Data in brief

Whole genome sequencing and analysis of *Campylobacter coli* YH502 from retail chicken reveals a plasmid-borne type VI secretion system

Sandeep Ghatak a, Yiping He b,⁎, Sue Reed b, Terence Strobaugh Jr b, Peter Irwin b

⁎ Division of Animal Health, ICAR Research Complex for NEH Region, Umiam, Meghalaya 793103, India

a Molecular Characterization of Foodborne Pathogens Research Unit, Eastern Regional Research Center, Agricultural Research Service, United States Department of Agriculture, 600 East Mermaid Lane, Wyndmoor, PA 19038, USA

Abstract

*Campylobacter* is a major cause of foodborne illnesses worldwide. *Campylobacter* infections, commonly caused by ingestion of undercooked poultry and meat products, can lead to gastroenteritis and chronic reactive arthritis in humans. Whole genome sequencing (WGS) is a powerful technology that provides comprehensive genetic information about bacteria and is increasingly being applied to study foodborne pathogens: e.g., evolution, epidemiology/outbreak investigation, and detection. Herein we report the complete genome sequence of *Campylobacter coli* strain YH502 isolated from retail chicken in the United States. WGS, de novo assembly, and annotation of the genome revealed a chromosome of 1,718,974 bp and a mega-plasmid (pCOS502) of 125,964 bp. GC content of the genome was 31.2% with 1931 coding sequences and 53 non-coding RNAs. Multiple virulence factors including a plasmid-borne type VI secretion system and antimicrobial resistance genes (beta-lactams, fluoroquinolones, and aminoglycoside) were found. The presence of T6SS in a mobile genetic element (plasmid) suggests plausible horizontal transfer of these virulence genes to other organisms. The *C. coli* YH502 genome also harbors CRISPR sequences and associated proteins. Phylogenetic analysis based on average nucleotide identity and single nucleotide polymorphisms identified closely related *C. coli* genomes available in the NCBI database. Taken together, the analyzed genomic data of this potentially virulent strain of *C. coli* will facilitate further understanding of this important foodborne pathogen most likely leading to better control strategies. The chromosome and plasmid sequences of *C. coli* YH502 have been deposited in GenBank under the accession numbers CP018900.1 and CP018901.1, respectively.

Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

2. Experimental design, materials and methods

Among all of the species of *Campylobacter*, *C. jejuni* and *C. coli* are responsible for a large majority of cases of campylobacteriosis [1]. Considering the importance of *Campylobacters* and the advantages of whole genome sequencing (WGS) for studying foodborne bacteria, many researchers have reported genome sequences of these organisms in public databases. However, high-quality complete genomes are relatively rare: e.g., there are only 17 complete genomes of *C. coli* against a total submission of 797 genomes (https://www.ncbi.nlm.nih.gov/genome/genomes/1145 accessed on Dec. 12th, 2016). In the present study, we report the complete and analyzed genome of *C. coli* strain YH502, recently isolated from retail chicken, using PacBio single molecule real-time and Illumina MiSeq sequencing technologies.

The strain *C. coli* YH502 was isolated from chicken collected from a local supermarket using a passive filtration method [2]. Briefly, a 450-g chicken carcass was rinsed with 200 mL of 0.1% buffered peptone water. The chicken rinse was concentrated and enriched in Bolton broth containing horse blood and antibiotic supplement (20 mg/L.
ceuoperazone, 20 mg/L trimethoprim, 20 mg/L vancomycin, and 50 mg/L cyclohexamide) for 24 h. Passive filtration was performed for the isolation of Campylobacter cells. The strain was routinely grown in Brucella broth or on Brucella agar plates under microaerophilic conditions (5% O₂, 10% CO₂ and 85% N₂) at 42 °C.

Genomic DNA was extracted from the isolate using the Qiagen Genomic-tip 100/G kit (Qiagen, Valencia, CA) as per manufacturer’s instruction and quantified with a Qubit 3.0 fluorometer (Thermo Fisher Scientific, Waltham, MA). The identity of the strain was initially established by 16S rRNA sequencing [3] and was further confirmed by a multiplex real-time PCR assay targeting the hipO and cadA genes [4].

Campylobacter has a wide genetic diversity due to frequent inter- and intra-species DNA recombination [5,6]. To determine if the C. coli pan-genome was open (As per Heap’s Law a pan-genome is the set of non-redundant genes obtained from already sequenced genomes and the pan-genome is considered open if sequencing of new genomes are expected to add new genes to the existing pan-genome; conversely a pan-genome is considered closed if sequencing of new genomes does not add to the growth of the existing pan-genome.), all of the 17 complete genomes of C. coli were downloaded from GenBank in NCBI. These genomes were analyzed for their gene content using the Micropan R-package which estimated the openness of the population following Heap’s law (https://www.r-project.org/) [7,8]. Results of the analysis indicated an open pan-genome for C. coli (decay parameter, α = 0.579; if ‘α < 1.0’ pan-genome is open, if ‘α > 1.0’ the pan-genome is closed), which indicated the possibility of finding new gene(s) in newly sequenced C. coli genomes. Therefore, a recently isolated C. coli strain YH502 was processed for WGS analysis.

To generate high quality sequence data from the strain, two parallel next-generation sequencing technologies were employed—Single Molecule Real-Time (SMRT) sequencing (Pacific Bioscience, Menlo Park, CA) and the MiSeq System (Illumina, San Diego, CA). SMRT library preparation and sequencing were provided by the University of Delaware Sequencing & Genotyping Center performed using the SMRTbell template prep kit and PacBio RS II. For MiSeq sequencing, a genomic DNA library was prepared using the Nextera XT sample preparation kit and subsequent reactions were run as per manufacturer’s guidelines (Illumina, San Diego, CA). PacBio reads were de novo assembled using software Canu v 1.3 [9] and Illumina data were assembled with Spades v 3.7.1 [10]. The contigs from both assemblies were joined and corrected using Pilon to produce a draft genome (https://sepsis-omics.github.io/tutorials/modules/cmdline_assembly/). The genome was manually edited, ordered, oriented, and then confirmed by mapping reads back using CLC Genomics Workbench 9.5 (Qiagen Bioinformatics, Redwood, CA). The depth coverages of SMRT sequencing and MiSeq reads were 506 x and 48 x, respectively.

3. Data description

The complete genome of C. coli YH502 contains a 1,718,974 bp chromosome and 125,964 bp plasmid with a GC content of 31.2%, which corresponded well with that of other C. coli genomes available in the NCBI database. Annotation of the genome via Rapid Annotation using Subsystems Technology (RAST, http://rast.nmpdr.org/) [11] revealed 318 subsystems, 1931 coding sequences and 53 RNAs. Among the subsystems identified by RAST there were 67 genes associated with virulence, disease causation and defense of the organism (Fig. 1). Four phage related genes were also present. In view of the large number of virulence associated genes in C. coli YH502, the genome was assessed for its pathogenic potential by the PathogenFinder tool [12]. The analysis showed the probability of being a human pathogen was 80.5%, indicating a high potential for this organism to cause human disease.

Among the virulence factors identified by RAST, there were a number of genes associated with antimicrobial resistance. These included beta-lactamase (1), CmeABC multidrug efflux pump (4), various families of multidrug resistance efflux pumps (11; RND, MATE, MFS), and fluoroquinolone resistance (2). Further investigation of the genome with the ResFinder tool [13] identified the signatures of beta-lactam resistance mediated by blaOXA-61 (encoding beta-lactamase) and aminoglycoside resistance mediated by aph(3′)-VIIa (encoding aminoglycoside kinase).

Interestingly, of the 66 genes identified as part of a membrane transport subsystem, 15 were involved in the Type VI Secretion System (T6SS), and 4 were associated with the Type IV Secretion System (T4SS). This strain harbored hallmark T6SS genes namely, vgrG (encoding valine-glycine repeats) and hcp (encoding hemolysin correlated protein). Recently, T6SS has been reported in C. jejuni strains [14,15], and to our knowledge this is the first report of a plasmid-borne putative T6SS in C. coli. The presence of T6SS in a mobile genetic element (plasmid) facilitates horizontal transfer of these virulence genes to strains.

![Subsystem Coverage](image1)

**Fig. 1.** Subsystem categories of the C. coli YH502 genome annotated by RAST. Bar diagram on the left shows subsystem coverage. The pie diagram shows the distribution of subsystem features.
other organisms. Other important virulence factors identified by RAST included genes encoding cytolethal distending toxins (cdtA, cdtB, and cdtC), which were confirmed by real-time PCR experiments during identity establishment of the strain.

In addition to the virulence factors, C. coli YH502 also harbored a distinct CRISPR (clustered regularly interspaced palindromic repeats) system between 1,469,763 and 1,476,435 bp, including 14 CRISPR repeats (GTTTTAGTCCCTTTTTAAATTTCTTTATGGTAAAAT) and genes encoding Cas1, Cas2 and Csn family proteins.

RAST identified C. coli RM2228 as the closest neighbor of the YH502 strain. However, analysis of Average Nucleotide Identity of YH502 against other complete C. coli genomes in the NCBI database revealed the highest nucleotide identity with C. coli strain 14983A isolated from a turkey farm house in North Carolina (Table 1) [18]. Therefore, we performed a phylogenetic analysis based on single nucleotide polymorphisms (SNPs) among these complete genomes of C. coli including our strain YH502. Results indicated that the closest neighbor of the YH502 strain was C. coli strain 14983A (Fig. 2) as was revealed by ANI calculation. Moreover, multilocus sequence typing (http://pubmlst.org/) analysis of the complete genome showed that C. coli YH502 and 14983A belonged to the same clonal complex (ST-828), confirming the close relatedness of these two strains [19].

Taken together, we believe that the genomic data of this potentially virulent strain of C. coli YH502 with T6SS on a plasmid and multidrug resistant genes will facilitate further understanding of this important foodborne pathogen likely leading to better control strategies.

4. Accession number of nucleotide sequence

The complete genome and plasmid sequences of C. coli YH502 have been deposited to NCBI under the accession CP018900.1 and CP018901.1, respectively.

Conflict of interest

The authors declare no conflict of interest about the work published in this paper.

Acknowledgement

This research was supported by the U.S. Department of Agriculture, Agricultural Research Service (USDA-ARS), Current Research Information System number 8072-42000-084. The first author (SG) is thankful to Department of Biotechnology, New Delhi, India for Overseas Associateship Grant for NER (2015-16) and to ICAR for necessary support. We thank Dr. Pina Fratamico at the USDA-ARS-ERRC for her constructive comments on the manuscript.

![Fig. 2. Phylogenetic tree of C. coli strains based on whole genome single nucleotide polymorphism analysis. The tree was drawn from single nucleotide polymorphism data generated by the comparison of complete C. coli genome sequences using CSI Phylogeny 1.4 (https://cge.cbs.dtu.dk/services/CSIPhylogeny/).](image-url)
[1] A.H. van Vliet, J.M. Ketley, Pathogenesis of enteric Campylobacter infection. Symp Ser Soc Appl Microbiol. 2001, pp. 455–565.

[2] Y. He, S. Reed, A.K. Bhunia, A. Gehring, L.H. Nguyen, P.L. Irwin, Rapid identification and classification of Campylobacter spp. using laser optical scattering technology. Food Microbiol. 47 (2015) 28–35.

[3] P. Irwin, T.L. Nguyen, C.Y. Chen, Binding of nontarget microorganisms from food washes to anti-Salmonella and anti-E. coli O157 immunomagnetic beads: minimizing the errors of random sampling in extreme dilute systems. Anal. Bioanal. Chem. 391 (2008) 515–524.

[4] Y. He, X. Yao, N.W. Gunther IV, Y. Xie, S.-I. Tu, X. Shi, Simultaneous detection and differentiation of Campylobacter jejuni, C. coli, and C. lari in chickens using a multiplex real-time PCR assay. Food Anal. Methods 3 (2010) 321–329.

[5] A.B. Vidal, F.M. Colles, J.D. Rodgers, N.D. McCarthy, R.H. Davies, M.C. Maiden, F.A. Clifton-Hadley, Genetic diversity of Campylobacter jejuni and Campylobacter coli isolates from conventional broiler flocks and the impacts of sampling strategy and laboratory method. Appl. Environ. Microbiol. 82 (2016) 2347–2355.

[6] S. Koren, B.P. Walenz, K. Berlin, J.R. Miller, N.H. Bergman, A.M. Phillippy, Canu: Scalable and Accurate Long-read Assembly Via Adaptive k-mer Weighting and Repeat Separation. bioRxiv, 2017 071282.

[7] K. Lertpiriyapong, E.R. Gamazon, Y. Feng, D.S. Park, J. Pang, G. Botka, M.E. Graffam, Z. Ge, J.G. Fox, Campylobacter jejuni type VI secretion system: roles in adaptation to deoxycholic acid, host cell adherence, invasion, and in vivo colonization. PLoS One 7 (2012), e42842.

[8] W.G. Miller, S. Huynh, C.T. Parker, J.A. Niedermeyer, S. Kathariou, Complete genome sequences of multidrug-resistant Campylobacter jejuni strain 14980A (turkey feces) and Campylobacter coli strain 14983A (housefly from a turkey farm), harboring a novel gentamicin resistance mobile element. Genome Announc. 4 (2016).