Check for updates

DATA ARTICLE

Molecular data analysis of selected housekeeping and informational genes from nineteen *Campylobacter jejuni* genomes [version 1; peer review: 2 not approved]

Vathsala Mohan¹, Mark Stevenson²

¹AgResearch Ltd, Grasslands, Palmerston North, New Zealand

²Epicentre, Institute of Veterinary and Biomedical Sciences, Massey University, Palmerston North, New Zealand

v1 First published: 14 Mar 2013, 2:87
<https://doi.org/10.12688/f1000research.2-87.v1>

Latest published: 14 Mar 2013, 2:87
<https://doi.org/10.12688/f1000research.2-87.v1>

Abstract

Campylobacter jejuni (*C. jejuni*) is a rapidly evolving bacterial species with massive genetic recombination potential to generate niche specific genotypes. Generally the housekeeping gene lineage has been evidenced to undergo lateral gene transfer and recombination fairly frequently compared to the information processing gene lineage. During such exchanges, gene amelioration takes place over time acquiring the host genomes' molecular characteristics. In this study fifty genes that comprised twenty five housekeeping lineage genes and twenty five information processing lineage genes from nineteen *C. jejuni* genomes were studied. These nineteen genomes included seven *C. jejuni* isolates that belonged to the same genotype or multilocus sequence type ST-474. The genes from both lineages were tested for recombination and the guanine-cytosine (GC) variation. This paper details about the data collected and the analyses performed in the corresponding research article entitled "

***Campylobacter jejuni* genomes exhibit notable GC variation within housekeeping genes**". Further, this paper provides details on the results that are not included in the research paper to provide completeness to the study conducted. The gene sequences from the seven *C. jejuni* ST-474 isolates were submitted to the GenBank and the corresponding gene IDs are provided for referencing purposes.

Open Peer Review

Approval Status **XX**

	1	2
version 1	X	X
14 Mar 2013	view	view

1. **David Ussery**, The Technical University of Denmark, Lyngby, Denmark
2. **Daniel Falush**, Max Planck Institute, Leipzig, Germany

Any reports and responses or comments on the article can be found at the end of the article.

Associated Research Article

Mohan V and Stevenson M » *Campylobacter jejuni* genomes exhibit notable GC variation within housekeeping genes, F1000Research 2013, 2:89 (<https://doi.org/10.12688/f1000research.2-89.v1>)

Corresponding author: Vathsala Mohan (Vathsala.Mohan@agresearch.co.nz)

Competing interests: No competing interests were disclosed.

Grant information: This work was part of Vathsala Mohan's PhD work and this project was supported by: 1. Building Research Capability in Strategically Relevant Areas (BRCSRA) fund; 2. Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North (RM12767); and 3. The Royal Society of New Zealand Marsden Fund (MAU0802). This was an original research work and the funders had no role in any of the processes or in the preparation of the manuscript.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2013 Mohan V and Stevenson M. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

How to cite this article: Mohan V and Stevenson M. **Molecular data analysis of selected housekeeping and informational genes from nineteen *Campylobacter jejuni* genomes [version 1; peer review: 2 not approved]** F1000Research 2013, 2:87 <https://doi.org/10.12688/f1000research.2-87.v1>

First published: 14 Mar 2013, 2:87 <https://doi.org/10.12688/f1000research.2-87.v1>

Background

Multilocus sequence typing (MLST) is a technique devised to sub-divide bacterial populations¹. MLST exploits the internal fragments that are approximately 470 to 500 base pairs of the housekeeping genes and constructs allelic profiles based on variations in the nucleotide sequences of (usually) seven housekeeping genes. Bacterial isolates are grouped into different clusters². *C. jejuni* is a zoonotic pathogen that causes gastroenteritis in humans worldwide and the natural competency and plasticity of *C. jejuni* have not been investigated outside the MLST housekeeping genes³⁻⁸. This technique has subsequently been exploited to structure and investigate the association of *C. jejuni* populations where, previous researches inferred that a mutation at any site can occur somewhere in the population within the space of a week⁷⁻⁹. Hybrid alleles were shown to be more common in some of the MLST loci (*tkt*, *aspA*, *gltA* and *glyA*) and in the MOMP gene of *C. jejuni* (interchangeably designated as *porA*) as a result of recombination and, has been interpreted as a consequential measure of convergence of *C. jejuni* and *C. coli*^{7,8,10}.

Housekeeping alleles of *C. jejuni* strains (genotypes) share 86% nucleotide sequence identity with each other however, the housekeeping alleles cover lesser than < 0.2% of the entire genome¹¹. The molecular differences at the genome level and/or at the full gene level may not be detected even when MLST is combined with other antigenic genes such as *porA*. Moreover, the housekeeping alleles or genes are involved in specific metabolic function which may not reflect the evolutionary events in other genes involved in a different function in a genome. Further, there is no evidence to suggest that MLST alleles are identical at their full length level and it is not known if a detailed knowledge of MLST alleles can capture the true evolutionary history of the other genes in a genome.

Objectives

Based on these background information and knowledge gap, the nucleotide sequence data from two different lineages of genes, the metabolic housekeeping and the informational genes were collected from 19 *C. jejuni* genomes. The main objectives were

- (1) To select genes those are categorised as housekeeping genes.
- (2) To ensure the presence of these genes in all of the *C. jejuni* genomes compared without segmentation.
- (3) To categorise them according to the functions they are involved in.
- (4) To create two different nucleotide datasets for comparison and,
- (5) To test different genetic hypotheses.

Materials and methods

Selection of metabolic and informational housekeeping genes

Although genes have been classified as housekeeping and informational genes, the list of genes that can be categorised as housekeeping genes have not been defined very clearly. Hence in this study we collected the set of genes that have been used in routine multilocus sequence typing (MLST) schemes for different bacterial species from the PubMLST database and from previous phylogenetic studies that used housekeeping genes for analyses¹²⁻¹⁴. The gene names were matched with the *C. jejuni* NCTC 11168 genome (GenBank accession number NC002163) that is fully annotated and has been used as reference in many other genome comparative studies¹⁵⁻²⁰. Following the identification of genes, a total of 50 genes were selected and these were further examined for their presence in other *C. jejuni* reference genomes to be analysed in this study. Full length nucleotide sequences of the selected set of housekeeping genes were retrieved in FASTA format from the GenBank from all of the selected *C. jejuni* genomes. Further the genes were classified into operational genes (metabolic housekeeping genes) and informational genes based on their function by referring to the KEGG pathway and the gene function websites (http://www.genome.jp/dbget-bin/www_bget?ko+K03495 and http://www.igs.cnrs-mrs.fr/mgdb/cgi/www_gene_catalog?rpr.ann.) The positions of the categorised selected subset of genes (metabolic housekeeping and informational genes) are marked on the reference *C. jejuni* NCTC 11168 (GenBank accession number NC002163) circular genome and are shown in Figures 1A and 1B, respectively. Here after in this paper the

Figure 1A Represents the housekeeping genes (n = 25) on the circular NCTC 11168 *C. jejuni* genome and the MLST housekeeping genes used in typing of *C. jejuni* are in red fonts.

Figure 1B Represents the informational genes (n = 25) located on the circular NCTC 11168 *C. jejuni* genome.

genes are referred to as housekeeping genes (metabolic housekeeping genes) and informational genes for the purpose of plain comparison and interpretation.

Reference *C. jejuni* genomes and nucleotide sequences from MLST ST-474 isolates

Details of the reference *C. jejuni* genomes, housekeeping and informational genes, their names and their functions are provided in Table 1, Table 2 and Table 3, respectively. Twelve fully sequenced *C. jejuni* genomes were used to compare 50 selected genes. The gene sequences were downloaded from the GenBank database (<http://www.ncbi.nlm.nih.gov/genbank/>). Seven *C. jejuni* MLST ST-474 isolates were sequenced at the Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand. The gene predictions were kindly provided by Dr. Biggs Patrick from Massey Genome Service, Massey University, Palmerston North for the seven *C. jejuni* MLST ST-474 isolates. These genomes were studied in partial fulfilment of the PhD project by the first author of this paper. The nucleotide sequences for the selected genes (n = 50) were retrieved from the gene predictions from the seven *C. jejuni* ST-474 genomes in FASTA format and gene sequences for five of the four ST-474 isolates P179a, H704a, H73020, P569a and P694a were submitted to the GenBank

and their sequence ids are given in Tables 3 and 5 for metabolic and informational or repair genes, respectively. In addition this data is accessible below:

Analysis of Guanine-Cytosine content and recombination

The length of the house keeping and the informational genes and sites of recombination can be found below:

Gene lengths and sites of recombination in metabolic housekeeping and informational genes

4 Data Files

<http://dx.doi.org/10.6084/m9.figshare.96318>

The results of the guanine-cytosine (GC) analysis with the GC range for all of the genes from 19 *C. jejuni* genomes can be found below:

Guanine-cytosine (GC) and GC3 contents of the individual metabolic housekeeping and informational genes

4 Data Files

<http://dx.doi.org/10.6084/m9.figshare.96319>

The overall GC and GC3 contents of individual genes were compared using DnaSP v5²¹. The frequency distribution graphs of GC contents from all the 19 genomes compared were generated in R programming language²². Inferences on recombination within each gene under investigation were drawn using Dual-Brothers within Geneious v5.3.4²³. This function uses a double change point model that detects spatial variation in the phylogenetic tree topology and spatial variation of the nucleotide substitution process²³. The gene sequences were aligned using Geneious v5.3.4 and the recombination detection functionality was applied on each gene alignment to infer changes in the topologies and substitution processes. In addition, the aligned sequences were examined using DnaSP v5 to identify the sites involved in recombination. The number of recombination events, referred to as Rm was estimated using DnaSP v5. The GC variance for the genes were estimated using the variance function in Microsoft Office Excel and the relationship between the GC variance and the recombination events was analysed using linear models by having Rm as a dependent variable, and the log GC variance and the length as independent variables.

Figures 2–9 describe the guanine-cytosine (GC) variation in the housekeeping and the informational genes. The plots describe the frequency distribution of GC content across 19 *C. jejuni* genomes and the number inside the parenthesis denotes the number of recombination sites for the respective genes.

Table 1 *Campylobacter jejuni* subsp. *jejuni* reference genomes.

Species	Strain	Size (Mb)	%GC	ORFs	Disease/source	Year and place of isolation	GenBank
<i>Campylobacter jejuni</i> subsp. <i>jejuni</i>	NCTC 11168	1.64	30.5	1643	Clinical Food poisoning	1977, UK	AL111168
<i>Campylobacter jejuni</i> subsp. <i>jejuni</i>	RM1221	1.78	30.3	1838	Chicken	2005*, USA	CP000025
<i>Campylobacter jejuni</i> subsp. <i>jejuni</i>	81-176	1.6	30.6	1653	Clinical Food poisoning	1981, USA	CP000538
<i>Campylobacter jejuni</i> subsp. <i>jejuni</i>	81116	1.63	30.5	1626	Clinical Food poisoning	2007, UK	CP000814
<i>Campylobacter jejuni</i> subsp. <i>jejuni</i>	CG8421	1.6	30.4	1512	Clinical Food poisoning	ns	ABGQ000000000
<i>Campylobacter jejuni</i> subsp. <i>jejuni</i>	HB93-13	1.7	30.6	1710	Clinical GBS	2006*, USA	AANQ000000000
<i>Campylobacter jejuni</i> subsp. <i>jejuni</i>	CG8486	1.65	30.4	1425	Clinical Food poisoning	ns	AASY000000000
<i>Campylobacter jejuni</i> subsp. <i>jejuni</i>	CF93-6	1.67	30.5	1757	Clinical MFS	2006*, Japan	AANJ000000000
<i>Campylobacter jejuni</i> subsp. <i>jejuni</i>	84-25	1.67	30.4	1748	Clinical Meningitis	2006*, USA	AANT000000000
<i>Campylobacter jejuni</i> subsp. <i>jejuni</i>	260.94	1.65	30.5	1716	Clinical GBS	ns	AANK000000000
<i>Campylobacter jejuni</i> subsp. <i>jejuni</i>	IA3902	1.64	30.5	1718	Sheep abortion	2010, USA	CP001876
<i>Campylobacter jejuni</i> subsp. <i>doylei</i>	269.97	1.85	30.6	2094	Human blood	2007, UK	CP000768
<i>Campylobacter jejuni</i> subsp. <i>Jejuni-P179a</i>	ST-474	–	–	–	Poultry	New Zealand	–
<i>Campylobacter jejuni</i> subsp. <i>Jejuni-P110b</i>	ST-474	–	–	–	Poultry	New Zealand	AEIO000000000.1
<i>Campylobacter jejuni</i> subsp. <i>Jejuni-P694a</i>	ST-474	–	–	–	Poultry	New Zealand	–
<i>Campylobacter jejuni</i> subsp. <i>Jejuni-P569a</i>	ST-474	–	–	–	Poultry	New Zealand	–
<i>Campylobacter jejuni</i> subsp. <i>Jejuni-H22082</i>	ST-474	–	–	–	Human	New Zealand	AEIP01000008.1
<i>Campylobacter jejuni</i> subsp. <i>Jejuni-H704a</i>	ST-474	–	–	–	Human	New Zealand	–
<i>Campylobacter jejuni</i> subsp. <i>Jejuni-H73020</i>	ST-474	–	–	–	Human	New Zealand	–

ST: Multilocus sequence type; GBS: Guillain-Barré syndrome; ORFs: Open reading frames; Mb: Mega bases; GC: Guanine; Cytosine; ns: not stated; *:Date of start of project.

Table 2 Metabolic genes selected from the MLST schemes of different bacterial species for comparison.

Gene	Gene ID	Name of the gene	Function
<i>argF</i>	Cj0994c	Probable ornithine carbamoyltransferase	Amino acid biosynthesis – Glutamate family
<i>aroE</i>	Cj0405	Probable shikimate 5-dehydrogenase	Amino acid biosynthesis – Aromatic amino acid family
<i>atpD</i>	Cj0107	Probable ATP synthase F1 sector beta subunit	Energy metabolism
<i>dapE</i>	Cj1048c	Probable succinyl-diaminopimelate desuccinylase	Amino acid biosynthesis – Aspartate family
<i>ftsZ</i>	Cj0696	Probable cell division protein	Cell division
<i>fumC</i>	Cj1364c	Probable fumarate hydratase	Energy metabolism TCA
<i>gapA</i>	Cj1403c	Glyceraldehyde 3-phosphate dehydrogenase	Energy metabolism
<i>gltB</i>	Cj0007	Probable glutamate synthase (NADPH) large subunit	Energy metabolism TCA
<i>groEL</i>	Cj1221	60 kD chaperonin	Heat Shock
<i>hemN</i>	Cj0580c	Probable oxidoreductase	Oxygen-independent coproporphyrinogen-III oxidases
<i>ilvD</i>	Cj0013	Probable dihydroxy-acid dehydratase	Amino acid biosynthesis – Glutamate family
<i>infB</i>	Cj0136	Probable translation initiation factor IF-2	Protein translation
<i>lysA</i>	Cj0314	Probable diaminopimelate decarboxylase	Amino acid biosynthesis – Aspartate family
<i>nuoD</i>	Cj1576c	Probable NADH dehydrogenase I chain D	Energy metabolism – Respiration – Aerobic
<i>pycA</i>	Cj1037c	Possible pyruvate carboxylase A subunit	Central intermediary metabolism – Gluconeogenesis
<i>sdhA</i>	Cj0437	Probable succinate dehydrogenase flavoprotein subunit	Energy metabolism – Tricarboxylic acid cycle
<i>trpB</i>	Cj0348	Probable tryptophan synthase beta chain	Amino acid biosynthesis
<i>trpC</i>	Cj0498	Probable indole-3-glycerol phosphate synthase	Amino acid biosynthesis – Aromatic amino acid family
<i>aspA</i>	Cj0087	Aspartase	Central intermediary metabolism
<i>glnA</i>	Cj0699c	Glutamine synthetase	Central intermediary metabolism
<i>gltA</i>	Cj1682c	Central intermediary metabolism	Citrate synthase
<i>glyA</i>	Cj0402	Serine hydroxy methyl transferase	Central intermediary metabolism
<i>glmM</i>	Cj0360	Phospho glucosamine mutase	Central intermediary metabolism
<i>tkt</i>	Cj1645	Transketolase	Central intermediary metabolism
<i>uncA/atpA</i>	Cj0105	ATP synthase a subunit	Central intermediary metabolism

Table 3 GenBank sequence ids for metabolic genes for the five ST-474 isolates.

Isolates	Sequence_ids	Genes	Isolates	Sequence_ids	Genes
H704a_1	KC405082	<i>argF</i>	H73020_17	KC405123	<i>infB</i>
H704a_2	KC405083	<i>aroE</i>	H73020_18	KC405124	<i>lysA</i>
H704a_3	KC405084	<i>aspA</i>	H73020_19	KC405125	<i>nuoD</i>
H704a_4	KC405085	<i>atpD</i>	H73020_20	KC405126	<i>pycA</i>
H704a_5	KC405086	<i>dapE</i>	H73020_22	KC405127	<i>sdhA</i>
H704a_6	KC405087	<i>ftsZ</i>	H73020_21	KC405128	<i>tkt</i>
H704a_7	KC405088	<i>fumC</i>	H73020_23	KC405129	<i>trpB</i>
H704a_8	KC405089	<i>gapA</i>	H73020_24	KC405130	<i>trpC</i>
H704a_9	KC405090	<i>glmM</i>	H73020_25	KC405131	<i>atpA</i>
H704a_10	KC405091	<i>glnA</i>	P179a_1	KC405132	<i>argF</i>
H704a_11	KC405092	<i>gltA</i>	P179a_2	KC405133	<i>aroE</i>
H704a_12	KC405093	<i>gltB</i>	P179a_3	KC405134	<i>aspA</i>
H704a_13	KC405094	<i>glyA</i>	P179a_4	KC405135	<i>atpD</i>
H704a_14	KC405095	<i>groEL</i>	P179a_5	KC405136	<i>dapE</i>
H704a_15	KC405096	<i>hemN</i>	P179a_6	KC405137	<i>ftsZ</i>
H704a_16	KC405097	<i>ilvD</i>	P179a_7	KC405138	<i>fumC</i>
H704a_17	KC405098	<i>infB</i>	P179a_8	KC405139	<i>gapA</i>
H704a_18	KC405099	<i>lysA</i>	P179a_9	KC405140	<i>glmM</i>
H704a_19	KC405100	<i>nuoD</i>	P179a_10	KC405141	<i>glnA</i>
H704a_20	KC405101	<i>pycA</i>	P179a_11	KC405142	<i>gltA</i>
H704a_21	KC405102	<i>sdhA</i>	P179a_12	KC405143	<i>gltB</i>
H704a_22	KC405103	<i>tkt</i>	P179a_13	KC405144	<i>glyA</i>
H704a_23	KC405104	<i>trpB</i>	P179a_14	KC405145	<i>groEL</i>
H704a_24	KC405105	<i>trpC</i>	P179a_15	KC405146	<i>hemN</i>
H704a_25	KC405106	<i>atpA</i>	P179a_16	KC405147	<i>ilvD</i>
H73020_1	KC405107	<i>argF</i>	P179a_17	KC405148	<i>infB</i>
H73020_2	KC405108	<i>aroE</i>	P179a_18	KC405149	<i>lysA</i>
H73020_3	KC405109	<i>aspA</i>	P179a_19	KC405150	<i>nuoD</i>
H73020_4	KC405110	<i>atpD</i>	P179a_20	KC405151	<i>pycA</i>
H73020_5	KC405111	<i>dapE</i>	P179a_21	KC405152	<i>sdhA</i>
H73020_6	KC405112	<i>ftsZ</i>	P179a_22	KC405153	<i>tkt</i>
H73020_7	KC405113	<i>fumC</i>	P179a_23	KC405154	<i>trpB</i>
H73020_8	KC405114	<i>gapA</i>	P179a_24	KC405155	<i>trpC</i>
H73020_9	KC405115	<i>glmM</i>	P179a_25	KC405156	<i>atpA</i>
H73020_10	KC405116	<i>glnA</i>	P569a_1	KC405157	<i>argF</i>
H73020_11	KC405117	<i>gltA</i>	P569a_2	KC405158	<i>aroE</i>
H73020_12	KC405118	<i>gltB</i>	P569a_3	KC405159	<i>aspA</i>
H73020_13	KC405119	<i>glyA</i>	P569a_4	KC405160	<i>atpD</i>
H73020_14	KC405120	<i>groEL</i>	P569a_5	KC405161	<i>dapE</i>
H73020_15	KC405121	<i>hemN</i>	P569a_6	KC405162	<i>ftsZ</i>
H73020_16	KC405122	<i>ilvD</i>			

Isolates	Sequence_ids	Genes
P569a_7	KC405163	<i>fumC</i>
P569a_8	KC405164	<i>gapA</i>
P569a_9	KC405165	<i>glmM</i>
P569a_10	KC405166	<i>glnA</i>
P569a_11	KC405167	<i>gltA</i>
P569a_12	KC405168	<i>gltB</i>
P569a_13	KC405169	<i>glyA</i>
P569a_14	KC405170	<i>groEL</i>
P569a_15	KC405171	<i>hemN</i>
P569a_16	KC405172	<i>ilvD</i>
P569a_17	KC405173	<i>infB</i>
P569a_18	KC405174	<i>lysA</i>
P569a_19	KC405175	<i>nuoD</i>
P569a_20	KC405176	<i>pycA</i>
P569a_21	KC405177	<i>sdhA</i>
P569a_22	KC405178	<i>tkt</i>
P569a_23	KC405179	<i>trpB</i>
P569a_24	KC405180	<i>trpC</i>
P569a_25	KC405181	<i>atpA</i>
P694a_1	KC405182	<i>argF</i>
P694a_2	KC405183	<i>aroE</i>
P694a_3	KC405184	<i>aspA</i>

Isolates	Sequence_ids	Genes
P694a_4	KC405185	<i>atpD</i>
P694a_5	KC405186	<i>dapE</i>
P694a_6	KC405187	<i>ftsZ</i>
P694a_7	KC405188	<i>fumC</i>
P694a_8	KC405189	<i>gapA</i>
P694a_9	KC405190	<i>glmM</i>
P694a_10	KC405191	<i>glnA</i>
P694a_11	KC405192	<i>gltA</i>
P694a_12	KC405193	<i>gltB</i>
P694a_13	KC405194	<i>glyA</i>
P694a_14	KC405195	<i>groEL</i>
P694a_15	KC405196	<i>hemN</i>
P694a_16	KC405197	<i>ilvD</i>
P694a_17	KC405198	<i>infB</i>
P694a_18	KC405199	<i>lysA</i>
P694a_19	KC405200	<i>nuoD</i>
P694a_20	KC405201	<i>pycA</i>
P694a_21	KC405202	<i>sdhA</i>
P694a_22	KC405203	<i>tkt</i>
P694a_23	KC405204	<i>trpB</i>
P694a_24	KC405205	<i>trpC</i>
P694a_25	KC405206	<i>atpA</i>

Nucleotide sequences of metabolic housekeeping genes from 19 *C. jejuni* genomes

25 Data Files

<http://dx.doi.org/10.6084/m9.figshare.96407>

Table 4 Genes of repair mechanism, their gene ID, name, function and the pathways they are involved in.

Genes	Gene ID	Name of the gene	Function	Pathway
<i>dnaE_</i>	Cj0718	DNA polymerase III subunit alpha	It is a main replicative polymerase(Catalyses DNA-template-directed extension of the 3'- end of a DNA strand by one nucleotide at a time)	Purine metabolism, pyrimidine metabolism, metabolic pathways, DNA replication, mismatch repair and homologous recombination
<i>glaA_</i>	Cj1188c	tRNA uridine 5-carboxymethylaminomethyl modification enzyme	Involved in the modification of the wobble third base in tRNAs (5-carboxymethylaminomethyl modification (mm(5)s(2)U) of the wobble uridine base in some tRNAs; also known as a glucose-inhibited cell division protein A)	tRNA modification and regulation
<i>guaA_</i>	Cj1248	Probable GMP synthase (glutamine-hydrolysing)	Purine ribonucleotide biosynthesis	Purine metabolism
<i>gyrA_</i>	Cj1027c	DNA gyrase subunit A	Negatively supercoils closed circular double-stranded DNA or prevents from super-coiling that is deleterious to bacterial survival	Direct DNA repair mechanism
<i>gyrB_</i>	Cj0003	DNA gyrase subunit B	Negatively super-coils closed circular double-stranded DNA	Direct DNA repair mechanism
<i>ligA</i>	Cj0586	NAD-dependent DNA ligase	Catalyses the formation of phosphodiester linkages and is essential for DNA replication and repair of damaged DNA (between 5'-phosphoryl and 3'-hydroxyl groups in double-stranded DNA using NAD as a coenzyme and as the energy source for the reaction)	Direct DNA repair mechanism, DNA replication, base excision repair, nucleotide excision repair and mismatch repair
<i>mfd</i> repair	Cj1085c	Transcription-repair coupling factor	Prevents or corrects mutagenic nucleotides by removing them from DNA strands during replication	Nucleotide excision repair – DNA
<i>mutS</i>	Cj1052c	Recombination and DNA strand exchange inhibitor protein	Involved in blocking homologous recombination and inhibits DNA strand exchange (has ATPase activity stimulated by recombination intermediates)	Mismatch excision repair – DNA repair
<i>mutY_</i>	Cj1620c	Probable A/G-specific adenine glycosylase	Corrects incorrectly paired bases during DNA replication and involved in recombinational repair	Base excision repair – DNA repair
<i>ogt</i>	Cj0836	Methylated-DNA–protein–cysteine methyltransferase	Direct DNA repair by alkylation reversal	Direct DNA repair

Genes	Gene ID	Name of the gene	Function	Pathway
<i>polA</i>	Cj0338c	DNA polymerase I	Has 3'-5' exonuclease, 5'-3' exonuclease and 5'-3' polymerase activities, primarily functions to fill gaps during DNA replication and repair	Replication and DNA repair
<i>pyrC_</i>	Cj0259	Dihydroorotase Pyrimidine	ribonucleotide biosynthesis: catalyses the formation of N-carbamoyl-L-aspartate from (S)-dihydroorotate in pyrimidine biosynthesis	Pyrimidine nucleotide biosynthesis
<i>pyrG_</i>	Cj0027c	CTP synthetase	Catalyses the ATP-dependent amination of UTP to CTP with either Lglutamine or ammonia as the source of nitrogen	Pyrimidine nucleotide biosynthesis
<i>recA_</i>	Cj1673c	Recombinase A	Involved in recombinational repair of DNA damage. Catalyses the hydrolysis of ATP in the presence of single-stranded DNA, the ATP-dependent uptake of single stranded DNA by duplex DNA, and the ATP-dependent hybridisation of homologous single-stranded DNAs	Recombinational repair
<i>recJ_</i>	Cj0028	Putative single-stranded-DNA-specific exonuclease	Synthesis and modification of macromolecules – DNA replication, restriction/modification, recombination and repair	Recombinational repair
<i>recN</i>	Cj0642	Putative DNA repair protein	DNA repair protein	Recombinational repair
<i>recR</i>	Cj1263	Recombination protein	Involved in a recombinational process of DNA repair	Recombinational repair
<i>rplB_</i>	Cj1704c	50S ribosomal protein L2	One of the primary rRNA-binding proteins; required for association of the 30S and 50S subunits to form the 70S ribosome, for tRNA binding and peptide bond formation	Ribosomal gene
<i>rpoB_</i>	Cj0479	DNA-directed RNA polymerase subunit beta'	RNA synthesis, RNA modification and DNA transcription	RNA polymerase
<i>rpoD_</i>	Cj1001	RNA polymerase sigma factor	Sigma factors are initiation factors that promote the attachment of RNA polymerase to specific initiation sites and are then released; this is the primary sigma factor of bacteria	RNA polymerase

Genes	Gene ID	Name of the gene	Function	Pathway
<i>ruvA</i>	Cj0799c	Holliday junction DNA helicase	Plays an essential role in ATPdependent branch migration of the Holliday junction	Branch migration repair mechanism
<i>ssb</i>	Cj1071	Single-stranded DNA-binding protein	Binds to single stranded DNA and may facilitate the binding and interaction of other proteins to DNA	Other repair mechanisms
<i>uvrA_</i>	Cj0342c	Excinuclease ABC subunit A <i>uvrA</i> is an ATPase and a DNAbinding protein.	A damage recognition complex composed of two <i>uvrA</i> and two <i>uvrB</i> subunits scans DNA for abnormalities	Nucleotide excision repair
<i>uvrB</i> Nucleotide excision repair	Cj0680c	Excinuclease ABC subunit B	The UvrABC repair system catalyzes the recognition and processing of DNA lesions	The beta hairpin of the Uvr-B subunit is inserted between the strands, where it probes for the presence of a lesion
<i>xseA</i>	Cj0325	Exo-deoxyribonuclease VII large subunit	Bi-directionally degrades single stranded DNA into large acid insoluble oligonucleotides	Mismatch excision repair: Repair, ribosomal and nucleotide metabolic genes employed for MLST typing of other bacterial species

Nucleotide sequences of informational genes from 19 *C. jejuni* genomes

25 Data Files

<http://dx.doi.org/10.6084/m9.figshare.96406>

Table 5 GenBank sequence ids of the genes involved in the repair mechanism for the five ST-474 isolates.

Isolate_number	Sequence Ids	Genes
H704axseA	KC408804	<i>xseA</i>
H73020xseA	KC408805	<i>xseA</i>
P179a_xseA	KC408806	<i>xseA</i>
P569axseA	KC408807	<i>xseA</i>
P694axseA	KC408808	<i>xseA</i>
H704auvrB	KC408809	<i>uvrB</i>
H73020uvrB	KC408810	<i>uvrB</i>
P179auvrB	KC408811	<i>uvrB</i>
P569auvrB	KC408812	<i>uvrB</i>
P694auvrB	KC408813	<i>uvrB</i>
H704auvrA	KC408814	<i>uvrA</i>
H73020uvrA	KC408815	<i>uvrA</i>
P179auvrA	KC408816	<i>uvrA</i>
P569auvrA	KC408817	<i>uvrA</i>
P694auvrA	KC408818	<i>uvrA</i>
H704assb	KC408819	<i>ssb</i>
H73020ssb	KC408820	<i>ssb</i>
P179assb	KC408821	<i>ssb</i>
P569assb	KC408822	<i>ssb</i>
P694assb	KC408823	<i>ssb</i>
H704aruvA	KC408824	<i>ruvA</i>
H73020ruvA	KC408825	<i>ruvA</i>
P179aruvA	KC408826	<i>ruvA</i>
P569aruvA	KC408827	<i>ruvA</i>
P694aruvA	KC408828	<i>ruvA</i>
H704arpoD	KC408829	<i>rpoD</i>
H73020rpoD	KC408830	<i>rpoD</i>
P179arpoD	KC408831	<i>rpoD</i>
P569arpoD	KC408832	<i>rpoD</i>
P694arpoD	KC408833	<i>rpoD</i>
H704arpoB	KC408834	<i>rpoB</i>
H73020rpoB	KC408835	<i>rpoB</i>
P179arpoB	KC408836	<i>rpoB</i>
P569arpoB	KC408837	<i>rpoB</i>
P694arpoB	KC408838	<i>rpoB</i>
H704arplB	KC408839	<i>rpIB</i>
H73020_rplB	KC408840	<i>rpIB</i>
P179arplB	KC408841	<i>rpIB</i>
P569arplB	KC408842	<i>rpIB</i>
P694arplB	KC408843	<i>rpIB</i>
H704arecR	KC408844	<i>recR</i>

Isolate_number	Sequence Ids	Genes
H73020recR	KC408845	<i>recR</i>
P179arecR	KC408846	<i>recR</i>
P569arecR	KC408847	<i>recR</i>
P694arecR	KC408848	<i>recR</i>
H704arecN	KC408849	<i>recN</i>
H73020recN	KC408850	<i>recN</i>
P179arecN	KC408851	<i>recN</i>
P569arecN	KC408852	<i>recN</i>
P694arecN	KC408853	<i>recN</i>
H704arecJ	KC408854	<i>recJ</i>
H73020recJ	KC408855	<i>recJ</i>
P179arecJ	KC408856	<i>recJ</i>
P569arecJ	KC408857	<i>recJ</i>
P694arecJ	KC408858	<i>recJ</i>
H704arecA	KC408859	<i>recA</i>
H73020recA	KC408860	<i>recA</i>
P179arecA	KC408861	<i>recA</i>
P569arecA	KC408862	<i>recA</i>
P694arecA	KC408863	<i>recA</i>
H704apyrG	KC408864	<i>pyrG</i>
H73020pyrG	KC408865	<i>pyrG</i>
P179apyrG	KC408866	<i>pyrG</i>
P569apyrG	KC408867	<i>pyrG</i>
P694apyrG	KC408868	<i>pyrG</i>
H704apyrC	KC408869	<i>pyrC</i>
H73020pyrC	KC408870	<i>pyrC</i>
P179apyrC	KC408871	<i>pyrC</i>
P569apyrC	KC408872	<i>pyrC</i>
P694apyrC	KC408873	<i>pyrC</i>
H704apolA	KC408874	<i>polA</i>
H73020polA	KC408875	<i>polA</i>
P179apolA	KC408876	<i>polA</i>
P569apolA	KC408877	<i>polA</i>
P694apolA	KC408878	<i>polA</i>
H704aogt	KC408879	<i>ogt</i>
H73020ogt	KC408880	<i>ogt</i>
P179aogt	KC408881	<i>ogt</i>
P569aogt	KC408882	<i>ogt</i>
P694aogt	KC408883	<i>ogt</i>
H704amutY	KC408884	<i>mutY</i>
H73020mutY	KC408885	<i>mutY</i>

Isolate_number	Sequence Ids	Genes
P179amutY	KC408886	<i>mutY</i>
P569amutY	KC408887	<i>mutY</i>
P694amutY	KC408888	<i>mutY</i>
H704amutS	KC408889	<i>mutS</i>
H73020mutS	KC408890	<i>mutS</i>
P179amutS	KC408891	<i>mutS</i>
P569amutS	KC408892	<i>mutS</i>
P694amutS	KC408893	<i>mutS</i>
H704amfd	KC408894	<i>mfd</i>
H73020mfd	KC408895	<i>mfd</i>
P179amfd	KC408896	<i>mfd</i>
P569amfd	KC408897	<i>mfd</i>
P694amfd	KC408898	<i>mfd</i>
H704aligA	KC408899	<i>ligA</i>
H73020ligA	KC408900	<i>ligA</i>
P179aligA	KC408901	<i>ligA</i>
P569aligA	KC408902	<i>ligA</i>
P694aligA	KC408903	<i>ligA</i>
H704agyrB	KC408904	<i>gyrB</i>
H73020gyrB	KC408905	<i>gyrB</i>
P179agyrB	KC408906	<i>gyrB</i>

Isolate_number	Sequence Ids	Genes
P569agyrB	KC408907	<i>gyrB</i>
P694agyrB	KC408908	<i>gyrB</i>
H704agyrA	KC408909	<i>gyrA</i>
H73020gyrA	KC408910	<i>gyrA</i>
P179agyrA	KC408911	<i>gyrA</i>
P569agyrA	KC408912	<i>gyrA</i>
P694agyrA	KC408913	<i>gyrA</i>
H704aguaA	KC408914	<i>guaA</i>
H73020_guaA	KC408915	<i>guaA</i>
P179aguaA	KC408916	<i>guaA</i>
P569aguaA	KC408917	<i>guaA</i>
P694aguaA	KC408918	<i>guaA</i>
H704agidA	KC408919	<i>gidA</i>
H73020gidA	KC408920	<i>gidA</i>
P179agidA	KC408921	<i>gidA</i>
P569agidA	KC408922	<i>gidA</i>
P694agidA	KC408923	<i>gidA</i>
H704adnaE	KC408924	<i>dnaE</i>
H73020dnaE	KC408925	<i>dnaE</i>
P179adnaE	KC408926	<i>dnaE</i>
P569adnaE	KC408927	<i>dnaE</i>
P694adnaE	KC408928	<i>dnaE</i>

Figure 2 Guanine-cytosine (GC) variation in the metabolic housekeeping genes – A. The plots describe the distribution of GC across 19 *C. jejuni* genomes and the number inside the parenthesis denotes the number of recombination sites.

Figure 3 Guanine-cytosine (GC) variation in the metabolic housekeeping genes – A. The plots describe the distribution of GC across 19 *C. jejuni* genomes and the number inside the parenthesis denotes the number of recombination sites.

Figure 4 Guanine-cytosine (GC) variation in the metabolic housekeeping genes – A. The plots describe the distribution of GC across 19 *C. jejuni* genomes and the number inside the parenthesis denotes the number of recombination sites.

Figure 5 Guanine-cytosine (GC) variation in the MLST housekeeping genes – A. The plots describe the distribution of GC across 19 *C. jejuni* genomes and the number inside the parenthesis denotes the number of recombination sites.

Figure 6 Guanine-cytosine (GC) distribution in the ribosomal and repair genes – A. This plot describes the GC distribution across all 19 *C. jejuni* genomes. The number of recombination sites is provided in parentheses.

Figure 7 Guanine-cytosine (GC) distribution in the ribosomal and repair genes – A. This plot describes the GC distribution across all 19 *C. jejuni* genomes. The number of recombination sites is provided in parentheses.

Figure 8 Guanine-cytosine (GC) distribution in the ribosomal and repair genes – A. This plot describes the GC distribution across all 19 *C. jejuni* genomes. The number of recombination sites is provided in parentheses.

Figure 9 Guanine-cytosine (GC) distribution in the ribosomal and repair genes – A. This plot describes the GC distribution across all 19 *C. jejuni* genomes. The number of recombination sites is provided in parentheses.

Author Contributions

VM did the data collection, analyses and wrote the paper in partial fulfillment of her PhD thesis. MS was VM's supervisor and worked alongside her to get this manuscript out.

Grant Information

This work was part of Vathsala Mohan's PhD work and this project was supported by:

1. Building Research Capability in Strategically Relevant Areas (BRCSRA) fund;
2. Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North (RM12767); and
3. The Royal Society of New Zealand Marsden Fund (MAU0802).

This was an original research work and the funders had no role in any of the processes or in the preparation of the manuscript.

Competing Interests

No competing interests were disclosed.

Acknowledgements

I thank Prof. Nigel French and Dr. Patrick Biggs, mEpiLab, IVABS, Massey University, for giving their comments on this piece of work. I thank Dr. Patrick Biggs for kindly providing me with the gene predictions for my PhD work. I thank Massey University Doctoral Scholarship and New Zealand International Doctoral Scholarship for funding me. This work was carried out in mEpiLab, IVABS, Massey University, New Zealand.

References

1. Maiden MCJ: Multilocus Sequence Typing of Bacteria. *Annu Rev Microbiol.* 2006; **60**: p. 561–588.
2. Maiden MCJ, Bygraves JA, Feil E, *et al*: Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A.* 1998; **95**: p. 3140–3145.
3. Colles FM, Dingle KE, Cody AJ, *et al*: Comparison of *Campylobacter* populations in wild geese with those in starlings and free-range poultry on the same farm. *Appl Environ Microbiol.* 2008; **74**: p. 3583–3590.
4. Dingle KE, Colles FM, Wareing DR, *et al*: Multilocus sequence typing system for *Campylobacter jejuni*. *J Clin Microbiol.* 2001; **39**: p. 14–23.
5. McCarthy ND, Colles FM, Dingle KE, *et al*: Host-associated genetic import in *Campylobacter jejuni*. *Emerg Infect Dis.* 2007; **13**: p. 267–272.
6. Sheppard SK, Colles F, Richardson J, *et al*: Host association of *Campylobacter* genotypes transcends geographic variation. *Appl Environ Microbiol.* 2010; **76**: p. 5269–5277.
7. Sheppard SK, McCarthy ND, Falush D, *et al*: Convergence of *Campylobacter* Species: implications for Bacterial Evolution. *Science.* 2008; **320**: p. 237–239.
8. Sheppard SK, McCarthy ND, Jolley KA, *et al*: Introgression in the genus *Campylobacter*: Generation and spread of mosaic alleles. *Microbiology.* 2011; **157**: p. 1066–1074.
9. Wilson DJ, Gabriel E, Leatherbarrow AJ, *et al*: Rapid Evolution and the Importance of Recombination to the Gastroenteric Pathogen *Campylobacter jejuni*. *Mol Biol Evol.* 2009; **26**: p. 385–397.
10. Clark CG, Beeston A, Bryden L, *et al*: Phylogenetic relationships of *Campylobacter jejuni* based on *porA* sequences. *Can J Microbiol.* 2007; **53**: p. 27–38.
11. Dingle KE, Maiden MC: In *Campylobacter: Molecular and Cellular Biology*. ed. M.E.K.E.p. Ketley JM, Ketley JM, Konkel ME, Eds Horizon Bioscience, Wymondham, Norfolk, UK, 2005.
12. Wertz JE, Goldstone C, Gordon DM, *et al*: A molecular phylogeny of enteric bacteria and implications for a bacterial species concept. *J Evol Biol.* 2003; **16**: p. 1236–1248.
13. Christensen H, Kuhnert P, Olsen JE, *et al*: Comparative phylogenies of the housekeeping genes *atpD*, *infB* and *rpoB* and the 16S rRNA gene within the Pasteurellaceae. *Int J Syst Evol Microbiol.* 2004; **54**: p. 1601–1609.
14. Margos G, Gatewood AG, Aanensen DM, *et al*: MLST of housekeeping genes captures geographic population structure and suggests a European origin of *Borrelia burgdorferi*. *Proc Natl Acad Sci U S A.* 2008; **24**: p. 8730–8735.
15. Fouts DE, Mongodin EF, Mandrell RE, *et al*: Major structural differences and novel potential virulence mechanisms from the genomes of multiple *Campylobacter* species. *PLoS Biol.* 2005; **3**: e15.
16. Gundogdu O, Bentley SD, Holden MT, *et al*: Re-annotation and re-analysis of the *Campylobacter jejuni* NCTC11168 genome sequence. *BMC Genomics.* 2007; **8**: p. 162.
17. Hofreuter D, Tsai J, Watson RO, *et al*: Unique features of a highly pathogenic *Campylobacter jejuni* strain. *Infect Immun.* 2006; **74**: p. 4694–4707.
18. Lefebvre T, Stanhope MJ: Pervasive, genome-wide positive selection leading to functional divergence in the bacterial genus *Campylobacter*. *Genome Res.* 2009; **19**: p. 1224–1232.
19. Parker CT, Quiñones B, Miller WG, *et al*: Comparative genomic analysis of *Campylobacter jejuni* strains reveals diversity due to genomic elements similar to those present in *C. jejuni* strain RM1221. *J Clin Microbiol.* 2006; **44**: p. 4125–4135.
20. Parkhill J, Wren BW, Mungall K, *et al*: The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature.* 2000; **403**: p. 665–668.
21. Librado P, Rozas J: DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics.* 2009; **25**: p. 1451–1452.
22. Zhou BB, Elledge SJ: The DNA damage response: putting checkpoints in perspective. *Nature.* 2000; **408**: p. 433–439.
23. Minin VN, Dorman KS, Fang F, *et al*: Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics.* 2005; **21**: p. 3034–3042.

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 07 February 2014

<https://doi.org/10.5256/f1000research.1102.r3557>

© 2014 Falush D. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Daniel Falush

Department of Evolutionary Genetics, Max Planck Institute, Leipzig, Germany

I do not understand either from the abstract or reading the earlier paper why this data was collected. Neither the strain collection nor the quantities calculated have any obvious special features. The information presented is of a similar character to statistics that are automatically generated for example by the pubMLST database. There is no reason to clog up the scientific literature with such reports.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 02 May 2013

<https://doi.org/10.5256/f1000research.1102.r921>

© 2013 Ussery D. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

David Ussery

Center for Biological Sequence Analysis, Department of Systems Biology, The Technical University of Denmark, Lyngby, Denmark

This short article describes a comparison of 'housekeeping genes' with 'information processing genes', across 19 *Campylobacter jejuni* genomes. I cannot fully evaluate this paper as the data is not available. I understand that this is a new journal, and as such perhaps *F1000Research* has not

agreed that all sequences should be deposited to GenBank or EMBL before publication, but I find this frustrating that the results are published, but the data is missing.

Since currently there are only 11 *C. jejuni* genomes listed as 'complete', I was curious as to what additional genomes were used in this study. See:<http://www.ncbi.nlm.nih.gov/genome/browse/>

Unfortunately, it seems that only about half of the finished genomes were used (that is, 6 out of 11), and an additional set of apparently randomly chosen 6 unfinished, draft genome sequences were used, in addition to the apparently draft sequences of another 7 genomes of sequence type ST-474. Although it is stated that seven of the genomes have been deposited to GenBank, they are still not present in the database of complete genomes, as of 1 May, 2013 (the article was published about 6 weeks ago, on 14 March, 2013). In Table 1, I can see accession numbers for DRAFT sequences for two of the seven ST-474 genomes, and the remaining five do not appear to be listed in the NCBI genomes project pages, which includes even planned projects with no data yet. But the two ST-474 genomes listed in Table 1 do not appear to have the sequence type (ST-474) in the GenBank file; instead, all that is listed is "rare MLST strain". This missing meta-data makes it difficult to know if the other five genomes are perhaps there, but I cannot find them, even if I search for the strain name (e.g., P179a) provided for in Table 1.

Figure 1 shows the location of the housekeeping genes in the NCTC 11168 genome (which I recognise because I know that the GenBank accession number is for that strain - but wouldn't it be much more clear to just label this chromosome as '*C. jejuni*', and then put the strain name (NCTC 11168)? I am curious as to whether there's a significant AT content difference in the genes in the top half (e.g., closer to the replication origin), compared to the lower half (replication terminus). How does this difference compare to that found between the two gene sets?

From my perspective, Figures 2 - 9 are not very informative at all. These could be collapsed perhaps into a SINGLE box and whisker's plot, where the distributions could easily be compared with each other.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research