Cronfa URL for this paper:
http://cronfa.swan.ac.uk/Record/cronfa29439

Paper:
Large sequence diversity within biosynthesis locus and common biochemical features of Campylobacter
colilipoooligosaccharides. *Journal of Bacteriology*, JB.00347-16
http://dx.doi.org/10.1128/JB.00347-16

This article is brought to you by Swansea University. Any person downloading material is agreeing to abide by the
terms of the repository licence. Authors are personally responsible for adhering to publisher restrictions or conditions.
When uploading content they are required to comply with their publisher agreement and the SHERPA RoMEO
database to judge whether or not it is copyright safe to add this version of the paper to this repository.
http://www.swansea.ac.uk/iss/researchsupport/cronfa-support/
Title: Large sequence diversity within biosynthesis locus and common biochemical features of Campylobacter coli lipooligosaccharides

Running title: Campylobacter coli LOS

Alejandra Culebro\textsuperscript{1*}, Joana Revez\textsuperscript{1}, Ben Pascoe\textsuperscript{2,4}, Yasmin Friedmann\textsuperscript{2}, Matthew D Hitchings\textsuperscript{2}, Jacek Stupak\textsuperscript{3}, Samuel K Sheppard\textsuperscript{2,4}, Jianjun Li\textsuperscript{3}, Mirko Rossi\textsuperscript{1}

\textsuperscript{1}Department of Food Hygiene and Environmental Health, Faculty of Veterinary Medicine, University of Helsinki, Agnes Sjöbergin katu 2, Helsinki, FI-00014, Finland.

\textsuperscript{2}College of Medicine, Institute of Life Science, Swansea University, Singleton Park, Swansea SA2 8PP, United Kingdom.

\textsuperscript{3}Institute for Biological Sciences, National Research Council, Ottawa, Ontario K1A 0R6, Canada.

\textsuperscript{4}Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom

* Corresponding author: e-mail: aleja.culebro@helsinki.fi; Phone: +3580294157428; Mobile: +3580504486622

Keywords: Campylobacter; LOS; glycoconjugate; Qui\textsubscript{3}pNAcyl;

Word count: Abstract: 227; Main text (excluding table and figure legends, references): 4907.

Depositories (where applicable): The whole genome sequences of C. coli are publicly available on the RAST server (http://rast.nmpdr.org) with guest account (login and password 'guest') under IDs: 195.91, 195.96-195.119, 195.124-195.126, 195.128-195.130, 195.133, 195.134, 6666666.9432

Abbreviations LOS, lipooligosaccharides; RAST, Rapid Annotation using Subsystem Technology; GOs, groups of orthologues; EA-OTLC-MS, electrophoresis-assisted open-tubular liquid chromatography-electrospray mass spectrometry; ESI, electrospray ionization; oligosaccharide (OS); GlcN, 2-amino-2-deoxy-D-glucose; GlcN3N, β-1’-6 linked 3-diamino-2, 3-dideoxy-D-glucopyranose; PEtn, phosphoethanolamine; Hep, L-glycero-D-manno-heptose; Kdo, 3-deoxy-D-manno-octulosonic residue; Qui\textsubscript{3}pNAcyl, 3-acetamido-3,6-dideoxy-D-glucose; HexNac, hexosamine; deoxyHex, deoxyhexose; Hex, hexose; Qui\textsubscript{3}NAc, 3-acetamido-3,6-dideoxy-D-glucose; LPS, lipopolysaccharide.
Despite the importance of lipooligosaccharides (LOS) in the pathogenicity of campylobacteriosis, little is known about the genetic and phenotypic diversity of LOS in \textit{C. coli}. In this study, we investigated the distribution of LOS locus classes among a large collection of unrelated \textit{C. coli} isolates sampled from several different host species. Furthermore, we paired \textit{C. coli} genomic information and LOS chemical composition for the first time to investigate possible associations between LOS locus classes sequence diversity and biochemical heterogeneity. After identifying three new LOS locus classes, only 85\% of the 144 isolates tested were assigned to a class, suggesting higher genetic diversity than previously thought. This genetic diversity is at the basis of a completely unexplored LOS structure heterogeneity. Mass spectrometry analysis of the LOS of nine isolates, representing four different LOS classes, identified two features distinguishing \textit{C. coli} LOS from \textit{C. jejuni}'s. GlcN-GlcN disaccharides were present in the lipid A backbone in contrast to the GlcN3N-GlcN backbone observed in \textit{C. jejuni}. Moreover, despite that many of the genes putatively involved in QuipNAcyl were apparently absent from the genomes of various isolates, this rare sugar was found in the outer core of all \textit{C. coli}. Therefore, regardless the high genetic diversity of LOS biosynthesis locus in \textit{C. coli}, we identified species-specific phenotypic features of \textit{C. coli} LOS which might explain differences between \textit{C. jejuni} and \textit{C. coli} in terms of population dynamics and host adaptation.

Despite the importance of \textit{C. coli} to human health and its controversial role as a causative agent of the Guillain–Barré syndrome, little is known about the genetic and phenotypic diversity of \textit{C. coli} LOS. Therefore, we paired \textit{C. coli} genomic information and LOS chemical composition for the first time to address this paucity of information. We identified two species-specific phenotypic features of \textit{C. coli} LOS, which might contribute to elucidating the reasons behind the differences between \textit{C. jejuni} and \textit{C. coli} in terms of population dynamics and host adaptation.
INTRODUCTION

Campylobacteriosis is the most common bacterial food-borne disease in developed countries, with over 200,000 human cases reported annually in the European Union alone (1). The true burden of the disease in the population is likely underestimated, as many infections result in mild gastroenteritis (1). Approximately ~80% of reported infections are caused by Campylobacter jejuni and 7-18% of cases are attributed to C. coli. Therefore, C. coli is among the five most important bacterial aetiological agents of human gastroenteritis (2, 3).

As in other Gram-negative bacteria, Campylobacter spp. cell surface glycoconjugates, including lipooligosaccharides (LOS), play an important role in serum and bile resistance, resistance to phagocytic killing, adhesion, invasion, and survival in host cells (4-8). Current knowledge on LOS diversity has been based primarily on work in C. jejuni and its role in promoting severe clinical symptoms (9-12). C. jejuni LOS is a potent TLR4 agonist and the subsequent immune response is affected by changes in LOS structure and composition (10-14). Additionally, due to molecular mimicry between human gangliosides and certain LOS structures, C. jejuni has been identified as one of the causative agents of the Guillain–Barré syndrome (GBS) (15). Contrarily, the little knowledge on C. coli LOS variability has limited our understanding of the pathogenesis of GBS in patients infected with C. coli, as it remains unclear whether C. coli is able to mimic human ganglioside structures (16-18).

Valuable insights into the genetic origins of significant strain variable traits have been gained by studying the effect that C. jejuni LOS genotypes have on phenotype (19-24). However, so far, only two studies have addressed the variation in gene composition in C. coli LOS biosynthesis locus. Until now, nine genetic classes composed of a variable combination of 10 to 20 genes have been described in C. coli (25, 26), but no chemical analysis of their LOS structures was executed. A couple of decades ago the LOS structure of a single C. coli strain was described (27). Additionally, three other studies have
explored the chemical composition of *C. coli* LOS in a few strains (28-30), but no genetic information of the strains is available to our knowledge.

In this study, we investigated the diversity and distribution of LOS locus classes among a large collection of unrelated *C. coli* isolates sampled from several different host species. We expanded the current *C. coli* LOS classification by describing three additional LOS locus classes (25, 26). Moreover, by analysing genomic data with the LOS chemical composition of selected isolates, we identified possible associations between gene content in the LOS biosynthesis locus and observed differences in LOS phenotype. Despite the extensive introgression between *C. coli* and *C. jejuni* (31, 32), only negligible levels of recombination were detected in LOS biosynthesis genes, which might explain the distinctive species-specific chemical features observed herein.

**METHODS**

**Bacterial isolates, cultivation, and DNA extraction.** In total, 144 *C. coli* isolates, including 90 isolated from swine, 34 from humans, 18 from poultry, and two from wild birds, were chosen for LOS locus screening. The selection comprised 133 *C. coli* isolates from previous studies collected between 1996 and 2012 from Finnish human patients, chicken and pigs reared in Finland, and wild birds sampled in Helsinki region (25, 33-39). This collection was supplemented with 11 *C. coli* isolates from the Campynet (CNET) collection (hosted by DSMZ GmbH, https://www.dsmz.de/). Isolate selection was based on genotype (PFGE, MLST), host, country of origin, and year of isolation to encompass the largest possible diversity. Cultivation and DNA isolation were carried out as previously described (25), unless otherwise stated.

**PCR.** The length of LOS biosynthesis loci was determined by amplifying the region between orthologue 10 (LOS biosynthesis glycosyltransferase, *waaV*) and orthologue 16 (uncharacterized glycosyltransferase) (ID numbers according to Richards and colleagues (26)). PCR reactions were set
up as follows: 25 µl reactions containing 0.5 U Phusion high-fidelity (Thermo Scientific), 200 µM of each dNTP (Thermo Scientific), 0.4 µM of each primer (ORF3F2 and waaV; Table 1), 1 X Phusion GC buffer (Thermo Scientific), 700 µM of MgCl2 (Thermo Scientific), and 50 ng of template. Cycling conditions were as follows: one cycle at 98 ºC for 30 s followed by 30 cycles of denaturation at 98 ºC for 10 s, annealing at 62.4 ºC for 30 s, extension at 72 ºC for 6 min, and a final elongation at 72 ºC for 6 minutes. The size of the LOS locus was estimated by gel electrophoresis with 1 kb-plus (Thermo Scientific) and long-range (Thermo Scientific) molecular weight markers. Specific primers for each class, based on the previously described C. coli LOS locus classes (I to IX), were designed (25, 26). Primer pairs and their amplicon size for each LOS class are shown in Table 1, and a graphic representation of the primers annealing positions within the LOS locus is shown in Supplementary Figure 2. Since global alignment using progressiveMauve (40) revealed that LOS locus class IV and V (26) differ by only 3 single nucleotide polymorphism (which resulted in fragmentation of orthologue 1959 in class V), hereafter the two LOS locus classes are considered as a single class named IV/V. The specificity of each primer pair was verified in silico. All primers were designed on specific features characterizing each LOS locus class using, when possible, multiple sequence alignments of homologous sequences to improve sensibility and specificity. A preliminary gradient PCR was performed for each primer pair to select the most stringent conditions to minimize artefacts. Additionally, same results were obtained when primers of PCR-2 to -12 were tested on both genomic DNA or as a nested PCR using PCR-1 as template. PCRs were carried out in a semi-high-throughput manner, thus isolates were classified into a LOS class based on the results of all PCRs (Table 1). Isolates with unexpected LOS size, negative to all tested orthologues, or with unexpected combinations of orthologues, were classified as untypable.

Genome sequencing and annotation. For ascertaining the LOS locus classes, 35 isolates were chosen for genome sequencing (Supplementary Table 1) using either HiSeq or MiSeq. For HiSeq, NGS library
preparation, enrichment, sequencing, and sequence analyses were performed by the Institute for Molecular Medicine Finland (FIMM Technology Centre, University of Helsinki, Finland). MiSeq sequencing was performed by Institute of Life Science, Swansea University (Swansea, United Kingdom). Reads were filtered and assembled using SPAdes Assembler v. 3.3.0 (41). Primary annotation of all the genomes was performed using Rapid Annotation using Subsystem Technology (RAST) (42). Sequences were manually curated using Artemis (43) and LOS locus classes were aligned and compared with ACT (44). The whole genome sequences of \textit{C. coli} are publicly available on the RAST server (http://rast.nmpdr.org) with guest account (login and password 'guest') under IDs: 195.91, 195.96-195.119, 195.124-195.126, 195.128-195.130, 195.133, 195.134, and 6666666.94320

**Orthologue clustering and phylogenetic analysis.** A database including all the translated coding sequences of \textit{C. jejuni} and \textit{C. coli} LOS biosynthesis was assembled using Richards and colleagues (26) orthologues nomenclature. Reciprocal all-versus-all BLASTp search was performed (threshold \( E \leq 1e^{-10} \)) (45) and orthologous groups were determined by orthAgogue and MCL (ignoring E-values, percent match length \( \geq 80\% \) and inflation value of 5 (46, 47)). The groups of orthologues (GOs) were then aligned using MUSCLE and back-translated to nucleotide sequence using Translatorx perl script (48-50). Maximum likelihood phylogenetic reconstruction of each GO was performed in MEGA6.06 (51) using Kimura-2 as nucleotide substitution model and a discrete Gamma distribution (4 categories) to model evolutionary rate differences among sites. A total of 100 bootstrap runs were performed and summarized in a 95\% consensus tree.

**LOS silver staining.** LOS profiles were assessed by silver staining as described earlier (52), with some modifications. In brief, the absorbance of the biomass obtained from a 16 h Nutrient broth no2 (Oxoid) culture (100 rpm, microaerobic atmosphere, 37 \(^\circ\)C) was adjusted to an \(OD_{600} \) of 0.5. Cells were digested with 20 mg/ml proteinase K (Thermo Scientific), and incubated at 55 \(^\circ\)C for 1 h followed by
boiling for 10 min. Samples were then diluted 1:5 in loading buffer, and resolved in 15% SDS-PAGE gels. Gels were silver stained for visualization as previously described (53).

**CE-MS and EA-OTLC-MS analyses.** Biomass was produced in broth as indicated above and LOS was prepared with the rapid method applying microwave irradiation as previously described (54). In short, the lyophilized biomass was suspended in 50 μl of 20 mM ammonium acetate buffer (pH 7.5) containing DNase (100 μg/ml) and RNase (200 μg/ml) and heated by direct microwave irradiation. Proteinase K was then added to a final concentration of 60 μg/ml and heated under the same conditions.

Solutions were allowed to cool at room temperature and subsequently dried using a Speed Vac (vacuum centrifuge concentrator; Savant). LOS samples were washed three times with methanol (100 μl) with vigorous stirring. Insoluble residues were collected by centrifugation and resuspended in 30 μl water for electrophoresis-assisted open-tubular liquid chromatography-electrospray MS (EA-OTLC-MS) analysis. A sheath solution (isopropanol-methanol, 2:1) was delivered at a flow rate of 1.0 μL/minute. Separation was performed using 30 mM morpholine in deionized water, pH 9.0. A separation voltage of 20 kV, together with a pressure of 500 mbar, was applied for the EA-OTLC-MS analysis. The electrospray ionization (ESI) voltage applied on the sprayer was set at −5.2 kV. Data acquisition was performed for an m/z range of 600 to 2000 at a 2s/spectrum scan rate.

**Statistical analysis.** Fisher's exact test was used to assess host-LOS locus class association. P values equal to or less than 0.05 were considered significant.

**RESULTS**

**PCR typing method for *C. coli* LOS locus diversity.** We explored the genetic diversity of the LOS biosynthesis loci in 144 *C. coli* isolates (Supplementary table 1) using a PCR typing scheme based on published LOS locus class definitions (25, 26). Isolates were classified into putative LOS locus classes.
according to their PCR-profile and LOS locus size as described in Table 1. The LOS PCR typing scheme was validated by genome sequencing of 35 isolates (isolates marked in yellow in Supplementary table 1). Typing results are summarised in Table 2. We were able to classify 68% of the isolates into one of the nine previously published LOS locus classes (25, 26). Most of the isolates were assigned to LOS locus class II (17%) with the remaining isolates assigned to LOS classes IV/V (15%), III (13%), VI (13%), VIII (7%), I (2%), VII (1%), and IX (0.7%). The final 46 (out of 144, ~32%) isolates remained untypable by this method.

Six untypable isolates, with a LOS locus length of ~11.5 kbp, were sequenced (45, 63, 114, 125, 149, and 153). All isolates belong to a novel LOS locus class X. This new class shares 12 (out of 15) orthologues with other LOS locus classes (see below), and is characterised by the presence of three unique genes (Supplemental Fig. 2). A blastp search of the NCBI database, revealed sequence similarity with: (i) hypothetical protein of *Helicobacter* sp. MIT 05-5293 (e-value 1e^{-98}; identity 45%); (ii) hypothetical protein of *Helicobacter hepaticus* (e-value 3e^{-108}; identity 53%); (iii) UDP-N-acetylglucosamine 2-epimerase of *H. hepaticus* (e-value 3e^{-165}; identity 63%). Following this finding, primers were designed (Table 1) for LOS locus class X which further identified 15% of the isolates (Table 2). The genomes of isolates 138 and 99, which have a similar LOS size to class X but a different PCR profile (Supplementary Table 1) were also sequenced. Analysis of these genomes revealed two additional LOS locus classes, defined as XI (isolate 138) and XII (isolate 99). In total, we were able to assign a LOS locus class to 85% of the isolates in our collection by incorporating these additional classes. LOS profile diversity was high, suggesting that further LOS locus classes may be described in the future.

**Origin of the novel LOS locus classes X, XI, and XII.** As in *C. jejuni*, *C. coli* exhibits a mosaic LOS loci (22) with several classes containing similar orthologous loci. LOS locus classes X and XI are very
similar to each other, diverging only at a single locus (1967 vs 1920; Fig. 1). Additionally, these two classes also have similarity in gene content and organisation to LOS locus classes I, III, IV/V, VI, and VII (Fig. 1). To infer evolutionary relationships between these classes, phylogenetic analyses were performed for each shared GOs. Phylogenetic reconstruction revealed LOS class I and LOS class III as the two possible origins for the region encompassing orthologue 16 to orthologue 1668 in LOS locus class X (Fig. 1). Specifically, in the phylogenetic tree of orthologues 16, 1850, and 1668, C. coli isolates 45, 63, and 114 are monophyletic with strains from LOS locus class III, while C. coli isolates 125 and 149 formed a separate clade with LOS locus class I isolates (Supplemental Fig. 1A, B, and C).

Orthologues 8 and 1821 in LOS class X and both IV/V and VI share the same origin. Contrarily, the origin of the region including orthologues 1967, 1742, and 1743 is less clear. In the phylogenetic tree of orthologue 1967 (Supplemental Fig. 1D), C. coli isolates 63 and 114 are grouped with LOS locus class VI isolates, while the other strains form a separate clades. In addition, the star-like phylogeny inferred for orthologues 1742 and 1743 hampered any kind of conclusion. These results suggest that extensive recombination and gene reorganisation between LOS locus classes took place, masking the origin of common shared loci. Excepting for orthologue 1920, LOS locus class XI orthologues are closely related to those found in LOS locus class X (Supplemental Fig. 1). LOS locus class XII shares orthologues with LOS locus classes I, IV/V, VII, and IX. Yet, in our phylogenetic analysis LOS locus class XII orthologues are distantly related to those found in other LOS classes, forming a separate branch in the phylogenetic trees. Additionally, LOS locus class XII is characterized by the presence of a set of unique genes having the best BLASTp hit against NCBI nr with: (i) methyltransferase type 12 of Helicobacter hepaticus (e-value 6e-75; identity 58%); (ii) hypothetical protein of Anaerovibrio lipolyticus (e-value 5e-102; identity 65%); (iii) phosphoserine phosphatase of Helicobacter sp. MIT 05-5293 (e-value 3e-92; identity 63%) (Fig. 1). Proposed functions for each ORF of the herein newly identified LOS locus classes are described in Supplemental Table 2.
Cluster analysis of the LOS locus classes. Both species share a total of 19 LOS orthologues (26) and with previous evidence of introgression between C. coli and C. jejuni in mind (31, 32) we attempted to quantify the level of interspecies recombination in C. coli LOS diversity. We compared individual gene descriptions of the LOS loci rather than the original gene family ontologies used by Richards and colleagues (26). Out of the 19 shared orthologues, 16 gene locus descriptions split into species-specific clusters while only three were common in both species (orthologues 10, 16 and 1821). Interspecies gene transfer was investigated by comparing the topology of individual gene trees with the overall population structure (25). Evidence of interspecies gene transfer was only observed for orthologue 10 (lipooligosaccharide biosynthesis glycosyltransferase, waaV) where all C. coli loci of LOS locus class II formed a monophyletic clade with C. jejuni genes (Fig. 2). Thus, interspecies recombination is likely to have a limited effect on the LOS loci diversity observed in C. coli.

Host-LOS locus class association. The non-random distribution of LOS locus classes between hosts was investigated further by supplementing our isolate collection with Richards and colleagues data (26). The distribution of LOS locus classes by source of isolation is represented in Figure 3. All LOS locus classes, except class XII, were present among strains isolated from humans. More than half (57%) of the clinical isolates were LOS locus classes II, III, and VIII, while LOS locus classes VI, VII, and X were less commonly found in clinical cases. Most pig isolates were of LOS locus class X, but also frequently found among LOS locus classes II, III, IV/V, and VI. Only one pig isolate belonged to LOS locus class VIII and no pig strain was from classes I, IX, or XII. Poultry isolates were also found among all LOS locus classes, except for classes VII, IX, and XII. Most poultry isolates were classified as LOS locus class II.

There was a positive association (p <0.05) of class VIII to human clinical infections, while class VI was negatively associated with clinical cases. Swine was positively associated with classes VI and X, but negatively associated with classes I and VIII. Poultry was positively associated only with LOS
locus class I. Bovine and wild-bird isolates were underrepresented in the dataset. However, some association was observed in bovine (class IV/V) and wild bird isolates (class XII). Isolates classified as LOS locus classes II and III were equally distributed among humans, pigs, and poultry.

Chemical analysis of *C. coli* LOS composition. The LOS phenotype of nine selected isolates was investigated. This selection included strains from classes overrepresented in clinical isolates, II and VIII, as well as isolates from two of the newly described LOS classes (X and XI) and which are uncommon in clinical isolates. Silver staining SDS-PAGE gels of LOS extracts provided migration profiles for the selected isolates (Fig. 4A). A complimentary mass spectroscopy approach was used (CE-MS and EA-OTLC-MS) to explore inter- and intra-LOS class structural diversity. Example spectra is shown in Supplemental Fig. 3. The oligosaccharide (OS) composition of each of the nine isolates was predicted based on the fragment ions and components of the previously reported *C. coli* OS (27).

Size and composition of the lipid A group was defined for each glycoform by tandem mass spectrometry. For example, the fragment ion at m/z 1063.2 (doubly charged ion) in *C. coli* 137 (Supplemental Fig. 3), which was produced from the glycoform detected as triply charged ion at m/z 1422.8, corresponds to a lipid A with a 2-amino-2-deoxy-D-glucose (GlcN) disaccharide backbone carrying negative charged groups, two PPEtn, substituted by six fatty acid chains and with a calculated mass of ~2128 Da. Additionally, the fragment ion at m/z 1001.7 corresponds to a second lower mass lipid A species (~2006 Da) as it carries P and PPEtn instead. All analyzed *C. coli* isolates exhibited a hexa-acylated lipid A containing four tetradecanoic (14:0) and two hexadecanoic (16:0) acid chains, modified with two phosphate residues (55-57). Only GlcN disaccharides were detected in *C. coli* isolates, in contrast to the hybrid backbone of β-1’-6 linked 3-diamino-2, 3-dideoxy-D-glucopyranose (GlcN3N) and GlcN observed in *C. jejuni* (55, 57). Thus, *C. coli* synthesizes a lipid A with two ester- and two amide-linked acyl chains, while *C. jejuni* has a lipid A containing mainly three amide-linked acyl chains and one ester-linked acyl chain. The lower mass lipid A was detected in all samples, while
LOS locus class II isolates (except for isolate 65, Supplemental Fig. 3) had an additional lipid A species as exemplified by strain 137 in the Supplemental Fig. 3. Like in *C. jejuni*, *C. coli* exhibited a conserved inner core consisting of two L-glycero-D-manno-heptose (Hep) residues attached to a 3-deoxy-D-manno-octulosonic residue (Kdo) which is linked to the lipid A through a Kdo linker (20, 57). In the variable outer core region at least one residue of an average molecular weight of 231 Da was detected in all isolates. Based on available structures, this residue was predicted to be Qui<sub>3</sub>NAcy1 (where Qui<sub>3</sub>NAc represents 3-acylamino-3,6-dideoxy-D-glucose in which the N-acyl residue was a 3-hydroxybutanoyl). Although more than one OS was detected by MS in all isolates (Fig. 4B), only isolates from LOS locus classes X and XI exhibited visible high-Mr and low-Mr LOS on SDS-PAGE (Fig. 4A). Intra-LOS class diversity was observed in both LOS class II and class X. Isolate 65 displayed a LOS composition like other LOS class II isolates but with the addition of two hexosamines (HexNac) and one deoxyhexose (deoxyHex), and absence of PEtn residues (Fig. 4B). Likewise, isolates 45 and 63 shared similar LOS composition, with the exception of a variable Qui<sub>3</sub>NAcy1 residue in isolate 63. In contrast, isolate 114 exhibited a very different LOS composition compared with other isolates of the same class, including the presence of a third Hep and a deoxyHex as well as a reduced number of hexoses (Hex) (Fig. 4B). The LOS of isolates 38, 45, and 138 have similar core size and proposed composition, yet they are classified into three different LOS locus classes. However, our biochemical analysis is not able to identify saccharide sequence, stereochemistry, absolute configuration (D or L), anomeric configurations (α or β), and linkage positions. Thus, further studies would be required to determine whether these three different LOS classes indeed produce the same LOS structure.

**Genetic and phenotypic diversity within C. coli LOS locus class II.** The four strains with LOS locus class II shared 99.64% DNA sequence similarity and from 99.39% to 99.98% pairwise alignment identity. Isolate 65 was the most dissimilar among strains with LOS locus class II due to large fragments deletions. Deletions resulted in shorter 2400 and 2473 orthologues, as one pseudogene (Fig. 12).
Orthologues 2470 and 2471 were also truncated as one pseudogene (re-annotated as 2470-1), as evidenced by isolate 151. The remainder of the class II isolates had an insertion of 68 nt in 2470-1, disrupting the orthologue (Fig. 5). Despite the differences observed in orthologue 2470-1 isolates 73, 137, and 151 were predicted to have identical LOS chemical compositions.

Amino acid sequences of orthologues 6, 1541, 1501, 2472, and 10 were identical (100%) in all four class II strains, while orthologues 9004 and 16 exhibited a single amino acid difference in isolate 65. All isolates, with the exception of 65, exhibited differences in the C-terminal of orthologue 1715 and were variable in the number of Hep and/or PEtn residues observed. However, no GC homopolymeric tracts or other possible genetic signals associated with phase variation were identified within the LOS loci.

Genetic and phenotypic diversity within *C. coli* LOS locus class X. In LOS locus class X the overall sequence identity among strains was 99.31%, with percentage identity ranging from 98.96% to 99.94% in pairwise alignments, with strain 45 being the most distantly related. Although some minor gaps were observed, single point mutations were largely responsible for the diversity observed at nucleotide level. The largest insertion (69 nt) was seen in strain 63 between orthologues 2 and 3. Between strains, 100% amino acid identity was observed in orthologues 16, 8, and 2, while one or two amino acid substitutions were present in orthologues 1668, 1, 1821, 1967, and 1743. The most prominent difference was observed in orthologue 1742 in the form of a deleted A base at position 668, resulting in premature translational termination in isolates 114 and 63. Furthermore, several single amino acid substitutions were detected in orthologue 1742 in strain 45, while 100% identity was observed between isolates 63 and 114. In spite of dissimilar LOS composition, the only difference observed within the LOS locus between isolates 63 and 114 was in eight amino acids at the C-terminal of orthologue 3.
DISCUSSION

Campylobacter LOS is a fundamental feature involved in the pathogenesis of gastroenteritis and post-infection sequelae (10-14, 58, 59). However, despite the burden imposed by C. coli and the importance of this structure in campylobacteriosis, little is known about the LOS diversity in this species (26-29, 60). Therefore, we sought to contribute to the paucity of information by investigating the variability and distribution of C. coli LOS locus genetic classes in a large collection of isolates and by coupling genomic and LOS chemical composition data for the first time.

We developed a PCR methodology which was able to classify 85% of the isolates into a LOS class (25, 26). Among them, we described three additional LOS locus classes, named X, XI, and XII, which accounted for 17% of the isolates in our collection. The remaining untypable isolates (15%) suggests that further new classes will likely be described in the future and that C. coli LOS biosynthesis is more diverse than previously observed (26).

This genetic diversity is at the basis of a completely unexplored LOS structure heterogeneity which might contribute substantially to the population dynamics of C. coli, including host specificity. We combined our 144 isolates with 33 C. coli previously studied (26) to investigate the non-random distribution of LOS locus classes among different hosts. All hosts were significantly associated with at least one LOS locus class. In particular, isolates possessing LOS locus classes VI and X were predominantly isolated from swine, which have very high prevalence of C. coli (up to 99%) (61). Both of these classes were rarely detected in human isolates, which is supported by a previous source attribution study in Scotland in which pigs are a relatively unimportant source of C. coli human infections (61). The majority of human cases in our study were assigned to LOS locus classes II or III, which were also found in swine and poultry isolates. However, human isolates were overrepresented among LOS locus class VIII, which was rarely detected in the sources included in this study. This indicates the presence of other, unknown potential reservoirs contributing to human infections, which corroborates with a previous study where 54% of human C. coli strains were attributed to other sources.
than poultry and pig (61). In opposition to previous findings (26), we did not observe partitioning between bovine and poultry sourced strains, and LOS locus classes previously shown to be associated with bovine hosts were populated by isolates of poultry and swine origin. Due to the limited number of isolates available from alternative sources, the host-LOS class associations found in this study may not necessarily represent the true *C. coli* population structure in various hosts. However, our findings suggest that generalist isolates possessing LOS locus class II and III might be more successful at colonizing multiple species and, as seen in generalist lineages of *C. jejuni* ST-45 and ST-21 clonal complexes, being largely responsible for human infections (32).

Mosaic *C. coli* LOS classes appear to have arisen by the insertion and/or deletion of genes or gene cassettes through homologous recombination, as previously described in *C. jejuni* (22). In spite of substantial genome-wide introgression between agricultural *C. coli* and *C. jejuni* (25, 31), very limited interspecies recombination was detected among LOS biosynthesis loci. Only orthologue 10 (waaV) in *C. coli* LOS locus class II may have originated as result of recombination with *C. jejuni*. These results confirmed previous studies (31), and are supported by the species-specific features detected in the chemical composition of *C. coli* LOS.

GlcN disaccharide backbones, which is the most common structure among members of the family *Enterobacteriaceae* (57), were predicted in the lipid A of all analysed *C. coli* strains. This result is in contrast to the hybrid GlcN3N-GlcN backbone observed in *C. jejuni*. The genes *gnnA* and *gnnB*, located outside the LOS biosynthesis locus, are associated with the synthesis of GlcN3N-substituted lipid A (9, 62). Inactivation of either of these genes in *C. jejuni* resulted in the substitution of an N-linked with an O-linked acyl chain and an increased LOS biological activity in humans (9). *C. coli* contains in a similar genomic location both genes, having approximately 70% BLASTp score ratios against *C. jejuni* orthologues (9). Yet, *C. coli* *gnnA* and *gnnB* are separated by a putative cobalamin...
independent methionine synthase II in the same gene orientation. We suggest therefore three possible explanations for the absence of GlcN3N in *C. coli* lipid A backbone: (i) single or multiple mutations in the putative active sites of GnnA and GnnB have rendered one or both enzymes inactive, as observed in functional studies in other bacteria (62, 63); (ii) *gnnB-gnnA* operon transcription might be hampered by the presence of the putative methionine synthase II (9); (iii) GnnA and GnnB may be involved in the biosynthesis of alternative glycoconjugates in *C. coli* (62). Nevertheless, the substitution of an N-linked with an O-linked acyl chain in *C. coli* might have an impact in host-bacterial interaction and adaptation (9).

A second species-specific feature, common among all our analysed isolates, was the presence of at least one putative Quip3Nacyl residue. Quip3N is an unusual deoxysugar, which has been observed in the O-antigen of various Gram negative bacteria and in the S-layer of glycoprotein glycans of some Gram positives (64-66). Although rarely studied, Quip3N has also been found in the OS of LOS class E, H, and P isolates in *C. jejuni* exclusively as an N-acetyl derivative (Quip3NAc) (54, 67-69). Conversely, Quip3N has only been reported in *C. coli* as an N-acyl derivative with two possible substituents; 3-hydroxybutanoyl or 3-hydroxy-2, 3-dimethyl-5-oxoprolyl (30). The presence of Quip3NAcyl in *C. coli* was first described by Seltmann and Beer (30), and later on it was reported in several *C. coli* (28). However, the molecular basis behind the biosynthesis of this sugar and associated glycoconjugate in *C. coli* remains unknown. The dTDP-D-Quip3NAc biosynthesis pathway has, to our knowledge, only been described in the Gram positive *Thermoanaerobacterium thermosaccharolyticum* (70). This pathway involves five enzymes; a thymidyltransferase (RmlA), a 4, 6-dehydratase (RmlB), a 3, 4-isomerase (QdtA), a transaminase (QdtB), and a transacetylase (QdtC). Genome comparison of *T. thermosaccharolyticum* and *C. coli* identified homologs of *rmlA* (GO 1743), *rmlB* (GO 1742), *qdtA* (GOs 1920 and 1967), and *qdtB* (GO 8) in a subset of strains. However, no homologue for *qdtC* was found in *C. coli*. This may be expected as *C. coli* Quip3N is an N-acyl derivative instead of the N-acetyl
derivative found in *T. thermosaccharolyticum* (27, 30). Moreover, these results are in agreement with previous studies in which *C. jejuni* isolates carrying the aforementioned orthologues in the LOS locus have been found to express Quip3NAc in their LOS (26, 54, 67-69). Despite the presence of this sugar in all *C. coli* investigated in this study, as described above, the putative dTDP-D-Quip3NAc biosynthesis genes are only present in a subset of strains all belonging to LOS classes IV/V, VI, VII, X, and XI (Supplemental Fig. 2). Furthermore, truncation of orthologue 1742 due to a single base deletion should have resulted in the loss of Quip3NAcyl in isolates 114 and 63, which was not the case. Cross talk between different glycosylation pathways have been previously observed in *C. jejuni* (67, 71). Thus, due to Quip3NAcyl being predicted to be ubiquitously found in *C. coli* LOS structures, we hypothesize that the synthesis of this residue might be carried out by genes in conserved glycosylation pathways. Because of structural similarity between Quip3NAc and bacillosamine precursors, it is tempting to speculate that the *pgl* system may play a role in the biosynthesis of Quip3NAc in *C. coli*.

In all *C. coli*, phenotypic variation was observed affecting at least one sugar residue, as strains exhibit different numbers of Hep, Quip3NAcyl, HexNac, or PEtn (Fig 4B). Phenotypic variation in *C. jejuni* has been mainly associated with phase variation of genes containing repeats of GC homopolymeric tracts (23). However, no GC tracts were detected in the LOS locus of the chemically analysed *C. coli* isolates. Further inspection of all the LOS locus sequences generated in this and previous studies (25, 26) revealed that G-tracts are uncommon in *C. coli* LOS. Only isolates from LOS class IV/V and VI had G-tracts longer than 5 bases in their LOS biosynthesis locus. It is therefore unlikely that the observed phenotypic variation in our analysed samples was caused by slipped strand mispairing due to homopolymeric tracts within the LOS locus. These data suggest that other mechanisms, such as post-transcriptional regulation or epigenetic methylation of DNA, might be responsible for phenotypic variation in LOS composition in *C. coli*. 

17
Among LOS locus class II isolates, strain 65 exhibited the most divergent composition. Orthologue 424 1715 (wlaTB) has been associated with a HexNac residue in C. jejuni 81116 (67) and the diversity observed in the C-terminal of this orthologue may be responsible for the absence of HexNac residues in isolates 73, 137, and 151. However, further research is required to confirm the exact role of 1715 in LOS biosynthesis. Similarly to LOS locus class II, strains with LOS locus class X isolates minor genetic dissimilarities resulted in major differences in LOS chemical composition. Isolates 65 and 114 also contained a deoxyHex residue in the LOS. No orthologues potentially involved in deoxyHex synthesis were identified within the LOS region in isolates 65, suggesting that genes outside the LOS locus may play a bigger role in LOS biosynthesis than previously thought. Deoxyhexoses, such as 6-deoxy-β-l-altrose, fucose, or rhamnose have been frequently detected in the O-chain of the lipopolysaccharide (LPS) of several Gram-negative species (72, 73). Nevertheless, in the genus Campylobacter, these sugars have been described as components of C. jejuni capsule (74) and C. fetus LPS (75).

In conclusion, the genetic and biochemical diversity of C. coli is greater than expected. C. coli LOS is characterised by a lipid A consisting of GlcN-GlcN disaccharides and an outer core substituted with at least one Qui3Nacyl residue. Our results hint at cross talk between different glycosylation pathways, which has not been generally considered to play a role in LOS diversity. The relevance of these characteristic features for the ecology and virulence of C. coli is yet to be explored.

ACKNOWLEDGEMENTS
The authors wish to thank Ann-Katrin Llarena for her comments and Marja-Liisa Hänninen for providing the strains. This research project was supported by the University of Helsinki research grant n. 313/51/2013. AC was supported by the Microbiology and Biotechnology graduate program (MBDP) from the University of Helsinki. SS and BP were supported by the Biotechnology and Biological
Sciences Research Council (BBSRC) grant BB/I02464X/1, and the Medical Research Council (MRC) grants 473 MR/M501608/1 and MR/L015080/1.
REFERENCES


71. Bernatchez S, Szymanski CM, Ishiyama N, Li J, Jarrell HC, Lau PC, Berghuis AM, Young NM, Wakarchuk WW. 2005. A single bifunctional UDP-GlcNAc/Glc 4-epimerase supports the


FIGURE LEGENDS

Figure 1. LOS locus classes related to X, XI, and XII. Arrows represent ORFs. Genes coloured white are common to all LOS classes. Genes coloured green are present in class I and/or III. Genes coloured blue are present in classes IV/V and VI. Grey genes are common among classes X and XI. The orange genes are particular of the class XII. Striped genes are fragmented. Lines connect closely related orthologues. Strains are identified if more than one origin was observed in the LOS locus class (see text). Gene size is not drawn to scale.

Figure 2. Consensus cladogram representing the evolutionary relationship among orthologues belonging to GO 10 (nomenclature from Richard et al. 26). C. jejuni strains are highlighted in green. C. coli with the exception of LOS locus class II strains are shown in red. C. coli LOS locus class II strains are highlighted in yellow. The 95% bootstrap consensus tree was built from 100 replicates. Strains LOS locus class is indicated after the strain’s ID.

Figure 3. Host-LOS locus class association. Circos diagram shows the distribution of LOS locus classes of C. coli strains isolated from different hosts, from both our collection and those from Richards and colleagues (26). Ribbon ends represent links between host and LOS locus class while the width of the ribbon correlates with the percentage of strains belonging to a specific LOS locus class in a certain host. Segments in the outer ring indicate the percentage of strains representing a certain LOS locus class or host while the inner ring indicates the number of strains. Human strains are shown in orange, bovine in red, poultry in green, and swine in cyan.

Figure 4. C. coli LOS biochemical profiles. A) Silver-stained LOS. B) Proposed chemical composition based on MS and MS/MS results analysis of intact LOS (Supplemental Figure 3).

Figure 5. Comparison of nucleotide sequence of LOS locus class II strains 151 and 65. Genes coloured white are common to all LOS classes. Genes coloured blue are present in LOS locus classes IV/V, VI,
and VII. Yellow coloured genes are particular to LOS locus class II. Lines between orthologues represent sequence similarity.
Table 1. List of primers used in the present study and expected sizes of the amplicons.

<table>
<thead>
<tr>
<th>PCR Primers</th>
<th>Sequence</th>
<th>LOS locus class</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong></td>
<td>GFF3-F: AAA AGC TTT TGG CCG TGG GCC TGA TCA GAC</td>
<td>I  7.1  9.9  7.2  12.6  13.2  15.3  18.2  7.1  11.5  11.4  11.1</td>
</tr>
<tr>
<td>wa5V-R</td>
<td>AAA AGC TTT GCA AAT CTC ATG TAT AAA TCA GAG</td>
<td></td>
</tr>
<tr>
<td><strong>2</strong></td>
<td>2209-L: TCC AGG TCT TTA TGA TTT GTT ATC</td>
<td>+ (355) - - - - - - - - -</td>
</tr>
<tr>
<td>2209-R: GCT TGT GCC TTT GGT ATA AGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>3</strong></td>
<td>CstV-F: TCT CCA GCA GCT AT AAT GGA</td>
<td>- - - - - - - - -</td>
</tr>
<tr>
<td>CstV-R: TTT CAT TTC CAA AAT CCA TGC</td>
<td>+ (190) - - - - - - - - -</td>
<td></td>
</tr>
<tr>
<td><strong>4</strong></td>
<td>1541-L: TGG CAA YTA TGG TTT CTT GAA</td>
<td>- - + + + + + + +</td>
</tr>
<tr>
<td>1541-R: TGC CTT TCC AAA AGC AAA AAA TTC</td>
<td>(327) (327) (327) (327) - - - + +</td>
<td></td>
</tr>
<tr>
<td><strong>5</strong></td>
<td>2210-L: AAA TTT GCG TGG TCA GCT TG</td>
<td>- - - - - - - - -</td>
</tr>
<tr>
<td>2210-R: GCT GAA GCC AAT TGA TGA TG</td>
<td>- - - - - - - - -</td>
<td></td>
</tr>
<tr>
<td><strong>6</strong></td>
<td>1790-L: CCA AGC CAG ATT TTC CAA GA</td>
<td>- - - - - - - - -</td>
</tr>
<tr>
<td>1790-R: TGC GTA TCT TGC TTA CAC</td>
<td>- - - - - - - - -</td>
<td></td>
</tr>
<tr>
<td><strong>7</strong></td>
<td>1920-L: CCA AGC CAG ATT TTC CAA GA</td>
<td>- - - - - - - - -</td>
</tr>
<tr>
<td>1920-R: TGG TTA TAG AAA TCA CTT GCC AAT</td>
<td>(229) (229) - - - - - + +</td>
<td></td>
</tr>
<tr>
<td><strong>8</strong></td>
<td>2344-L: AAA GAA AGA GAA GCC AAA GGA G</td>
<td>- - - - - - - - -</td>
</tr>
<tr>
<td>2344-R: TCT GCG TTT AAT TCT GCC ATA TTC</td>
<td>- - - - - - - - -</td>
<td></td>
</tr>
<tr>
<td><strong>9</strong></td>
<td>1790-L: CCA AGC CAG ATT TTC CAA GA</td>
<td>- - - - - - - - -</td>
</tr>
<tr>
<td>1790-R: TGC GTA TCT TGC TTA CAC</td>
<td>- - - - - - - - -</td>
<td></td>
</tr>
<tr>
<td><strong>10</strong></td>
<td>38_3454: AGC CCT AGC GTG TAA ACC AT</td>
<td>- - - - - - - - -</td>
</tr>
<tr>
<td>38_2031: ATC GTC CTA TGC CCG GTG AT</td>
<td>- - - - - - - - -</td>
<td></td>
</tr>
<tr>
<td><strong>11</strong></td>
<td>CstV-F: TGC CTT TGC AAC AGC AAA TAA</td>
<td>- - - - - - - - -</td>
</tr>
<tr>
<td>CstV-R: GCT TGT GAG GTA CTA CCC AAT TA</td>
<td>- - - - - - - - -</td>
<td></td>
</tr>
<tr>
<td><strong>12</strong></td>
<td>45_3: GTC CTT TGC GAC CCG CTA CTT CT</td>
<td>- - - - - - - - -</td>
</tr>
<tr>
<td>45_1: GAG GGG CCT TAT GGA GCA AA</td>
<td>- - - - - - - - -</td>
<td></td>
</tr>
</tbody>
</table>

* the amplicons of this PCR are expressed in kb, while all others are in bp.
### TABLE 2. Distribution of LOS classes among hosts

<table>
<thead>
<tr>
<th>LOS class</th>
<th>Total (%)</th>
<th>Human</th>
<th>Swine</th>
<th>Poultry</th>
<th>Wild birds</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>3 (20)</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>II</td>
<td>24 (17)</td>
<td>7</td>
<td>13</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>III</td>
<td>18 (13)</td>
<td>4</td>
<td>13</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>IV/V</td>
<td>22 (15)</td>
<td>3</td>
<td>16</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>VI</td>
<td>18 (13)</td>
<td>1</td>
<td>15</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>VII</td>
<td>2 (1)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>VIII</td>
<td>10 (7)</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>IX</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>X</td>
<td>22 (15)</td>
<td>3</td>
<td>18</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>XI</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>XII</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Untypable</td>
<td>22 (15)</td>
<td>5</td>
<td>12</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Isolate</td>
<td>Detected Lipid A (Da)</td>
<td>Lipid A phosphorylation</td>
<td>Detected core (Da)</td>
<td>Hex</td>
<td>Hep</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------</td>
<td>-------------------------</td>
<td>-------------------</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>65</td>
<td>2006.6</td>
<td>PPEtn</td>
<td>1896.1</td>
<td>4</td>
<td>3/2</td>
</tr>
<tr>
<td>73</td>
<td>2005.6</td>
<td>PPEtn</td>
<td>1828.1</td>
<td>4</td>
<td>3/2</td>
</tr>
<tr>
<td>137</td>
<td>2005.6</td>
<td>PPEtn</td>
<td>1828.1</td>
<td>4</td>
<td>3/2</td>
</tr>
<tr>
<td>151</td>
<td>2005.6</td>
<td>PPEtn</td>
<td>1828.1</td>
<td>4</td>
<td>3/2</td>
</tr>
<tr>
<td>38</td>
<td>2005.6</td>
<td>PPEtn</td>
<td>2189.8</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>45</td>
<td>2005.6</td>
<td>PPEtn</td>
<td>2189.8</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>63</td>
<td>2005.6</td>
<td>PPEtn</td>
<td>2392.3</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>114</td>
<td>2005.6</td>
<td>PPEtn</td>
<td>2392.3</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>138</td>
<td>2005.6</td>
<td>PPEtn</td>
<td>2189.8</td>
<td>7</td>
<td>5</td>
</tr>
</tbody>
</table>

**Notes:**
- LOS class: II, VIII, X, XI.
- ~4 KDa marker.