S7N 5A2. ²⁰Instituto de Higiene, Facultad de Medicina, Universidad de la República, Montevideo 11600, Uruguay. ²¹Instituto de Bioquímica y Microbiología, Facultad de Ciencias, Universidad Austral de Chile, 5090000 Valdivia, Chile. ²²Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian EH25 9RG, UK. ²³EpiLab, Infectious Disease Research Centre, Massey University, Palmerston North 4442, New Zealand. ²⁴Department of Microbiology, Faculty of Medicine, Selçuk University, Selçuklu 42250, Turkey. ²⁵Departamento de Producción Animal y Pasturas, Facultad de Agronomía, Universidad de la República, 12900 Montevideo, Uruguay