

The Complete Genome of the Human Clinical Isolate *Campylobacter upsaliensis* RM3195.

Miller, W. G., Yee, E. and Parker, C. T.

Agricultural Research Service, U.S. Department of Agriculture, Albany, CA, USA.

ABSTRACT

Introduction: *Campylobacter upsaliensis* is a catalase negative/weakly catalase positive thermotolerant campylobacter that was isolated originally from dogs. *Campylobacter upsaliensis* is also a human pathogen that is isolated, albeit infrequently, from human diarrheal stool samples. In this study we present the complete genome of the human clinical isolate *C. upsaliensis* strain RM3195.

Methods: The genome of strain RM3195 was sequenced previously to draft level; however, numerous repeat regions prevented total closure. A combination of large-fragment library shotgun sequencing and combinatorial long PCR/sequencing was used to span these repeat regions and close the genome. In addition, to complete the genome, low coverage regions were amplified and sequenced and hypervariable GC tracts were cloned and sequenced. The RM3195 genome was re-annotated, based on the closed and completed genome sequence.

Results: The genome of strain RM3195 is approx. 1.68 mb with a G+C content of 34.6 mol%. The genome of strain RM3195 is predicted to encode 1593 coding sequences and at least 34 pseudogenes. *Campylobacter upsaliensis* strain RM3195 also possesses two plasmids: a 110 kb megaplasmid and a 3 kb cryptic plasmid. The *C. upsaliensis* RM3195 genome is distinguished by the presence of a large suite of methylases and restriction enzymes and by the presence of 75 homopolymeric tracts greater than seven bases in length; at least 40 of these 75 tracts were demonstrated to be hypervariable. One of these hypervariable GC tracts is located within *licA*, the first gene in an operon involved in both phosphorylcholine biosynthesis and decoration of outer surface structures with phosphorylcholine moieties. The *licABCD* operon is present only within a phylogenetically-distinct clade of *C. upsaliensis*.

Conclusions: The completion of the *C. upsaliensis* genome will provide further insights into both *Campylobacter* biology and pathogenicity. Additionally, this genome will prove useful in enhancing *C. upsaliensis* culturing, detection and strain identification methods.

Campylobacter upsaliensis strain RM3195 was isolated in Sept. 1994 from a 4 year-old boy with clinically-confirmed GBS. The patient had loose stools and was infected also with *Ascaris* and *Tricuris*. The genome of strain RM3195 was sequenced to completion. >22,000 reads were assembled, to a final coverage of 6.6x. Repeat regions and hypervariable tracts were amplified independently, sequenced and inserted manually into the assembly. The final assembly was verified by comparing an in silico *MluI* map of the completed sequence to an experimental *MluI* bacterial optical map (generated by OpGen - see Figure 1 below).

Figure 1. Comparison of RM3195 in silico *MluI* restriction map with RM3195 bacterial optical *MluI* restriction map. Note: optical restriction map technology cannot resolve fragments smaller than 2 kb.

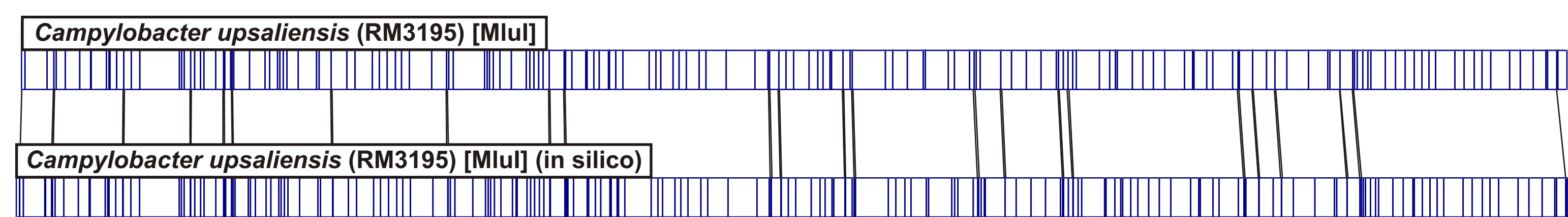


Table 1. General features of the *C. upsaliensis* strain RM3195 genome

Genome size (bp)	1,678,691
% G+C	34.6
CDS numbers	
Assigned	784
General / unknown function	784
Pseudo/contingency genes	92
Plasmids	
pCU3195-1; 3 kb	
pCU3195-2; 110 kb	
Prophage / genomic islands	0/1
IS elements / CRISPRs	0/0
Mono-/di-nucleotide tracts	see Table 2
Chemotaxis proteins	
Che/Mot proteins	8
MCP domain proteins	11
Two-component systems	
Response regulator	8
Sensor histidine kinase	4
Transcriptional regulators	
Regulatory proteins	10
Sigma factors	FliA, RpoN, RpoD
Proteases/peptidases	32
R/M systems	see Table 3

In general, the predicted gene content of the *C. upsaliensis* strain RM3195 genome is similar to the predicted gene content of other *Campylobacter* genomes. However, strain RM3195 does contain some noteworthy features related to both gene content and genome organization.

- Similar to *C. lari* strain RM2100, *C. upsaliensis* strain RM3195 is multiply auxotrophic. *Campylobacter upsaliensis* strain RM3195 does not encode the genes necessary to synthesize proline, the sulfur-containing amino acids (methionine and cysteine) and the branched-chain amino acids (isoleucine, leucine and valine). Like *C. lari*, the genome of strain RM3195 contains a large number of predicted proteases/peptidases. It is possible that this strain obtains the amino acids listed above via protein degradation.
- Campylobacter upsaliensis* strain RM3195 contains no discrete LOS or capsular loci. Instead, genes involved in biosynthesis of these structures are spread in small clusters over the entire genome. Strain RM3195 contains 11 predominantly-hypervariable glycosyltransferase genes that are unlinked to other surface-structure genes; it is unclear if these genes are involved in LOS or capsular biosynthesis.
- Campylobacter upsaliensis* strain RM3195 also contains five motility accessory factor genes unlinked to other flagellar genes.
- Although strain RM3195 contains three copies of each of the ribosomal RNA genes, two of the 16S genes in strain RM3195 are unlinked to 23S/5S genes. In the third ribosomal RNA locus, the *metK* gene is present between the 16S and 23S genes.
- The *C. upsaliensis* strain RM3195 genome encodes a number of proteins with high similarity (>95%) to those predicted to be encoded by the genome of *Helicobacter cinaedi*.

Table 2. Location of all mononucleotide GC tracts (≥ 8 nt) and all hypervariable tracts in *C. upsaliensis* strain RM3195

Gene category	Genes	All tracts	Hypervariable tracts
Surface structures (LOS/capsule)	25	26	19
Motility	7	7	3
DNA restriction/modification	6	6	2
General function	14	16	9
Conserved hypothetical proteins	19	20	16
Intergenic	N/A	6	6
Totals	71	81	55

- For comparison, *C. jejuni* strain NCTC 11168 contains 29 homopolymeric tracts (23 hypervariable) and *C. lari* strain RM2100 contains 15 tracts (5 hypervariable).
- With two exceptions, putative contingency genes in strain RM3195 contain hypervariable GC tracts. In contrast, the hypervariable intergenic tracts are all AT.
- One glycosyltransferase gene contains a hypervariable dinucleotide (AG) tract.
- The average tract length in *C. upsaliensis* strain RM3195 is 13.6 nt (average tract lengths in other campylobacters are 9-10 nt). Tracts in RM3195 are up to 19 nt in length.

Table 3. Restriction/modification loci in *C. upsaliensis* strain RM3195

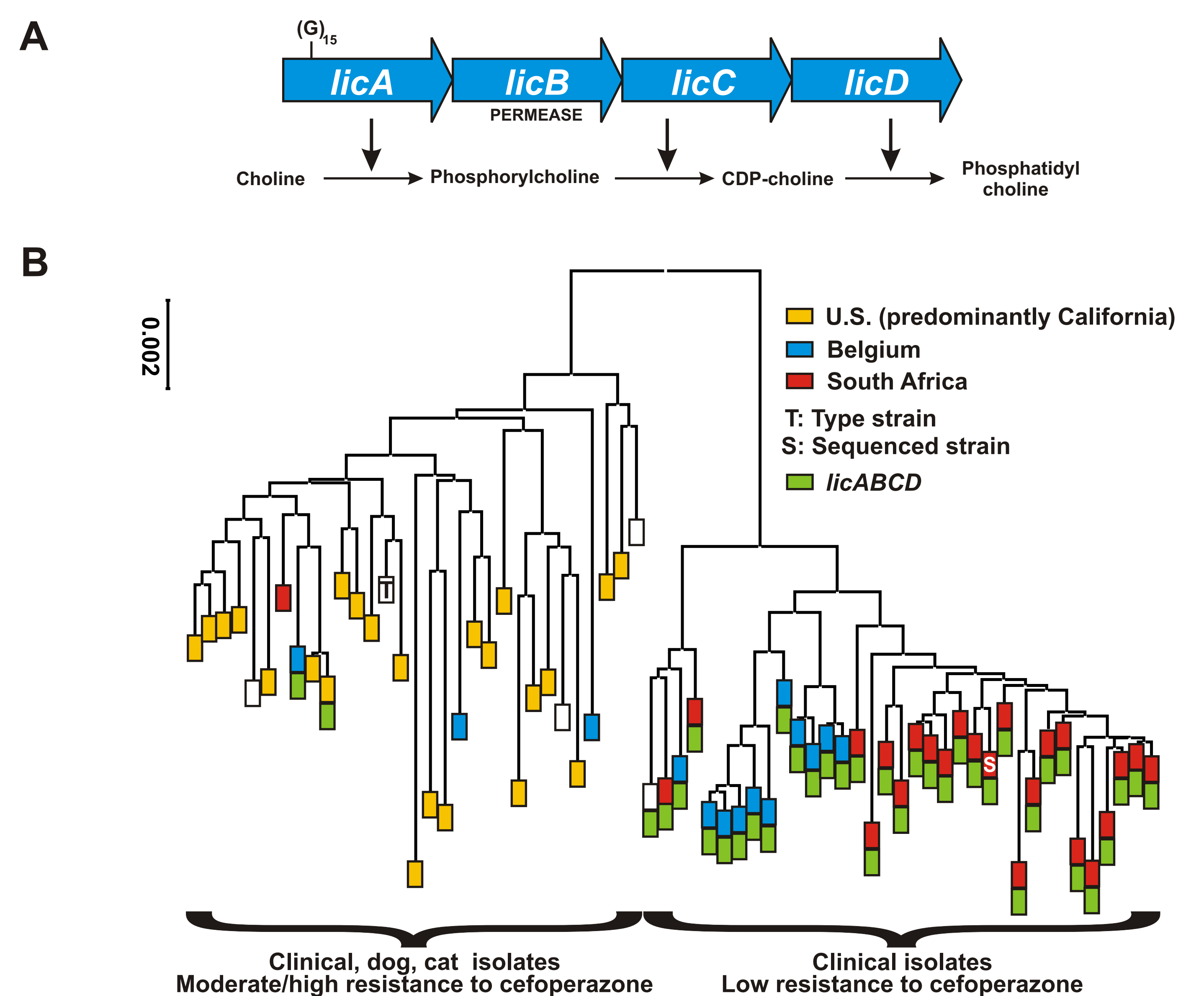
Type	Subtypes	Number
Type I (<i>hdsR/hdsI/hdsM</i>)	Type IB, Type IC	3
Type II (<i>res/mod</i>)	Type IIP, Type IIS	6
Type III (<i>res/mod</i>)		2
Individual adenine/cytosine-specific DNA methylases (<i>mod</i>)		21

- The genomic DNA of *C. upsaliensis* RM3195 (and other related *upsaliensis* strains) cannot be digested by several commonly-used PFGE enzymes. The large suite of DNA modification methylases predicted to be encoded by this strain is likely the cause.
- For comparison, the other *Campylobacter* genomes contain 2-5 Type I, II and III loci in total.

Figure 2. The *licABCD* operon of *C. upsaliensis* strain RM3195.

A. Operon organization.

B. Distribution of *licABCD* among US, European and South African *C. upsaliensis* strains.



- In *Haemophilus influenzae*, incorporation of phosphorylcholine (P-Cho) into the LPS is encoded by the *lic* operon.
- The *H. influenzae licA* gene contains CAAT tandem repeats; LPS phase variation is caused by changes in the copy number of these repeat units. The *licA* gene of RM3195 contains a hypervariable GC tract. Antibodies against P-Cho also demonstrate phase-variable addition of P-Cho in *C. upsaliensis*.
- In *H. influenzae*, *Streptococcus pneumoniae*, *Pseudomonas aeruginosa* and *Neisseria gonorrhoeae*, cell surface P-Cho plays a role in pathogenicity.
- In *C. upsaliensis*, *licABCD* is primarily present in a phylogenetically-distinct group of human clinical isolates.