2014

Comparative Genomics of the *Campylobacter lari* Group

William G. Miller
*Produce Safety and Microbiology Research Unit, william.miller@ars.usda.gov*

Emma Yee
*Produce Safety and Microbiology Research Unit*

Mary H. Chapman
*Produce Safety and Microbiology Research Unit*

Timonthy P. L. Smith
*USDA-ARS, tim.smith@ars.usda.gov*

James L. Bono
*Meat Safety and Quality Research Unit, jim.bono@ars.usda.gov*

*See next page for additional authors*

---

Follow this and additional works at: [http://digitalcommons.unl.edu/hruskareports](http://digitalcommons.unl.edu/hruskareports)

---

Miller, William G.; Yee, Emma; Chapman, Mary H.; Smith, Timonthy P. L.; Bono, James L.; Huynh, Steven; Parker, Craig T.; Vandamme, Peter; Luong, Khai; and Korfach, Jonas, "Comparative Genomics of the *Campylobacter lari* Group" (2014). *Roman L. Hruska U.S. Meat Animal Research Center*. 351.
[http://digitalcommons.unl.edu/hruskareports/351](http://digitalcommons.unl.edu/hruskareports/351)

This Article is brought to you for free and open access by the U.S. Department of Agriculture: Agricultural Research Service, Lincoln, Nebraska at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Roman L. Hruska U.S. Meat Animal Research Center by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
Comparative Genomics of the *Campylobacter lari* Group

William G. Miller¹*, Emma Yee¹, Mary H. Chapman¹, Timothy P.L. Smith², James L. Bono², Steven Huynh¹, Craig T. Parker¹, Peter Vandamme³, Khai Luong⁴, and Jonas Korlach⁴

¹Produce Safety and Microbiology Research Unit, Agricultural Research Service, U.S. Department of Agriculture, Albany, California
²Meat Safety and Quality Research Unit, Agricultural Research Service, U.S. Department of Agriculture, Clay Center, Nebraska
³Laboratory of Microbiology, Department of Biochemistry and Microbiology, Ghent University, Belgium
⁴Pacific Biosciences, Menlo Park, California

*Corresponding author: E-mail: william.miller@ars.usda.gov.

Accepted: November 4, 2014

Data deposition: All genome sequences, including plasmids, have been deposited at GenBank under the accessions CP007766 through CP007778 (details provided in table 1).

Abstract

The *Campylobacter lari* group is a phylogenetic clade within the epsilon subdivision of the Proteobacteria and is part of the thermotolerant *Campylobacter* spp., a division within the genus that includes the human pathogen *Campylobacter jejuni*. The *C. lari* group is currently composed of five species (*C. lari*, *Campylobacter insulaenigrae*, *Campylobacter volucris*, *Campylobacter subantarcticus*, and *Campylobacter peloridis*), as well as a group of strains termed the urease-positive thermophilic *Campylobacter* (UPTC) and other *C. lari*-like strains. Here we present the complete genome sequences of 11 *C. lari* group strains, including the five *C. lari* group species, four UPTC strains, and a *lari*-like strain isolated in this study. The genome of *C. lari* subsp. *lari* strain RM2100 was described previously. Analysis of the *C. lari* group genomes indicates that this group is highly related at the genome level. Furthermore, these genomes are strongly syntenic with minor rearrangements occurring only in 4 of the 12 genomes studied. The *C. lari* group can be bifurcated, based on the flagella and flagellar modification genes. Genomic analysis of the UPTC strains indicated that these organisms are variable but highly similar, closely related to but distinct from *C. lari*. Additionally, the *C. lari* group contains multiple genes encoding hemagglutination domain proteins, which are either contingency genes or linked to conserved contingency genes. Many of the features identified in strain RM2100, such as major deficiencies in amino acid biosynthesis and energy metabolism, are conserved across all 12 genomes, suggesting that these common features may play a role in the association of the *C. lari* group with coastal environments and watersheds.

Key words: flagella, hemagglutination, methylome, UPTC.

Introduction

*Campylobacter lari* (formerly *Campylobacter laridis*) strains were originally described as nalidixic acid-resistant thermophilic *Campylobacter* (NARTC; Skirrow and Benjamin 1980). Subsequently, the urease-positive thermophilic *Campylobacter* (UPTC), the nalidixic-acid susceptible (NASC) group, and the urease-producing NASC were identified as *C. lari*-like strains (Endtz et al. 1997; Megraud et al. 1988; Owen et al. 1988). Additional *lari*-like species were described, such as *Campylobacter insulaenigrae* (Foster et al. 2004) and *Campylobacter subantarcticus* (Debruyne et al. 2010b). Together, these taxa comprise the *C. lari* group.

Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution 2014. This work is written by US Government employees and is in the public domain in the US.

GBE

### Table 1

<table>
<thead>
<tr>
<th>Strain</th>
<th>Source</th>
<th>Location</th>
<th>Type</th>
<th>Coverage</th>
<th>Optical Map</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. insulaenigrae</td>
<td>Marine mammal</td>
<td>Scotland (GB)</td>
<td>Y</td>
<td>107 × H; 98 × H</td>
<td>H; 1,347</td>
<td>N/A</td>
</tr>
<tr>
<td>C. lari concheus</td>
<td>Human</td>
<td>United Kingdom (CA)</td>
<td>Y</td>
<td>61 × H; 1,429</td>
<td>H; 1,287</td>
<td>N/A</td>
</tr>
<tr>
<td>C. lari</td>
<td>River water</td>
<td>United States (CA)</td>
<td>Y</td>
<td>121 × M; 708</td>
<td>M; 1,478</td>
<td>N/A</td>
</tr>
<tr>
<td>C. peloridis</td>
<td>River water</td>
<td>The Netherlands</td>
<td>Y</td>
<td>35 × H; 1,146</td>
<td>H; 1,175</td>
<td>N/A</td>
</tr>
<tr>
<td>C. subantarcticus</td>
<td>Marine mammal</td>
<td>United States (CA)</td>
<td>N</td>
<td>110 × H; 708</td>
<td>M; 1,438</td>
<td>N/A</td>
</tr>
<tr>
<td>C. volucris</td>
<td>Shellfish</td>
<td>Sweden</td>
<td>Y</td>
<td>110 × H; 708</td>
<td>M; 1,438</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Although the *C. lari* group is a phylogenetically distinct clade within the genus *Campylobacter* (supplementary fig. S1, Supplementary Material online), taxa within this clade are highly related, and this similarity may reflect, in part, the similar hosts and environments from which these strains are isolated. *Campylobacter lari* was isolated originally from gulls (Skirrow and Benjamin 1980; Benjamin et al. 1983). Other members of the *C. lari* group are also isolated from gulls (UPTC, *C. volucris* [Kaneko et al. 1999; Debruyne et al. 2010b]) and other shorebirds, such as plovers, redshanks, dunlins, sandpipers, skuas, albatrosses, and penguins (*C. lari*, UPTC, *C. subantarcticus* [Waldenstrom et al. 2002, 2007; Lectta et al. 2006; Debruyne et al. 2010a; Ryu et al. 2014]). The *C. lari* group strains are also isolated from marine mammals (*C. insulaenigrae*, *C. lari* [Foster et al. 2004; Stoddard et al. 2007; Garcia-Pena et al. 2010; Gonzalez et al. 2011]), shellfish (*C. lari*, *C. peloridis* [Endtz et al. 1997; Van Doorn et al. 1998; Debruyne et al. 2009]), and seawater/fresh water (*C. lari*, UPTC [Obiri-Danso and Jones 1999; Obiri-Danso et al. 2001; Meinersmann et al. 2013; Khan et al. 2014]). Even though members of the *C. lari* group have been isolated from livestock (Tresierra-Ayala et al. 1994; Aarestrup et al. 1997; Harvey et al. 1999; Scanlon et al. 2013), this clade is primarily associated with coastal regions and watersheds.

Within *Campylobacter*, *C. jejuni* is the primary agent in nearly all *Campylobacter*-related human illnesses. However, *C. lari* has been associated occasionally with human illness, causing gastroenteritis with abdominal pain, fever, and diarrhea (Broczyk et al. 1987; Lin et al. 1998; Prasad et al. 2001; Debruyne et al. 2009), and seawater/fresh water (*C. lari*, UPTC [Obiri-Danso and Jones 1999; Obiri-Danso et al. 2001; Meinersmann et al. 2013; Khan et al. 2014]). Even though members of the *C. lari* group have been isolated from livestock (Tresierra-Ayala et al. 1994; Aarestrup et al. 1997; Harvey et al. 1999; Scanlon et al. 2013), this clade is primarily associated with coastal regions and watersheds.

The genome of *C. lari* subsp. *lari* strain RM2100 (ATCC-BAA 1060; CDC strain D67, “case 6” [Tauxe et al. 1985]), isolated from an 8-month-old girl with watery diarrhea, has been sequenced to completion (Miller et al. 2008). Although 90% of the genes predicted in the strain RM2100 genome are similar to genes present in the genomes of other thermotolerant *Campylobacter* spp., a substantial number of genes, especially those associated with amino acid biosynthesis and energy metabolism, and identified previously in other *Campylobacter* genomes, are absent from the *C. lari* subsp. *lari* strain RM2100 genome. Citrate synthase, a key component of the TCA cycle, is not encoded by the *C. lari* subsp. *lari* strain RM2100 genome. Polymerase chain reaction (PCR) analysis indicated that citrate synthase is not encoded by...
C. insulanigrae (Stoddard et al. 2007) or other members of the C. lari group (data not shown), suggesting that the genomic features identified in C. lari subsp. lari may apply to the clade as a whole. To determine whether the genomic features identified in C. lari subsp. lari strain RM2100 are common to the C. lari group and can help to explain the host and environmental association of this clade, the genomes of the remaining validly named taxa within the C. lari group (table 1) were sequenced to completion. Also, to provide further evidence for the taxonomic placement of the UPTC strains, the genomes of four UPTC strains were also sequenced. Here, we present the comparative genomic analysis of 12 C. lari group strains.

Materials and Methods

Growth Conditions and Chemicals

Campylobacter strains were cultured at 37 °C on Brain Heart Infusion agar (Becton Dickinson, Sparks, MD) amended with 5% (v/v) laked horse blood (Hema Resource & Supply, Aurora, OR). The incubation atmosphere was 5% H2, 10% CO2, and 85% N2. PCR enzymes and reagents were purchased from New England Biolabs (Beverly, MA) or Epicentre (Madison, WI). All chemicals were purchased from Sigma-Aldrich Chemicals (St. Louis, MO) or Fisher Scientific (Pittsburgh, PA). DNA sequencing reagents and capillaries were purchased from Applied Biosystems (Foster City, CA), Roche Life Science (Indianapolis, IN), Illumina Inc. (San Diego, CA), or Pacific Biosciences (Menlo Park, CA).

Isolation of Campylobacter Strains from River Water Samples

In a six-well plate (Corning, Corning, NY), 7.2 ml of a river water sample was added to 0.8 ml 10× ABB (Anaerobe Basal Broth, Oxoid [Remel], Lenexa, KS) + 10× Preston supplement (ampicillin B [10 μg/ml], rifampicin [10 μg/ml], trimethoprim lactate [10 μg/ml], and polymyxin B [5 μU/ml]; Oxoid) to achieve a final 1× ABB + Preston concentration. These six-well plates were placed inside a plastic zip-loc bag containing 1–2% O2 + Bioblend gas (10% CO2, 10% H2, 80% N2; Praxair, Danbury, CT) and incubated for 24 h at 37 °C and 40 rpm. For each enriched sample, a 10-μl loop was struck onto an ABA (Anaerobe Basal Agar) plate (Oxoid) amended with 5% laked horse blood (Hema) and CAT supplement (cefoperazone [8 μg/ml], amphotericin B [10 μg/ml], and teicoplanin [4 μg/ml]; Oxoid). Plates were incubated in a microaerobic gas jar (Anaerolab 3.5 L System; Oxoid) at 37 °C for 1–2% O2 + Bioblend gas for 24–48 h. All positive cultures were examined under a 1,000× microscope. Cultures positive for Campylobacter were then filtered through a 0.6-μm mixed cellulose filter onto an ABA plate. After growth for 24 h, single colonies were picked onto a new ABA plate, then incubated 24–48 h. Pure cultures were stored at −80 °C; species identification of the isolates was achieved by 16S rRNA gene sequencing. All strains were tested for urease and nitrate reductase activity.

Urease and Nitrate Tests

Urease and nitrate reductase activities were detected using the API-Campy system (bioMerieux, France). All tests were repeated at least twice. Campylobacter jejuni subsp. doylei (Cjd) strain SSI 5384 (ure−, nap−) was used as a negative control.

Polymerase Chain Reactions

Genomic DNA was prepared as described previously (Miller et al. 2005). Standard amplifications were performed on a Tetrad thermocycler (Bio-Rad, Hercules, CA) with the following settings: 94 °C for 30 s, 53 °C for 30 s, and 72 °C for 2 min (30 cycles). Each amplification mixture contained 50 ng genomic DNA, 1× PCR buffer (Epicentre), 1× PCR enhancer (Epicentre), 2.5 mM MgCl2, 250 μM each dNTP, 50 pmol each primer, and 1 U polymerase (New England Biolabs). Amplicons were purified using ExoSAP-IT (Affymetrix, Santa Clara, CA). Sequencing and PCR oligonucleotides were designed using Primer Premier (v 5.0; Premier Biosoft, Palo Alto, CA) and purchased from Eurofins (Huntsville, AL).

Sanger Sequencing

Sanger cycle sequencing reactions were performed on a 96-well Tetrad thermocycler (Bio-Rad) using the ABI PRISM BigDye terminator cycle sequencing kit (version 3.1) and standard protocols. Cycle sequencing extension products were purified using BigDye XTerminator (Applied Biosystems). DNA sequencing was performed on an ABI PRISM 3730 DNA Analyzer (Applied Biosystems), using POP-7 polymer and ABI PRISM Genetic Analyzer Data Collection and ABI PRISM Genetic Analyzer Sequencing Analysis software. Sequences were trimmed, assembled, and analyzed in SeqMan (v 8.0.2; DNASTAR, Madison, WI).

Roche and Illumina Next-Generation Sequencing

Shotgun and paired-end (8–12 kb) 454 reads were obtained on a Roche 454 GS-FLX+ Genome Sequencer with Titanium chemistry using standard protocols. Illumina libraries were prepared using the KAPA Low-Throughput Library Preparation Kit with Standard PCR Amplification Module (Kapa Biosystems, Wilmington, MA), following manufacturer’s instructions except for the following changes: 750 ng DNA was sheared at 30 psi for 40 s and size selected to 700–770 bp following Illumina protocols. Standard desalting TruSeq LT and PCR Primers were ordered from Integrated DNA Technologies (Coralville, IA) and used at 0.375 and 0.5 μM final concentrations, respectively. PCR was reduced to 3–5 cycles. Libraries were quantified using the KAPA Library Quantification Kit.
(Kapa), except with 10 μl volume and 90-s annealing/extension PCR, then pooled and normalized to 4 nM. Pooled libraries were requantified by ddPCR on a QX200 system (Bio-Rad), using the Illumina TruSeq ddPCR Library Quantification Kit and following manufacturer’s protocols, except with an extended 2-min annealing/extension time. The libraries were sequenced 2 x 250 bp paired end v2 on a MiSeq instrument (Illumina) at 13.5 pM, following manufacturer’s protocols. Illumina HiSeq reads were obtained from SeqWright (Houston, TX).

Single Molecule, Real-Time Sequencing

Single Molecule, Real-Time (SMRT) sequencing was performed on the Pacific Biosciences (PacBio) RS sequencing platform using 10- or 20-kb SMRTbell libraries, and C2/C2 (most strains) or P5/C3 (strain RM16704) sequencing chemistry. The two chemistries used the 90- or 120-min data collection protocols, respectively. The SMRTbell libraries were prepared from 5 to 10 μg of bacterial genomic DNA, using the standard protocol from Pacific Biosciences as described in the 10-kb Template Preparation and Sequencing and Procedure (http://www.smrtcommunity.com/SamplePrep), or the 20-kb procedure for strain RM16704, and processed for sequencing as recommended by the supplier. A FASTQ file was generated from the PacBio reads using SMRTanalysis (ver. 2.1), and the reads were error-corrected using pacBioToCA with self-correction (Koren et al. 2013). The longest 20-x of the corrected reads was assembled with Celera Assembler 7.0 (Koren et al. 2012). The resulting contigs were polished using Quiver (Chin et al. 2013).

Genome Sequencing and Assembly

The 11 genomes characterized in this study (table 1) were all sequenced initially using Roche 454 technology. Shotgun and paired-end 454 reads were assembled using the Roche Newbler assembler (v2.6) into one or two chromosomal scaffolds, providing draft genome sequences with a coverage of 35–136 x (table 1). Intrascaffold gaps were filled using the 454 repeat contigs and the Perl script contig_extender3 (Merga et al. 2013). Contig gaps were closed/validated using PCR amplification and Sanger sequencing, generating draft pseudomolecules for each genome. Illumina HiSeq or MiSeq reads were assembled within Newbler as reference assemblies using the draft pseudomolecules as templates: These Illumina contigs were assembled into the 454 assembly within Seqman (v. 8.0.2) to validate all 454 base calls. The Illumina reads also provided an additional 475–1,478 x coverage (table 1). The presence/absence of single nucleotide polymorphisms (SNPs) within the repeat contigs was assessed using Geneious (v. 7.1.2; Biomatters, Auckland, New Zealand) and the Illumina reads or by using the 454 paired-end reads to link SNPs to adjacent unique contigs. For five strains (table 1), the genomes could not be closed using standard 454/Sanger/ILLUMINA technology, due to a complex topography consisting of multiple repeat contigs, and two of these genomes could not be assembled into a single scaffold. Therefore, these genomes were sequenced using a PacBio RS sequencer (Pacific Biosciences), which produced sequencing reads long enough to uniquely span each of the repeat regions. Using only the PacBio data, all five strains were closed into single, circularly closed chromosomes. Illumina reads were used to validate the PacBio base calls as described above. Seven assemblies were verified (table 1) using a bacterial optical restriction map (OpGen, Gaithersburg, MD).

Genome Annotation and Analysis

Putative coding sequences (CDSs) were determined using GeneMark (Besemer and Borodovsky 2005). Initial annotation was accomplished by comparing the predicted proteins with the proteome of the C. lari subsp. lari strain RM2100 (Miller et al. 2008) and with the NCBI (National Center for Biotechnology Information) nonredundant (nr) database using BLASTP; positive matches had an identity of ≥50%, and an alignment length of ≥75% across both the query and match sequences. Annotation was also assisted through the detection of characteristic Pfam motifs (Punta et al. 2012). The list of putative CDSs was then used to create a preliminary GenBank-formatted (.gbk) file that was entered into Artemis (release 16.0; Rutherford et al. 2000). Annotation/curation within Artemis included the fusion of split CDSs into pseudogenes and the identification of genes overlooked in the initial GeneMark analysis. The start codon of each putative CDS was curated manually, either through visual inspection within Artemis or through BLAST comparison of each CDS to its orthologs within the C. lari group, where present. Transfer RNAs (tRNAs) were annotated using tRNAscan-SE (v 1.23; Lowe and Eddy 1997). rRNA loci were identified through RNAmmer (Lagesen et al. 2007). Comparative genomic analysis was performed through a pairwise BLASTP analysis of the C. lari group proteome against itself. For each protein, a custom Perl script was used to identify the top match within the other proteomes, where present, using the match parameters described above. This Perl script also determined the core proteome for the C. lari group and calculated the pairwise average amino acid identities (AAI) between the core proteomes of any two strains. Further comparative analyses were performed using 1) the BLAST Ring Image Generator (BRIG v0.95; Alkhani et al. 2011) and BLASTN, with a default minimum threshold of 50% and an E value of 0.0001; or 2) JSpecies (v. 1.2.1; Richter and Rossello-Mora 2009), using default parameters, to determine average nucleotide identity (ANI) values.

Determination of Variability at the Homopolymeric G:C Tracts

A set of Illumina HiSeq or MiSeq reads were available for each of the 12 genomes characterized in this study, with a coverage...
ranging from 475× to 1,478×. A custom Perl script was designed to 1) identify the position of all G:C tracts (≥8 bp) within a genome, 2) determine the genome sequences that flank each identified G:C tract 15 nt upstream and downstream, 3) bin each Illumina read as positive if the flanking sequences within the read match 15/15 nt to the genomic flanking sequences and if the G:C tract within the read is homopolymeric, and 4) tabulate the tract lengths within the positive reads. This script was run using the entire Illumina read set for each genome. If the results accurately reflected true variation of the genomic G:C tracts and were not due to, for example, poor Illumina read quality, then hypervariability would not be observed in the more stable homopolymeric A:T tracts. Thus, the script was modified to identify A:T tracts ≥9 bp (9 bp was chosen over 8 bp due to the extremely large number of 8 bp A:T tracts per genome); this modified script was tested on the C. insulaenigrae and C. subantarcticus genomes. As expected, the A:T tracts were quite stable and demonstrated an average variability of approximately 0.4% (data not shown), far lower than that observed in the G:C tracts (for C. insulaenigrae, average variability = 7%, range = 0–41%). Therefore, an arbitrary cutoff value of 2% (±2 SD) was selected as a base value: Tract lengths whose proportion was less than 2% were not included in the determination of hypervariability.

Phylogenetic Analysis

Sequence alignments were performed using CLUSTALX (ver. 2.1). Dendrograms were constructed using the neighbor-joining method and Poisson correction. Bootstraps were conducted with 500 replicates. Phylogenetic analyses were performed using MEGA version 6.05 (Tamura et al. 2013).

Methylome Analysis Using SMRT Sequencing

As described in previous publications, the kinetic information contained in the same SMRT sequencing data used for the de novo assembly can be utilized to characterize the methylobe of bacteria (Flusberg et al. 2010; Clark et al. 2012; Murray et al. 2012). The methylobmes of the four strains sequenced and closed on the PacBio RS were determined using the RS Modification and Motif Analysis.1 protocol included in SMRT Portal. The de novo PacBio-only assemblies of the corresponding strains from this study were used as the reference genomes in the base modification analysis. The methylome results are deposited in New England Biolab’s REBASE (http://rebase.neb.com/cgi-bin/pacbiolist).

Accession Numbers

The complete nucleotide sequences and annotations of the strains characterized in this study were deposited in GenBank (table 1). The complete nucleotide sequence and annotation of C. lari subsp. lari strain RM2100 (Miller et al. 2008) was deposited previously in GenBank under the accession numbers CP000932 (chromosome) and CP000933 (megaplasmid pCL2100).

Results and Discussion

General Features

The genome of C. lari subsp. lari strain RM2100 was determined to contain 1,525,460 bp (Miller et al. 2008). The genomes of 11 other strains, representing at least five additional taxa within the C. lari group, were sequenced in this study. A summary of the features of these 11 genomes, with inclusion of the previously sequenced C. lari subsp. lari strain RM2100, is presented in table 2. The genome sizes of these 11 strains ranged from 1,465 (C. insulaenigrae) to 1,853 Mb (C. subantarcticus strain LMG 24377). The size differential between the genomes can be explained in part by the presence of genomic islands and prophage (see below). Consistent with the %G+C content of C. lari subsp. lari strain RM2100 (29.7), the %G+C content of the 11 genomes ranges from 28.19 to 29.94.

The C. lari strain RM2100 genome is predicted to contain 1,495 CDSs with an additional 18 fragmented CDSs, containing frameshift mutations or other point mutations, designated as putative pseudogenes (table 2). Excluding those CDSs contained in prophage or genomic islands, similar numbers of CDSs (1,384–1,499) were identified within the C. lari group. The number of pseudogenes per genome was also generally consistent (9–22; table 2); however, a higher number of pseudogenes was identified in Campylobacter spp. strain RM16704 (30) and in both C. subantarcticus strains (51 and 56). The percentages of CDSs annotated with an assigned function, general function or as hypothetical are similar between all 12 genomes (table 2).

With the exception of Campylobacter spp. strain RM16704, all of the genomes characterized here contain at least one genomic island (C. lari subsp. concheus, C. subantarcticus, C. peloridis, UPTC) or putative prophage (C. lari subsp. lari, C. insulaenigrae, C. subantarcticus, C. volucris) (supplementary table S1 , Supplementary Material online). All of the genomic islands/prophage are inserted immediately adjacent to a tRNA, with two-thirds located next to a leucyl-tRNA. Additionally, each of the C. lari group genomes (with the exception of Campylobacter spp. strain RM16704) contains a genomic insertion, ranging in size from 2 to 192 kb, at the same location, adjacent to flgL. At flgL, Campylobacter spp. strain RM16704 also contains several predicted pseudogenes and a CRISPR (clustered regularly interspaced short palindromic repeats)-Cas locus; therefore, it is possible that this strain contains a degenerate genomic insertion in this region. Although the insertion points of the integrated elements within the C. lari group are strongly conserved, each of the integrated elements is unique with regard to gene content and size. Five of the integrated elements encode putative type VI secretion systems (T6SS). Two such elements are contained within the
Table 2
General Features of the *Campylobacter lari* Group Genomes

<table>
<thead>
<tr>
<th></th>
<th><em>C. lari</em></th>
<th><em>C. lari</em></th>
<th>UPTC</th>
<th>UPTC</th>
<th>UPTC</th>
<th>C. subantarcticus</th>
<th>C. subantarcticus</th>
<th>Campylobacter</th>
<th>C. peloridis</th>
<th>C. volucris</th>
<th>C. insulaenigrae</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General features</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size (kb)</td>
<td>1,525.46</td>
<td>1,502.10</td>
<td>1,791.51</td>
<td>1,523.03</td>
<td>1,516.46</td>
<td>1,782.54</td>
<td>1,853.00</td>
<td>1,557.54</td>
<td>1,711.37</td>
<td>1,517.95</td>
<td>1,465.08</td>
</tr>
<tr>
<td>% G+C content</td>
<td>29.70</td>
<td>29.73</td>
<td>29.36</td>
<td>29.86</td>
<td>29.74</td>
<td>29.94</td>
<td>29.75</td>
<td>28.47</td>
<td>28.51</td>
<td>28.57</td>
<td>28.19</td>
</tr>
<tr>
<td>CDS numbers</td>
<td>1,495</td>
<td>1,451</td>
<td>1,702</td>
<td>1,481</td>
<td>1,475</td>
<td>1,565.01</td>
<td>1,782.54</td>
<td>1,557.54</td>
<td>1,591</td>
<td>1,478</td>
<td>1,440</td>
</tr>
<tr>
<td>Assigned function</td>
<td>836 (56)</td>
<td>821 (57)</td>
<td>873 (49)</td>
<td>835 (56)</td>
<td>822 (56)</td>
<td>884 (55)</td>
<td>840 (50)</td>
<td>832 (47)</td>
<td>817 (55)</td>
<td>827 (52)</td>
<td>830 (56)</td>
</tr>
<tr>
<td>Pseudogenes</td>
<td>18</td>
<td>20</td>
<td>21</td>
<td>9</td>
<td>20</td>
<td>15</td>
<td>20</td>
<td>30</td>
<td>19</td>
<td>19</td>
<td>22</td>
</tr>
<tr>
<td>Hypothetical</td>
<td>286 (19)</td>
<td>252 (17)</td>
<td>439 (26)</td>
<td>256 (17)</td>
<td>266 (18)</td>
<td>377 (23)</td>
<td>453 (26)</td>
<td>285 (19)</td>
<td>360 (23)</td>
<td>265 (18)</td>
<td>268 (19)</td>
</tr>
<tr>
<td>Prophage/genetic islands</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Ribosomal RNA operons</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>CRISPR</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>G+C tracts ≥8 nt (# HV)</td>
<td>15 (19)</td>
<td>15 (14)</td>
<td>23 (22)</td>
<td>10 (18)</td>
<td>14 (14)</td>
<td>18 (16)</td>
<td>50 (46)</td>
<td>44 (43)</td>
<td>41 (39)</td>
<td>26 (26)</td>
<td>17 (17)</td>
</tr>
<tr>
<td>Plasmids (size kb)</td>
<td>46.2</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>47.8; 3.6</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Gene classes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Signal transduction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Che/Mot proteins</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>MCP</td>
<td>12</td>
<td>11</td>
<td>14</td>
<td>14</td>
<td>11</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>13</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>2CS response regulator</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>2CS histidine kinase</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Other</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>R/M systems</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type I (his)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Type II/III</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Type III</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Other (mobC)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>DNA methylases</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Transcription</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regulatory proteins</td>
<td>18</td>
<td>17</td>
<td>16</td>
<td>16</td>
<td>17</td>
<td>17</td>
<td>13</td>
<td>18</td>
<td>15</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>r factors</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Hap proteins</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>10</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>Motility</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>flaAB (class; orient.)</td>
<td>1;→→↑</td>
<td>1;→→↑</td>
<td>2;→→↑</td>
<td>2;→→↑</td>
<td>2;→→↑</td>
<td>2;→→↑</td>
<td>1;→→↑</td>
<td>2;→→↑</td>
<td>2;→→↑</td>
<td>2;→→↑</td>
<td>1;→→↑</td>
</tr>
<tr>
<td>MAF (class)</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Pse/Leg</td>
<td>Y/Y</td>
<td>Y/Y</td>
<td>Y/N</td>
<td>Y/N</td>
<td>Y/N</td>
<td>Y/N</td>
<td>Y/Y</td>
<td>Y/N</td>
<td>Y/N</td>
<td>Y/Y</td>
<td>Y/Y</td>
</tr>
</tbody>
</table>

**Note:**—HV, hypervariable; Che/Mot, chemotaxis/motility; MCP, methyl-accepting chemotaxis protein; 2CS, two component system; MAF, motility accessory factor; Pse/Leg, pseudaminic acid/legionaminic acid.

*<sup>a</sup>* CDS numbers do not include pseudogenes.

<sup>b</sup>cas1 gene nonfunctional.
Genes Absent within the C. lari Group

Multiple genes encoding enzymes involved in amino acid/cofactor biosynthesis and energy metabolism/respiration were predicted to be absent in C. lari. These included genes involved in the biosynthesis of acetyl-coenzyme A (acs); arginine (argBCDFO), glutamate (gltBD); leucine (leuABCD); methionine (metABEF); pantothenate (panBCD); proline (proA); and tryptophan (trpACDEF) (Miller et al. 2008). Additionally, the respiratory enzyme-encoding genes gltA (citrate synthase), acn (aconitase), ict (isocitrate dehydrogenase), sucCD (succinyl-CoA synthetase), sdhABC (succinate dehydrogenase), and ciaAB (terminal oxidase) were not identified in strain RM2100 (Miller et al. 2008). Other genes, such as the quorum sensing-associated gene luxS and those encoding the CeuBCDE enterochelin and ChuABCD hemin ABC transporters, were also absent. With a few exceptions, these genes are absent also within the other 11 genomes analyzed in this study, making their absence a general characteristic of the C. lari group, and indicating that multiple auxotrophy is a general feature of this clade. The pantothenate biosynthetic genes were identified in both C. subantarcticus strains and argF was identified in Campylobacter spp. strain RM16704, C. lari subsp. concisus and UPTC strain NCTC 11845. As carA and carB are core to the C. lari group, those strains that encode ArgF would be able to synthesize arginine in the presence of ornithine; the remainder of the C. lari group could presumably synthesize arginine in the presence of citrulline, given the presence of argG and argH in all C. lari group strains.

Genomic Variation within the C. lari Group

Although, as described above, many genes are absent generally within the lari clade, the presence/absence of other genes is restricted to one or a few strains. cysE and cysK were not identified in C. insulaenigrae, and the thiH, thiG, thiS and thiN
**FIG. 1.**—BRIG plot of the *Campylobacter lari* group. The BRIG image was created using BLASTN and a minimum default threshold of 50%. The reference strain was *C. lari* subsp. *lari* strain RM2100. White areas correspond to sequences with similarity values below the minimum threshold.

**RINGS**

1. *C. lari* subsp. *conheus*
2. UPTC NCTC 11645
3. UPTC CCUG 22395
4. UPTC RM16701
5. UPTC RM16712
6. *C. subantarcticus* LMG 24374
7. *C. subantarcticus* LMG 24377
8. *Campylobacter* spp. RM16704
9. *C. peloridis*
10. *C. volucris*
11. *C. insulaeirae*

**GENOMIC FEATURES**

A. *add - sorAB*
B. hypothetical proteins
C. hypothetical proteins
D. capsule
E. *hsd*
F. *psbBACS-phosRS*
G. genomic island
H. efflux system
I. *napDLBHG*
J. **LOS**
K. flagellar modification / *psb* genes
L. *ldh* - TMAO reductase
M. *pppB* - alpha macroglobulin domain protein

*Fig. 1.*—BRIG plot of the *Campylobacter lari* group. The BRIG image was created using BLASTN and a minimum default threshold of 50%. The reference strain was *C. lari* subsp. *lari* strain RM2100. White areas correspond to sequences with similarity values below the minimum threshold.
genes were not identified in Campylobacter spp. strain RM16704, suggesting cysteine and thiamine auxotrophy, respectively. The NrdDG anaerobic nucleoside-triphosphate reductase activating system is not encoded by C. insulaenigrae, nor is the Mdh NAD-dependent malate dehydrogenase. The PhosSR-PstSCAB phosphate two-component system/ABC transporter was not encoded by C. peloridis or UPTC strain RM16701 (fig. 1). The napDLBGHA nitrate reductase gene cluster was not identified in strain Campylobacter spp. RM16704; additionally, the sorAB molybdenum-containing sulfite:cytochrome c oxidoreductase (Myers and Kelly 2005) was only in the C. insulaenigrae, C. lari subsp. lari, C. peloridis, C. volucris and UPTC NCTC 11845 genomes, and the lactate dehydrogenase/lactate permease gene cluster (Thomas et al. 2011) was identified only within the C. lari subsp. lari, C. insulaenigrae, C. volucris, and C. subantarcticus genomes (fig. 1). Finally, C. insulaenigrae appears to contain a gene cluster similar to the C. jejuni L-fucose utilization genomic island (cj0480c-cj0490) (Stahl et al. 2011).

Phylogenetic Placement of the UPTC Strains and Campylobacter spp. Strain RM16704

The genomes of four UPTC strains were examined to see whether the data could more accurately place the UPTC strains within the taxonomic structure of the C. lari group. The strains that we sequenced were two from culture collections and two isolated in 2013 from a watershed survey of the Salinas Valley of California. The genomes of these four strains, while not identical, were quite similar and several features observed here, for example, absence of nalidixic acid resistance, arsenate resistance (Nakajima et al. 2013), truncated flagellar subunits (Sekizuka et al. 2004, 2007), and the organization of the urease operon (Kakinuma et al. 2007), were
consistently with those characterized previously in other UPTC strains. Average AAI analysis of the C. lari group (table 3) places the UPTC strains in the same clade as the two C. lari subspecies; however, finer taxonomic distinctions could not be achieved. ANI analysis similarly indicated that the UPTC strains were most closely related to C. lari subsp. lari and C. lari subsp. concheus (supplementary table S3, Supplementary Material online). Here, however, ANI analysis also clearly distinguished the UPTC strains from both C. lari subsp. lari and C. lari subsp. concheus. Although UPTC strains have been shown to form two distinct clusters following AFLP analysis (Duim et al. 2004; Debruyne et al. 2009), in this study the UPTC genomes examined formed a single cluster in both AAI and ANI analyses. It is unknown whether the genomes of four strains within the same cluster were sequenced here or whether the two AFLP clusters are indistinguishable following genomic analysis. To provide a more definitive taxonomic placement of the UPTC strains, the genomes of additional strains will need to be sequenced. Genome sequencing of 24 additional UPTC strains, representing both AFLP clusters, is currently in progress.

Campylobacter spp. strain RM16704 can be more readily placed within the taxonomy of the C. lari group. Campylobacter spp. strain RM16704 was isolated during the 2013 watershed survey described above. AAI and ANI analyses (table 3 and supplementary table S3, Supplementary Material online) clearly place this strain as a species distinct from the other validly named taxa within this clade. These results were verified through AtpA and 16S rRNA phylogenetic analyses (data not shown). Campylobacter spp. strain RM16704 does not encode the Nap nitrate reductase, and the absence of nitrate reductase activity in this strain was confirmed through biochemical tests (data not shown). The nap phenotype is a useful defining characteristic, as the inability to reduce nitrate is unique to the C. lari group and nearly unique among Campylobacter. Cjd is the only other campylobacter unable to reduce nitrate (Steele and Owen 1988) and Campylobacter spp. strain RM16704 could be easily distinguished from Cjd through the use of a hippuricase test. Although it is likely that Campylobacter spp. strain RM16704 represents a novel taxon within Campylobacter, characterization of a single strain does not meet the current minimum standards for the definition of a new species; therefore, additional strains will need to be isolated and characterized before this taxon can be officially described.

The Flagellar and Flagellar Modification Loci

The flagella of the C. lari group clearly divide this clade into two distinct subgroups. Sekizuka et al. (2007) first identified flagellin genes in some UPTC isolates that were truncated relative to those of C. lari subsp. lari strain RM2100; furthermore, in these isolates flaA and flaB were divergently transcribed. With the exception of UPTC strain NCT 11845, whose truncated flagellar genes are transcribed in the same direction, the UPTC isolates characterized in this study all contain truncated, divergently transcribed flagellar genes (table 4), as do C. peloridis and Campylobacter spp. strain RM16704. The flagellar genes of C. subantarctica, C. volucris, C. insulaenigrae, and C. lari subsp. concheus are similar in size and orientation to those of C. lari subsp. lari strain RM2100 (table 4).

Division of the C. lari group into two subgroups is also maintained following analysis of the flagellar glycosylation loci. Campylobacter flagella are often modified through the addition of the nine carbon sugars pseudaminic acid (PseAm) and/or legionaminic acid (LegAm) and their derivatives (Thibault et al. 2001; Logan et al. 2008; Schoenhofen et al. 2009; Morrison and Imperiali 2014). The C. lari group in general is predicted to contain the complete PseAm biosynthetic pathway (pseBCHGIF; table 4). However, three strains (Campylobacter spp. strain RM16704, C. peloridis, and UPTC strain NCT 11845) do not contain pseA (table 4), indicating that these strains cannot synthesize the acetamidino derivative of PseAm (Thibault et al. 2001). Moreover, an ortholog of cj1295, necessary for the modification of PseAm with di-O-methyl glyceric acid (Hitchen et al. 2010), is found throughout the C. lari group, though in some strains this ortholog is a contingency gene, in others not, and in C. volucris the gene is present as a contingency gene in two copies. Therefore, it is likely that the C. lari group flagella are glycosylated by a wide variety of PseAm derivatives. Truncation of the UPTC flagellar subunits was predicted to remove most of the flagellar glycosylation sites (Sekizuka et al. 2007). Nevertheless, the UPTC subgroup is predicted to synthesize PseAm; thus, either glycosylation occurs at different sites in the C. lari group or the UPTC subgroup flagella maintain a lower glycosylation density. Although PseAm and its derivatives are synthesized by the C. lari group, only the C. lari subsp. lari flagellar subgroup is predicted to encode the LegAm biosynthetic pathway (table 4). In C. jejuni and Campylobacter coli, genes encoding both biosynthetic pathways are located immediately adjacent to flaAB. However, in the C. lari group, the PseAm genes are unlinked to flaAB, and the LegAm genes in the C. lari subsp. lari subgroup are located either within the LOS (C. lari subsp. lari, C. subantarctica, C. insulaenigrae, C. volucris) or capsular (CPS) (C. lari subsp. concheus) loci. The placement of the LegAm genes within these loci is interesting, suggesting perhaps that the flagellar subunits in these strains are not modified with LegAm, but rather the LOS or CPS structures. The C. lari subsp. concheus LegAm biosynthetic genes have much lower similarity to other Campylobacter LegAm genes; therefore, it is unknown whether LegAm or another nine-carbon sugar is synthesized in this strain. Also, no legC ortholog was identified in C. lari subsp. lari, C. subantarctica, C. insulaenigrae, and C. volucris; however, it is possible that the LegC function may be substituted by PglE. Finally, maf (motility accessory factor) genes in Campylobacter have been associated
with glycosylation of the flagellar subunits (Guerry et al. 2006; van Alphen et al. 2008). Consistent with this role, the maf genes also form two distinct clades within the C. lari group (table 4). Additionally, in the UPTC subgroup, the maf genes are present only downstream of flaAB in 6–7 copies, whereas in the C. lari subsp. lari subgroup, one maf gene is located upstream of flaAB and 2–4 copies are located downstream.

LOS and CPS Biosynthesis Loci

The LOS and CPS genomic loci encode enzymes involved in the biosynthesis of surface polysaccharide structures that function in the interactions of the bacteria with the environment. The LOS and CPS biosynthesis loci of the C. lari group strains are organized as previously observed in C. jejuni strains (Karlyshev et al. 2005; Parker et al. 2008). At either end of the LOS loci are the conserved heptosyltransferase genes, waaC and waaF, with additional LOS biosynthesis genes in between. The CPS loci possess the conserved CPS transporter genes (kpsMTEDF and kpsCS) flanking polysaccharide biosynthesis genes. Although the structure and content of these molecules is beyond the scope of this manuscript, these genomic loci of the C. lari group suggest diverse structures, based on variable oligo/polysaccharide biosynthesis genes. Indeed, all four of the UPTC strains and both C. subantarcticus strains possess their own unique set of LOS and CPS biosynthesis genes, suggesting interstrain variability. Interestingly, C. insulaenigrae and C. subantarcticus strain LMG 24374 possess genes (cstIIneuBCA) responsible for the production of sialylated LOS that are ganglioside mimics, and thus, certain C. lari group strains could lead to the development of Guillain–Barré and Fisher syndromes following infection.

Hag Proteins in the C. lari Group

A unique feature of the lari clade is the presence of multiple genes encoding Hag proteins that range in size from 96 to 166 kDa. These genes have been identified in other Campylobacter spp. (Fouts et al. 2005; Asakura et al. 2012); however, in many strains the Hag genes are highly fragmented and likely nonfunctional. All members of the C. lari group, with the exception of C. insulaenigrae and C. volucris, contain at least one presumably functional Hag gene. Notably, Campylobacter spp. strain RM16704 and C. peloridis strain LMG 23910 contain 10 and 15 Hag genes, respectively; analysis of an additional C. peloridis strain identified a similar number of Hag genes (data not shown), suggesting that the high number of Hag genes might be typical of this species. The Hag genes of the lari clade fall into two groups: “Stand-alone” Hag genes and those linked to genes encoding a hypothetical protein and a putative hemolysin activation/secrection protein, the latter group termed here a “Hag triad.” It is also noteworthy that most of the stand-alone Hag genes and Hag triads are associated with hypervariable G:C tracts: 15 of the 18 stand-alone Hag genes identified here contain G:C tracts and in all but one of the 27 Hag triads, the hypothetical protein-encoding gene contains a G:C tract at its 5’-end. Phylogenetic analysis identified a large amount of diversity within the C. lari group Hag proteins (supplementary fig. S2, Supplementary Material online). In comparison, the hypothetical protein and hemolysin activation/secrection protein

Table 3
Average Amino Acid Identities of the Campylobacter lari Group Core Proteome

<table>
<thead>
<tr>
<th></th>
<th>C. lari subsp. lari</th>
<th>C. lari subsp. concisus</th>
<th>UPTC CCGU 22395</th>
<th>UPTC RM16704</th>
<th>UPTC RM16712</th>
<th>UPTC NCTC 11845</th>
<th>C. subantarcticus LMG 24374</th>
<th>C. subantarcticus LMG 24377</th>
<th>Campylobacter spp. RM16704</th>
<th>C. peloridis</th>
<th>C. volucris</th>
<th>C. insulaenigrae</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. lari subsp. lari</td>
<td>100</td>
<td>96</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>92</td>
<td>92</td>
<td>92</td>
<td>90</td>
<td>89</td>
<td>85</td>
<td>83</td>
</tr>
<tr>
<td>C. lari subsp. concisus</td>
<td>96</td>
<td>100</td>
<td>97</td>
<td>97</td>
<td>96</td>
<td>93</td>
<td>93</td>
<td>93</td>
<td>91</td>
<td>89</td>
<td>85</td>
<td>83</td>
</tr>
<tr>
<td>UPTC CCGU 22395</td>
<td>95</td>
<td>97</td>
<td>100</td>
<td>99</td>
<td>98</td>
<td>97</td>
<td>94</td>
<td>94</td>
<td>91</td>
<td>90</td>
<td>85</td>
<td>83</td>
</tr>
<tr>
<td>UPTC RM16704</td>
<td>95</td>
<td>97</td>
<td>99</td>
<td>100</td>
<td>98</td>
<td>97</td>
<td>94</td>
<td>94</td>
<td>91</td>
<td>90</td>
<td>85</td>
<td>83</td>
</tr>
<tr>
<td>UPTC RM16712</td>
<td>95</td>
<td>96</td>
<td>98</td>
<td>98</td>
<td>100</td>
<td>97</td>
<td>93</td>
<td>93</td>
<td>91</td>
<td>89</td>
<td>85</td>
<td>83</td>
</tr>
<tr>
<td>UPTC NCTC 11845</td>
<td>94</td>
<td>96</td>
<td>97</td>
<td>97</td>
<td>100</td>
<td>93</td>
<td>93</td>
<td>93</td>
<td>91</td>
<td>89</td>
<td>85</td>
<td>83</td>
</tr>
<tr>
<td>C. subantarcticus LMG 24374</td>
<td>92</td>
<td>93</td>
<td>94</td>
<td>94</td>
<td>93</td>
<td>100</td>
<td>99</td>
<td>90</td>
<td>88</td>
<td>84</td>
<td>82</td>
<td>82</td>
</tr>
<tr>
<td>C. subantarcticus LMG 24377</td>
<td>92</td>
<td>93</td>
<td>94</td>
<td>94</td>
<td>93</td>
<td>99</td>
<td>100</td>
<td>90</td>
<td>88</td>
<td>84</td>
<td>82</td>
<td>82</td>
</tr>
<tr>
<td>Campylobacter spp. RM16704</td>
<td>90</td>
<td>91</td>
<td>91</td>
<td>91</td>
<td>91</td>
<td>89</td>
<td>89</td>
<td>89</td>
<td>100</td>
<td>88</td>
<td>84</td>
<td>84</td>
</tr>
<tr>
<td>C. peloridis</td>
<td>89</td>
<td>89</td>
<td>90</td>
<td>89</td>
<td>89</td>
<td>88</td>
<td>88</td>
<td>88</td>
<td>100</td>
<td>84</td>
<td>82</td>
<td>83</td>
</tr>
<tr>
<td>C. volucris</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>84</td>
<td>84</td>
<td>84</td>
<td>100</td>
<td>84</td>
<td>82</td>
<td>85</td>
</tr>
<tr>
<td>C. insulaenigrae</td>
<td>83</td>
<td>83</td>
<td>83</td>
<td>83</td>
<td>83</td>
<td>83</td>
<td>83</td>
<td>83</td>
<td>82</td>
<td>83</td>
<td>85</td>
<td>100</td>
</tr>
</tbody>
</table>

Note.—To more easily visualize amino acid similarities of the proteomes, we utilized a gradient heat map with black = 100%.
components of the Hag triads were nearly identical within any given strain (supplementary fig. S3, Supplementary Material online, and data not shown), suggesting that intrastrain Hag diversity is not due to acquisition by lateral transfer with subsequent homologous recombination at the flanking loci. The role of the Hag proteins, or for that matter the role of either of the Hag-associated proteins within the Hag triad, in C. lari group biology is unknown. Nevertheless, the Hag genetic diversity and the association of these genes, alone or within the context of the Hag triad, with hypervariable G:C tracts indicate diversity and the association of these genes, alone or within the context of the Hag triad, in C. lari group biology is warranted.

Methylyme Diversity in the C. lari Group

The methylomes of the five strains sequenced on the PacBio RS sequencer (C. peloridis, UPTC strain NCTC 11845, Campylobacter spp. strain RM16704, and both C. subantarcticus strains) reveal a diverse range of N6-methyladenine (m6A) methyltransferase (Mtase) activities within the C. lari group. As shown in supplementary table S4, Supplementary Material online, of the 24 different detected sequence motifs targeted by m6A Mtases, only one (5′-RA\textsuperscript{m6}ATT-3′, underlined base indicates methylation on the opposite DNA strand) is shared across all five strains and another (5′-G\textsuperscript{m6}ATC-3′) is shared across four strains but not detected in the Campylobacter spp. strain RM16704 genome. Campylobacter lari strain RM2100 is predicted to have several restriction-modification (R-M) systems but only one has a putative Type II N4-methylcytosine or m6A Mtase with a predicted recognition sequence of 5′-GAATT-3′ (http://tools.neb.com/~vinzke/genomes/). The absence of detection of N4-methylcytosine, which is typically easily detected with SMRT sequencing, in this study suggests that this putative Mtase more likely targets adenine and may have the slightly less restrictive recognition motif of 5′-RA\textsuperscript{m6}ATT-3′. Methylation of the 5′-RA\textsuperscript{m6}ATT-3′ motif is likely related to the presence of cj0208 orthologs within the five strains characterized here. The Cj0208 DNA methyltransferase is a core protein within the C. lari group and is encoded by several other Campylobacter species (Fouts et al. 2005). Noteworthy in the C. insulaenigrae and C. volucris genomes is the presence of the cognate EcoRI family endonuclease gene adjacent to the cj0208 orthologs. Methylation of the 5′-G\textsuperscript{m6}ATC-3′ motif is consistent with the presence of a DpnII family R-M system in the C. peloridis genome and in both C. subantarcticus genomes; the UPTC strain NCTC 11845 genome contains the DpnII family methyltransferase but not the cognate endonuclease. Although a thorough investigation into all of the R-M systems observed and their potential biological roles is beyond the scope of this study, this work provides the first comprehensive list of different sequence motifs targeted by Mtases in members of the C. lari group.

Conclusions

The C. lari group is a highly related phylogenetic clade within the genus Campylobacter, whose members often share similar hosts (e.g., shore birds, marine mammals, and shellfish) and environments (i.e., coastal regions and watersheds). Consistent with this association, genomic analyses of C. lari group members indicate that these strains are very similar in terms of gene content and organization. Seventy-seven percent of the genes identified in the previously sequenced C. lari subsp. strain RM2100 genome are conserved throughout the C. lari group. Furthermore, the genomic topography of this group is also quite conserved, with minor rearrangements...
identified within the *C. insulaenigrae*, *C. peloridis*, *C. volucris*, and *Campylobacter* spp. strain RM16704 genomes. Additionally, many features, such as profound defects in the amino acid biosynthetic and respiratory machinery, first identified in strain *C. lari* subsp. *lari* RM2100, are also conserved within the *C. lari* group as a whole. Many of the taxa characterized in this study are represented by a single strain, and it should be noted that variation identified in a single strain does not imply that the same variation exists in the species as a whole. Nevertheless, genes conserved or absent across the entire *C. lari* group should be taken seriously as candidates for future research into the host- or environmental-association of *C. lari* group strains.

The goal of this study was to identify genes related to adaptation to the coastal or watershed environment, and to provide additional data to determine whether UPTC strains represent a possible third subspecies within *C. lari*. Although putative halotolerance genes were identified in *C. lari* subsp. *lari* strain RM2100 in previous work, leading to the hypothesis that their presence might be critical to the colonization of shorebirds and shellfish, these genes were not consistently found in isolates characterized in this study. In fact, no obvious genes or pathways were identified within the *C. lari* group that could be implicated in the host- or environmental-association of members of the clade. The present data suggest the possibility that it is the absence of amino acid biosynthetic pathways in some strains that may be more important in determining host/environment range, rather than the presence of pathways for increased halotolerance. For example, defects in biosynthetic pathways might restrict colonization of these organisms to hosts that can supply the missing metabolites. With respect to the UPTC strains, the present data are consistent with the hypothesis that they represent a third subspecies within *C. lari*, but are not sufficient to prove the hypothesis. The 11 high-quality closed and completed genomes reported here form a critical basis for expanded studies to further illuminate the complex basis of *Campylobacter* speciation and ecology, especially for the *C. lari* group.

**Supplementary Material**

Supplementary figures S1–S3 and tables S1–S4 are available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals.org/).

**Acknowledgments**

This work was funded by the United States Department of Agriculture, Agricultural Research Service, CRIS project 5325-42000-230-047. The authors thank Birgitta Duim for critical reading of this manuscript.

**Literature Cited**


Associate editor: Rotem Sorek
Figure S1. Phylogenetic relationships among *Campylobacter* spp. The amino acid sequences of 20 core proteins (AtpA, GlnA, GlyA, Tkt, Eno, IleS, NrdB, OorA, SpoT, RecJ, AroD, Hom, LysA, GppA, HisS, OxaA, Tsf, Dxr, PrfB and PurA), extracted from completed or draft *Campylobacter* genomes, were concatenated and aligned. Bootstrap values (>70%) are depicted at the nodes. *C. lari* group taxa are highlighted in red. The core protein sets of *Sulfurospirillum deleyianum*, *Arcobacter butzleri* and *Helicobacter pylori* were included for comparison.
Figure S2. Phylogenetic tree of the hemagglutination domain proteins. Bootstrap values (>70%) are depicted at the nodes. Shaded locus tags represent stand-alone hemagglutination domain proteins. Putative contingency genes are labeled with a (C).
Figure S3. Phylogenetic tree of the hemagglutination domain-associated proteins. Bootstrap values (>70%) are depicted at the nodes. Putative contingency genes are labeled with a (C).
Table S1: Genomic islands/Integrated elements within the *C. lari* group

<table>
<thead>
<tr>
<th>#</th>
<th>Strain</th>
<th>Genomic position</th>
<th>Description</th>
<th>Size (kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><em>C. lari</em> subsp. <em>lari</em></td>
<td><em>flgL</em> - tRNA-Leu</td>
<td>prophage</td>
<td>61</td>
</tr>
<tr>
<td>2</td>
<td><em>C. lari</em> subsp. <em>concheus</em></td>
<td><em>flgL</em> - tRNA-Leu</td>
<td>hypothetical proteins</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td><em>C. lari</em> subsp. <em>concheus</em></td>
<td><em>rarA</em> - tRNA-Val</td>
<td>restriction/modification</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td><em>C. subantarcticus</em> LMG 24374</td>
<td><em>flgL</em> - tRNA-Leu</td>
<td>integrated element (T6SS)::μ</td>
<td>169</td>
</tr>
<tr>
<td>5</td>
<td><em>C. subantarcticus</em> LMG 24374</td>
<td>rRNA - tRNA-Leu</td>
<td>integrated element</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td><em>C. subantarcticus</em> LMG 24377</td>
<td><em>rodA</em> - tRNA-Met</td>
<td>integrated plasmid</td>
<td>81</td>
</tr>
<tr>
<td>7</td>
<td><em>C. subantarcticus</em> LMG 24377</td>
<td><em>flgL</em> - tRNA-Leu</td>
<td>integrated element (T6SS)::μ</td>
<td>192</td>
</tr>
<tr>
<td>8</td>
<td><em>C. subantarcticus</em> LMG 24377</td>
<td>rRNA - tRNA-Leu</td>
<td>integrated element</td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td><em>C. insulaenigrae</em></td>
<td><em>flgL</em> - tRNA-Leu</td>
<td>prophage</td>
<td>43</td>
</tr>
<tr>
<td>10</td>
<td><em>C. peloridis</em></td>
<td><em>rodA</em> - tRNA-Met</td>
<td>integrated element</td>
<td>62</td>
</tr>
<tr>
<td>11</td>
<td><em>C. peloridis</em></td>
<td><em>flgL</em> - tRNA-Leu</td>
<td>integrated element (T6SS)</td>
<td>115</td>
</tr>
<tr>
<td>12</td>
<td><em>C. volucris</em></td>
<td><em>flgL</em> - tRNA-Leu</td>
<td>prophage</td>
<td>43</td>
</tr>
<tr>
<td>13</td>
<td>UPTC NCTC 11845</td>
<td><em>flgS</em> - tRNA-Arg</td>
<td>integrated element</td>
<td>40</td>
</tr>
<tr>
<td>14</td>
<td>UPTC NCTC 11845</td>
<td><em>flgL</em> - tRNA-Leu</td>
<td>integrated element (T6SS)</td>
<td>187</td>
</tr>
<tr>
<td>15</td>
<td>UPTC CCUG 22395</td>
<td><em>flgL</em> - tRNA-Leu</td>
<td>integrated element</td>
<td>26</td>
</tr>
<tr>
<td>16</td>
<td>UPTC CCUG 22395</td>
<td><em>rarA</em> - tRNA-Val</td>
<td>restriction/modification</td>
<td>6</td>
</tr>
<tr>
<td>17</td>
<td>UPTC RM16701</td>
<td><em>flgL</em> - tRNA-Leu</td>
<td>integrated element</td>
<td>31</td>
</tr>
<tr>
<td>18</td>
<td>UPTC RM16712</td>
<td><em>flgL</em> - tRNA-Leu</td>
<td>integrated element (T6SS)</td>
<td>56</td>
</tr>
<tr>
<td>19</td>
<td>UPTC RM16712</td>
<td><em>rarA</em> - tRNA-Val</td>
<td>restriction/modification</td>
<td>11</td>
</tr>
</tbody>
</table>

Tables S2 & S4 are MS Excel files attached as supplementary or Additional Files
Table S3. Average nucleotide identity (ANI %) values of the *C. lari* group

<table>
<thead>
<tr>
<th></th>
<th><em>C. lari</em> subsp. <em>lari</em></th>
<th><em>C. lari</em> subsp. <em>concheus</em></th>
<th>UPTC CCUG 22395</th>
<th>UPTC RM16701</th>
<th>UPTC RM16712</th>
<th>UPTC NCTC 11845</th>
<th><em>C. subantarcticus</em> LMG 24374</th>
<th><em>C. subantarcticus</em> LMG 24377</th>
<th>Campylobacter spp. RM16704</th>
<th><em>C. peloridis</em></th>
<th><em>C. volucris</em></th>
<th><em>C. insulaenigrae</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>C. lari</em> subsp. <em>lari</em></td>
<td>---</td>
<td>93</td>
<td>94</td>
<td>92</td>
<td>92</td>
<td>92</td>
<td>89</td>
<td>89</td>
<td>---</td>
<td>89</td>
<td>86</td>
<td>83</td>
</tr>
<tr>
<td><em>C. lari</em> subsp. <em>concheus</em></td>
<td>93</td>
<td>---</td>
<td>94</td>
<td>94</td>
<td>94</td>
<td>93</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>88</td>
<td>86</td>
<td>82</td>
</tr>
<tr>
<td>UPTC CCUG 22395</td>
<td>92</td>
<td>94</td>
<td>---</td>
<td>98</td>
<td>97</td>
<td>95</td>
<td>90</td>
<td>90</td>
<td>98</td>
<td>86</td>
<td>86</td>
<td>82</td>
</tr>
<tr>
<td>UPTC RM16701</td>
<td>92</td>
<td>94</td>
<td>98</td>
<td>---</td>
<td>97</td>
<td>95</td>
<td>90</td>
<td>90</td>
<td>98</td>
<td>86</td>
<td>86</td>
<td>82</td>
</tr>
<tr>
<td>UPTC RM16712</td>
<td>92</td>
<td>94</td>
<td>97</td>
<td>97</td>
<td>---</td>
<td>95</td>
<td>90</td>
<td>90</td>
<td>98</td>
<td>86</td>
<td>86</td>
<td>82</td>
</tr>
<tr>
<td>UPTC NCTC 11845</td>
<td>91</td>
<td>93</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>---</td>
<td>90</td>
<td>90</td>
<td>98</td>
<td>86</td>
<td>86</td>
<td>82</td>
</tr>
<tr>
<td><em>C. subantarcticus</em> LMG 24374</td>
<td>89</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>98</td>
<td>---</td>
<td>86</td>
<td>85</td>
<td>82</td>
<td>80</td>
</tr>
<tr>
<td><em>C. subantarcticus</em> LMG 24377</td>
<td>89</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>98</td>
<td>---</td>
<td>86</td>
<td>85</td>
<td>82</td>
<td>80</td>
</tr>
<tr>
<td>Campylobacter spp. RM16704</td>
<td>87</td>
<td>88</td>
<td>88</td>
<td>88</td>
<td>88</td>
<td>88</td>
<td>86</td>
<td>86</td>
<td>---</td>
<td>85</td>
<td>82</td>
<td>81</td>
</tr>
<tr>
<td><em>C. peloridis</em></td>
<td>86</td>
<td>86</td>
<td>86</td>
<td>86</td>
<td>87</td>
<td>86</td>
<td>86</td>
<td>86</td>
<td>85</td>
<td>83</td>
<td>81</td>
<td>---</td>
</tr>
<tr>
<td><em>C. volucris</em></td>
<td>83</td>
<td>82</td>
<td>82</td>
<td>82</td>
<td>82</td>
<td>82</td>
<td>82</td>
<td>82</td>
<td>82</td>
<td>---</td>
<td>83</td>
<td>---</td>
</tr>
<tr>
<td><em>C. insulaenigrae</em></td>
<td>81</td>
<td>81</td>
<td>81</td>
<td>81</td>
<td>81</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>81</td>
<td>81</td>
<td>83</td>
<td>---</td>
</tr>
</tbody>
</table>