JCM
Journals.ASM.org

Real-Time Genomic Epidemiological Evaluation of Human *Campylobacter* Isolates by Use of Whole-Genome Multilocus Sequence Typing

Alison J. Cody,[a] Noel D. McCarthy,[b] Melissa Jansen van Rensburg,[a] Tomide Isinkaye,[b] Stephen D. Bentley,[c] Julian Parkhill,[c] Kate E. Dingle,[d,e] Ian C. J. W. Bowler,[f] Keith A. Jolley,[a] Martin C. J. Maiden[a]

Department of Zoology, University of Oxford, Oxford, United Kingdom[a]; Thames Valley Health Protection Unit, Centre for Radiation, Chemical, and Environmental Hazards, Oxfordshire, United Kingdom[b]; Wellcome Trust Sanger Institute, Cambridgeshire, United Kingdom[c]; Nuffield Department of Clinical Laboratory Sciences, Oxford University, John Radcliffe Hospital, Oxford, United Kingdom[d]; National Institute for Health Research, Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, United Kingdom[e]; Department of Microbiology, John Radcliffe Hospital, Oxford University Hospitals NHS Trust, Oxford, United Kingdom[f]

Sequence-based typing is essential for understanding the epidemiology of *Campylobacter* infections, a major worldwide cause of bacterial gastroenteritis. We demonstrate the practical and rapid exploitation of whole-genome sequencing to provide routine definitive characterization of *Campylobacter jejuni* and *Campylobacter coli* for clinical and public health purposes. Short-read data from 384 *Campylobacter* clinical isolates collected over 4 months in Oxford, United Kingdom, were assembled *de novo*. Contigs were deposited at the pubMLST.org/campylobacter website and automatically annotated for 1,667 loci. Typing and phylogenetic information was extracted and comparative analyses were performed for various subsets of loci, up to the level of the whole genome, using the Genome Comparator and Neighbor-net algorithms. The assembled sequences (for 379 isolates) were diverse and resembled collections from previous studies of human campylobacteriosis. Small subsets of very closely related isolates originated mainly from repeated sampling from the same patients and, in one case, likely laboratory contamination. Much of the within-patient variation occurred in phase-variable genes. Clinically and epidemiologically informative data can be extracted from whole-genome sequence data in real time with straightforward, publicly available tools. These analyses are highly scalable, are transparent, do not require closely related genome reference sequences, and provide improved resolution (i) among *Campylobacter* clonal complexes and (ii) between very closely related isolates. Additionally, these analyses rapidly differentiated unrelated isolates, allowing the detection of single-strain clusters. The approach is widely applicable to analyses of human bacterial pathogens in real time in clinical laboratories, with little specialist training required.

Rapid and cost-effective generation of whole-genome sequence (WGS) data for pathogen isolates will revolutionize treatment and public health control of infectious diseases (1). However, realization of this goal requires that the data generated be presented and analyzed in a way that is accurate, informative, and readily interpreted by clinicians and epidemiologists (2). Several studies have demonstrated the potential of WGS data, establishing relationships among clinical isolates at very high resolution; however, those studies examined isolates known to be similar by previous analysis with multilocus sequence typing (MLST) (3–5). Further, the single-nucleotide polymorphism-based methods employed to generate detailed consensus phylogenies are time-consuming, computationally intensive, and sensitive to the effects of horizontal genetic exchange (3). Consequently, these approaches cannot be readily applied to clinical data in real time, where it is necessary first to identify outbreaks by comparison of routine clinical isolates to each other and the known diversity of the pathogen, with subsequent extension of the analysis by the addition of data on new isolates as the data become available.

Campylobacteriosis, caused by *Campylobacter jejuni* (about 90% of cases) and *Campylobacter coli* (about 10% of cases), is the most common cause of bacterial gastroenteritis worldwide, costing an estimated £0.58 billion *per annum* in the United Kingdom alone (6). Despite its importance, it was recognized only in the 1970s and remains underreported globally (7). Molecular epidemiological approaches, especially MLST, have shown that *Campylobacter* isolates are highly diverse, with much of their diversity being generated by horizontal genetic exchange. This, combined with the infrequency of identified point source outbreaks, has hindered epidemiological studies of these widely distributed bacteria (8). Genetic attribution indicates that farm animals, especially chickens (via retail sale of poultry meat), are the most important sources of human infection in many countries (9–11), resulting in interventions in the food chain to reduce the prevalence of *Campylobacter* in broiler chickens (12). Improved understanding of the epidemiology of *Campylobacter* in humans, farm animals, and the environment is essential for assessment of the effectiveness of these interventions, and WGS analysis is a potentially invaluable tool for achieving this (13).

Here, reference-free *de novo* assembly of whole-genome sequence data from clinical specimens was combined with hierarchical gene-by-gene analysis (14) to investigate 379 *Campylobac-*

Received 15 January 2013 Returned for modification 27 March 2013
Accepted 10 May 2013
Published ahead of print 22 May 2013
Address correspondence to Martin C. J. Maiden, martin.maiden@zoo.ox.ac.uk.
Supplemental material for this article may be found at http://dx.doi.org/10.1128/JCM.00066-13.

doi:10.1128/JCM.00066-13
The authors have paid a fee to allow immediate free access to this article.

ter isolates from 4 months of disease surveillance in Oxfordshire, United Kingdom. This whole-genome multilocus sequence typing (wgMLST) approach is a rapid and efficient means of data analysis and presentation which is backward compatible, generalizable, and automatable. It is computationally nonintensive, is highly sensitive for the identification of outbreaks of infection and even likely laboratory contamination, and is publicly available at pubMLST.org/campylobacter.

MATERIALS AND METHODS

Isolates. The study included 384 *Campylobacter* isolates obtained from 361 individuals between 27 June and 26 October 2011 at the microbiology laboratory of the John Radcliffe Hospital, which has a geographically contiguous catchment of ~600,000 inhabitants (about 1% of the United Kingdom total). Surveillance since 2003 has demonstrated that *Campylobacter* genotypes isolated in this area are similar to those obtained elsewhere in the United Kingdom (15). Isolates identified as previously described (15) were cultured on *Campylobacter* blood-free selective agar (Oxoid Ltd., Basingstoke, United Kingdom) and incubated for 48 h at 42°C in a microaerophilic atmosphere, and a single colony was recultured on Columbia blood agar (Oxoid Ltd.) under the same conditions. Suspensions from both cultures were made in brain heart infusion broth containing 20% glycerol and were stored at −80°C. Genomic DNA was extracted from the growth of single-colony cultures with a Wizard genomic DNA purification kit (Promega, Southampton, United Kingdom).

Whole-genome sequencing. Illumina multiplex libraries were generated with 1 μg of genomic DNA acoustically sheared to 200 to 300 bp using a Covaris E210 device. DNA fragments were end repaired, and a 3′ nontemplate adenosine residue was ligated to the Illumina multiplexing adaptor oligonucleotide for sequencing. Up to 96 libraries were pooled and analyzed together, in equimolar amounts, in a flow cell lane of the Illumina HiSeq 2000, generating 76-bp paired-end reads. Genome sequence data were assembled using Velvet version 1.2.01 shuffle and optimization scripts to create contigs with optimal parameters, with *k*-mer lengths between 49 and 69 bp (16).

Assembled data were deposited in the pubmlst.org/campylobacter database, implemented with Bacterial Isolate Genome Sequence Database (BIGSdb) software (17); at the time of analysis, a total of 1,667 loci were defined in this database, including the MLST loci, ribosomal MLST (rMLST) loci, *porA* antigen-encoding genes, and other loci derived from a number of sources, including the reannotation of the sequence of *Campylobacter jejuni* isolate NCTC11168 (18). The BIGSdb autotagger automatically identified loci, assigned alleles, and tagged the sequences for future reference. The database automatically provided a report of the MLST alleles, sequence types (STs), and clonal complex assignments, along with *porA* alleles.

Genome Comparator analyses. Relationships among isolates were established using phylogenetic networks based on rMLST (19) sequences. The 52 *Campylobacter* ribosomal protein subunit loci identified in the automated annotation process were compared among all isolates using the BIGSdb Genome Comparator module. The distance matrix generated on the basis of shared alleles was visualized with the Neighbor-net algorithm (20), implemented in SplitsTree version 4.8 (21) within the BIGSdb Web interface. A subset of isolates that were indistinguishable by rMLST were further analyzed with the Genome Comparator at 1,643 defined loci to generate a distance matrix based on shared alleles, which was visualized with Neighbor-net. This algorithm does not assume a tree-like structure for the data and resolves interrelationships among isolates as a phylogenetic network where appropriate, thereby accommodating departures from tree-like phylogeny (which can result from horizontal gene transfer, for example).

The number of allelic differences among all possible isolate pairs was counted using the Genome Comparator. This analysis was limited to the 1,026 loci present in all 379 isolates, effectively a "core genome" analysis at the level of shared alleles. The numbers of differences among known isolates from the same patient were also identified. Inspection of the distribution of pairwise differences identified clusters, which were defined as groups of isolates differing at 20 loci or fewer, on the basis of the observed distribution of differences among isolates from the same or different patients. The largest cluster identified in the data set was chosen for further analysis. The isolates belonging to this cluster were compared again with the Genome Comparator at all 1,587 shared loci, and gene differences were analyzed. Common epidemiological features were determined retrospectively from routine laboratory and environmental health investigation notes for cases sharing closely related isolates. A set of instructions for the Genome Comparator analyses is provided in Fig. S1 in the supplemental material.

RESULTS

MLST and clonal complexes. WGS data were obtained from 379/384 (98.7%) *Campylobacter* isolates, without retesting; the isolates were annotated and MLST profiles were generated, giving 126 seven-locus STs, 20 of which were new (13 *C. jejuni* and 7 *C. coli* STs). Of the STs, 104/126 (82.5%) were assigned to 29 clonal complexes, accounting for 351/379 (92.6%) of the isolates, with 28 (7.4%) isolates having STs not assigned to a clonal complex (Fig. 1). There were 15 new MLST alleles (7.0% of the total, including one *aspA*, two *glnA*, two *gltA*, three *glyA*, three *pgm*, two *tkt*, and two *uncA* loci). A total of 48 (13%) isolates had STs typical of *C. coli*, mostly belonging to the ST-828 clonal complex (42 isolates), with one ST-1150 complex isolate and five isolates not assigned to a clonal complex. The rank abundance of the clonal complexes was similar to that reported in previous investigations employing seven-locus MLST (15, 22, 23), with a slightly elevated number of *C. coli* isolates (Fig. 1).

Ribosomal MLST. All 52 rMLST loci identified were variable, while a 53rd locus (*rpmD*; BACT000059) was absent in all isolates; this locus was absent in all *Campylobacter* genomes examined up to the time of analysis. Complete rMLST profiles for 52 *rps* loci were obtained for 376/379 (99.2%) isolates, giving 212 unique rMLST types, 144 of which occurred once. In three isolates, eight loci were truncated, i.e., were located at the end of a contig, resulting in incomplete rMLST profiles; these isolates were not analyzed further. A total of 478 new rMLST alleles were assigned at 51 of the 52 loci, with 1 to 48 new alleles being identified in 87 isolates (see Table S1 in the supplemental material). Neighbor-net phylogenies based on shared rMLST alleles grouped the 376 isolates into clusters that were congruent with MLST clonal complexes (Fig. 2A) and also indicated relationships among the clonal complexes that were not readily apparent from MLST, showing, for example, that the ST-21, ST-48, and ST-206 complexes were related. The *C. coli* ST-828 clonal complex isolates were grouped correctly with the single *C. coli* ST-1150 isolate, although the Neighbor-net phylogeny based on a distance matrix generated with allele numbers, rather than nucleotide sequences, inevitably did not represent the 15% nucleotide sequence divergence between *C. jejuni* and *C. coli* (Fig. 2A).

Whole-genome MLST. The rMLST phylogenies provided resolution within clonal complexes, e.g., the 89 ST-21 clonal complex isolates were resolved by rMLST into three groups, which were mostly congruent with ST designations (Fig. 2B). Among these isolates, 13 rMLST loci were identical and 39 were variable, with 34 unique rMLST types. The members of one of these types, designated strain 3, was analyzed by the Genome Comparator using 1,643 defined loci, 1,595 (97%) of which were present in all 10

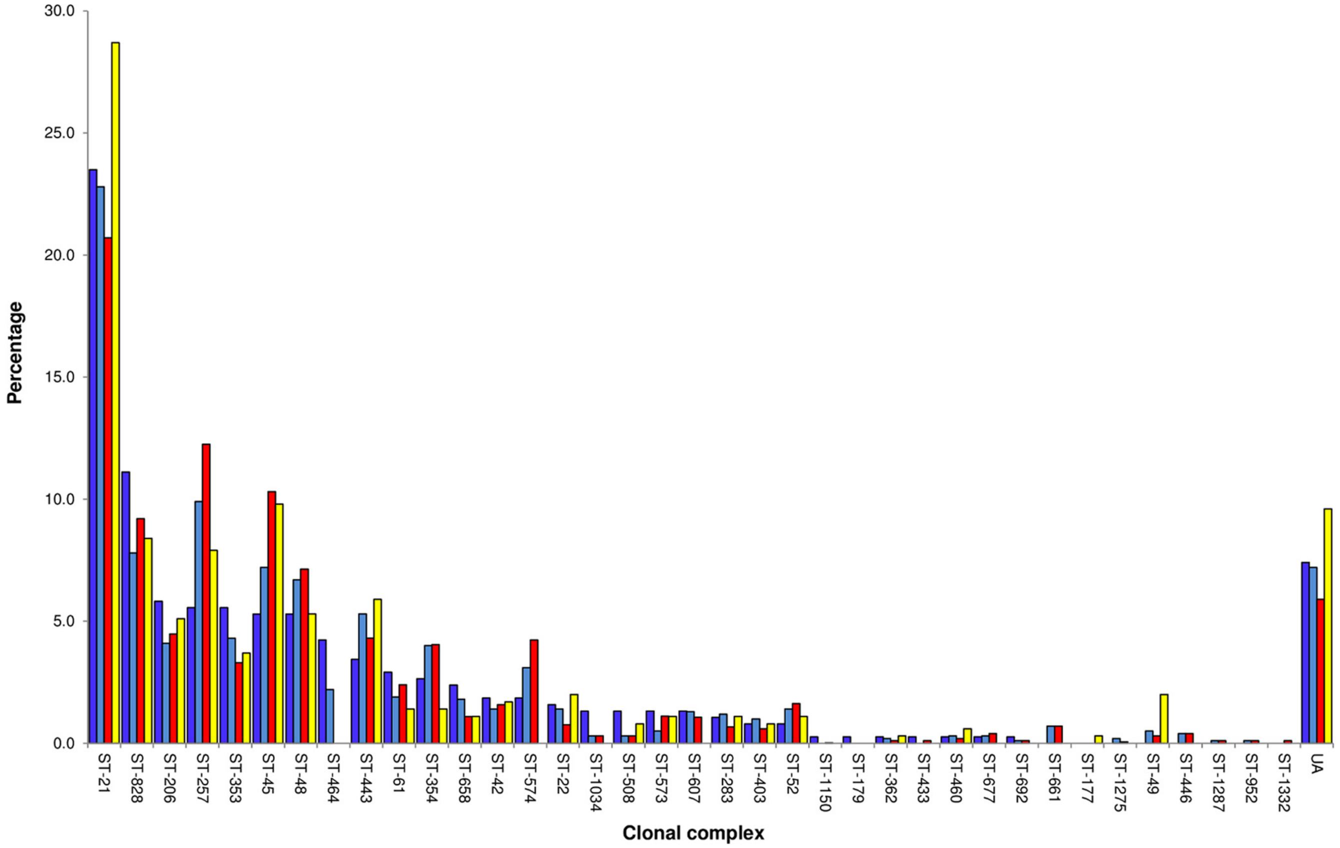

FIG 1 Rank abundance distributions of clonal complexes in genome sequence data from Oxfordshire in 2011 (dark blue) and MLST-typed samples from Oxfordshire from 2003 to 2009 (light blue) (15), Scotland from July 2005 to September 2006 (red) (22), and northwest England from April 2003 to April 2004 (yellow) (23).

strain 3 isolates. This analysis resolved strain 3 into four groups differing at 272 to 351 loci, i.e., one group of five isolates, two groups of two isolates, and one single isolate (Fig. 2C). One of the groups of two contained isolates (OXC6331 and OXC6347) originating from the same patient. Members of the group of five isolates were compared at all 1,605 shared loci; three of these isolates (OXC6266, OXC6286, and OXC6292) were closely related, with just 9 variable loci among them, and two were more distantly related, with 38 (OXC6571) and 61 (OXC6615) locus differences (Fig. 2D). Examination of laboratory and epidemiological records showed that two of the closely related isolates (OXC6266 and OXC6292) were obtained from the same patient on the morning and evening of the same day and the third sample (OXC6286) was obtained from a different patient and received by the microbiology laboratory 3 min after the morning sample from the first patient. These patients did not share other epidemiological features. The patient with two positive samples had returned from international travel and fell ill while visiting part of the catchment area of the hospital. The other patient was a local resident living in a different part of the hospital catchment area.

Whole-genome pairwise comparisons among all 379 isolates (71,631 whole-genome comparisons) using the 1,026 loci which they all shared showed allele differences at 0 to 1,026 loci, with a mean of 877 loci (Fig. 3). Only 0.34% of these comparisons differed at fewer than 12 loci, 0.4% at 20 or fewer, and 1.25% at 100 or fewer. Pairwise comparisons across the 1,026 shared loci among multiple isolates from the same patient produced 10 comparisons with no differences, 14 comparisons with fewer than 10 locus differences (eight comparisons with one difference, three comparisons with two differences, one comparison with three differences, and two comparisons with nine differences), and three comparisons with more than 850 differences (865, 892, and 919 differences).

Pairwise comparisons of the same-patient isolates were repeated using all 1,643 loci for the same-patient isolates with the same ST. This showed between 3 and 14 locus differences among isolates from the same patient (Table 1; see also Table S2 in the supplemental material). The occurrence of different alleles was not random across the genome; three loci showed differences in isolates from 9 patients, while 1,574 loci did not show any within-patient variation in this data set. Loci that recurrently resulted in differentiation between same-patient isolates often differed at homopolymeric tracts and represented loci identified as phase variable in the *C. jejuni* genome (24) or were paralogous gene sequences, i.e., genes that occurred more than once in the genome.

The largest cluster identified within the all-by-all comparison distance matrix, comprising 13 isolates, was further analyzed at the 1,643 defined loci; this analysis identified 1,529 identical (93%) and 58 (3.5%) variable loci (Fig. 4; Table 2), which formed three subclusters. This within-subcluster genome variation was similar to the within-patient variation observed for other samples in this data set, in both numbers and types of genetic differences,

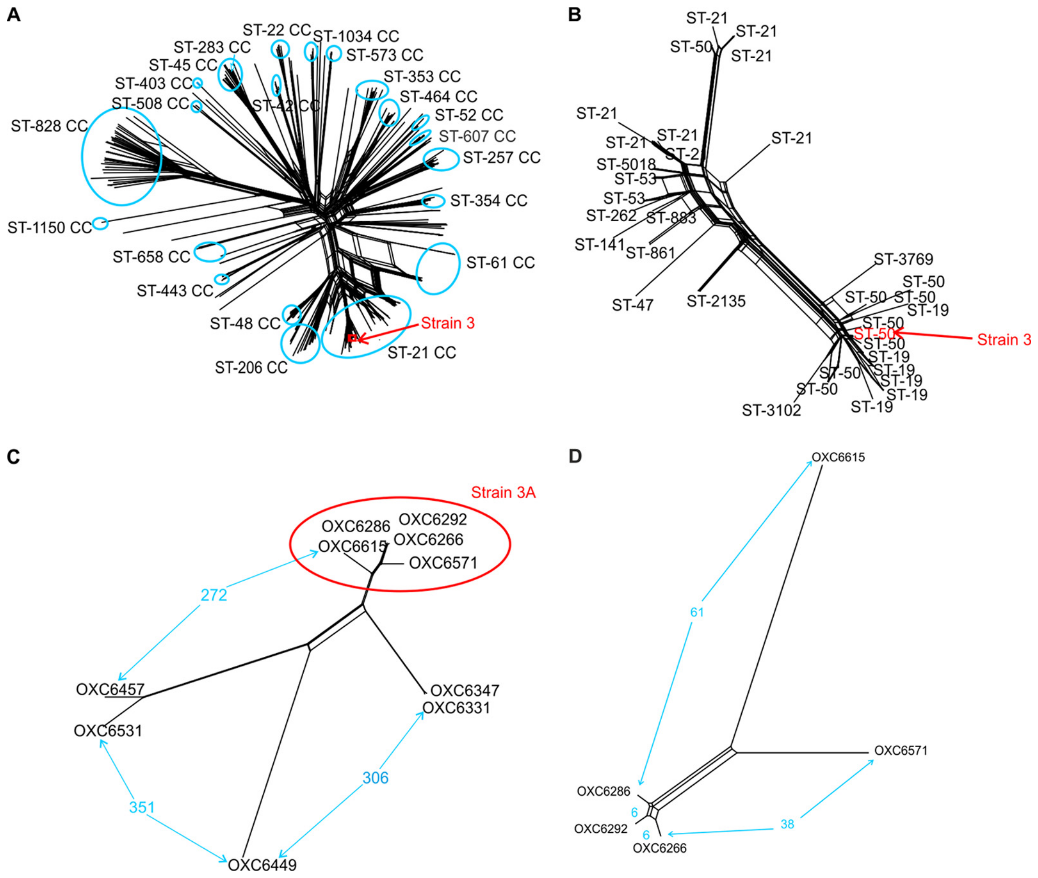

FIG 2 Neighbor-net phylogenies generated for alleles of rMLST loci for 376 *Campylobacter* isolates labeled by clonal complex (A), rMLST loci for 89 ST-21 complex isolates representing 34 rMLST types labeled by ST (B), 1,595 loci for 10 isolates belonging to ST-21 complex rMLST strain 3 (C), and 1,605 loci for a subset of five strain 3 isolates, two of which (OXC6266 and OXC6292) were isolated from the same patient on the same day (D). Numerals in blue type indicate the distances (numbers of allelic differences) between isolates.

while the between-subcluster differences involved a wider and different range of loci (Table 3). The dates of illness onset within subclusters were similar (2 cases were 22 days apart, 2 cases were 7 days apart, and 3 cases occurred within 13 days), compared to the average time differences in the data set. No epidemiological links were apparent on review of the available routine follow-up questionnaires.

For 10 isolates, repeat sequencing was undertaken with the same DNA specimens, to assess the reproducibility of the approaches used. Genome Comparator analyses at all of the 1,643 loci present in the isolates (1,478 to 1,586 loci, depending on the isolate) demonstrated high reproducibility between replicates (99.56 to 99.94% identical loci) (see Table S3 in the supplemental material); the few differences observed were likely due to different assemblies at paralogous loci. None of these differences would have altered the interpretation of any of the analyses described here.

DISCUSSION

The genetic analysis of human campylobacteriosis isolates presents a number of challenges, as it is caused by two related bacterial species which are highly diverse and frequently participate in horizontal genetic exchange (25, 26). These properties, which are common among bacterial pathogens, complicate conventional phylogenetic analyses because frequent recombination introduces multiple polymorphisms, while point mutations, which are rare, introduce only single polymorphisms. The wgMLST approach for the analysis of WGS data assembled *de novo* accommodates both the diversity and the population structure of such bacteria, as sequence variations are effectively summarized and each type of genetic change is treated as a single evolutionary event. The approach also enables hierarchical analyses, permitting the successive examination of different sets of loci according to different clinical and epidemiological questions (13). Here, genome sequences from nearly 400 isolates were examined across the entire range of resolutions required for clinical and epidemiological investigations, from placing them in the global context of *Campylobacter* diversity to defining the nature and extent of within-patient variations. The analyses were conducted rapidly with a Web-based interface, allowing the generation of reports that would be of immediate use to clinicians and which can be generated without specialist bioinformatics training (17).

The deposition of the assembled sequences in the pubMLST.org/campylobacter database, with automated generation of MLST information, ensured that the data were comparable to findings for the >18,000 *Campylobacter* isolates typed previously (8, 15). This demonstrated that the genotypes of the isolates were (i) similar to those seen in Oxfordshire in 2003 to 2009 and in high-income countries

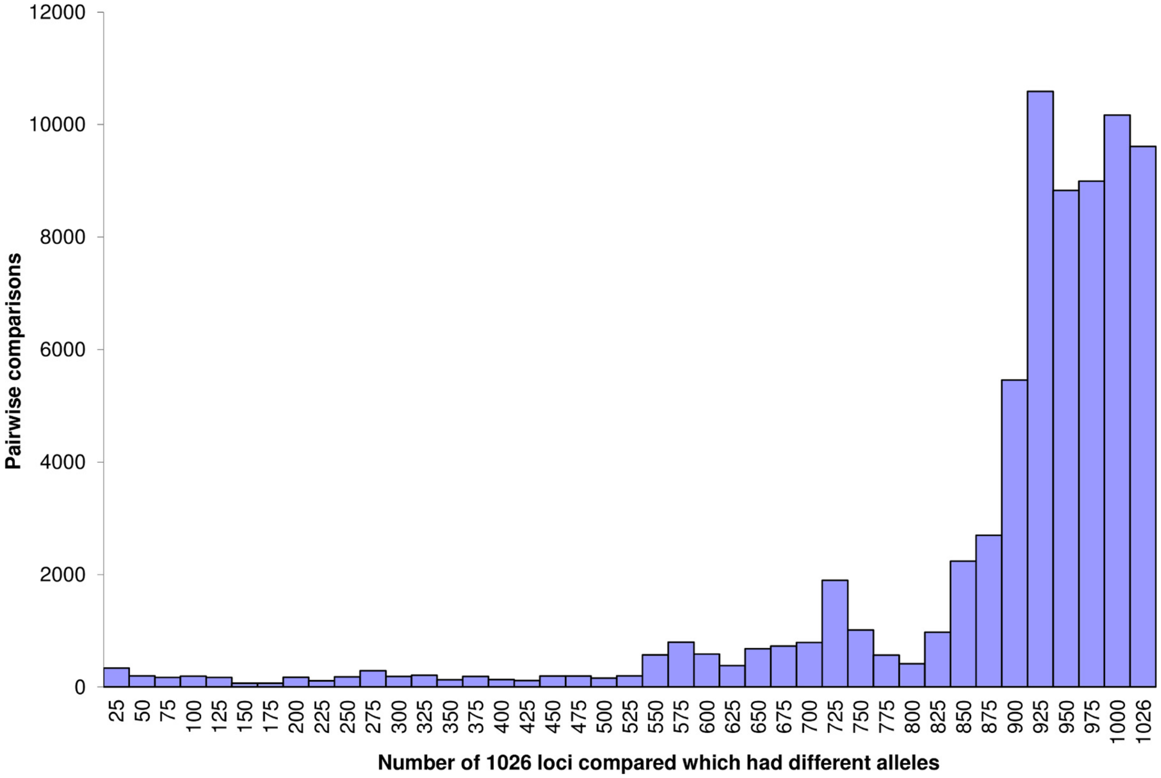

FIG 3 Distribution of the numbers of locus differences among pairwise comparisons of 379 *Campylobacter* isolates at 1,026 shared loci.

generally (Fig. 1) (8, 15) and (ii) consistent with the attribution of the majority of campylobacteriosis cases to chicken meat (9–11, 15, 27, 28). MLST has proved invaluable for understanding the epidemiology of many bacteria, including *Campylobacter* infections in humans and animals (8); however, it is most suited to assigning isolates to clonal complexes and, without additional loci, has little power to detect differences within and relationships among *Campylobacter* clonal complexes (8). The analysis of the 52 rMLST (19) loci confirmed that MLST clonal complexes were genealogically coherent and it provided further information in two respects, (i) improved resolution within clonal complexes, enabling resolution of groups with greater discrimination than achieved with MLST, and (ii) information on the relationships among a number of clonal complexes, for example, the ST-21, ST-48, and ST-206 clonal complexes. The latter clonal complexes typically are isolated from livestock as well as from human patients and include *Campylobacter* strains that are adapted to the agricultural environment (29) (Fig. 2). Therefore, these data provide the prospect of improved genetic attribution with respect to source, an important means of understanding the impact of interventions in the food chain on the reduction of campylobacteriosis (12).

The rMLST and wgMLST analyses confirmed the great diversity of campylobacteriosis isolates recovered in Oxfordshire over only 4 months (Fig. 2); indeed, this was not exceeded by equivalent data for a collection of 83 cattle, human disease, pig, and poultry isolates obtained from several countries over more than 10 years (see Fig. S2 in the supplemental material) (30). The diversity of the Oxfordshire isolates was also evident from the pairwise comparisons of all 379 isolates at the 1,026 loci which they all shared; a large number of isolates differed at all of these loci, and only a very small number differed at fewer than 500 loci. This showed that, notwithstanding the complexities of *Campylobacter* population structure, it is relatively straightforward to employ these data to determine whether *Campylobacter* isolates are likely to be part of a transmission network.

Two approaches were used to identify clusters, i.e., a hierarchical approach, using increasing numbers of loci to identify related isolates, and all-against-all pairwise comparison. These yielded comparable results and represent alternative means of identifying campylobacteriosis outbreaks, which are known to be rare and difficult to detect (31). The first approach is most appropriate when examining a group of isolates known or suspected to form an outbreak, as has been demonstrated for meningococcal disease (14), while the second approach is suited to identifying outbreaks without *a priori* information, such as the data set presented here; this can be automated to provide warnings of outbreaks from routine surveillance data.

Three patients with multiple isolates were shown to have infections involving more than one strain, at the level of seven-locus MLST. The remaining 17 patients with multiple samples yielded very similar isolates at the genomic level, with 3 to 14 locus differences within the same patient (see Table S2 in the supplemental material). Comparison with the data from the 10 repeat-sequencing experiments suggested that this was biological variation and that each patient was infected with a number of sequence variants at this level of resolution. Consequently, it would not be possible to estimate mutation rates from these data without more-extensive sampling from the same patient.

On the basis of these observations, a value of 20 locus differences between isolates was chosen as the cutoff value below which possible clusters were investigated. Cases both clustered in time

TABLE 1 Loci at which polymorphisms were identified between samples of the same sequence type, obtained from the same patient

Locus	Function	No. of patients with polymorphisms
CAMP0044 (*Cj0045c*)	Putative iron-binding protein	9
CAMP1223 (*Cj1305c*)[a]	Hypothetical protein	9
CAMP1224 (*Cj1306c*)[a]	Hypothetical protein	9
CAMP0631 (*cipA*)	Invasion protein CipA	5
CAMP0575 (*Cj0617*)[a]	Hypothetical protein	4
CAMP1413 (*Cj1506c*)	Putative MCP-type signal transduction protein	4
CAMP1228 (*Cj1310c*)	Hypothetical protein	4
CAMP1213 (*Cj1295*)	Conserved hypothetical protein	3
CAMP1214 (*Cj1296*)	Hypothetical protein	3
CAMP1236 (*maf1*)[b]	Motility accessory factor	3
CAMP1250 (*pseD*)[b]	PseD protein	3
CAMP1252 (*maf4*)[b]	Motility accessory factor	3
CAMP1253 (*pseE*)[b]	PseE protein	3
CAMP1257 (*maf6*)[b]	Motility accessory factor	3
CAMP0254 (*cheV*)	Chemotaxis protein	2
CAMP0637 (*Cj0691*)	Hypothetical protein	2
CAMP0751 (*Cj0816*)	Hypothetical protein	2
CAMP0974 (*cjeI*)	Restriction modification enzyme	2
CAMP1031 (*Cj1110c*)	Putative MCP-type signal transduction protein	2
CAMP1258 (*maf7*)[a]	Motility accessory factor	2
CAMP1331 (*Cj1420c*)	Putative methyltransferase	2
CAMP1256 (*Cj1340c*)[b]	Hypothetical protein	2
CAMP0010 (*rnhB*)	Probable RNase HII	1
CAMP0019 (*Cj0019c*)	Putative MCP-domain signal transduction protein	1
CAMP0052 (*fliM*)	Flagellar motor switch protein FliM	1
CAMP0068 (*cdtB*)	Cytolethal distending toxin B	1
CAMP0157 (*Cj0170*)	Hypothetical protein	1
CAMP0167 (*tonB1*)	Putative TonB-dependent outer membrane receptor	1
CAMP0202 (*nrdF*)	Ribonucleotide diphosphate reductase subunit beta	1
CAMP0208 (*cynT*)	Carbonic anhydrase	1
CAMP0219 (*Cj0249*)	Hypothetical protein	1
CAMP0252 (*cheW*)	Chemotaxis protein	1
CAMP0253 (*cheA*)	Chemotaxis histidine kinase	1
CAMP0363 (*gatC*)	Aspartyl/glutamyl-tRNA amidotransferase subunit C	1
CAMP0573 (*pstA*)	Putative phosphate-transport system permease protein	1
CAMP0586 (*Cj0628*)	Putative lipoprotein	1
CAMP0624 (*kdpB*)	Potassium-transporting ATPase subunit B	1
CAMP0687 (*Cj0741*)	Hypothetical protein	1
CAMP0714 (*tpx*)	Thiol peroxidase	1
CAMP0749 (*Cj0814*)	Hypothetical protein	1
CAMP0750 (*Cj0815*)	Hypothetical protein	1
CAMP0803 (*Cj0874c*)	Putative cytochrome *c*	1
CAMP0822 (*aroA*)	3-Phosphoshikimate 1-carboxyvinyltransferase	1
CAMP0849 (*cheR*)	Putative MCP methyltransferase	1
CAMP0851 (*rpiB*)	Ribose 5-phosphate isomerase	1
CAMP0898 (*Cj0975*)	Putative outer membrane protein	1
CAMP0899 (*Cj0976*)	Putative methyltransferase	1
CAMP0950 (*gyrA*)	DNA gyrase subunit A	1
CAMP1064 (*neuA1*)	Two-domain bifunctional protein	1
CAMP1097 (*Cj1178c*)	Highly acidic protein	1
CAMP1108 (*cetB*)	Bipartate energy taxis response protein CetB	1
CAMP1109 (*cetA*)	Bipartate energy taxis response protein CetA	1
CAMP1123 (*atpB*)	F_0F_1-ATP synthase subunit A	1
CAMP1124 (*radA*)	DNA repair protein RadA	1
CAMP1143 (*Cj1224*)	Putative iron-binding protein	1
CAMP1165 (*uvrC*)	Excinuclease ABC subunit C	1
CAMP1181 (*racS*)	Two-component sensor (histidine kinase)	1
CAMP1243 (*Cj1325*)	Putative methyltransferase	1
CAMP1255 (*flaA*)	Flagellin	1
CAMP1283 (*Cj1367c*)	Putative nucleotidyltransferase	1
CAMP1396 (*ccoQ*)	Cb-type cytochrome *c* oxidase subunit IV	1
CAMP1422 (*Cj1516*)	Putative periplasmic oxidoreductase	1
CAMP1442 (*Cj1542*)	Putative allophanate hydrolase subunit 1	1
CAMP1510 (*chuA*)	Hemin uptake system outer membrane receptor	1
CAMP1527 (*Cj1631c*)	Conserved hypothetical protein	1
CAMP1572 (*Cj1677*)	Putative lipoprotein	1
CAMP1581 (*secY*)	Preprotein translocase subunit SecY	1
CAMP1631 (*kdpA*)	Pseudogene (potassium-transporting ATPase A chain)	1

[a] Member of the 617 family of variable genes (24).

[b] Member of the 1318 family of variable genes (24).

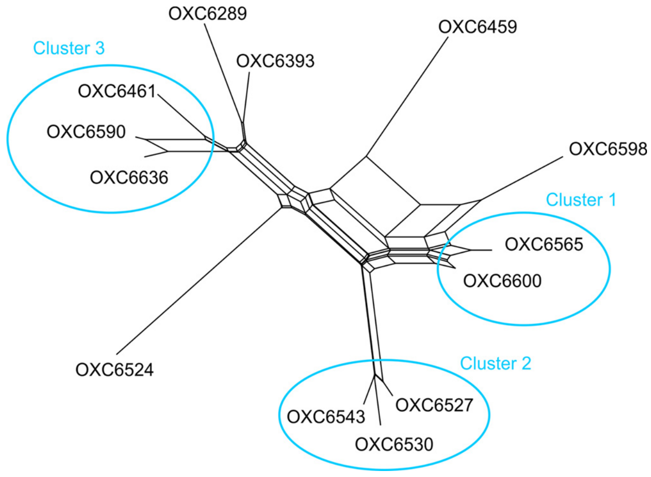

FIG 4 Neighbor-net phylogeny of a cluster of 13 isolates identified by pairwise comparisons of 1,026 shared loci. Blue circles, subclusters within the cluster, representing potential outbreaks.

and exhibiting minimal differences across the genome indicated how this approach can identify epidemiologically related clusters, while a case of likely laboratory cross-contamination demonstrated how this approach can provide an accessible quality assurance tool for clinical microbiology laboratories. Although a retrospective review of the limited epidemiological data collected routinely did not establish direct transmission in any of the clusters identified, in the context of the diversity present among the isolates as a whole it is highly likely that these closely related genotypes from temporally clustered cases shared a common infection source, which would support source identification if investigated in a timely way.

Routine analysis of clinical isolates must characterize samples affordably and rapidly and must identify both long-term trends in disease patterns and occasional disease outbreaks, as exemplified by the *Escherichia coli* hemolytic-uremic syndrome outbreak in Germany in 2011 (32). Such outbreaks are unpredictable and, although reference sequences of the causative agent are unlikely to be immediately available, it is essential to place isolates rapidly within the overall diversity of the species in question, for effective epidemiological investigation and intervention. A universal portable approach for all or most pathogenic bacteria is highly desirable, as demonstrated by the success of MLST (8).

The approach described here, which exploits open-source tools, can be combined with rapid sequence determination platforms (33) in a flexible and scalable analysis pipeline. Whole-genome sequence data can be generated *de novo* and assembled within 48 h (13). Deposition of these data in a BIGSdb database allows species identification with rMLST, which takes a matter of minutes (19). Once the species has been identified, relationships to other isolates, up to the level of whole genomes (wgMLST), can be established rapidly using the Genome Comparator and Neighbor-net, as described here. Again, this analysis takes a matter of minutes for small numbers of isolates and can be scaled hierarchically to larger numbers of comparisons. Therefore, this system can interpret the output of current high-throughput sequencing equipment in real time. Very high levels of diversity among *Campylobacter* isolates were handled without difficulty with this method, offering sufficient discrimination to identify both within- and among-host pathogen diversity and epidemiological clusters.

While this approach presents many opportunities, a number of challenges remain. For example, accepted means of encapsulating the very high levels of discrimination described here and elsewhere into nomenclature schemes have yet to be devised. In the future, it may be necessary to adopt nomenclatures based on 53-locus rMLST or wgMLST data, but the generation of meaningful plain-language interpretation of such data is challenging. However, for the majority of applications, it is not yet necessary to

TABLE 2 Loci at which polymorphisms were identified between samples within clusters of a group of 13 related isolates, as identified by whole-genome MLST

		Allelic variant present							
		Cluster 1		Cluster 2			Cluster 3		
Locus	Function	OXC6565	OXC6600	OXC6527	OXC6530	OXC6543	OXC6461	OXC6590	OXC6636
BACT000040 (*rplK*)	50S ribosomal protein	129	129	129	1332	129	1332	129	129
CAMP0117 (*Cj0129c*)	Outer membrane protein	26	26	26	26	26	26	100	100
CAMP0848 (*pebC*)	ABC-type amino acid transporter ATP-binding protein	1	1	1	1	1	1	55	55
CAMP0850 (*cheB*')	Putative MCP-glutamate methylesterase	1	1	1	1	1	48	1	1
CAMP0898 (*Cj0975*)[a]	Putative outer membrane protein	8	8	8	8	8	15	8	8
CAMP1031 (*Cj1110c*)[a]	Putative MCP-type signal transduction protein	1	1	55	56	1	55	55	1
CAMP1097 (*Cj1178c*)[a]	Highly acidic protein	31	51	51	31	43	31	51	43
CAMP1223 (*Cj1305c*)[a]	Hypothetical protein	39	39	39	102[b]	103[b]	56	56	39
CAMP1331 (*Cj1420c*)[a]	Putative methyltransferase	68[b]	40	40	40	40	40	40	40
CAMP1413 (*Cj1506c*)[a]	Putative MCP-type signal transduction protein	51	1	1	1	1	1	1	1
CAMP1558 (*Cj1663*)	Putative ABC transport system ATP-binding protein	21	21	21	21	21	21	60	60

[a] Locus also identified in same-patient isolates.

[b] Involves a change in the homopolymeric tract length of the assigned allele, compared with the consensus sequence for this open reading frame in strain NCTC11168.

TABLE 3 Loci at which polymorphisms were identified in samples between clusters in a group of 13 related isolates, as identified by whole-genome MLST

		Allelic variant present							
		Cluster 1		Cluster 2			Cluster 3		
Locus	Function	OXC6565	OXC6600	OXC6527	OXC6530	OXC6543	OXC6461	OXC6590	OXC6636
CAMP0061 (*Cj0069*)	Hypothetical protein	78	78	1	1	1	1	1	1
CAMP0064 (*Cj0074c*)	Putative iron-sulfur protein	1	1	1	1	1	79	79	79
CAMP0157 (*Cj0170*)[a]	Hypothetical protein	10[b]	10[b]	1	1	1	1	1	1
CAMP0448 (*Cj0486*)	Putative sugar transporter	6	6	24	24	24	6	6	6
CAMP0490 (*icd*)	Isocitrate dehydrogenase	25	25	25	25	25	86	86	86
CAMP0610 (*Cj0653c*)	Putative aminopeptidase	100	100	100	100	100	21	21	21
CAMP0631 (*cipA*)[a]	Invasion protein CipA	4	4	4	4	4	7[b]	7[b]	7[b]
CAMP0741 (*dapA*)	Putative dihydrodipicolinate synthase	1	1	74	74	74	1	1	1
CAMP0745 (*nadE*)	NAD synthetase	62	62	62	62	62	24	24	24
CAMP1006 (*mfd*)	Transcription-repair coupling factor	111	111	111	111	111	39	39	39
CAMP1019 (*pyrB*)	Aspartate carbamoyltransferase catalytic subunit	41	41	82	82	82	41	41	41
CAMP1026 (*smpB*)	SsrA-binding protein	1	1	53	53	53	1	1	1
CAMP1043 (*Cj1122c*)	Putative integral membrane protein	6	6	2	2	2	2	2	2
CAMP1056 (*Cj1135*)	Two-domain glucosyltransferase	1	1	1	1	1	70	70	70
CAMP1063 (*neuC1*)	Putative UDP-*N*-acetylglucosamine 2-epimerase	51[b]	51[b]	1	1	1	1	1	1
CAMP1133 (*Cj1214c*)	Putative exporting protein	1	1	1	1	1	57	57	57
CAMP1178 (*porA*)	Major outer membrane protein	42	42	145	145	145	42	42	42
CAMP1309 (*feoB*)	Ferrous iron transport protein	20	20	20	20	20	85	85	85
CAMP1432 (*Cj1532*)	Putative periplasmic protein	87	87	87	87	87	1	1	1

[a] Locus also identified in same-patient isolates.

[b] Involves a change in the homopolymeric tract length of the assigned allele, compared with the consensus sequence for this open reading frame in strain NCTC11168.

change the accepted nomenclature schemes (for example, those based on MLST), as these adequately meet the requirements of many practical applications. Even in the absence of more-discriminatory or whole-genome-based nomenclature, representations of genomic data in two-dimensional networks, as used here, are readily generated, are easily understood by practitioners, and are all that is necessary for management of individual disease clusters. In conclusion, with the imminent deployment of benchtop genome sequencers in the clinical laboratory, our approach represents a practical and effective means of exploiting WGS data for clinical benefit.

ACKNOWLEDGMENTS

This work was funded by the United Kingdom Department for Environment, Food, and Rural Affairs and Food Standards Agency (contract OZ0624) and the Wellcome Trust (grants 087622 [to M.C.J.M.] and 098051 [to S.D.B. and J.P.]). K.E.D. was supported by the NIHR Biomedical Centre, Oxford, United Kingdom. M.C.J.M. is a Wellcome Trust Senior Research Fellow.

J.P. has received funding for conference travel and accommodation from Illumina Inc.

We thank the staff of the Clinical Microbiology Laboratory, John Radcliffe Hospital, Oxford, United Kingdom, for routine collection and provision of *Campylobacter* isolates.

REFERENCES

1. **Fournier P-E, Drancourt M, Raoult D.** 2007. Bacterial genome sequencing and its use in infectious diseases. Lancet Infect. Dis. **7**:711–723.
2. **Köser CU, Holden MT, Ellington MJ, Cartwright EJ, Brown NM, Ogilvy-Stuart AL, Hsu LY, Chewapreecha C, Croucher NJ, Harris SR, Sanders M, Enright MC, Dougan G, Bentley SD, Parkhill J, Fraser LJ, Betley JR, Schulz-Trieglaff OB, Smith GP, Peacock SJ.** 2012. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. N. Engl. J. Med. **366**:2267–2275.
3. **Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS, Pichon B, Baker S, Parry CM, Lambertsen LM, Shahinas D, Pillai DR, Mitchell TJ, Dougan G, Tomasz A, Klugman KP, Parkhill J, Hanage WP, Bentley SD.** 2011. Rapid pneumococcal evolution in response to clinical interventions. Science **331**:430–434.
4. **Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, Edgeworth JD, de Lencastre H, Parkhill J, Peacock SJ, Bentley SD.** 2010. Evolution of MRSA during hospital transmission and intercontinental spread. Science **327**: 469–474.
5. **Eyre DW, Golubchik T, Gordon NC, Bowden R, Piazza P, Batty EM, Ip CL, Wilson DJ, Didelot X, O'Connor L, Lay R, Buck D, Kearns AM, Shaw A, Paul J, Wilcox MH, Donnelly PJ, Peto TE, Walker AS, Crook DW.** 2012. A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and *Clostridium difficile* for outbreak detection and surveillance. BMJ Open **2**:e001124. doi:10.1136/bmjopen-2012-001124.
6. **Strachan NJC, Forbes KJ.** 2010. The growing UK epidemic of human campylobacteriosis. Lancet **376**:665–667.
7. **Allos B.** 2001. *Campylobacter jejuni* infections: update on emerging issues and trends. Clin. Infect. Dis. **32**:1201–1206.
8. **Colles FM, Maiden MC.** 2012. *Campylobacter* sequence typing databases: applications and future prospects. Microbiology **158**:2695–2709.
9. **Mullner P, Spencer SE, Wilson DJ, Jones G, Noble AD, Midwinter AC, Collins-Emerson JM, Carter P, Hathaway S, French NP.** 2009. Assigning the source of human campylobacteriosis in New Zealand: a comparative genetic and epidemiological approach. Infect. Genet. Evol. **9**:1311–1319.
10. **Wilson DJ, Gabriel E, Leatherbarrow AJH, Cheesbrough J, Gee S, Bolton E, Fox A, Fearnhead P, Hart A, Diggle PJ.** 2008. Tracing the source of campylobacteriosis. PLoS Genet. **4**:e1000203. doi:10.1371/journal.pgen.1000203.
11. **Sheppard SK, Dallas JF, Strachan NJ, MacRae M, McCarthy ND, Wilson DJ, Gormley FJ, Falush D, Ogden ID, Maiden MC, Forbes KJ.** 2009. *Campylobacter* genotyping to determine the source of human infection. Clin. Infect. Dis. **48**:1072–1078.

12. **Sears A, Baker MG, Wilson N, Marshall J, Muellner P, Campbell DM, Lake RJ, French NP.** 2011. Marked campylobacteriosis decline after interventions aimed at poultry, New Zealand. Emerg. Infect. Dis. **17:**1007–1015.
13. **Sheppard SK, Jolley KA, Maiden MCJ.** 2012. A gene-by-gene approach to bacterial population genomics: whole genome MLST of *Campylobacter*. Genes **3:**261–277.
14. **Jolley KA, Hill DM, Bratcher HB, Harrison OB, Feavers IM, Parkhill J, Maiden MC.** 2012. Resolution of a meningococcal disease outbreak from whole genome sequence data with rapid web-based analysis methods. J. Clin. Microbiol. **50:**3046–3053.
15. **Cody AJ, McCarthy NM, Wimalarathna HL, Colles FM, Clark L, Bowler IC, Maiden MC, Dingle KE.** 2012. A longitudinal six-year study of the molecular epidemiology of clinical *Campylobacter* isolates in Oxfordshire, UK. J. Clin. Microbiol. **50:**3193–3201.
16. **Zerbino D.** 2010. Using the Velvet de novo assembler for short-read sequencing technologies. Curr. Protoc. Bioinformatics **31:**11.5.1–11.5.12.
17. **Jolley KA, Maiden MC.** 2010. BIGSdb: scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics **11:**595. doi:10.1186/1471-2105-11-595.
18. **Gundogdu O, Bentley SD, Holden MT, Parkhill J, Dorrell N, Wren BW.** 2007. Re-annotation and re-analysis of the *Campylobacter jejuni* NCTC11168 genome sequence. BMC Genomics **8:**162. doi:10.1186/1471-2164-8-162.
19. **Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony CM, Colles FM, Wimalarathna HM, Harrison OB, Sheppard SK, Cody AJ, Maiden MC.** 2012. Ribosomal multi-locus sequence typing: universal characterization of bacteria from domain to strain. Microbiology **158:**1005–1015.
20. **Bryant D, Moulton V.** 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. **21:**255–265.
21. **Huson DH, Bryant D.** 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. **23:**254–267.
22. **Sheppard SK, Dallas JF, MacRae M, McCarthy ND, Sproston EL, Gormley FJ, Strachan NJ, Ogden ID, Maiden MC, Forbes KJ.** 2009. *Campylobacter* genotypes from food animals, environmental sources and clinical disease in Scotland 2005/6. Int. J. Food Microbiol. **134:**96–103.
23. **Sopwith W, Birtles A, Matthews M, Fox A, Gee S, Painter M, Regan M, Syed Q, Bolton E.** 2006. *Campylobacter jejuni* multilocus sequence types in humans, northwest England, 2003–2004. Emerg. Infect. Dis. **12:**1500–1507.
24. **Parkhill J, Wren BW, Mungall K, Ketley JM, Churcher C, Basham D, Chillingworth T, Davies RM, Feltwell T, Holroyd S, Jagels K, Karlyshev AV, Moule S, Pallen MJ, Penn CW, Quail MA, Rajandream MA, Rutherford KM, van Vliet AH, Whitehead S, Barrell BG.** 2000. The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. Nature **403:**665–668.
25. **Wilson DJ, Gabriel E, Leatherbarrow AJ, Cheesbrough J, Gee S, Bolton E, Fox A, Hart CA, Diggle PJ, Fearnhead P.** 2009. Rapid evolution and the importance of recombination to the gastroenteric pathogen *Campylobacter jejuni*. Mol. Biol. Evol. **26:**385–397.
26. **Sheppard SK, McCarthy ND, Falush D, Maiden MC.** 2008. Convergence of *Campylobacter* species: implications for bacterial evolution. Science **320:**237–239.
27. **Mullner P, Collins-Emerson JM, Midwinter AC, Carter P, Spencer SEF, van der Logt P, Hathaway S, French NP.** 2010. Molecular epidemiology of *Campylobacter jejuni* in a geographically isolated country with a uniquely structured poultry industry. Appl. Environ. Microbiol. **76:**2145–2154.
28. **Mickan L, Doyle R, Valcanis M, Dingle KE, Unicomb L, Lanser J.** 2007. Multilocus sequence typing of *Campylobacter jejuni* isolates from New South Wales, Australia. J. Appl. Microbiol. **102:**144–152.
29. **Sheppard SK, Colles FM, McCarthy ND, Strachan NJ, Ogden ID, Forbes KJ, Dallas JF, Maiden MC.** 2011. Niche segregation and genetic structure of *Campylobacter jejuni* populations from wild and agricultural host species. Mol. Ecol. **20:**3484–3490.
30. **Lefebure T, Pavinski Bitar PD, Suzuki H, Stanhope MJ.** 2010. Evolutionary dynamics of complete *Campylobacter* pan-genomes and the bacterial species concept. Genome Biol. Evol. **2:**646–655.
31. **Clark CG, Taboada E, Grant CC, Blakeston C, Pollari F, Marshall B, Rahn K, Mackinnon J, Daignault D, Pillai D, Ng LK.** 2012. Comparison of molecular typing methods useful for detecting clusters of *Campylobacter jejuni* and *C. coli* isolates through routine surveillance. J. Clin. Microbiol. **50:**798–809.
32. **Frank C, Werber D, Cramer JP, Askar M, Faber M, an der Heiden M, Bernard H, Fruth A, Prager R, Spode A, Wadl M, Zoufaly A, Jordan S, Kemper MJ, Follin P, Muller L, King LA, Rosner B, Buchholz U, Stark K, Krause G, HUS Investigation Team.** 2011. Epidemic profile of Shiga-toxin-producing *Escherichia coli* O104:H4 outbreak in Germany. N. Engl. J. Med. **365:**1771–1780.
33. **Loman NJ, Constantinidou C, Chan JZM, Halachev M, Sergeant M, Penn CW, Robinson ER, Pallen MJ.** 2012. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat. Rev. Microbiol. **10:**599–606.