Origin, evolution, and distribution of the molecular machinery for biosynthesis of sialylated lipooligosaccharide structures in Campylobacter coli

Alejandra Culebro, Miguel P. Machado, João André Carriço & Mirko Rossi

Campylobacter jejuni and Campylobacter coli are the most common cause of bacterial gastroenteritis worldwide. Additionally, C. jejuni is the most common bacterial etiological agent in the autoimmune Guillain-Barré syndrome (GBS). Ganglioside mimicry by C. jejuni lipooligosaccharide (LOS) is the triggering factor of the disease. LOS-associated genes involved in the synthesis and transfer of sialic acid (glycosyltransferases belonging to family GT-42) are essential in C. jejuni to synthesize ganglioside-like LOS. Despite being isolated from GBS patients, scarce genetic evidence supports C. coli role in the disease. In this study, through data mining and bioinformatics analysis, C. coli is shown to possess a larger GT-42 glycosyltransferase repertoire than C. jejuni. Although GT-42 glycosyltransferases are widely distributed in C. coli population, only a fraction of C. coli strains (1%) are very likely able to express ganglioside mimics. Even though the activity of C. coli specific GT-42 enzymes and their role in shaping the bacterial population are yet to be explored, evidence presented herein suggest that loss of function of some LOS-associated genes occurred during agriculture niche adaptation.

Glycan mimicry is a strategy utilized by pathogens to evade detection by the host innate immune system. Campylobacter jejuni, the most commonly reported cause of gastroenteritis in the world, boasts a large repertoire of human glycans. Molecular mimicry between sialylated C. jejuni lipooligosaccharides (LOS) and gangliosides may result in the onset of Guillain-Barré syndrome (GBS); an autoimmune acute progressive polyradiculoneuropathy with approximately 5% mortality rate. To express ganglioside-like LOS, C. jejuni synthesizes cytidine-5′-monophospho-N-acetylneuraminic acid (CMP-Neu5Ac) from uridine-5′-diphosphate-N-acetylglucosamine (UDP-GlcNAc) by the consecutive actions of an N-acetylglucosamine-6-phosphate 2-epimerase (NeuC), a sialic acid synthase (NeuB), and a CMP-Neu5Ac synthase (NeuA). Then, CMP-Neu5Ac is transferred by either of the LOS associated sialyltransferases; CstII (α2,3/8-sialyltransferase) or CstIII (α2,3-sialyltransferase). Both sialyltransferases belong to the, so far, monospecific CAZy (Carbohydrate-active enzymes database) glycosyltransferase (GT) family 42. Although the presence of GT-42 and N-acetyleneuraminate biosynthesis genes (neuABC) is insufficient for expressing molecular mimics, all C. jejuni strains containing this set of genes (LOS locus classes A, B, C, M, and R) have been shown to synthesize ganglioside-like structures. Therefore, the presence of GT-42 and neuABC genes has been used as proxy for identifying C. jejuni strains capable of producing human glycan mimics.

Campylobacter coli is the second most common cause of campylobacteriosis contributing, depending on the geographical region, to as many as 25% of all the infections. Although C. coli has also been isolated from GBS patients, its role in promoting this autoimmune disease remains controversial. Additionally, despite the pervasive introgression with C. jejuni, C. coli containing C. jejuni-like LOS classes linked to ganglioside mimicry activity have not been identified in GBS strains.
have not been reported so far. Based on genomic data analysed hitherto, *C. coli* LOS locus appears to be marginally affected by horizontal gene transfer (HGT) or homologous recombination.

Discovery of alternative orthologues of GT-42 encoding genes and associated LOS locus classes has been hindered by the very limited availability of genomic data. Consequently, it was only recently that *C. coli* LOS locus classes containing putative sialyltransferases, distantly related to those found in *C. jejuni*, were described. The *C. coli* LOS locus class IX contains a GT-42 (*cstV*) and *neuABC* genes, LOS class II harbours an orphan GT-42 (*cstIV*), and LOS class III has a pseudogenized orphan GT-42.

At present, the decreasing costs of next generation sequencing has driven a mass production of genomic sequences of several bacterial pathogens including *Campylobacter* spp. At the time of writing, the approximately 12,000 *C. jejuni* and 3,000 *C. coli* genome sequences found in public repositories offer unforeseeable opportunities. Thus, we took advantage of the large number of sequenced *Campylobacter* spp. strains to comprehensively investigate presence, frequency, and distribution of the molecular machinery for the biosynthesis of sialylated LOS structures in *C. coli* population.

### Results

**C. coli** GT42 genes. Of the 45 *C. coli* GT42 protein sequences retrieved from NCBI nr database, six were partial sequences (i.e. incomplete coding sequences). Thus, they were excluded from further analysis (Supplementary Table S1). Based on BlastP Score Ratio (BSR), the remaining 39 sequences clustered into 7 different groups (Supplementary Tables S2 and S3), with average BSR values ranging from 0.80 to 0.98 (Table 1). Group 1, 2, and 4 contain proteins showing the highest similarity to CstI, CstIII, and CstII, respectively, while the other groups described CstV in LOS class IX of *C. coli* (Supplementary Tables S2 and S3), with average BSR values ranging from 0.80 to 0.98 (Table 1). Group 1, 2, and 4 contain proteins showing the highest similarity to CstI, CstIII, and CstII, respectively, while the other groups show limited homology to *C. jejuni* GT-42 enzymes (Table 1). Group 5 comprises orthologues to the previously described CstV in LOS class IX of *C. coli* 7633929, while Group 7 includes CstIV, the GT-42 within *C. coli* LOS locus class II20,28. Group 6 contains a novel group of orthologous proteins (named herein CstVI) showing high similarity to the pseudogenized GT-42 described as part of LOS locus class III20,28. Similarly, Group 3 includes a single novel protein sequence named herein CstVII.

Furthermore, evolutionary analysis revealed that the 7 BSR groups form monophyletic clades and are divided into two clusters (Fig. 1). Cluster A is comprised of cstI, cstII, cstIII, and cstVII, while cluster B includes cstIV, cstV, and cstVI.

### Prevalence of GT-42 encoding genes in *C. coli* population.

Raw reads from 2,582 genomes submitted as *C. coli* were retrieved from European Nucleotide Archive (ENA) and classified into one of the three major *C. coli* phylogenetic clades based on *atpA* phylogeny and hierBAPS clustering (Supplementary Fig. S1). A total of 29 genomes were excluded from further analyses, as *atpA* phylogenetic analysis confirmed them to be *C. jejuni*. Altogether, 2,432 (95%) genomes belonging to Clade 1, 40 (1.6%) to Clade 2 and 81 (3.2%) to Clade 3 were mapped against all the sequences classified into the 7 *C. coli* GT-42 groups. A total of 818 (32%) *C. coli* genomes were positive for at least one GT-42 encoding gene (Table 2; Supplementary Table S4). GT-42 genes were found in approximately one third of *C. coli* Clade 1 (774/2,432; 31.8%). Furthermore, GT-42 genes were underrepresented in *C. coli* Clade 2 (2/40; 5%; *P* < 0.0001), while overrepresented in Clade 3 (42/81; 52%; *P* < 0.001). Overall, cluster B GT-42 genes (*cstIV, cstV* and *cstVI*) were the most abundant GT-42 detected in the *C. coli* population, accounting for 84.2% of the alleles. Conversely, cluster A GT-42 genes (*cstI, cstII, cstIII* and *cstVII*) only represented 15.8% of the alleles (Table 2). The most abundant *C. coli* GT-42 was *cstVI*, whereas *cstIII* was the rarest. *C. coli* Clade 1 strains were overrepresented in *cstVII* and *cstVI*, and underrepresented in *cstV* (*P* < 0.01). Conversely, Clade 3 strains were underrepresented in *cstVII* and *cstVI*, and overrepresented in *cstV* (*P* < 0.01).

### *C. coli* LOS classes contain GT-42 encoding genes.

To predict the LOS locus composition of GT-42 positive strains, genomes were mapped against all genes from all known LOS locus classes. Results are available in Supplementary Table S4. The presence of GT-42 gene alleles from cluster B was strongly concordant with predicted LOS locus classes. For *cstVI* positive strains, 99% were predicted to have a LOS locus class III-like. Similarly, 93% of *cstV* positive *C. coli* possessed a LOS locus class IX-like, and 68% of *cstIV* positive strains harboured a LOS locus class II-like. Contrastingly, genomes exclusively positive for cluster A GT-42 genes had no significant match to any of the previously defined LOS locus classes (Supplementary Table S4).

---

| Groups | Gene(s) | BSR | Cov (%) | Id (%) | BSR | Cov (%) | Id (%) |
|--------|---------|-----|---------|-------|-----|---------|-------|
| 1      | cstI21  | 0.87| 93      | 79    | 89  | 71      | 83    |
| 2      | cstII21 | 0.80| 56      | 56    | 98  | 51      | 100   |
| 3      | cstVII  | 1   | 61      | 51    | 98  | 52      | 86    |
| 4      | cstII31 | 0.87| 60      | 52    | 99  | 89      | 95    |
| 5      | cstVIII | 0.98| —      | —     | 98  | 48      | 94    |
| 6      | cstVII  | 0.95| —      | —     | 97  | 37      | 96    |
| 7      | cstIV29 | 0.86| —      | —     | 97  | 40      | 94    |

**Table 1.** Average Blastp Score Ratio (BSR) of the *C. coli* GT-42 homologs. *Singleton.* No significant hits. *Gene name proposed in this study.*
To determine the exact genetic composition and synteny of the LOS loci, 261 GT-42 positive genomes were assembled and manually inspected. The data set included all Clade 2 and 3 strains and a selection of Clade 1 strains comprising all \( \text{cstI} \), \( \text{cstII} \), \( \text{cstIII} \), and \( \text{cstVII} \) positive strains, and a subset of randomly selected \( \text{cstIV} \) and \( \text{cstVI} \) positive strains (Supplementary Table S4). Annotation of the identified LOS locus classes is available in Supplementary Table S5. Apart from \( \text{cstI} \) and \( \text{cstVII} \), all GT-42 genes were found within the LOS locus. Among assembled genomes, 61.3% (160) were found to contain a LOS-associated GT-42 gene. Besides the three previously described LOS locus classes containing GT-42 genes (i.e. classes II, III, and IX), 23 novel classes were identified (Fig. 2). LOS class III was the most abundant accounting for 72 isolates, followed by II (39), XXIII (27), XXV (26), XXVI (25), XXX (23), XVI (21), XXVIII (14), and XXXIII (14). The rest of the classes (17) were represented by a single strain. A strong association between LOS locus composition, \( \text{C. coli} \) Clade, and GT-42 gene alleles, was observed. In general, \( \text{C. coli} \) Clade 1 exhibited lower LOS locus diversity compared to the other clades. In Clade 1, genomes positive for \( \text{cstIV} \) and \( \text{cstVI} \) (88.9% of the total) possess LOS locus classes II and III, respectively, with 99% nt sequence identity. In all cases \( \text{cstVI} \) was present as a pseudogene. Contrastingly, Clade 3 \( \text{C. coli} \) evince a larger genetic variability in LOS locus classes containing \( \text{cstIV} \) (8 classes), \( \text{cstV} \) (3), or \( \text{cstVI} \) (2). Interestingly, no pseudogenes were found.

Albeit the rarity of LOS associated cluster A GT-42 genes (\( \text{cstII} \) and \( \text{cstIII} \)) in \( \text{C. coli} \) population (1.2%), several distinct LOS locus classes were identified (Fig. 2a). Out of the ten LOS classes containing \( \text{cstII} \) (Fig. 2a), only XXIII was detected in multiple Clade 1 (7) and Clade 3 (3) strains. Meanwhile, \( \text{cstIII} \) was located in two different LOS locus classes in \( \text{C. coli} \) Clade 2 strains.

---

**Table 2.** Distribution of GT-42 genes among \( \text{C. coli} \) clades.

| \( \text{C. coli} \) | GT-42 Cluster A | GT-42 Cluster B | Total |
|-----------------|----------------|----------------|-------|
|                 | \( \text{cstI} \) (G1)  | \( \text{cstII} \) (G4)  | \( \text{cstIII} \) (G2)  | \( \text{cstVII} \) (G3)  | \( \text{cstIV} \) (G7)  | \( \text{cstV} \) (G5)  | \( \text{cstVI} \) (G6)  |       |
| Clade 1         | 2              | 22             | 2     | 91    | 267  | 0   | 414  | 798  (94%) |
| Clade 2         | 0              | 0              | 2     | 0     | 0    | 0   | 0    | 2     (0.2%) |
| Clade 3         | 10             | 5              | 0     | 0     | 14   | 15  | 5    | 49    (5.8%) |
| Total           | 12 (1.41%)     | 27 (3.19%)     | 4 (0.47%) | 91 (10.73%) | 281 (33.10%) | 15 (1.77%) | 419 (49.35%) | 849 (100%) |

To determine the exact genetic composition and synteny of the LOS loci, 261 GT-42 positive genomes were assembled and manually inspected. The data set included all Clade 2 and 3 strains and a selection of Clade 1 strains comprising all \( \text{cstI} \), \( \text{cstII} \), \( \text{cstIII} \), and \( \text{cstVII} \) positive strains, and a subset of randomly selected \( \text{cstIV} \) and \( \text{cstVI} \) positive strains (Supplementary Table S4). Annotation of the identified LOS locus classes is available in Supplementary Table S5. Apart from \( \text{cstI} \) and \( \text{cstVII} \), all GT-42 genes were found within the LOS locus. Among assembled genomes, 61.3% (160) were found to contain a LOS-associated GT-42 gene. Besides the three previously described LOS locus classes containing GT-42 genes (i.e. classes II, III, and IX), 23 novel classes were identified (Fig. 2). LOS class III was the most abundant accounting for 72 isolates, followed by II (39), XXIII (27), XXV (26), XXVI (25), XXX (23), XVI (21), XXVIII (14), and XXXIII (14). The rest of the classes (17) were represented by a single strain. A strong association between LOS locus composition, \( \text{C. coli} \) Clade, and GT-42 gene alleles, was observed. In general, \( \text{C. coli} \) Clade 1 exhibited lower LOS locus diversity compared to the other clades. In Clade 1, genomes positive for \( \text{cstIV} \) and \( \text{cstVI} \) (88.9% of the total) possess LOS locus classes II and III, respectively, with 99% nt sequence identity. In all cases \( \text{cstVI} \) was present as a pseudogene. Contrastingly, Clade 3 \( \text{C. coli} \) evince a larger genetic variability in LOS locus classes containing \( \text{cstIV} \) (8 classes), \( \text{cstV} \) (3), or \( \text{cstVI} \) (2). Interestingly, no pseudogenes were found.

Albeit the rarity of LOS associated cluster A GT-42 genes (\( \text{cstII} \) and \( \text{cstIII} \)) in \( \text{C. coli} \) population (1.2%), several distinct LOS locus classes were identified (Fig. 2a). Out of the ten LOS classes containing \( \text{cstII} \) (Fig. 2a), only XXIII was detected in multiple Clade 1 (7) and Clade 3 (3) strains. Meanwhile, \( \text{cstIII} \) was located in two different LOS locus classes in \( \text{C. coli} \) Clade 2 strains.
All LOS locus classes containing cstII, cstIII, or cstV were positive for neuABC genes. Contrastingly, only 6.44% and 4.51% of cstIV and cstVI positive strains, respectively, contained neuABC genes which were invariably located outside the LOS locus and frequently in association with cstI or cstVII.
the agricultural niche.

Generally, gene diversification processes (including pseudogenization) and clonally expanded as a consequence of adaptation to specific environmental pressures. Thus, LOS locus classes II and III plausibly originated from Clade 3 LOS classes and underwent gene pseudogenization was observed: the phosphoethanolamine transferase genes (eptC) showed ~88% nucleotide identity over ~99% of length. Likewise, the terminal part of LOS class III showed high similarity (>90% nucleotide identity) to LOS classes XXVII and XXVIII. Notably, in both Clade 1 LOS locus classes gene pseudogenization was observed: the phosphoethanolamine transferase genes (eptC) in class II, and cstV in class III. Thus, LOS locus classes II and III plausibly originated from Clade 3 LOS classes and underwent a diversification process (including pseudogenization) and clonally expanded as a consequence of adaptation to the agricultural niche.

Figure 3. Venn diagram illustrating the number of orthologues shared between C. coli major clades. 10 orthologue were found in all three C. coli clades.

Gene flow and evolution of GT-42 containing LOS locus classes in C. coli. Based on orthologous group delineation by Roary (>95% amino acid identity), strains belonging to different C. coli clades were shown to share LOS-associated orthologues (Fig. 3). Hence, suggesting gene flow of LOS genes across C. coli clades. Interestingly, most of the share orthologues between clades encode proteins putatively involved in sugar biosynthesis or sugar modification (Table 3).

Insights into the evolution of C. coli GT-42 containing LOS classes were gained by comparing Clade 1 with Clade 3 LOS classes. Reciprocal blastn analysis between LOS locus classes II (Clade 1) and class XXXIV (Clade 3) showed ~88% nucleotide identity over ~99% of length. Likewise, the terminal part of LOS class III showed high similarity (>90% nucleotide identity) to LOS classes XXVII and XXVIII. Notably, in both Clade 1 LOS locus classes gene pseudogenization was observed: the phosphoethanolamine transferase genes (eptC) in class II, and cstV in class III. Thus, LOS locus classes II and III plausibly originated from Clade 3 LOS classes and underwent a diversification process (including pseudogenization) and clonally expanded as a consequence of adaptation to the agricultural niche.

Prevalence of GT-42 genes in C. jejuni. Prevalence of GT-42 homologues in C. jejuni was investigated by mapping 12,391 genome sequences deposited as C. jejuni against the 7 C. coli GT-42 groups. A total of 61.15% of the putative C. jejuni genomes were positive for at least one gene. Unsurprisingly, cstII and cstIII were the most abundant representing 98.75% of the GT-42 sequences detected. The remaining gene groups were either present in a minority of the tested genomes (cstIV, 101; cstV, 211; cstVII, 2) or non-detected (cstV). Genomes positive to GT-42 sequences other than cstII and cstIII were assembled for species verification and to manually inspect the LOS locus gene composition. Only 52 (16.6%) genomes were confirmed as C. jejuni, 40 of which (77%) were positive for cstIV, 10 to cstVI (19.2%), and 2 to cstVII (3.8%). Similarly to C. coli, C. jejuni cstVII was located outside the LOS locus and downstream from neuABC genes.

Introgression between C. jejuni and C. coli affect GT-42 containing LOS classes. High similarity between C. coli and C. jejuni cstII-associated LOS locus was observed (i.e. >80% gene lengths and >95% of nucleotide identity), implying recent gene flow between the two species. In fact, C. coli LOS classes XVI, XVII, XVIII, XIX, XX, XXI, and XXII are a mosaic of C. jejuni LOS classes A, B, S/F, and I/D. C. coli LOS class XXIV is further evidence of admixture between the two species, as it includes the C. coli specific cstV and neuB orthologues, as well as the C. jejuni cstII and neuB copies (Supplementary Table S6, Fig. 2). Finally, cross-species mobilization of an entire LOS locus classes was also encountered. C. coli strain SRR5152313 carries C. jejuni 11168 LOS class C, and 35 out of 40 cstIV positive C. jejuni strains, 37.1% of which from MLST sequence type 459, harbour a C. coli LOS locus class II.

Discussion
The small number of GBS associated C. coli isolates and the supposedly absence of molecular machinery for ganglioside mimicry are the main reasons, so far, supporting the idea of no link between C. coli and GBS. In 1994 von Wulffen and colleagues reported the first C. coli isolated from a GBS patient in a comparative seroreactivity study. The C. coli strain in question exhibited a Lior type 11 phenotype, which had also been found in GBS-associated C. jejuni strains. Thus, in the following years C. coli was considered as a plausible GBS causing organism. However, after recognition of C. jejuni expressing ganglioside-like LOS as the infectious agent triggering GBS, testing for cross-reactivity with anti-ganglioside autoantibodies became critical in understanding GBS
In the present study, 16 C. coli LOS locus classes (Fig. 2) were shown to contain the essential molecular machinery to potentially express sialylated LOS (i.e. a cst homologue and neuABC). While genotype is generally insufficient to predict LOS structure,27,19, considerable evidence supporting the expression of ganglioside-like LOS in C. coli was found. In contrast to previous reports27,28, C. coli LOS locus may be substantially affected by introgression with C. jejuni. Herein, 10 C. coli LOS locus classes containing a cstII were demonstrated to be mosaics of C. jejuni LOS classes. C. jejuni strains carrying cstII containing LOS classes have hitherto rarely been found to express non-ganglioside sialylated LOS19,18,21. Furthermore, extreme introgression resulted in acquisition of the entire C. jejuni LOS class C in C. coli SRR5152313 (100% homology). Consequently, this strain, isolated from turkey in US in 2016, could potentially trigger GBS as most likely expresses a GM1a- or GM2-like LOS9.

However, it is to be noted that strains carrying C. jejuni-like LOS locus are a minority in the C. coli population (approximately 1% of sequenced strains). Most of the C. coli possessing GT-42 genes (i.e. cstIV and cstVI) carry LOS classes lacking neuABC genes (approximately 27% of the sequenced strains). Furthermore, genome-wise analysis failed to identify genes potentially linked to the synthesis of CstIV and CstVII sugar donors. Even though functional studies are needed to clarify the activity of CstIV and CstVII, it seems plausible to believe that these elements are not involved in LOS ganglioside mimicry based on the results presented here and the absence of Neu5Ac in the LOS of cstIV positive strains28. Thus, the infrequency of the genetic structures related to ganglioside mimicry in the population might be the reason behind C. coli little contribution to GBS incidence26.

Beside ganglioside mimicry and the pathogenesis of GBS, expression of sialylated structures has a strong impact on host-bacteria interaction10. In our broad-gauge screening, we have shown that a considerable proportion of C. coli strains carry GT-42 genes within the LOS locus (29% of C. coli deposited in ENA at the time of writing). Overall, 23 new GT-42 associated LOS classes were described, 15 of which were present exclusively in the non-agriculture C. coli belonging to Clade 3. Thus, underrepresentation of non-agricultural C. coli strains22 in studies characterizing the LOS loci of extensive strain collections28,29 probably hampered earlier identification of a wider diversity of LOS classes with GT-42 genes.

We also discovered that LOS locus classes II and III28, the most predominant among agriculture-adapted Clade 1 C. coli, most likely originated from non-agriculture Clade 3 LOS classes. Moreover, few genes in both classes, including the GT-42 cstVI, lost their function in Clade 1. Cell surface structural changes as result of natural selection is a dominant phenomenon in microbial evolution. In pneumococcus, for example, natural selection as a consequence of vaccination programs targeting polysaccharide structures has resulted in shifts in the population of nonvaccine-type strains26. Outer membrane or wall-associated structures in bacteria (i.e. oligo and polysaccharides and proteins) play also a fundamental role in host interaction. Thus, they are subjected to diversifying selective pressure to conform to distinct receptors in different host species25. Moreover, reductive evolution leading to functional loss of several genes through e.g. pseudogenization is a common feature of bacterial undergoing niche adaptation25. For example, a single naturally occurring nucleotide mutation responsible for

| Roary Orthologue | Prokka annotation                                      | 1 | 2 | 3 |
|------------------|--------------------------------------------------------|---|---|---|
| 4026             | dTDP-glucose 4,6-dehydratase                           | + | + | + |
| 4027             | Glucose-1-phosphate thymidylyltransferase               | + | + | + |
| 4028             | TDP-4-oxo-6-deoxy-alpha-D-glucose 3,4-oxoisomerase      | + | + | + |
| 4029             | UDP-N-acetylglucosamine 2-epimerase                    | + | + | + |
| 4030             | N,N-diaceotylenamic acid synthase                       | + | + | + |
| 4031             | Polyosyl acid O-acetyltransferase                      | + | + | + |
| 4032             | UDP-glucose 6-dehydrogenase                           | + | + | + |
| 4033             | UDP-glucose 4-epimerase                                | + | + | + |
| 4034             | UDP-galactopyranose mutase                             | + | + | + |
| 4035             | General stress protein A                               | + | + | − |
| 4036             | GalNAc-alpha-(1->4)-GalNAc-alpha-(1->3)-diNAcBac-PP-undecaprenol alpha-1,4-N-acetyl-D-galactosaminyltransferase | + | + | − |
| 4037             | putative glycosyltransferase Epel                      | + | + | − |
| 4038             | putative glycosyltransferase epsl                      | + | + | − |
| 4039             | N-acetyrneuraminate cytidylyltransferase               | + | + | − |
| 4040             | N-acetyrneuraminate cytidylyltransferase               | + | + | − |
| 4041             | hypothetical protein                                   | + | + | − |
| 4042             | hypothetical protein                                   | + | + | − |
| 4043             | hypothetical protein                                   | + | + | − |
| 4044             | hypothetical protein                                   | + | + | − |

Table 3. Group of orthologues shared among C. coli clades.
the inactivation of a gene essential for D-alanylation of teichoic acids, has been shown to be sufficient to convert a human-specific *Staphylococcus aureus* strain into one that could infect rabbits37. Introduction of the agricultural niche was key in the evolution of *C. coli* clades24; clade I expanded within this niche and underwent an extensive genome introgression with *C. jejuni*.

Therefore, it is tempting to speculate that gene loss within imported LOS classes II and III, may have played a significant role in the expansion of *C. coli* in the agricultural niche by shaping the outer membrane composition. This hypothesis is supported by two pieces of evidence: (i) the predominance of LOS classes II and III in *C. coli* Clade I generalist (i.e. multihost) strains28 and (ii) the strong purifying selection resulting in limited nucleotide variability in these LOS locus classes (>99% identity. The importance of LOS locus classes II and III in adaptation to the agricultural niche is further evidenced by the flow of these genetic elements between *C. coli* and agricultural *C. jejuni*. Although introgression between *C. jejuni* and *C. coli* has been considered to be unilateral until now27, we identified several *C. jejuni* strains carrying LOS classes typically detected in *C. coli* Clade 1. Most of the *C. jejuni* strains carried a LOS class II with 99% identity, while some other presented a mosaic of Clade 3 LOS classes containing *cstIV*. As described previously22, a strong association between MLST type and LOS class was observed, being the bovine associated ST-459 the most prevalent among the *C. jejuni* carrying *C. coli* LOS class II.

**Conclusion**

Although at extremely low frequencies, bacterial factors implicated in GBS aetiology can cross clade and species barriers. Furthermore, spreading of these factors in the population could potentially result in *C. coli* playing a more prominent role in GBS. *C. coli* also presents a larger GT-42 enzyme repertoire than *C. jejuni*. Nevertheless, the activity of these enzymes and their role shaping *C. coli* population is yet to be explored. Overall, *C. coli* glyco-biology is largely unknown in spite of being a major foodborne pathogen.

**Methods**

**Genome sequences mining, genes detection and allele calling.** All whole genome raw sequence reads of entries deposited in the ENA as either *Campylobacter coli* or *Campylobacter jejuni* at the time of analysis (August 2017) were mapped against a set of reference genes (see below) for performing variant calling and inferring presence or absence using the ReMatCh framework v3.2 (https://github.com/B-UMMI/ReMatCh). Briefly, ReMatCh interacts with ENA for extracting and downloading all publicly available raw Illumina reads in fastq format for a given taxon. Then, it maps the reads onto the desired target loci using Bowtie2, and performs variant calling with Samtools/Bcftools and ReMatCh Single Nucleotide Polymorphism call criteria. The minimum coverage depth to consider a position to be present in the alignment was fixed at 5 reads, and to perform allele calling the threshold was 10 reads. A locus was considered to be present if (1) at least 70% of the target reference gene sequence was successfully mapped and 2) if the consensus sequence was ≥80% identity at nucleotide level. When needed, the consensus sequence alignment was extracted using the script *combine_alignment_consensus.py* available in ReMatCh utilities.

**Identification and frequency of *C. coli* GT-42 homologues.** To collect a set of *C. coli* reference genes homologous to *C. jejuni* GT-42 encoding genes, amino acid sequences of *CstI* (Uniprot Q9RGF1), *CstII* (Uniprot Q9FK09M) and *CstIII* (Uniprot Q7BP25) were used to search non-redundant (nr) NCBI protein sequences collection using blast + V 2.7.1 for best *C. coli* blastp hits (>30% of amino acid identity; >50% query coverage). Partial sequences were discarded and the remaining ones were used for an all-versus-all blastp analysis. Sequences were then categorized in separate groups having >0.7 of BSR. A Minimum Evolution phylogenetic tree based on the back-translated nucleotide sequence alignments (built with MUSCLE with default parameters) of all detected *C. coli* GT-42 proteins and *C. jejuni* *cstI*, *cstII*, and *cstIII* was inferred using MEGA. Finally, the detected *C. coli* GT-42 nucleotide sequences were used as reference for calling orthologues in all *C. coli* and *C. jejuni* strains using ReMatCh as described above.

**Identification of *Campylobacter coli* clades.** To assign *C. coli* samples to one of the three previously described major phylogenetic clades27,29, population structure analysis and inferred phylogenetic relationships based on *atpA* gene43 were performed. The *atpA* sequence of *C. coli* strain RM2228 (KF855277) was used for allele calling in all *C. coli* strains using ReMatCh as described above. Based on the ReMatCh *atpA* consensus sequence alignment, samples were clustered using hierBAPS at first level and a Neighbor joining phylogenetic tree was inferred using MEGA. Representative strains from each *C. coli* clade were used as reference for classifying the clusters, and a set of *C. jejuni* strains were used as outgroup. The generated tree was visualized in iTOL.

**Classification into LOS classes.** To assign samples to one of the previously described *C. coli* LOS locus classes, nucleotide sequences of *cst* genes located at the "conserved putative two-domain glycosyltransferase" (orthologue 16 as described previously6,28) and the "LOS biosynthesis glycosyltransferase waaV" (orthologue 10 described previously6,28) from *C. coli* LOS locus classes I to XII were used for calling orthologues in all GT-42 positive *C. coli* using ReMatCh, as described above. Results were reported as percentage of genes present for a given LOS locus class.

**Pangenome analysis and gene flow investigation of LOS loci.** For a set of *C. coli* strains of interest, raw sequencing data were retrieved from ENA with getSeqENA (https://github.com/B-UMMI/getSeqENA). Then, the paired-end raw reads were assembled using the INNUca pipeline (https://github.com/INNUENDOCON/INNUca), which consists of several modules and QA/QC steps. In brief, INNUca starts by calculating if the sample raw data fulfill the expected coverage (min 15×). After subjecting reads to quality analysis using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and cleaning with Trimmomatic46, INNUca proceeds to de novo draft genome assembly with SPAdes 3.11.47 and checking assembly depth of coverage (min 30×).
Finally, Pilon improves the draft genome by correcting bases, fixing misassemblies, and filling gaps, prior species confirmation and MLST prediction with mlst software (https://github.com/tseemann/mlst).

Draft genomes passing INNUca QA/QC were annotated with Prokka, and pangenome analysis was executed using Roary (default parameters). To annotate novel LOS locus classes, assemblies were manually inspected with Artemis.

Horizontal Gene Transfer (HGT) among C. coli clades was inferred by mapping presence/absence of LOS associated group of orthologues into the atpA tree (see above). To infer possible gene transfer between C. coli and C. jejuni, representative sequences of LOS associated group of orthologues were blasted against nt NCBI database and HGT was detected if the best blast hit for C. jejuni was >90% nucleotide identity over >70% of the C. coli query length.

Statistical analysis. Fisher’s exact test was used to assess clade and GT-42 associations. P values of ≤0.05 were considered significant.

Data availability. Data are available in Supplementary Information.

References
1. Cress, B. F. et al. Masquerading microbial pathogens: capsular polysaccharides mimic host-tissue molecules. FEMS Microbiol. Rev. 38, 660–697 (2014).
2. Carlin, A. F. et al. Molecular mimicry of host sialylated glycans allows a bacterial pathogen to engage neutrophil Siglec-9 and dampen the innate immune response. Blood 113, 3333–3336 (2009).
3. Houlston, R. S. et al. Lipooligosaccharide of Campylobacter jejuni: similarity with multiple types of mammalian glycans beyond gangliosides. J. Biol. Chem. 286, 12361–12370 (2011).
4. Tam, C. C. et al. Incidence of Guillain-Barré syndrome among patients with Campylobacter infection: a general practice research database study. J. Infect. Dis. 194, 95–97 (2006).
5. Yuki, N. Carbohydrate mimicry: a new paradigm of autoimmune diseases. Curr. Opin. Immunol. 17, 577–582 (2005).
6. Goodellow, J. A. & Willison, H. J. Guillain-Barré syndrome: a century of progress. Nat. Rev. Neurol. 12, 723–731 (2016).
7. Gilbert, M. et al. Biosynthesis of ganglioside mimics in Campylobacter jejuni OH4384 identification of the glycosyltransferase genes, enzymatic synthesis of model compounds, and characterization of nanomole amounts by 600-MHz 1H AND 13C NMR analysis. J. Biol. Chem. 275, 3896–3906 (2000).
8. Godschalk, P. C. R. et al. The genetic bases for the variation in the lipo-oligosaccharide of the mucosal pathogen. Campylobacter jejuni biosynthesis of sialylated ganglioside mimics in the core oligosaccharide. J. Biol. Chem. 277, 327–337 (2002).
9. Lewis, A. L. et al. Innovations in host and microbial sialic acid biosynthesis revealed by phylogenomic prediction of nonulosonic acid structure. Proc. Natl. Acad. Sci. 106, 13552–13557 (2009).
10. Cheng, J. et al. Multifunctionality of Campylobacter jejuni sialyltransferase CstII: characterization of GD3/GT3 oligosaccharide synthase, GD3 oligosaccharide sialidase, and trans-sialidase activities. Glycobiology 18, 686–697 (2008).
11. Cheng, J. et al. Comparative genomics of unintrogressed Campylobacter coli clades 2 and 3. J. Mol. Biol. 328, 307–317 (2003).
12. Lombard, V., Golanczad Ramulu, H., Drula, E., Coutinho, P. M. & Henrisat, B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 42, D490–D495 (2014).
13. Campbell, J., Davies, G., Bulone, V. & Henrisat, B. A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. Biochem. J. 329, 719 (1998).
14. Coutinho, P. M., Deleury, E., Davies, G. J. & Henrisat, B. An evolving hierarchical family classification for glycosyltransferases. J. Biol. Chem. 277, 327–337 (2002).
15. Lu, J. F. et al. A plate based high-throughput activity assay for polysialyltransferase from Neisseria meningitidis. Anal. Biochem. 444, 67–74 (2014).
16. Parker, C. T. et al. Comparison of Campylobacter jejuni lipo-oligosaccharide biosynthesis loci from a variety of sources. J. Clin. Microbiol. 43, 2771–2781 (2005).
17. Zhang, P., Zuccolo, A. J., Li, W., Zheng, R. B. & Ling, C. C. Probing a sialyltransferase’s recognition domain to prepare α(2,8)-linked oligosialosides and analogs. Chem. Commun. 0, 4233–4235 (2009).
18. Yu, C.-C. et al. Masquerading microbial pathogens: capsular polysaccharides mimic host-tissue molecules. FEMS Microbiol. Rev. 38, 660–697 (2014).
31. Cohen, T., Colijn, C. & Murray, M. Modeling the effects of strain diversity and mechanisms of strain competition on the potential performance of new tuberculosis vaccines. Proc. Natl. Acad. Sci. USA 105, 16302–16307 (2008).

32. Guinane, C. M. et al. Evolutionary genomics of Staphylococcus aureus reveals insights into the origin and molecular basis of ruminant host adaptation. Genome Biol. Evol. 2, 454–466 (2010).

33. Viana, D. et al. A single natural nucleotide mutation alters bacterial pathogen host-tropism. Nat. Genet. 47, 361–366 (2015).

34. Sheppard, S. K. et al. Evolution of an agriculture-associated disease causing Campylobacter coli clade: evidence from national surveillance data in Scotland. PLOS ONE 5, e15708 (2010).

35. Cha, W. et al. Comparing the genetic diversity and antimicrobial resistance profiles of Campylobacter jejuni recovered from cattle and humans. Front. Microbiol. 8, 818 (2017).

36. Machado, M. P., Ribeiro-Gonçalves, B., Silva, M., Ramirez, M. & Carriço, J. A. Epidemiological surveillance and typing methods to track antibiotic resistant strains using high throughput sequencing. Methods Mol. Biol. Clifton NJ 1520, 331–355 (2017).

37. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).

38. Li, H. et al. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).

39. Camacho, C. et al. BLAST+: architecture and applications. BMC Bioinformatics 10, 421 (2009).

40. Rasko, D. A., Myers, G. S. & Ravel, J. Visualization of comparative genomic analyses by BLAST score ratio. BMC Bioinformatics 6, 2 (2005).

41. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797 (2004).

42. Kumar, S., Stecher, G. & Tamura, K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870–1874 (2016).

43. Miller, W. G., Yee, E., Jolley, K. A. & Chapman, M. H. Use of an improved atpA amplification and sequencing method to identify members of the Campylobacteraceae and Helicobacteraceae. Lett. Appl. Microbiol. 58, 582–590 (2014).

44. Cheng, L., Connor, T. R., Sirén, J., Aanensen, D. M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. Mol. Biol. Evol. 30, 1224–1228 (2013).

45. Letunic, I. & Bork, P. Interactive tree of life (iTOL)v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 44, W242–W245 (2016).

46. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120 (2014).

47. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. J. Comput. Mol. Cell Biol. 19, 455–477 (2012).

48. Walker, B. J. et al. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLOS ONE 9, e112963 (2014).

49. Seemann, T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069 (2014).

50. Page, A. J. et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31, 3691–3693 (2015).

51. Rutherford, K. et al. Artemis: sequence visualization and annotation. Bioinforma. Oxf. Engl. 16, 944–945 (2000).

Acknowledgements

This study was funded by the following grants; University of Helsinki three years research grant 313/51/2013, ONEIDA project (LISBOA-01-0145-FEDER-016417) co-funded by FEEI: "Fundos Europeus Estruturais e de Investimento" from "Programa Operacional Regional Lisboa 2020" and by national funds from FCT - "Fundação para a Ciência e a Tecnologia" and BacGenTrack (TÜBİTAK/0004/2014) [FCT/ Scientific and Technological Research Council of Turkey (Türkiye Bilimsel ve Teknolojik Araştırma Kurumu, TÜBİTAK)]. A. C was supported by the Microbiology and Biotecetogenate graduate program from the University of Helsinki. The authors wish to thank CSC- Tieteen tietotekniikan keskus Oy for providing access to cloud computing resources.

Author Contributions

A.C. designed and coordinated the study. A.C. and M.R. performed data analysis, prepared figures, and wrote the manuscript. J.A.C. and M.P.M. design and developed INNUca and ReMatch. All authors have contributed to data interpretation, have critically reviewed the manuscript, and approved the final version as submitted.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-21438-2.

Competing Interests: The authors declare no competing interests.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.