Genome-resolved metagenomics of *Campylobacter* at a wildlife–livestock–human interface in Uganda reveals novel species

Supplementary materials

Table of Contents

Supplementary methods

Literature search

We searched PubMed from Jan 1, 2000, to Jan 31, 2026, without language restrictions, using combinations of terms related to *Campylobacter* diversity, wildlife–livestock–human interfaces, metagenomics, and clinical infection. We identified several thousand studies investigating *Campylobacter* infection in humans and livestock, primarily focusing on *C. jejuni* and *C. coli*. Studies examining *Campylobacter* diversity at wildlife–livestock–human interfaces were fewer and largely relied on culture-based or PCR approaches. Hundreds of publications described whole-genome sequencing or metagenomic analyses of *Campylobacter*, but few applied genome-resolved approaches across multiple host species in high-contact interfaces in low-income settings. Only 12 publications specifically investigated *Candidatus* Campylobacter infans, with inconsistent evidence regarding its association with human disease. We found no published reports describing *C.* sp900539255 in humans or assessing its clinical relevance. Heterogeneity in study design and diagnostic methods precluded quantitative meta-analysis.

DNA extraction and amplification

Approximately 150 mg of faeces was processed for nucleic acid extraction using the Zymo Quick-DNA™ Faecal/Soil Microbe Miniprep Kit (Zymo Research, CA, USA). For downstream sequencing, whole genome amplification was performed using the illustra™ GenomiPhi V3 DNA Amplification Kit (GE Healthcare) to increase DNA yield. DNA quantity was measured with the Qubit® 2.0 Fluorometer using the dsDNA HS Qubit kit (ThermoFisher, MA, USA).

Genome sequencing and library preparation

Two sequencing approaches were employed: individual and pooled samples. The sequencing data associated with this study are available under BioProject accession number PRJNA1213876 on NCBI.

Individual samples: A total of 553 DNA samples were sequenced at Orion Integrated Biosciences (KS, USA) using the Illumina® MiSeq with Nextera XT library preparation. RNAlater® procedural blanks were included as negative controls. Sequences were demultiplexed, trimmed (Q > 25, minimum length 50 bp), and quality-filtered using CLC Genomic Workbench (v7.5). Sequencing data were used for read mapping to MAGs to estimate individual-level abundances and for taxonomic classification and quantification of *Campylobacter* using Kraken2 (v2.1.3)[1] and Bracken (v2.7)[1].

Pooled samples: Samples were pooled by host type, with final pool sizes ranging from 14 to 20 samples. Sequencing was conducted at the Australian Genome Research Facility (AGRF, Melbourne, Australia) using the xGen™ cfDNA & FFPE DNA Library Preparation Kit (IDT, IA, USA) on an Illumina® NovaSeq platform (2×150 bp). Demultiplexed reads yielded 31 paired-end FASTQ files, totalling ~900 Gb and ~3 billion reads. The pooled datasets were used for MAG assembly, coverage estimation, taxonomic classification, mapping, multiple sequence alignment with GTDB-Tk:2.4.0[2], phylogenetic analysis, gene prediction, and functional annotation.

Bioinformatics pipeline and code availability

All bioinformatic analyses were performed on New Zealand eScience Infrastructure (NeSI) high-performance computing platforms. The software, versions, and key parameters used for each step are detailed below.

All custom code and scripts used for the bioinformatic analyses described in this manuscript are publicly available on GitHub and can be found in the repository titled *Campylobacter*-at-a-wildlife-livestock-human-interface-in-Uganda:

(https://github.com/ValterAlmeida/Campylobacter-at-a-wildlife-livestock-human-interface-in-Uganda).

Initial read processing

Raw sequencing data were first assessed for quality using FastQC (v0.11.9).[3] Reads were then quality-filtered and adapter-trimmed with Trimmomatic (v0.39),[4] retaining reads with a minimum Phred score of 10 for leading and trailing bases (LEADING:10 and TRAILING:10), a minimum length of 80 bp (MINLEN:80), and an initial crop of the first 10 bases (HEADCROP:10). Post-trimming quality was re-assessed with FastQC (v0.11.9).[3] To minimise contamination, host-derived reads were filtered by mapping all reads to their respective host reference genomes using BBMap (v38.95).[5] The key parameters used for filtering were: a minimum sequence identity of 95% (minid=0.95), a maximum of 3 insertions or deletions (maxindel=3), and a minimum of 2 hits to the reference genome (minhits=2). Low-quality bases at the ends of reads were also trimmed prior to mapping (qtrim=rl and trimq=10). The specific host reference genomes used were:

Gorilla: GCF_008122165.1_Kamilah_GGO_v0_genomic.fna
Human: GCF_000001405.39_GRCh38.p13_genomic.fna
Cattle: GCF_002263795.1_ARS-UCD1.2_genomic.fna
Goat: GCF_001704415.1_ARS1_genomic.fna

Metagenomic assembly and binning

The remaining reads were assembled into contigs using MEGAHIT (v1.2.9).[6] Contigs were then filtered to a minimum length of 1,000 bp using seqmagick (v0.8.4) (GitHub) with the (--min-length 1000) parameter. For binning, a mapping index was created from the assemblies with Bowtie2 (v2.4.5)[7] and used for read mapping with the same software. Reads were mapped back to their respective assemblies, and the resulting files were sorted and compressed with SAMtools (v1.15.1).[8] Contigs were then binned using three independent methods: MetaBAT2 (v2.17),[9] MaxBin2 (v2.2.7),[10] and CONCOCT (v1.1.0).[11] The best bins from each method were consolidated into a single, high-quality set of non-redundant bins using DAS Tool (v1.1.5).[12]

Genome quality assessment and dereplication

The quality and general characteristics of the consolidated bins were assessed using CheckM2 (v1.0.1).[13] *Campylobacter* genomes with >50% completeness and <6.5% contamination were retained for further analysis. Genomes were then dereplicated using dRep (v3.4.2)[14] with a 99% average nucleotide identity (ANI) threshold, ensuring a final set of unique MAGs.

Taxonomic classification

Taxonomic assignment of the MAGs was performed using GTDB-Tk (v2.4.0)[2] with the GTDB (Release 10-RS226). Genomes with ≥95% ANI were classified as the same species. For an independent classification, raw reads were classified using Kraken2 (v2.1.3)[1] with the standard database (12/2024), and abundances were refined with Bracken (v2.7).[15] For comparative purposes, a representative genome for each *Campylobacter* species was selected from GTDB and downloaded from NCBI RefSeq or GenBank. The *Helicobacter pylori* reference genome (strain CHC155; NCBI accession GCA_025998455.1) was included as an outgroup for phylogenetic analysis.

Phylogenetic and comparative genomics

Publicly available genomes were downloaded from NCBI using a custom script, which first identified unique *Campylobacter* species-level accessions from the GTDB (Release 10-RS226). Phylogenetic tree inference was performed using GTDB-Tk (v2.4.0),[2] which identifies a standard set of 122 (bacteria) single-copy marker genes, aligns them to reference HMM profiles, and generates a concatenated multiple sequence alignment for phylogenetic placement within the Genome Taxonomy Database (Release 10-RS226) and IQ-TREE2 (v2.2.2.2).[16] The best-fit substitution model was automatically selected by IQ-TREE2 using ModelFinder included in the IQ-TREE2 pipeline, and branch support was assessed with 1,000 ultrafast bootstrap replicates. The resulting tree was visualised and annotated using the Interactive Tree of Life (iTOL v6).[17] The *cpn60* gene from public database reference genomes[18] was aligned with MAFFT (v7.450),[19] and primers[18] were mapped in silico using Geneious (v10.6.2)[20] to test for cross-reactivity. ANI comparisons between all MAGs and public genomes were performed using FastANI (v1.33).[21]

Comparative genome alignments were performed using MAGs generated in this study against the *C. jejuni* reference genome (GCA_000009085.1; ASM908v1) and the *C. coli* reference genome (GCF_009730395.1;

ASM973039v1) using MUMmer (v4.0).[22] The same reference genomes were used for read and contig mapping with Bowtie2 (v2.4.5)[7] and SAMtools (v1.15.1).[8] Host-filtered reads were mapped against *C. jejuni* and *C. coli* reference genomes, as well as against the highest-quality human-associated *Campylobacter* MAGs, selected based on genome completeness and contamination. In a second mapping step, reads that did not map to *C. infans* or *C.* sp900539255 were mapped to the *C. jejuni* and *C. coli* reference genomes to assess whether reads initially assigned to these species could instead be explained by locally assembled *C. infans* and *C.* sp900539255 genomes. Reads mapping to *C. jejuni* were subsequently assembled using MEGAHIT (v1.2.9)[6] and queried against the BLAST database (BLASTDB/2025-08) using BLAST (v2.16.0).[23]

Host-filtered paired-end reads were mapped using Bowtie2 (v2.4.5) in combination with SAMtools (v1.15.1). To evaluate whether reads initially attributed to *C. jejuni* by *k*-mer-based classification could instead be explained by locally assembled *C. infans* and *C.* sp900539255 MAGs, two complementary mapping strategies were applied. First, a sequential two-step approach was used as described above, in which reads were mapped to the MAG index before unmapped reads were remapped to *C. jejuni*. Second, to confirm results under competitive conditions, all host-filtered reads were simultaneously mapped to a combined reference index containing *C. infans*, *C.* sp900539255 (concatenated unique MAG sequences), and the *C. jejuni* reference genome (GCA_000009085.1). Prior to index construction, *C. jejuni* contig headers were prefixed with a unique identifier (JEJUNI_) to enable unambiguous post-mapping classification of reads by reference origin. Bowtie2 was run in --sensitive mode with paired-end insert size constraints (--minins 200, --maxins 800) for both strategies. Only mapped reads were retained (-F 4) and converted to BAM format using SAMtools. Mapped read counts were calculated using samtools view -c. Bowtie2 alignment scores (AS tag) were extracted from BAM files using SAMtools and used to compare alignment quality between references across all samples. Mean AS values per sample were compared between *C. infans*/sp900539255 and *C. jejuni* using a paired Wilcoxon signed-rank test.

Functional annotation and abundance

Gene prediction was performed using Prodigal (v2.6.3),[24] and genomes were annotated with Prokka (v1.14.5).[25] Virulence and antibiotic resistance genes were identified using Abricate (v1.0.0) (GitHub) against the ResFinder[26,27] and VFDB[28] databases. Metabolic functions were analysed with DRAM (v1.3.5)[29], KofamScan[30,31] and KEGG-Decoder[32]. Genome coverage abundance was determined by mapping reads to MAGs using Bowtie2 (v2.4.5),[7] with coverage statistics calculated by MetaBAT2 (v2.15)[9] and reads counted with SAMtools (v1.15.1).[8] The abundance of specific *Campylobacter* species was normalised using a custom Python script that corrected for sequencing depth variation.

Statistical analysis

Statistical analyses were conducted in R:4.5.1[33], including ANOVA and Tukey's HSD.

Supplementary figures and tables

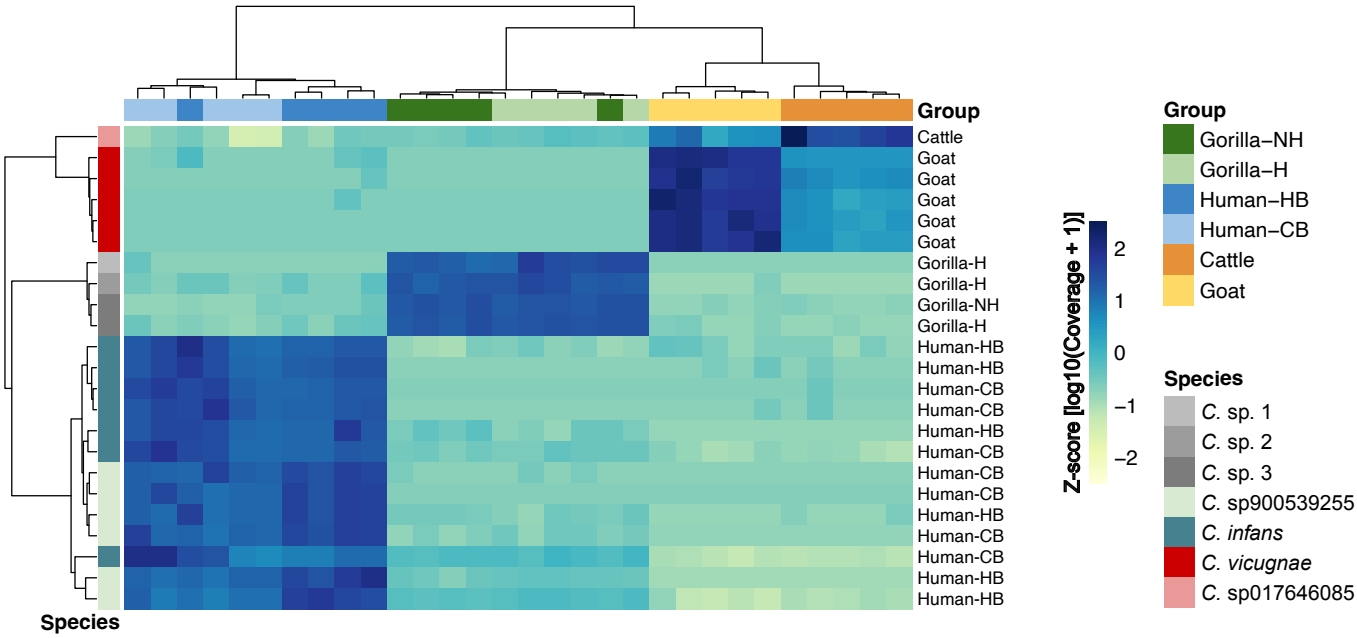


Figure S1. Sequence coverage heatmap of *Campylobacter* MAGs across host-associated samples.

A) Relative coverage of dereplicated *Campylobacter* MAGs across gut microbiome samples from humans, livestock, and gorillas. Average sequence coverage is shown for samples pooled by host group (top – samples per group). Sequence coverage values were normalised per MAG to account for sequencing depth, log-transformed ($\log_{10}[x + 1]$), and z-score scaled to highlight relative abundance differences. MAG identifiers indicate the pooled host group in which each MAG was assembled (right): Gorilla non-habituated (Gorilla-NH) and habituated (Gorilla-H), human healthcare-based (Human-HB) or community-based (Human-CB), cattle, and goat. The *Campylobacter* species represented by each MAG are shown on the left.

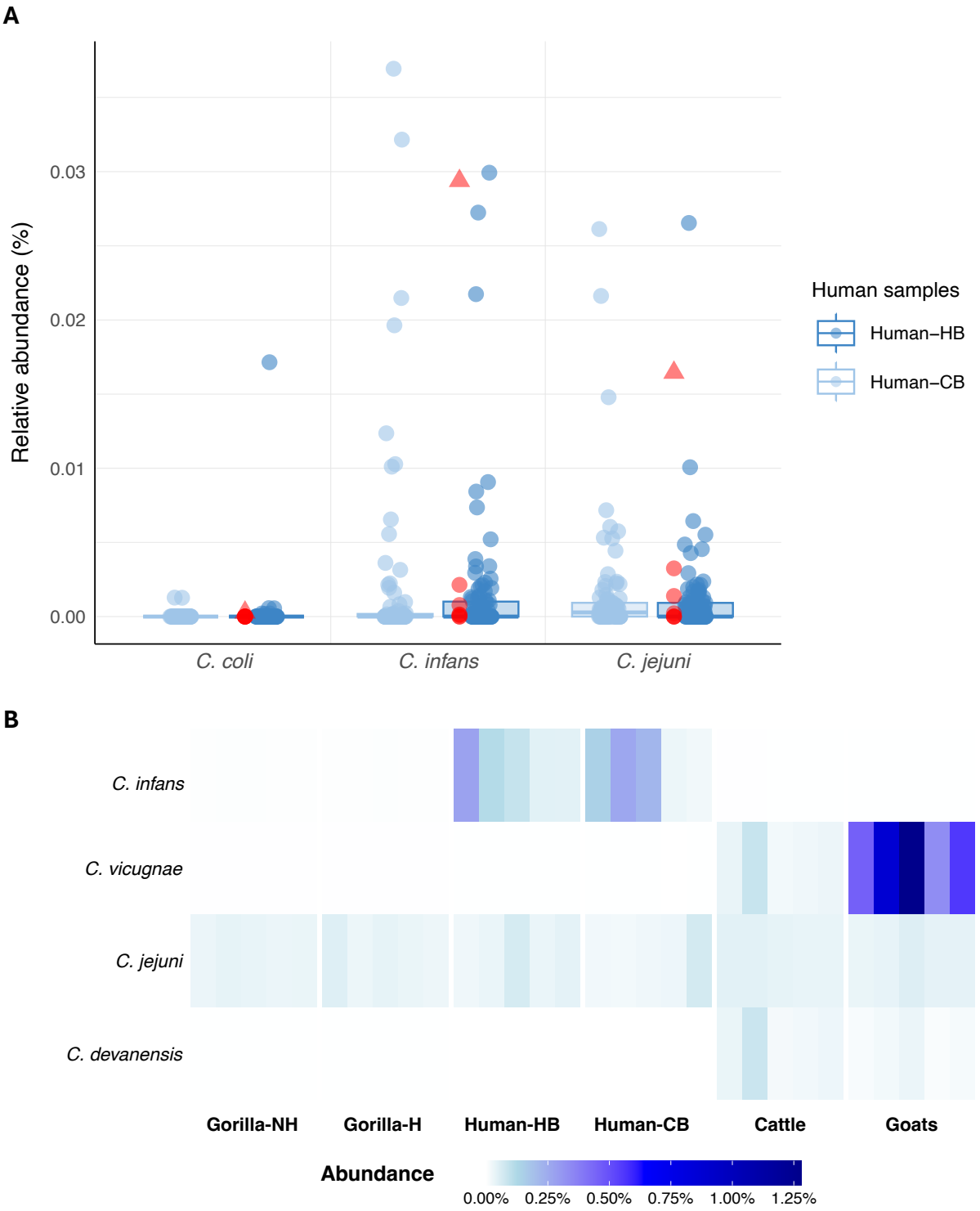


Figure S2. *K*-mer-based metagenomic classification of *Campylobacter* species in human, gorilla and livestock gut microbiomes.

A) *K*-mer–based metagenomic analysis revealed a *Campylobacter* species distribution in human samples (individual samples dataset) that differed from the MAG-based classification (figure 1), with *C. jejuni* identified as one of the predominant species. B) Species for which MAGs were recovered in this study are highlighted. The panel also shows detectable reads assigned to *C. jejuni* across all pooled samples and to *C. devanensis* among predominant livestock-associated taxa, despite no corresponding MAGs being assembled. The *k*-mer approach classified *C. infans* in agreement with the MAG-based results and identified *C. jejuni* as an additional dominant species, although no *C. jejuni* MAGs were assembled by the MAG-derived approach. In contrast, *C.* sp900539255 was not detected by *k*-mer-based classification, despite the samples being collected six years prior to the most recent database update used in this study.

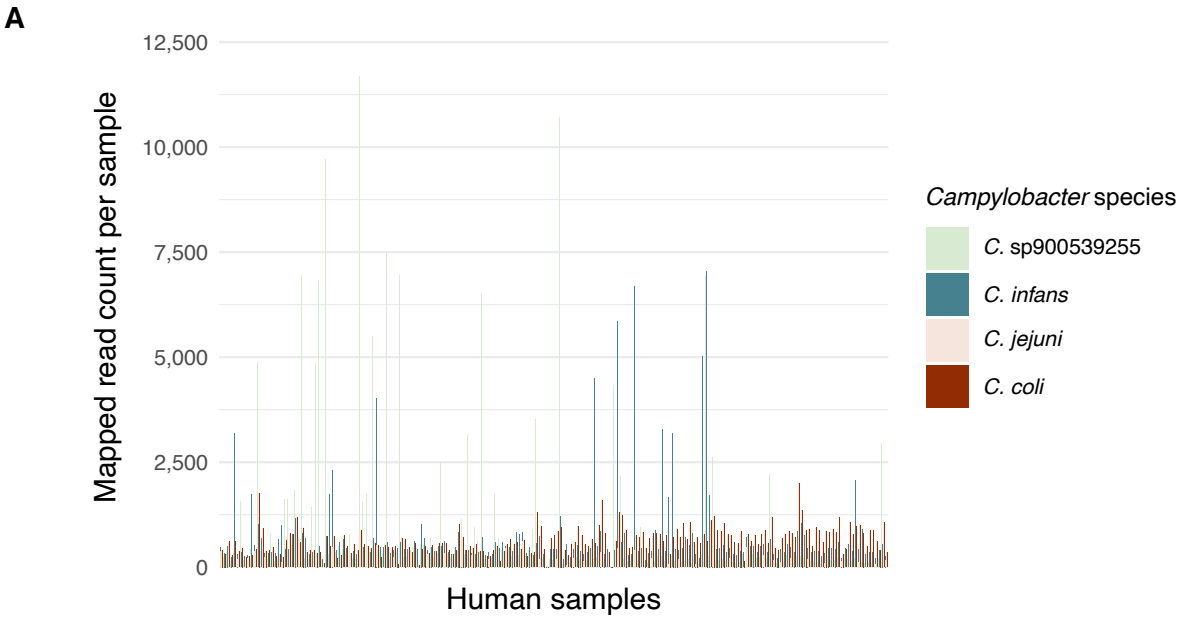
Supplementary Table 1: Prevalence of human-associated *Campylobacter* species in healthcare-based and community-based samples (*k*-mer–based analysis).

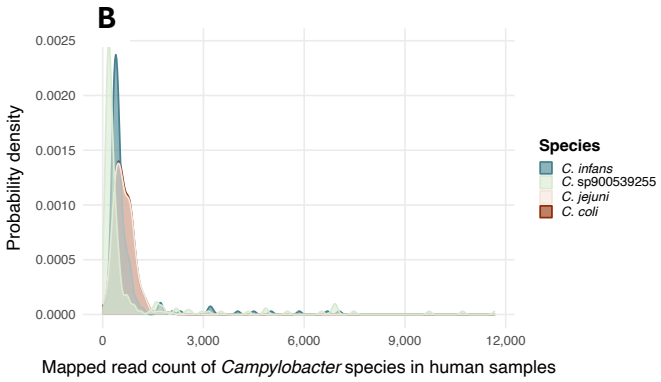
	Healthcare-based		Community-based			
Species	Positive	Prevalence	Positive	Prevalence	p-value	OR (95% CI)
C. coli	06/99	6·1%	03/97	3·1%	0·50	2·01 (0·42-12·81)
C. jejuni	47/99	47·5%	57/97	58·8%	0·12	0·64 (0·35-1·16)
C. infans	41/99	41·4%	32/97	33·0%	0·24	1·43 (0·77-2·68)

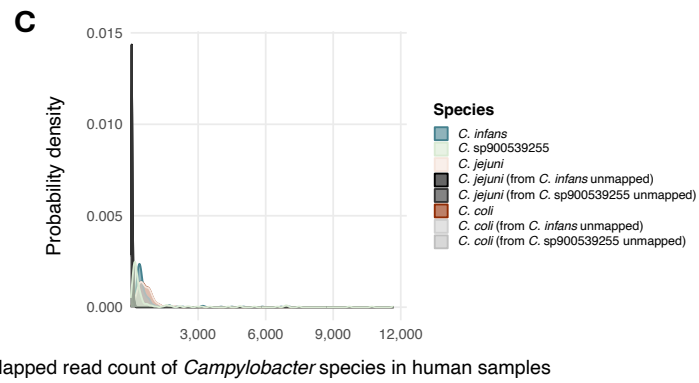
Statistical tests: Two-sided Fisher's exact tests compared prevalence between groups, with odds ratios (OR) and 95% confidence intervals (CI) as effect measures.

Supplementary Table 2: Genomic alignment of study-derived MAGs to a *Campylobacter jejuni* reference genome.

Host	MAG species	Aligned to *C. jejuni* (bp)	Average identity (%)	Genes with sequence alignments
Human	*C. infans*	7,227	85·5	*tuf, dsbI, dba, nifJ, tRNA (Leu, Arg, His, Pro)*
	C. sp. 900539255	2,310	83·3	*tuf, tRNA (Thr, Tyr, Gly, Leu, Met)*
Gorilla	*Campylobacter* sp. 1	823	95·2	*rRNA-5S, tRNA (Ser, Arg, Thr)*
	Campylobacter sp. 2	1,374	85·5	*tuf, tRNA (Thr, Leu, Lys, Asp)*
	Campylobacter sp. 3	11,782	80·8	*ribE, kdsA, napA, gatA, accD, dnaK, tuf, rpmG, secE, nusG, rplK, rplA, rplP, rpmC, rpsC, rplV, nifJ, rpsU, tRNA (Thr, Trp, Leu, Cys, Ser, His, Pro)*
Cattle	*C.* sp. 017646085	12,607	89·4	*tuf, secE, rpmG, rpsI, rplM, gyrB, rRNA-23S tRNA (Trp, Thr, Gly, Pro, His, Arg, Lys, Val, Ser)*
Goat	*C. vicugnae*	15,563	82·4	*aspA, fusA, ssrA, rplO, secY, rpsI, rplM, tRNA (Arg, Gly, Leu, Cys, Ser, Phe, Val, Glu, Lys, Asp, His, Pro, Asn)*


A
12,500
10,000
7,500
5,000
2,500
0
Mapped read count per sample
Human samples
Campylobacter species
C. sp900539255
C. infans
C. jejuni
C. coli


B
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
Probability density
0
3,000
6,000
9,000
12,000
Species
C. infans
C. sp900539255
C. jejuni
C. coli
Mapped read count of Campylobacter species in human samples


C
0.015
0.010
0.005
0.000
Probability density
3,000
6,000
9,000
12,000
Species
C. infans
C. sp900539255
C. jejuni
C. jejuni (from C. infans unmapped)
C. jejuni (from C. sp900539255 unmapped)
C. coli
C. coli (from C. infans unmapped)
C. coli (from C. sp900539255 unmapped)
Mapped read count of Campylobacter species in human samples

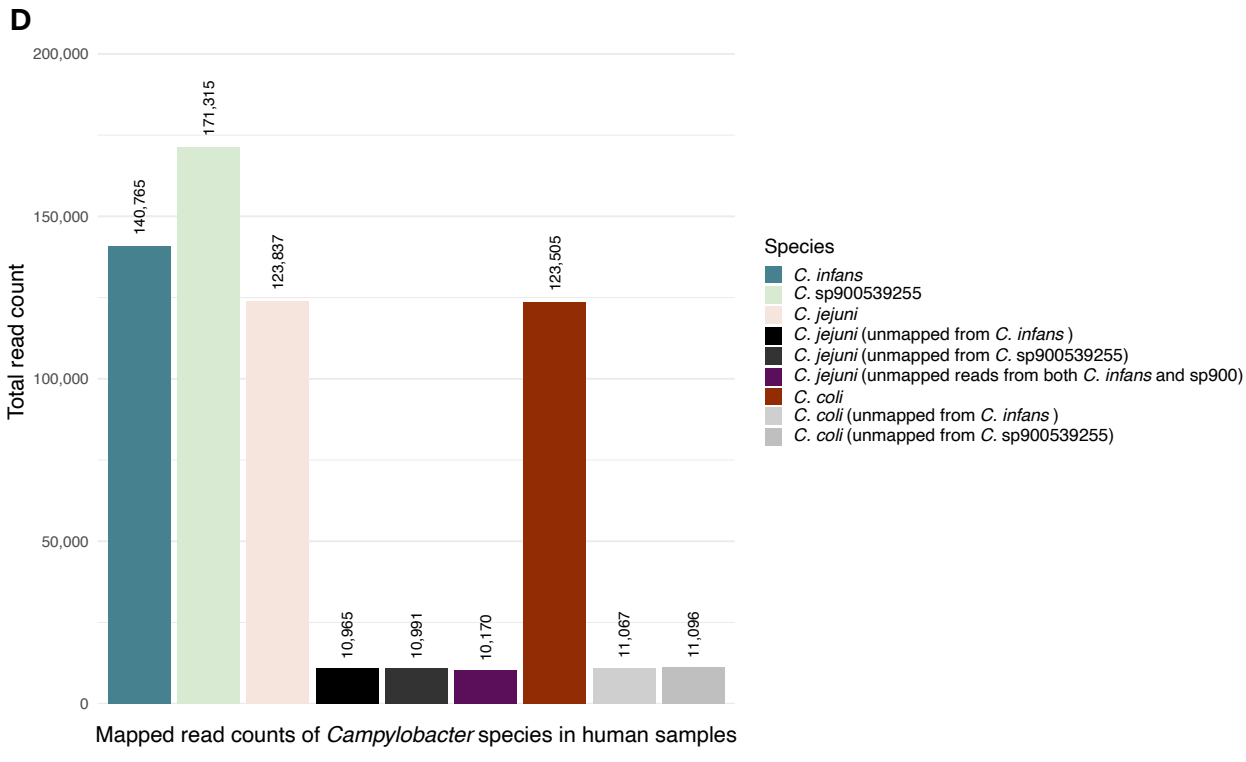

D
200,000
150,000
100,000
50,000
0
Total read count
140,765
171,315
123,837
10,965
10,991
10,170
123,505
11,067
11,096
Species
C. infans
C. sp900539255
C. jejuni
C. jejuni (unmapped from C. infans)
C. jejuni (unmapped from C. sp900539255)
C. jejuni (unmapped reads from both C. infans and sp900)
C. coli
C. coli (unmapped from C. infans)
C. coli (unmapped from C. sp900539255)
Mapped read counts of Campylobacter species in human samples

Figure S3. Assessment of *k*-mer-based *C. jejuni* taxonomic classification using read-mapping approaches.

Evaluation of *C. jejuni* taxonomic assignments across read-mapping approaches. A *C. coli* reference genome was included for comparison as a closely related species and a common human pathogen. A) Read mapping from individual human samples showing higher read counts for *C.* sp900539255 and *C. infans* compared with *C. jejuni* and *C. coli*, which were consistently detected at lower levels across samples. B) Density distribution of mapped reads, showing that read densities were largely concentrated below 1500 reads, with higher-density distributions observed primarily for *C.* sp900539255 and *C. infans*. C) Read density distribution following sequential mapping, in which reads unmapped from *C.* sp900539255 and *C. infans* were subsequently mapped to *C. jejuni* and *C. coli*, resulting in near-zero read densities for these species. D) Total mapped reads before and after sequential mapping, demonstrating a marked reduction in reads assigned to *C. jejuni* and *C. coli* when accounting for MAG-based species identified in this study. Collectively, these results indicate that conserved genomic regions among *Campylobacter* species can lead to ambiguous read assignments, inflating detection of well-characterised species by *k*-mer-based classification approaches, particularly in underexplored settings where novel taxa, such as those identified in this study, are present.

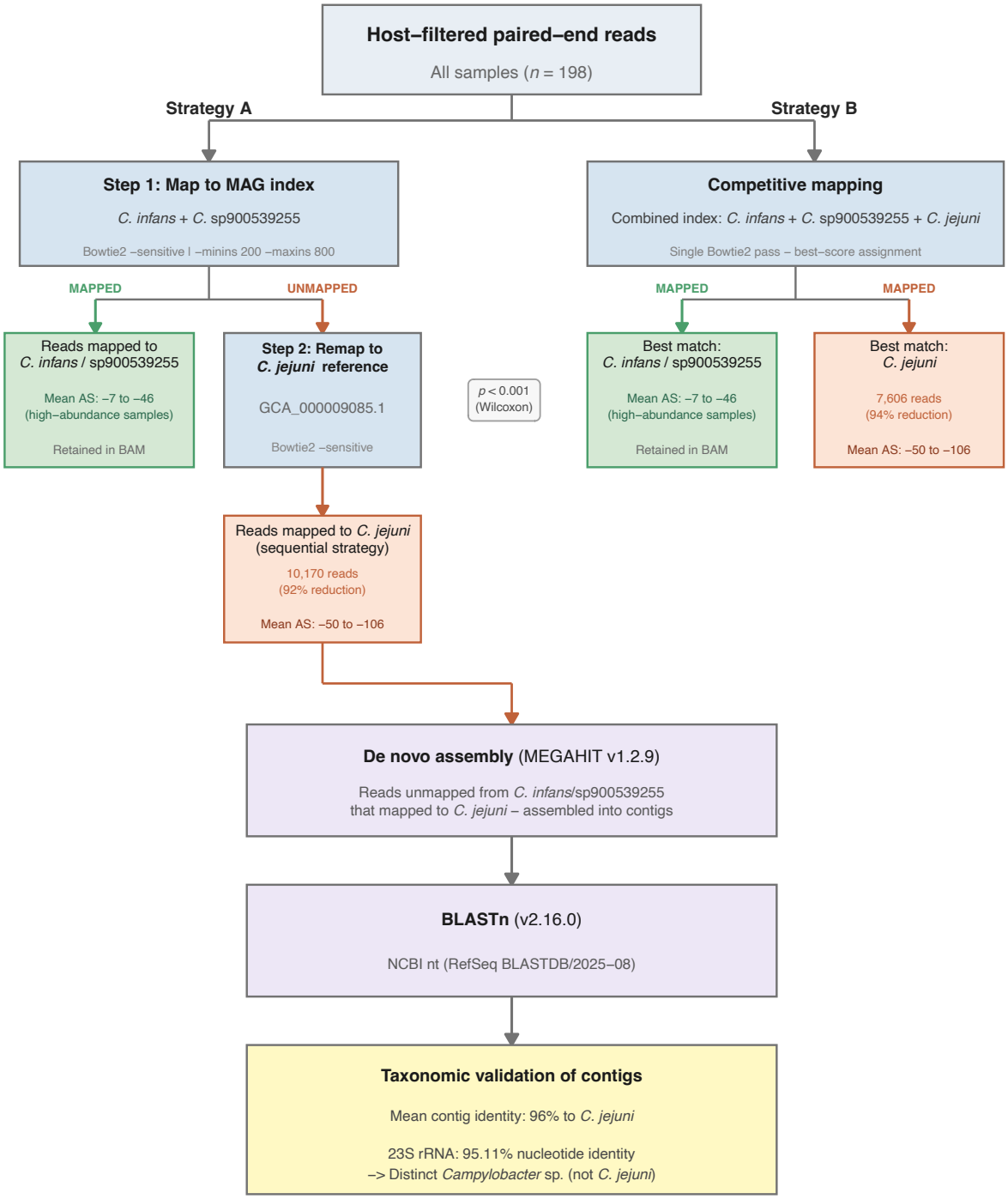


Figure S4. Schematic of the two complementary read mapping strategies used to discriminate locally assembled *Campylobacter* MAGs from *C. jejuni*. Strategy A (left): a sequential two-step approach in which host-filtered paired-end reads were first mapped to a combined *C. infans* and *C.* sp900539255 MAG index using Bowtie2 (--sensitive mode; --minins 200, --maxins 800). Reads that failed to align were subsequently remapped to the *C. jejuni* reference genome (GCA_000009085.1), reducing *C. jejuni*-assigned reads from 123,837 to 10,170 (92% reduction). Strategy B (right): a competitive mapping approach in which all reads were simultaneously aligned to a combined index containing *C. infans*, *C.* sp900539255, and *C. jejuni* sequences in a single Bowtie2 pass, with each read assigned to the reference achieving the highest alignment score. Under competitive conditions, only 7,606 reads were assigned to *C. jejuni* (94% reduction), confirming that the sequential approach was conservative. In high-abundance samples (>500 reads assigned to *C. infans*/sp900539255), mean Bowtie2 alignment scores were substantially better for the locally assembled MAGs (−7 to −46) than for *C. jejuni* (−50 to

−106; Wilcoxon signed-rank test, p<0·001). Reads assigned to *C. jejuni* under strategiy A were assembled de novo using MEGAHIT (v1.2.9) and queried against the NCBI nucleotide database using BLASTn (v2.16.0), confirming that assembled contigs represent a distinct *Campylobacter* species rather than *C. jejuni*.

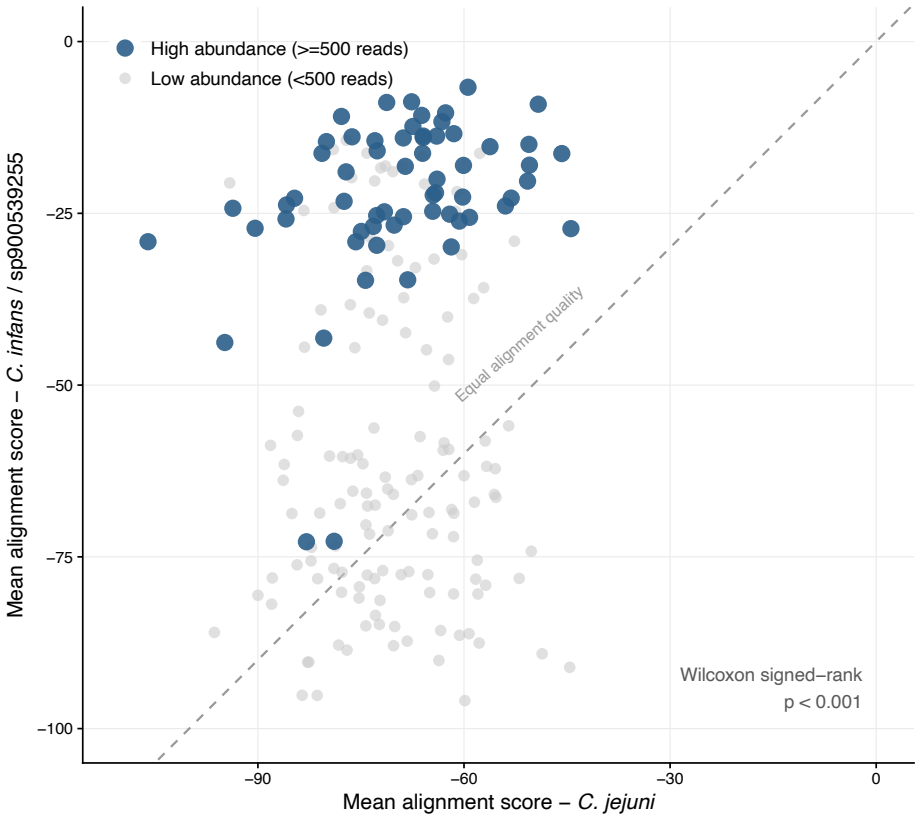


Figure S5. Competitive alignment score (AS) comparison between locally assembled *Campylobacter* MAGs and *C. jejuni* reference genome. Mean AS from competitive mapping, in which all host-filtered reads were simultaneously aligned to a combined index containing *C. infans*, *C.* sp900539255, and *C. jejuni* sequences in a single pass. Each point represents one sample; the y-axis shows mean AS for reads assigned to *C. infans*/sp900539255 MAGs and the x-axis shows mean AS for reads assigned to *C. jejuni*. AS of 0 indicates a perfect alignment; increasingly negative values indicate progressively poorer sequence identity to the reference. The dashed line indicates equal alignment quality between the two references; points above the line indicate greater sequence identity to *C. infans*/sp900539255. High-abundance samples (>500 reads assigned to *C. infans*/sp900539255; dark blue) cluster consistently above the diagonal, confirming preferential alignment to the locally assembled MAGs under true competitive conditions. Low-abundance samples (grey) show comparable alignment quality to both references, consistent with reads from conserved genomic regions shared across *Campylobacter* species. Wilcoxon signed-rank test, p<0·001.

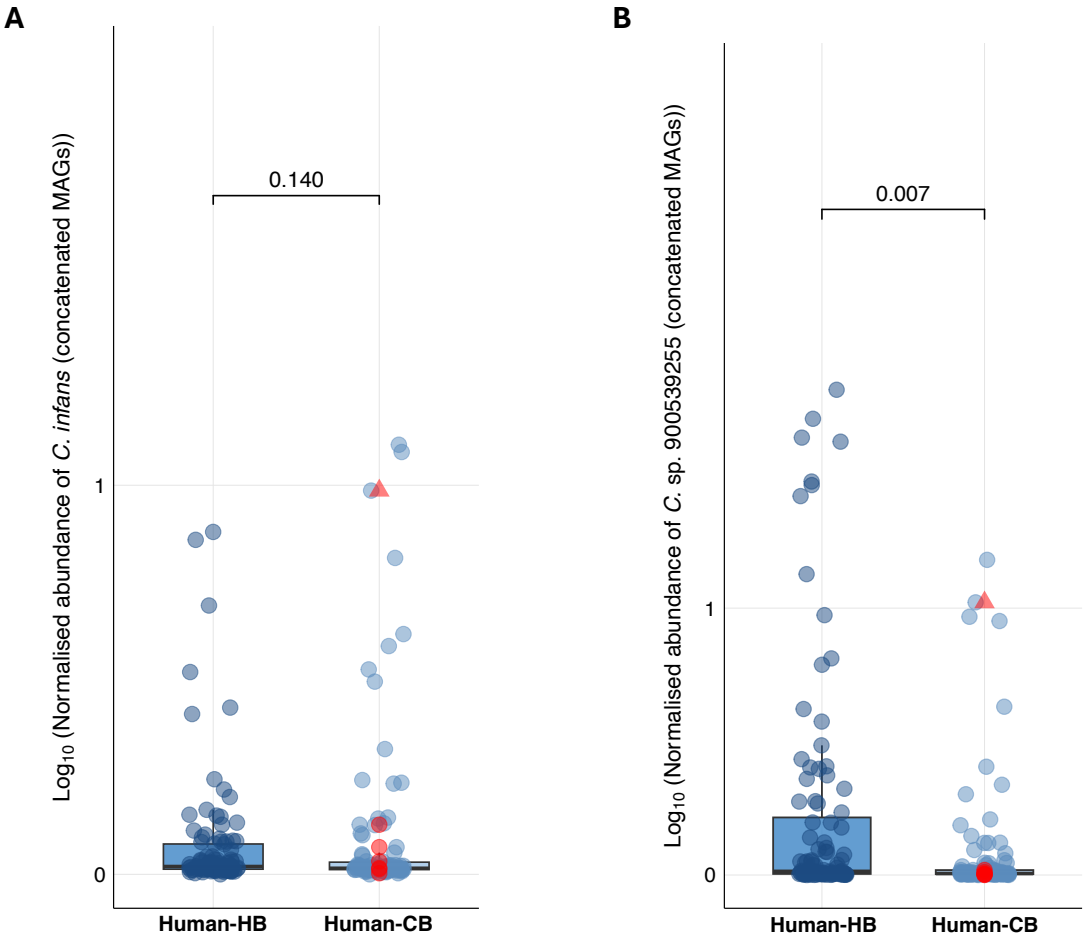


Figure S6. Abundance of human-associated *Campylobacter* species based on read mapping to all unique MAGs.

Individuals reporting diarrhoea are highlighted in red; the individual reporting severe diarrhoea is indicated by a triangle. A) Normalised read abundances of *C. infans* based on mapping to all unique *C. infans* MAGs assembled in this study, showing no significant differences between healthcare-based and community-based samples. B) In contrast, normalised read abundances of *C.* sp900539255 based on mapping to all unique MAGs were higher in healthcare-based samples than in community-based samples. Collectively, these findings support a specific association of *C.* sp900539255 with healthcare-seeking individuals, independent of the MAG mapping strategy used.

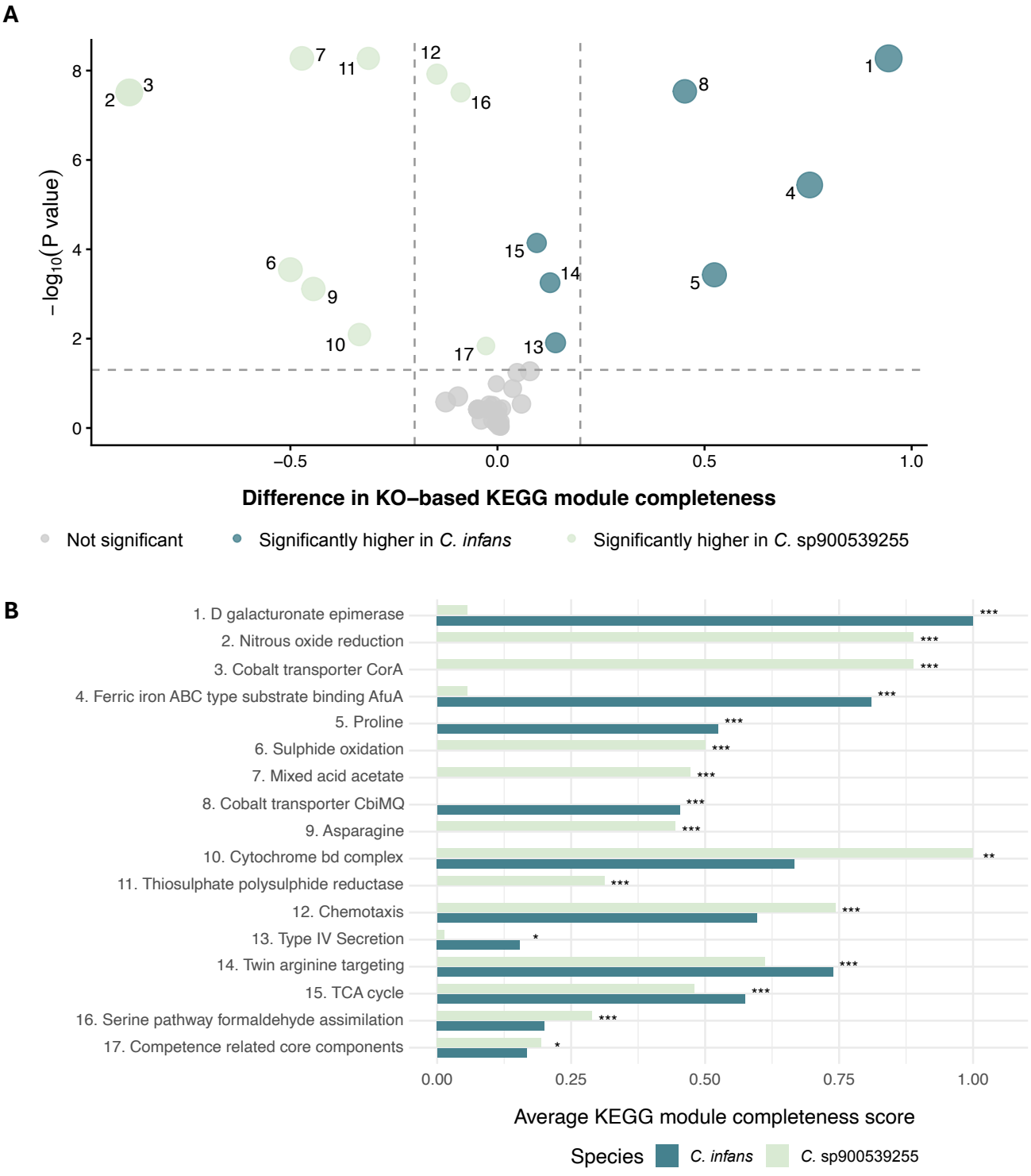


Figure S7. Genomic divergences between *C.* sp900539255 and *C. infans*.

Comparison of KEGG module completeness between *C. infans* and *C.* sp900539255 across 178 functional pathways. A) Statistical comparison of KEGG module completeness between *C. infans* and *C.* sp900539255 based on KEGG Orthology (KO) annotations. Each point represents a KEGG module, plotted by the difference in completeness between the two genomes (x-axis) and the $-\log_{10}$(p-value) of that difference (y-axis). Modules significantly more complete in *C. infans* are shown to the right, while those enriched in *C.* sp. 900539255 appear to the left. Modules without significant differences cluster near zero. B) Comparison of KEGG module completeness between *C. infans* and *C.* 900539255 across 17 functional pathways. Numbers correspond to the modules indicated in panel A. Of the pathways identified, 17 differed significantly in prevalence or completeness

between the two species, indicating overlapping metabolic repertoires with some species-specific differences. For example, a sulphur metabolism pathway involving thiosulphate–polysulphide reductase was present in 17 of 18 *C.* sp900539255 genomes but absent from all *C. infans* genomes ($p<0{\cdot}001$). Similarly, a nitrogen metabolism pathway involving nitrous oxide reduction was detected in 16 of 18 *C.* sp900539255 genomes and in none of the *C. infans* genomes ($p<0{\cdot}001$).

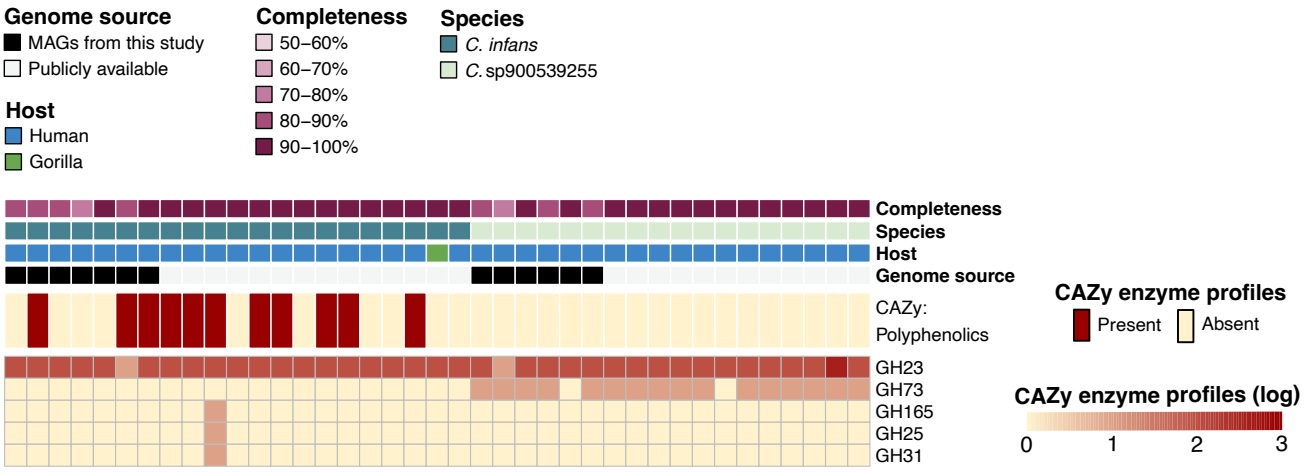


Figure S8. Functional diversity of carbohydrate-active enzymes (CAZymes) across *Campylobacter* genomes.

Heatmap showing the functional diversity of CAZymes across human-associated *Campylobacter* species MAGs assembled in this study and publicly available genomes. Pathways associated with polyphenolic compound degradation were identified in 11 of 21 *C. infans* genomes and were not detected in *C.* sp900539255 ($p<0{\cdot}001$), whereas GH73 was detected in 16 of 18 *C.* sp900539255 genomes but was absent from all *C. infans* genomes ($p<0{\cdot}001$). Together, these patterns indicate species-specific CAZyme repertoires consistent with functional divergence between human-associated *Campylobacter* species.

Supplementary Table 3: Prevalence of antibiotic resistance genes in *C. infans* genomes from this study and public databases.

	Detected from		
ARG	Genomes from this study	Public databases	Total
bla$_{OXA-471\_1}$	2/7	6/14	8/21
bla$_{OXA-193\_1}$	0/7	2/14	2/14
bla$_{OXA-489\_1}$	1/7	0/14	1/21
tet(W)$_{\_5}$	0/7	1/14	1/21

Supplementary Table 4: Prevalence of virulence genes in *Campylobacter* genomes from this study and public databases.

Virulence gene	*Campylobacter* species	Detected from		
		Genomes from this study	Public databases	Total
cheY	*C. infans*	6/7	14/14	20/21
cheY	*C. vicugnae*	4/5	20/21	24/26
fliN	*C.* sp900539255	0/6	1/11	1/17

References

1. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* 2019; **20**:257.
2. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics*. 2022; **38**: 5315-16.
3. Andrews S. FastQC: A quality control tool for high throughput sequence data. *Babraham Bioinformatics*. 2010.

   Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
4. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014; **30**: 2114-20.
5. Bushnell B. BBMap: a fast, accurate, splice-aware aligner. Lawrence Berkeley National Laboratory. 2014.
6. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*. 2015; **31**: 1674-76.
7. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012; **9**: 357-59.
8. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009; **25**: 2078-79.
9. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*. 2019; **7**: e7359.
10. Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*. 2016; **32**: 605-07.
11. Alneberg J, Bjarnason BS, De Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. *Nat Methods*. 2014; **11**: 1144-46.
12. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol*. 2018; **3**: 836-43.
13. Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat Methods*. 2023; **20**: 1203-12.
14. Olm MR, Brown CT, Brooks B, Banfield JF. DRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J*. 2017; **11**: 2864-68.
15. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput Sci*. 2017; **3**: e104.

16. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*. 2020; **37**: 1530-34.

17. Letunic I, Bork P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res*. 2024; **52**: W78-82.

18. Ojeda A, Deblais L, Mummed B, Brhane M, Hassen KA, Ahmedo BU, et al.; CAGED Research Team. Determinants of *Campylobacter* species diversity in infants and association with family members, livestock, and household environments in rural Eastern Ethiopia. *Gut Pathog*. 2025; **17**: 51.

19. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013; **30**: 772-80.

20. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. 2012; **28**: 1647-49.

21. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun*. 2018; **9**: 5114.

22. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput Biol*. 2018; **14**: e1005944.

23. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990; **215**: 403-410.

24. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. 2010; **11**: 119.

25. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014; **30**: 2068-69.

26. Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, Cattoir V, et al. ResFinder 4.0 for predictions of phenotypes from genotypes. *J Antimicrob Chemother*. 2020; **75**: 3491-500.

27. Florensa AF, Kaas RS, Clausen PTLC, Aytan-Aktug D, Aarestrup FM. ResFinder – an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. *Microb Genom*. 2022; **8**: 000748.

28. Liu B, Zheng D, Zhou S, Chen L, Yang J. VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Res*. 2022; **50**: D912-17.

29. Shaffer M, Borton MA, McGivern BB, Zayed AA, La Rosa SL, Solden LM, et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res*. 2020; **48**: 8883-900.

30. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*. 2020; **36**: 2251-2252.

31. Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol*. 2011; **7**: e1002195.

32. Graham ED, Heidelberg JF, Tully BJ. Potential for primary productivity in a globally-distributed bacterial phototroph. *ISME Journal*. 2018; **12**: 1861-1866.

33. Team RC. R Core Team 2023 R: A language and environment for statistical computing. R foundation for statistical computing. https://www.R-project.org/. R Foundation for Statistical Computing. 2023.