Preterm Infant-Associated *Clostridium tertium*, *Clostridium cadaveris*, and *Clostridium paraputrificum* Strains: Genomic and Evolutionary Insights

Raymond Kiu1,2, Shabhonam Caim1, Cristina Alcon-Giner1, Gusztav Belteki3, Paul Clarke4, Derek Pickard5, Gordon Dougan5, and Lindsay J. Hall1,*

1The Gut Health and Food Safety Programme, Quadram Institute Bioscience, Norwich Research Park, Norwich, United Kingdom
2Norwich Medical School, Norwich Research Park, University of East Anglia, Norwich, United Kingdom
3Neonatal Intensive Care Unit, The Rosie Hospital, Cambridge University Hospitals NHS Foundation Trust, United Kingdom
4Neonatal Intensive Care Unit, Norfolk and Norwich University Hospitals NHS Foundation Trust, Norwich, United Kingdom
5Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom

*Corresponding author: E-mail: lindsay.hall@quadram.ac.uk.

Accepted: September 28, 2017

Data deposition: This project has been deposited at European Nucleotide Archive (EMBL-EBI) under the accession PRJEB22142. Bacterial strain deposition: Newly sequenced strains are deposited at National Collection of Type Cultures (NCTC; a culture depository of Public Health England).

Abstract

*Clostridium* species (particularly *Clostridium difficile, Clostridium botulinum, Clostridium tetani* and *Clostridium perfringens*) are associated with a range of human and animal diseases. Several other species including *Clostridium tertium, Clostridium cadaveris*, and *Clostridium paraputrificum* have also been linked with sporadic human infections, however there is very limited, or in some cases, no genomic information publicly available. Thus, we isolated one *C. tertium* strain, one *C. cadaveris* strain and three *C. paraputrificum* strains from preterm infants residing within neonatal intensive care units and performed Whole Genome Sequencing (WGS) using Illumina HiSeq. In this report, we announce the open availability of the draft genomes: *C. tertium* LH009, *C. cadaveris* LH052, *C. paraputrificum* LH025, *C. paraputrificum* LH058, and *C. paraputrificum* LH141. These genomes were checked for contamination in silico to ensure purity, and we confirmed species identity and phylogeny using both 16S rRNA gene sequences (from PCR and in silico) and WGS-based approaches. Average Nucleotide Identity (ANI) was used to differentiate genomes from their closest relatives to further confirm speciation boundaries. We also analysed the genomes for virulence-related factors and antimicrobial resistance genes, and detected presence of tetracycline and methicillin resistance, and potentially harmful enzymes, including multiple phospholipases and toxins. The availability of genomic data in open databases, in tandem with our initial insights into the genomic content and virulence traits of these pathogenic *Clostridium* species, should enable the scientific community to further investigate the disease-causing mechanisms of these bacteria with a view to enhancing clinical diagnosis and treatment.

Key words: *Clostridium*, functional annotation, whole genome sequencing, virulence.

Medical Relevance

*Clostridium*, which means “a small spindle” in Greek (due to its rod-shaped morphology), is classified as a genus under the phylum Firmicutes and class Clostridia, and comprises 221 species to date (September 2017) (Parte 2014). *Clostridium* spp. are Gram-positive spore-forming anaerobes found ubiquitously in the environment (soil and water) and the intestinal tract of humans and animals (Yamagishi et al. 1964; Miwa 1975; de Vos et al. 1982). There are several significant human and animal disease causing *Clostridium* species including *Clostridium difficile* (pseudomembranous colitis), *Clostridium botulinum* (infant botulism), *Clostridium tetani* (tetanus), and *Clostridium perfringens* (acute watery diarrhea/necrotising enterocolitis [NEC]), with associated pathology ascribed to the...
There are also several less well-studied species including Clostridium tertium, Clostridium paraputrificum, and Clostridium cadaveris, which have been sporadically reported in the literature to be associated with human infection. C. cadaveris (formerly Clostridium capitovale), is thought to be a key tissue-decomposing bacterium in dead carcasses, and is generally not considered pathogenic in living individuals (Poduval et al. 1999). However, this bacterium has infrequently been associated with human systemic diseases, including intraperitoneal infection (Leung et al. 2009) and bacteremia (Poduval et al. 1999). There is currently no sequenced genome of C. tertium or C. cadaveris available to the research community (apart from 16S rRNA gene sequences) and only four genomes of C. paraputrificum accessible on NCBI databases as of September 2017 (Geer et al. 2010). In this study, we sequenced one C. cadaveris isolate, one C. tertium isolate and three C. paraputrificum isolates from preterm infant faecal samples obtained from two neonatal intensive care units (NICUs) units in England. We identified these using their 16S rRNA gene sequences (both full-length PCR and in silico) and WGS-based k-mer phylogenetic assignment, thus contributing new genomic data on these pathogenic bacteria. We also verified their phylogenetic positions using WGS data, measured genetic distances via Average Nucleotide Identity (ANI), and performed genome-wide functional annotation (COG classification). These genomic data and analyses increase our understanding of the virulence potentials and functionalities of these pathogenic bacteria, with a future view to unraveling disease-causing mechanisms.

### Genome Description

Here, we report the release of draft genomes sequenced on Illumina HiSeq 2500 platform as stated in table 1. C. paraputrificum isolates have a genome size between 3.6 and 3.7 million bases and a stable GC content from 29.6 to 29.9%, which is in line with the four public genomes (Geer et al. 2010). C. tertium has a larger genome (3.9 million bases) and relatively lower GC content of 27.8%, whilst C. cadaveris has a smaller genome (3.4 million bases) compared with C. paraputrificum, and a significantly higher GC content of 31.2%. All draft genomes were assembled using Prokka de novo assembler and 80% (four out of five) of the genomes analyzed were <50 contigs, except for C. paraputrificum LH058 with 84 contigs.
These five strains were isolated from preterm infants residing at two different NICUs (table 1), which is in line with previous findings that report frequent detection of *C. paraputrificum* (16–22%) and *C. tertium* (4–9%) in infant cohorts (Tonooka et al. 2005; Ferraris et al. 2012). However, to date there are no reports of *C. cadaveris* isolation from infants.

**Phylogenetic Positions**

To assign phylogenetic position, and identify these isolates, we computationally extracted 16S rRNA sequences from genomes to construct a *Clostridium* 16S rRNA phylogeny (based on 19 isolates in the NCBI nucleotide database) as in figure 1A. Here, we coupled three genomic approaches to confirm taxonomic position of these newly released genomes. We firstly, performed a PCR targeting almost the full length of the 16S rRNA gene, and predicted the whole 16S rRNA gene sequence in silico. Secondly, we employed Average Nucleotide Identity (ANI) to confirm species boundaries; ANI cut-offs for species discrimination is known to be approximately 95%, and this value has been reported to mirror the traditional taxonomic gold standard method DNA–DNA hybridization (DDH) to define species (Richter and Rossello 2009). Lastly, we performed CVTree—an alignment-free whole genome-based phylogenetic construction method, which is known for speed and accuracy for taxonomic assignment (Xu and Hao 2009).

At a 16S rRNA level, LH058, LH141, and LH025 fall in the same lineage as *C. paraputrificum* DSM2630, indicating species-level relatedness (fig. 1A), with LH052 clustering with *C. cadaveris* JCM1392, and LH009 within the same lineage as *C. tertium* ATCC14573 and *Clostridium sartagoforme* KAR69. CVTree phylogenetic analysis, providing greater resolution based on sequence comparison, showed similar relationships. (fig. 1B); all *C. paraputrificum* isolates grouped within the same lineage as *C. paraputrificum* DSM2630, when compared with other *Clostridium* species, indicating correct species assignment for isolates LH058, LH141, and LO25. *C. cadaveris* LH052 is most closely related to *C. perfringens* ATCC13124, and LH009 (*C. tertium* as assigned according to 16S data) is closely related to *C. sartagoforme* AAU1 (fig. 1B).

We next used ANI analysis to provide higher phylogenetic resolution (fig. 2A). *C. paraputrificum* AGR2156 are identical to LH025, LH141, and LH058 in terms of nucleotide sequences, sharing ANI of >95.7%, thus determined to be the same species. Although LH009 is closely related to *C. sartagoforme* AAU1, the ANI calculation does not allocate these two within the same species (ANI = 83.6%, < 95% as species cut-off), which indicates LH009 is distinct from its closest relatives, and may be identified as the species *C. tertium*. LH052 is also evolutionarily distant (based on ANI calculation, 68.5%) from other *Clostridium*, indicating this isolate is a separate species, *C. cadaveris*.

**Virulence Traits and Genome-Wide Functional Analyses**

Using genome annotations, we performed a thorough search on virulence-related terms including “phospholipase,” “hemolysin,” “resistance,” “lactamase,” “drug,” and “toxin” to provide initial insights into the potential
virulence-linked genes encoded within these genomes (Table 2).

*C. tertium* LH009, *C. cadaveris* LH052, and *C. paraputrificum* LH141 harbour phospholipase genes (*ytpA*) that are homologous to phospholipases encoded in other pathogen genomes including *Bacillus subtilis*, *Pseudomonas aeruginosa*, and *Streptococcus pneumoniae*. Phospholipases are known to possess hydrolytic activity against eukaryotic cell membranes, and are thus considered key virulence factors. *C. perfringens* produces homologous phospholipase C (also known as alpha toxin) that has previously been reported to damage epithelial cells (Verherstraeten et al. 2013), and which shares >58% protein sequence identity with the phospholipase encoded by gene *ytpA*. Importantly, LH052 and LH141 also

---

**Fig. 2.**—(A) Average Nucleotide Identity (ANI) values in representative *Clostridium* genomes. (B) Comparison of functional annotations based on COG classifications on three representative genomes.
and carbapenems) precursor (inactive protein sequence that could potentially be activated via posttranslational modification) was encoded in all genomes (Marciano et al. 2007). The prevalence of multiple antimicrobial resistance genes in these clinical strains may correspond to the environment in which they were isolated; preterm infants residing in NICUs where antimicrobial usage is extensive (Albrich et al. 2004).

From the COG-based genome-wide annotation, most genes (>40% in each genome) did not map with any known functional orthologs, which highlights the limitation of genomic tools and current databases, for understanding these bacteria at a functional level. Gene counts in most categories of these three genomes did not differ significantly from one another (fig. 2B). However, the number of genes involved in carbohydrate metabolism and transport is lower in C. cadaveris LH052 (n = 159), than encoded in C. tertium LH009 (n = 300) and C. paraputrificum LH141 (n = 269), whereas LH052 possesses more genes (n = 249) involved in amino acid metabolism and transport as compared with LH009 (n = 203) and LH141 (n = 212). These functional differences may correspond to divergent modes of metabolism and nutritional substrates for C. cadaveris, which is distinct from C. tertium and C. paraputrificum (correlates to WGS phylogeny positions), and may link to previous isolations of this species from additional environmental niches, i.e. dead carcasses. Therefore, we conclude that these three Clostridium species are similar in terms of genomic functionalities, however due to the high number of function-unknown genes, this somewhat reduces in-depth comparison between genomes and will require further experimental work.

### Materials and Methods

**Faecal Sample Collection**

Fecal sample collection was performed under an on-going preterm infant study (BAMBI) which is approved by University of East Anglia (UEA) Faculty of Medical and Health Sciences (FMH) Ethics Committee. Sample collection was done in accordance with the procedures outlined by National Research Ethics Service (NRES) approved UEA Biorepository (Licence no.: 11208). Participating infants were given written consent by their parents for fecal sample collection at Norfolk and Norwich University Hospital (Norwich, UK) and Rosie Hospital (Cambridge, UK). Fecal samples were routinely collected from infant nappies in the NICUs into sterile stool containers and stored at 4 °C. 

### Bacterial Isolates and Preliminary 16S rRNA PCR Identification

A total of five Clostridium isolates (including C. tertium, C. cadaveris, and C. paraputrificum) were analyzed in this study. Isolates were preliminarily identified using 16S rRNA full-length PCR (Weisburg et al. 1991). Primers used as in table 3.

---

**Table 2**  
**Virulence-Related Genes Detected in Selected Clostridium Genomes**

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Gene Names</th>
<th>Gene Description and Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. tertium LH009</td>
<td>ytpA</td>
<td>Phospholipase</td>
</tr>
<tr>
<td></td>
<td>vanW</td>
<td>Vancomycin B-type resistance</td>
</tr>
<tr>
<td></td>
<td>stp</td>
<td>Multidrug resistance</td>
</tr>
<tr>
<td></td>
<td>mdtK</td>
<td>Multidrug resistance</td>
</tr>
<tr>
<td></td>
<td>tetM</td>
<td>Tetracycline resistance</td>
</tr>
<tr>
<td></td>
<td>marA</td>
<td>Multiple antibiotic resistance</td>
</tr>
<tr>
<td></td>
<td>mecR1</td>
<td>Methicillin resistance</td>
</tr>
<tr>
<td></td>
<td>norM</td>
<td>Multidrug resistance</td>
</tr>
<tr>
<td></td>
<td>mepA</td>
<td>Multidrug export protein</td>
</tr>
<tr>
<td></td>
<td>hcpC</td>
<td>Beta-lactamase precursor</td>
</tr>
<tr>
<td></td>
<td>sme-1</td>
<td>Carbapenem-hydrolyzing beta-lactamase precursor</td>
</tr>
<tr>
<td>C. tertium LH141</td>
<td>ytpA</td>
<td>Phospholipase</td>
</tr>
<tr>
<td></td>
<td>marA</td>
<td>Multiple antibiotic resistance</td>
</tr>
<tr>
<td></td>
<td>vanW</td>
<td>Vancomycin B-type resistance</td>
</tr>
<tr>
<td></td>
<td>norm</td>
<td>Multidrug resistance protein</td>
</tr>
<tr>
<td></td>
<td>tetM</td>
<td>Tetracycline resistance</td>
</tr>
<tr>
<td></td>
<td>marR</td>
<td>Multiple antibiotic resistance</td>
</tr>
<tr>
<td></td>
<td>mdtA</td>
<td>Multidrug resistance</td>
</tr>
<tr>
<td></td>
<td>mepA</td>
<td>Multidrug export protein</td>
</tr>
<tr>
<td></td>
<td>(n/a)</td>
<td>Beta-lactamase precursor</td>
</tr>
<tr>
<td>C. paraputrificum LH141</td>
<td>toxA</td>
<td>Toxin A</td>
</tr>
<tr>
<td></td>
<td>marR</td>
<td>Multiple antibiotic resistance</td>
</tr>
<tr>
<td></td>
<td>tetM</td>
<td>Tetracycline resistance</td>
</tr>
<tr>
<td></td>
<td>norM</td>
<td>Multidrug resistance</td>
</tr>
<tr>
<td></td>
<td>vanW</td>
<td>Vancomycin B-type resistance</td>
</tr>
<tr>
<td></td>
<td>mdtK</td>
<td>Multidrug resistance</td>
</tr>
<tr>
<td></td>
<td>mepA</td>
<td>Multidrug export protein</td>
</tr>
<tr>
<td></td>
<td>(n/a)</td>
<td>Beta-lactamase precursor</td>
</tr>
<tr>
<td></td>
<td>sme-1</td>
<td>Carbapenem-hydrolyzing beta-lactamase precursor</td>
</tr>
</tbody>
</table>

possess toxA gene, which encodes C. difficile-associated Toxin A, known to be one of the main virulence factors during infection having cytotoxic and proinflammatory activities (Awad et al. 2014).

Notably, antimicrobial resistance genes are encoded in all three genomes, including vancomycin (vanW) and tetracycline resistance (tetM) (Evers and Courvalin 1996; Donhofer et al. 2012). Other resistance traits include multidrug efflux pumps; that is, those encoded by mdtK and norM (fluoroquinolones) (Horiyama et al. 2011; Golparian et al. 2014), mepA (aminocoumarin) (Guerrero et al. 2012), and efflux pump transcriptional regulators marA and marR (Maira-Litran et al. 2000). In addition, methicillin resistance gene mecR1 was detected in LH009 (Shore et al. 2011), whilst beta-lactamase (penicillins
Near 1kbp PCR products were subsequently sequenced (Eurofins, Luxembourg) and compared with 16S rRNA bacteria sequence database on NCBI using BLASTn (optimized for megablast) search algorithm (Camacho et al. 2009).

**Genomic DNA Extraction**

Overnight 10 ml pure cultures in BHI were harvested for phenol-chloroform DNA extraction. Briefly, bacterial pellets were resuspended in 2 ml 25% sucrose in 10 mM Tris and 1 mM EDTA at pH 8.0. Cells were lysed using 50 µl 100 mg/ml lysozyme (Roche). 100 µl 20 mg/ml Proteinase K (Roche), 30 µl 10 mg/ml RNase A (Roche), 400 µl 0.5 M EDTA (pH 8.0) and 250 µl 10% Sarkosyl NL30 (Fisher) were added subsequently into the lysed bacterial suspension. This follows by 1-h ice incubation and 50 °C overnight water bath.

Second-day protocol comprises three rounds of phenol-chloroform-isoamyl alcohol (Sigma) extraction using 15 ml gel-lock tubes (Qiagen). Chloroform-isoamyl alcohol (Sigma) extraction was performed to remove residual phenol, followed by ethanol precipitation and 70% ethanol wash. DNA pellets were finally resuspended in 200–300 µl of 10 mM Tris (pH 8.0). DNA concentration was quantified using Qubit dsDNA BR assay kit (Invitrogen) and DNA quality assessed by Nanodrop spectrophotometer.

**Whole Genome Sequencing, Genome Assembly and Annotation**

Isolated DNA of pure cultures was subjected to multiplex standard illumina library preparation protocol followed by sequencing via illumina HiSeq 2500 platform with read length 2 × 125 bp (paired-end reads) and an average sequencing coverage of 60×. Draft genome assemblies were generated using an assembly and annotation pipeline as described previously (Page et al. 2016). All genomes were annotated using Prokka v1.11 (Seemann 2014).

**Contamination Estimation**

Webtool ContEst16S was used to check for potential contamination of the draft genomes based on Genbank database (Lee et al. 2017).

**Alignment-Free WGS Phylogeny**

Selected *Clostridium* genome sequences were retrieved from NCBI genome database. Annotated multiple protein sequences were used as input for CVTree v5.0 to generate alignment-free WGS-based phylogeny using the optimized six as the k-tuple length (Xu and Hao 2009). Tree was edited using iTOL as described in previous section.

**Average Nucleotide Identity (ANI)**

OrthoANI Tool v.093 (OAT) was employed to calculate the ANI (both directions) between genomes (Lee et al. 2016). Identity >95% was used as cut-off for species delineation.

**Genome-Wide Functional Assignment (COG)**

Functional assignments were implemented using eggNOG-mapper v0.99.3 (Huerta-Cepas et al. 2017), based on eggNOG orthology data (Huerta-Cepas et al. 2016). Sequence searches were performed using HMMER3 (Eddy 2011). Data were extracted using Shell scripts (https://github.com/raymond.kiu/eggnog-mapper_COExtraction, last accessed September 20, 2017) and visualized in GraphPad PRISM v5.04.

**Ethics Approval and Consent for Participation**

This study was approved by the University of East Anglia (UEA) Faculty of Medical and Health Sciences (FMH) Ethics Committee. Sample collection follows the protocols outlined by NRES approved UEA Biorepository (License no.: 11208). Written consent was given by the parents for their infants for participation in this study.

**Acknowledgments**

This work was supported by a Wellcome Trust Investigator Award to LJH (grant number 100974/C/13/Z). RK is a PhD student at the Norwich Medical School of University of East Anglia (UEA), partially funded by UEA international bursary.
**Literature Cited**


**Associate editor:** Howard Ochman