Comparative phylogenomics of the food-borne pathogen *Campylobacter jejuni* reveals genetic markers predictive of infection source

Olivia L. Champion*, Michael W. Gaunt*, Ozan Gundogdu*, Abdi Elmi*, Adam A. Witney†, Jason Hinds†, Nick Dorrell*, and Brendan W. Wren*‡

*Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom; and †Bacterial Microarray Group, Medical Microbiology, Department of Cellular and Molecular Medicine, St. George's, University of London, Cranmer Terrace, London SW17 0RE, United Kingdom

Edited by Stanley Falkow, Stanford University, Stanford, CA, and approved September 12, 2005 (received for review May 12, 2005)

***Campylobacter jejuni* is the predominant cause of bacterial gastroenteritis worldwide, but traditional typing methods are unable to discriminate strains from different sources that cause disease in humans. We report the use of genomotyping (whole-genome comparisons of microbes using DNA microarrays) combined with Bayesian-based algorithms to model the phylogeny of this major food-borne pathogen. In this study 111 *C. jejuni* strains were examined by genomotyping isolates from humans with a spectrum of *C. jejuni*-associated disease (70 strains), chickens (17 strains), bovines (13 strains), ovines (5 strains), and the environment (6 strains). From these data, the Bayesian phylogeny of the isolates revealed two distinct clades unequivocally supported by Bayesian probabilities ($P = 1$); a livestock clade comprising 31/35 (88.6%) of the livestock isolates and a "nonlivestock" clade comprising further clades of environmental isolates. Several genes were identified as characteristic of strains in the livestock clade. The most prominent was a cluster of six genes (*cj1321* to *cj1326*) within the flagellin glycosylation locus, which were confirmed by PCR analysis as genetic markers in six additional chicken-associated strains. Surprisingly these studies show that the majority (39/70, 55.7%) of *C. jejuni* human isolates were found in the nonlivestock clade, suggesting that most *C. jejuni* infections may be from nonlivestock (and possibly nonagricultural) sources. This study has provided insight into a previously unidentified reservoir of *C. jejuni* infection that may have implications in disease-control strategies. The comparative phylogenomics approach described provides a robust methodological prototype that should be applicable to other microbes.**

microarray analysis | gastrointestinal pathogen | Bayesian-based algorithm

The bacterium *Campylobacter jejuni* is the principal cause of human gastroenteritis worldwide, but it can present as a spectrum of disease from asymptomatic carriage to severe bloody diarrhea. In rare instances this infection may be followed by sequelae including septicaemia (1) and neuropathies such as Guillain-Barré syndrome (GBS) (2). Traditionally, poultry has been considered the major source of infection, but other sources identified include cattle (3), water (4), and milk (5). Birds also act as a reservoir that may account for the isolation of *C. jejuni* from diverse environmental samples (6). The proportion of human disease attributable to these different sources of infection is unknown, because traditional typing methods such as Penner serotyping and phage typing have generally failed to identify strains with phenotypic characteristics associated with different ecological habitats. Thus, despite the medical and socioeconomic importance of *C. jejuni*, the proportion of human disease caused by different sources of infection remains unclear, which has hindered effective control strategies to reduce *C. jejuni* from the food chain.

Microarray technology, allied to complex mathematical analysis to determine phylogeny, promises a sensitive and robust method to examine the genetic relatedness of bacterial populations. Examples where whole-genome microarray analysis has been used include the investigation of *Vibrio cholerae* strains associated with the current seventh cholera pandemic (7). A *V. cholerae* microarray based on the El Tor O1 strain N16961 was used to analyze a collection of nine strains of diverse global origin isolated between 1910 and 1992, and it was possible to differentiate classical biotype strains from El Tor biotype strains, and two putative chromosomal islands (VSP-I and VSP-II) with a deviant G+C content were identified in El Tor biotypes (7). Similarly, the pathological outcome of *Neisseria* species infections has been investigated with microarrays (8). DNA microarray studies revealed a series of relatively small sequences scattered throughout the genome that were either specific to *Neisseria meningitidis* or shared with *Neisseria gonorrhoeae*, but absent from the commensal *Neisseria lactamica*. This study confirmed that the capsule biosynthesis loci and the RTX toxin family were meningococcal-specific (8). Further DNA microarray studies have indicated that the loss of 11 DNA loci in *Yersinia pseudotuberculosis* may have contributed to the rapid emergence of the plague pathogen *Yersinia pestis* (9).

Previous comparative genomic studies of *C. jejuni* using DNA microarrays have focused on the identification of the genetic functional core of this pathogen (10, 11). Initially, at least 21% of the genes present in the sequenced strain appeared dispensable, as they were absent or highly divergent in one or more of the isolates tested, defining 1,300 of 1,654 predicted coding sequences (CDSs) as *C. jejuni* species-specific (10). These core genes mainly encode housekeeping functions such as metabolic, biosynthetic, cellular, and regulatory processes. However, many virulence determinants are also conserved, indicating that they are indispensable for *C. jejuni* to cause disease in humans. These include the cytolethal distending toxin, the flagellar structural proteins, the PEB antigenic surface proteins, and the general protein glycosylation locus (10). In another study, the genomic diversity of 18 *C. jejuni* strains from diverse sources was investigated (12). Seven hypervariable plasticity regions (PRs) were identified in the genome (PR1 to PR7). PR1 contains genes important in the utilization of alternative electron acceptors for respiration and may confer a selective advantage to strains in restricted oxygen environments. PR2, PR3, and PR7 contain many outer membrane and periplasmic proteins and hypothetical proteins of unknown function that might be linked to phenotypic variation and adaptation to different ecological niches. PR4, PR5, and PR6 contain genes involved in the production and

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: GBS, Guillain-Barré syndrome; CDS, coding sequence; PR, plasticity region; MLST, multilocus sequence typing.

Data deposition: Fully annotated microarray data have been deposited in the ArrayExpress database, www.ebi.ac.uk/arrayexpress (accession no. E-BUGS-22).

‡To whom correspondence should be addressed. E-mail: brendan.wren@lshtm.ac.uk.

modification of antigenic surface structures including the flagellin glycosylation locus. In this particular study, algorithms were used that selected a dynamic boundary between the conserved and variable genes similar to the GACK algorithm (13). More recently, genomic comparisons of 51 strains isolated from food and clinical sources have been integrated with data from the three previous *C. jejuni* DNA microarray studies (10–12) to perform a metaanalysis that included 97 strains from the four separate data sets (14). In that study (14), a large proportion of the variable genes were found to be absent or divergent in single strains only, and these uniquely variable genes could be mapped to previously defined variable loci. Thus Taboada *et al.* (14) propose large regions of the *C. jejuni* genome are genetically stable. Of the highly divergent genes that were identified 117 of 122 genes had divergent neighbors and showed high levels of intraspecies variability (14). Another recent study used DNA microarrays to address the issue of whether genetic markers specific to strains causing GBS could be identified. Strains associated with GBS and strains associated with enteritis were investigated; however, no GBS genetic markers were found (15). To date, these studies have examined only *C. jejuni* strains from limited ecological niches or clinical outcomes and have not used phylogenetic approaches to identify clonal groups of strains that may be related to strain source or disease severity.

In this study we have carried out whole-genome analysis of 111 isolates of *C. jejuni* from humans with a range of disease outcomes and clinical presentations and from diverse animal and environment sources by using a gene-specific *C. jejuni* NCTC11168 microarray. The whole-genome DNA microarray data have been combined with sensitive Bayesian-based algorithms to gain insight into the population structure of *C. jejuni*. We identify a distinct livestock clade and also a nonlivestock-associated clade where surprisingly the majority of human isolates are found. Additionally, several genes/genetic islands were identified as characteristic of strains associated with clades. The most prominent was a cluster of six genes (*cj1321* to *cj1326*) within the flagellin glycosylation locus, which we suggest encodes a distinct flagellin glycoform specifically maintained in livestock isolates.

Methods

Strain Selection and Growth Conditions. *C. jejuni* strains were cultured on Columbia blood agar plates in a variable atmosphere incubator under microaerobic conditions (5% O_2, 85% N_2, 10% CO_2) at 37°C for 48 h. DNA isolation was carried out by using Wizard (Promega) or Puregene (Gentra Systems) genomic DNA purification kits, and the DNA retrieved was quantified by using a GeneQuant spectrophotometer (Amersham Pharmacia). All DNA was stored at 4°C to minimize shearing caused by freeze thawing.

Microarray Design and Construction. All 1,654 annotated gene sequences from NCTC11168 were included in the design process (16). Ten pairs of gene-specific primers were designed to each sequence by using PRIMER3 (17). Primers were designed to have a length of 20–25 bp, T_m between 50°C and 80°C, and an amplicon ranging in size from 100 to 800 bp in length with an optimum size of 600 bp. A single pair of PCR primers was selected based on the BLAST similarity of the predicted PCR product to other genes on the microarray; products with no similarity or least similarity were selected to ensure the least possible cross-hybridization on the microarray. Array designs are available from the ArrayExpress database (www.ebi.ac.uk/arrayexpress, accession nos. A-BUGS-8 and A-BUGS-9).

Labeling, Hybridization, and Data Acquisition. Four micrograms of chromosomal DNA from the test strain and the control strain (NCTC11168) was labeled with Cy5 and Cy3, respectively, as described (10). Microarray slides were prehybridized in 3.5 × SSC, 0.1% SDS, and 10 mg/ml BSA at 65°C for 20 min before washing in distilled water for 1 min and a subsequent 1-min wash in isopropanol. Each test strain Cy5-labeled DNA was mixed with control strain Cy3-labeled DNA, purified with a MiniElute kit (Qiagen, Crawley, U.K.), denatured, and mixed to achieve a final 45-μl hybridization solution of 4 × SSC, 0.3% SDS. Microarrays were hybridized overnight under 22 × 22-mm LifterSlips (Erie Scientific, Portsmouth, NH), sealed in a humidified hybridization chamber (Telechem International, Sunnyvale, CA), immersed in a water bath at 65°C for 16–20 h. Slides were washed once in 400 ml of 1 × SSC, 0.06% SDS at 65°C for 2 min and twice in 400 ml of 0.06 × SSC for 2 min. Slides were scanned by using an Affymetrix 418 scanner (MWG Biotech, High Point, NC), and signal data were extracted by using IMAGENE 5.2 (BioDiscovery, EL Segundo, CA). For each strain tested, two microarray experiments were performed.

Comparative Phylogenomics. We carried out the whole-genome comparisons with GENESPRING 6.1 software as described (10) and by using GACK analysis (13), which allows a dynamic signal value to be used to discriminate genes that were absent (or divergent) from those that were present. Using a Nexus-format matrix the relationship of strains with Bayesian-based algorithms implemented through MR BAYES 3.0 software was determined (18). With samples and saves from every 40th tree, 1,100,000 generations of four incrementally heated Markov chain Monte Carlos (MCMC) were performed on the DNA–DNA microarray data by using the default annealing temperature of 0.5, a burn-in of 100,000 MCMC generations, and a 16-category gamma distribution. Ninety five percent majority rule consensus trees and clade credibility values were obtained by using PAUP4.0 software.

The deepest split within the *C. jejuni* phylogeny was determined by including a single *Campylobacter coli* strain (K8) in the initial data set (data not shown), which was used to root the phylogeny, i.e., resolve the "order of evolutionary divergence." The *C. coli* root was then removed and the analysis was repeated, first because there was concern over transspecies hybridisation and second to minimize saturation that can disrupt the stability of the phylogeny. The resulting phylogeny excluding *C. coli* strains was then rooted, assuming the *C. coli* out-group, at the point where the root lineage dissects the *C. jejuni* in-group, and determined in the initial analysis. Finally, to ensure the phylogenetic signal within the data was not caused by any single genetic island, the *cj1321–cj1326* genetic island that was identified by the data analysis was removed and the analysis was repeated. All Bayesian analyses were replicated to ensure the stability of the phylogeny.

Identification of Predictive Genes. MACCLADE 4 (19) was used to identify source predictive genes. If a particular gene is absent or divergent in a strain, that branch of the tree representing that strain is colored yellow. If the same gene is present in another strain, the branch was colored blue. By visualizing the distribution of each gene in the genome, genes that were specific for a particular clade have been identified.

Confirmatory PCR Analysis of Source Predictive Genes. PCR analysis was used for the validation of microarray results. Oligonucleotide primers were designed by using the NCTC11168 sequence (Sigma, Genosys) (Table 1, which is published as supporting information on the PNAS web site). PCRs were carried out in a volume of 50 μl. This reaction consisted of 0.2 mM of each dNTP (Amersham Pharmacia), 1 unit of TaqDNA polymerase (Promega), 0.1 nM of the downstream and upstream primer, and 1–100 ng of template DNA. PCR was carried out with an Omn-E thermal cycler (Hybaid, Middlesex, U.K.).

Results

Strain Selection and Initial Microarray Analysis. A well characterized collection of 111 *C. jejuni* strains were selected based on phenotype (Table 2, which is published as supporting information on the PNAS

web site). Seventy *C. jejuni* clinical isolates were selected comprising 47 from patients with gastroenteritis, seven from patients with septicaemia, nine from patients who developed GBS, and seven from asymptomatic carriers. Seventeen *C. jejuni* isolates from live chickens or chicken portions, 13 bovine (including two from ox liver portions), five ovine, and six environmental isolates from sand samples were also selected for genomic comparisons.

All isolates were competitively hybridized in duplicate with a *C. jejuni* DNA microarray that contained duplicate reporter elements of all of the 1,654 predicted CDSs in the genome of strain NCTC11168. The sequenced strain NCTC11168, originally isolated from a patient with gastroenteritis in the United Kingdom, was used as the control in all experiments. Whole-genome comparisons were carried out by using GENESPRING version 6.1 (Agilent Technologies, Palo Alto, CA). A signal value of 0.5 was selected to discriminate absent or divergent genes from those that were present. This empirically determined cut-off value to classify absence or divergence was compared with GACK analysis, a method based on an algorithm that selects a dynamic cut-off value based on the shape of the signal ratio distribution. The GACK analysis produced a near-identical list of absent or divergent genes, confirming the accuracy of the empirically determined cut-off. The *C. jejuni* functional core was determined by calculating the number of genes that were present in each of the test strains as well as the control strain. This species-specific functional core consisted of 979 predicted CDSs comprising 59.2% of the genome and was involved with regulatory, metabolic, cellular, and biosynthetic processes. Loci encoding surface antigens including capsular polysaccharide, flagella, and lipooligosaccharide did not form part of the functional core and are designated accessory genes.

Comparative Phylogenomic Analysis Based on Strain Source. The 111 *C. jejuni* strains isolated from humans (70 strains), chickens (17 strains), bovines (13 strains), ovines (5 strains), and the environment (6 strains) were examined by genomotyping. From these data, the Bayesian phylogeny of the *C. jejuni* isolates revealed two distinct clades unequivocally supported by Bayesian probabilities ($P = 1$): a livestock-associated clade that comprised 31/35 (88.6%) livestock isolates (16/17 chicken isolates, 12/13 bovine isolates, 3/5 ovine isolates), and a nonlivestock-associated clade. The nonlivestock-associated clade comprised further clades of environmental isolates, which were unequivocally supported by Bayesian probabilities ($P = 1$) (Fig. 1). The chicken, bovine, and ovine isolates were of diverse origin. For example, the chicken-associated strains were isolated between 1991 and 2001 from either broiler flocks, abattoirs, or poultry portions from different locations in the United Kingdom. Serotyping would fail to group the chicken-associated strains, as chicken strains in the livestock clade comprised four serotypes (HS2, HS5, HS44, and HS50), and three of these (HS2, HS5, and HS50) were also found in the nonlivestock clade (Table 2). Additionally, the chicken-associated strains consisted of 11 different phage types (PT1, PT2, PT5, PT15, PT19, PT25, PT34, PT35, PT36, PT44, and PT59), suggesting that phage typing would also be unable to distinguish between chicken and nonchicken-associated strains (Table 2).

Comparative Phylogenomic Analysis Based on Disease Severity. The livestock clade contained 55.9% (62/111) of the total number of strains including 44.3% (31/70) of clinical isolates. The nonlivestock clade comprised 44.1% (49/111) of the total number of strains including 55.7% (39/70) clinical isolates. The clinical isolates included in this study comprised potentially nonpathogenic strains from asymptomatic carriers that were previously identified from the UK Infectious Intestinal Disease study (20). A second set of possibly nonpathogenic strains, identified by multilocus sequence typing (MLST), was included in the study. Two clonal complexes associated with the sand of bathing beaches, and not with human disease or livestock, have been identified by using MLST (21). Having

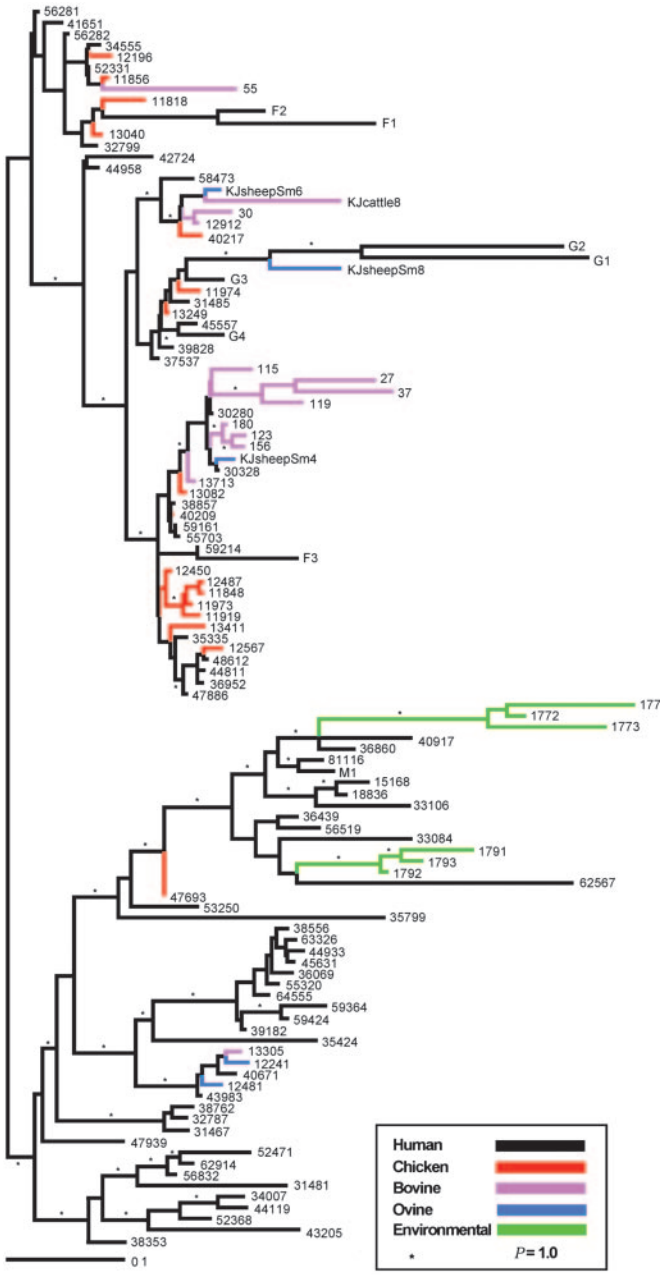

Fig. 1. Phylogenomic relationship of strains associated with different ecological niches. $P = 1.0$ represents 100% of all phylogenies showing a given topology. Strains are designated at the end of branches and are colored according to the ecological niche from which the *C. jejuni* strain was isolated.

established two possible sources of nonpathogenic strains, these strains were compared with strains of known different clinical presentations, including three cohorts of patients presenting with distinct clinical symptoms (diarrhea, bloody diarrhea, and vomiting). Additionally, 16 strains associated with *C. jejuni* sequelae were selected, including seven isolates from the blood samples of patients with *C. jejuni* septicaemia and nine *C. jejuni* strains isolated from patients who subsequently developed GBS. Phylogenetic analysis of *C. jejuni* strains from different clinical outcomes indicated that strains associated with asymptomatic carriage, diarrhea, bloody diarrhea, vomiting, septicaemia, and GBS did not form distinct clonal populations, related through a common ancestor (Fig. 2). By contrast, *C. jejuni* strains associated with the six environmental

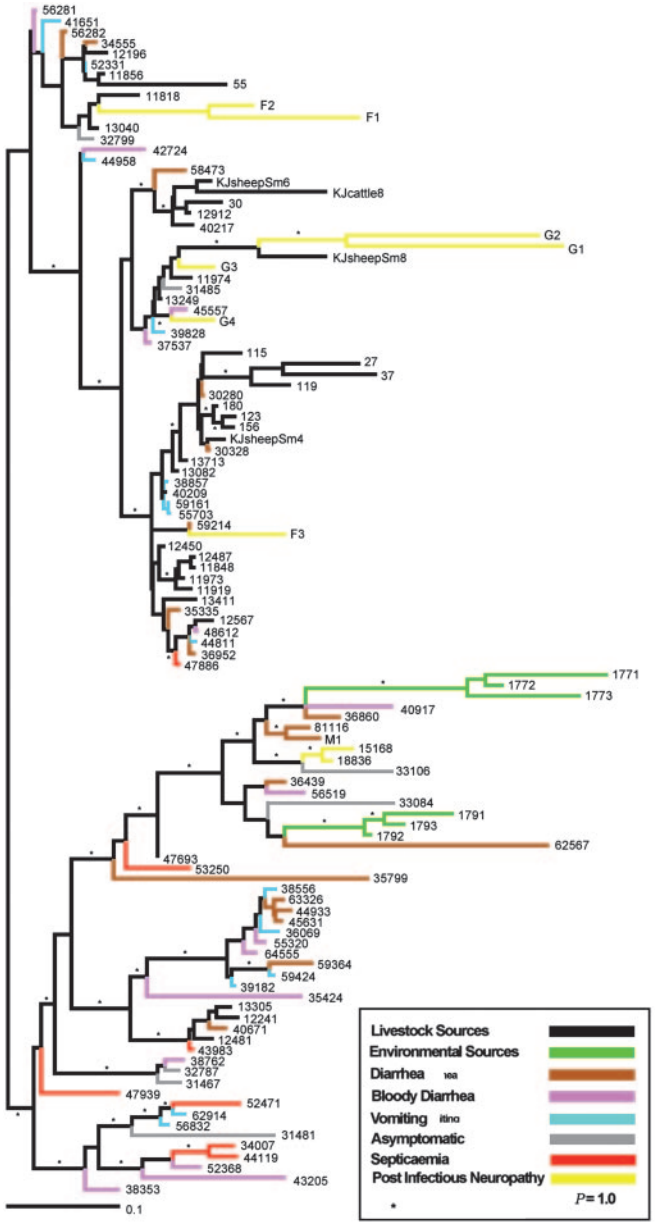

Fig. 2. Phylogenomic relationship of strains associated with different clinical outcomes. Strains are designated at the end of branches and are colored according to the clinical outcome of the *C. jejuni* infection or livestock/environmental source.

isolates (MLST ST-177 and ST-179) from beaches formed distinct clonal populations (Fig. 2).

Identification of Genes Specific to Clades. Genes contributing to the phylogenetic relationships of strains and moreover the formation of clades were identified by using parsimony-based methods implemented through MACCLADE 4 software (19). Character evolution within the hypothesized phylogenetic tree was traced. Tracing shows the most parsimonious hypothesis of ancestral states. This analysis indicates how the presence and absence or divergence of certain genes in an ancestral strain has led to the formation of a new clade. Patterns of deletions may be indicative of the selective pressures that caused a gene(s) to be lost or divergent or to remain stable. Such analysis identified a cluster of six genes (*cj1321* to *cj1326*) within the O-linked flagellin glycosylation locus as the most

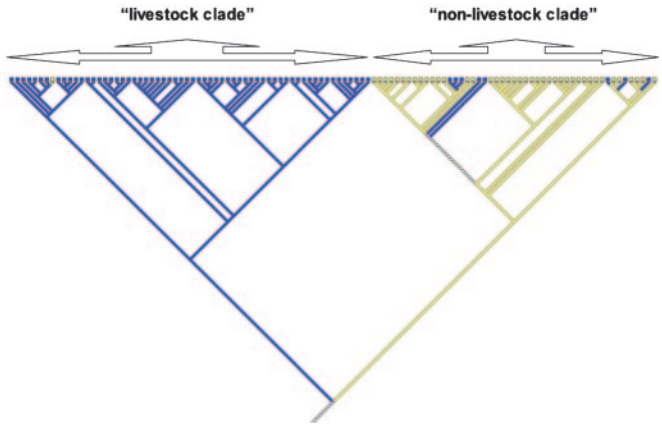

Fig. 3. Distribution of *cj1321* among *C. jejuni* strains. Parsimony-based gene analysis for determining the distribution of individual CDS *cj1321* throughout the phylogenetic tree. Strains in which *cj1321* are absent are colored yellow; strains in which *cj1321* are present are colored blue. Strains in the livestock clade all contain *cj1321* with one exception. *cj1321* is absent in strains in the nonlivestock clade with eight exceptions.

prominent loci present in livestock-associated strains (see Fig. 3 for MACCLADE 4 analysis of *cj1321* and Table 3, which is published as supporting information on the PNAS web site, for additional CDSs associated with the livestock clade). The percentage identity of CDSs *cj1321* to *cj1326* to the other sequenced bacteria was calculated by using BLAST (www.sanger.ac.uk/cgi-bin/blast/submitblast/c_jejuni). From the predicted amino acid sequences, Cj1321 is similar to an acetyl transferase from many bacterial species; Cj1322 and Cj1323 have similarity to hydroxyacyl dehydrogenases; Cj1324 has similarity to WbpG, a LPS biosynthesis protein found in several bacteria; Cj1325 is similar to Cj1330 involved in the synthesis of pseudaminic acid; and Cj1326 has no obvious similarity to any known protein. PCR analysis confirmed the presence of these genes in 16/17 chicken isolates from this study (O.L.C., unpublished data). Additionally, six chicken isolates, which were not examined in the comparative phylogenomics study, were tested for the presence of genes *cj1321* to *cj1326* by PCR (see Fig. 4 for PCR analysis of three of the strains). This result strongly suggests genes *cj1321* to *cj1326* as a genetic marker for chicken/livestock-associated strains.

Within the nonlivestock clade, strains isolated from the sand on

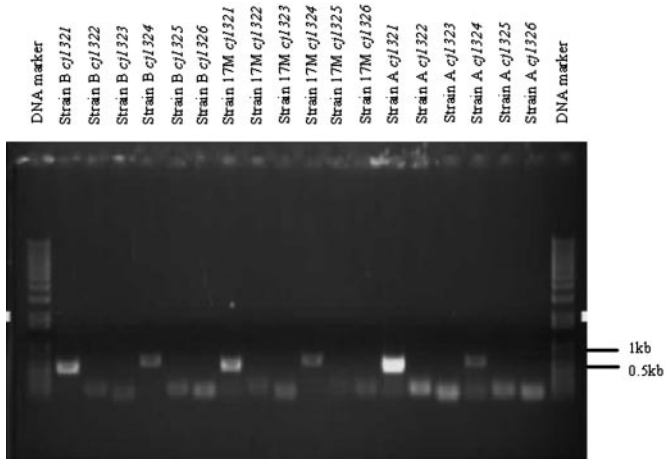

Fig. 4. Detection of *cj1321–cj1326* in additional chicken isolates B, 17M, and A. PCR products from DNA of chicken isolates B, 17M, and A were amplified with primers for genes *cj1321* to *cj1326* inclusive.

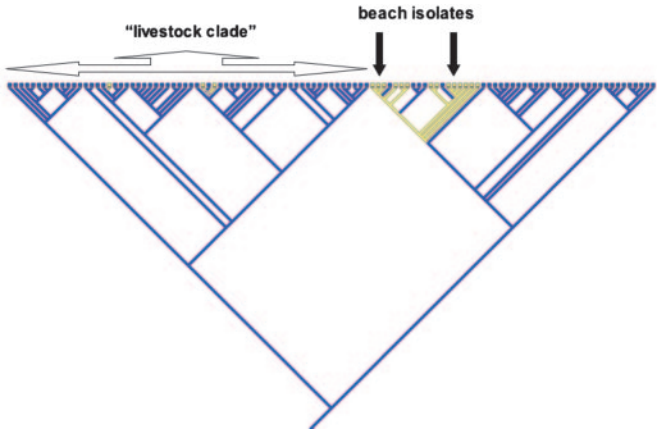

Fig. 5. Distribution of *cj1365* among *C. jejuni* strains. Parsimony-based gene analysis for determining the distribution of individual CDS *cj1365* throughout the phylogenetic tree. Strains in which *cj1365* are absent are colored yellow. Strains in which *cj1365* are present are colored blue. *cj1365* is absent from all six beach isolates, eight clinical isolates, and three livestock isolates.

bathing beaches formed distinct clades. The beach isolates were all lacking *cj1365*, encoding a putative secreted serine protease. This predicted CDS was present in 62/70 (88.6%) of the clinical isolates (Fig. 5). PCR analysis confirmed that *cj1365* was absent from the beach isolates, validating this microarray data (Fig. 6, which is published as supporting information on the PNAS web site).

Discussion

Bayesian algorithms have been used to generate robust hypotheses of phylogenetic relationships of *C. jejuni* based on genetic differences detected through whole-genome comparisons by using DNA microarrays to gain insight into host specificity and thus potential routes of transmission. We have demonstrated a population structure comprising two distinct clades, a livestock-associated and a nonlivestock-associated clade. It is possible that the presence of gene cluster *cj1321-26* is not fundamentally related to an underlying phylogenetic framework. We therefore repeated the Bayesian analysis after removing the *cj1321-26* island, which exactly recovers the same livestock and nonlivestock clades, albeit at a lower probability ($P = 0.56$) (Fig. 7, which is published as supporting information on the PNAS web site). This result confirms the distinct evolutionary lineage of the two clades irrespective of the loss or gain of a genetic island. Although it could be argued that the livestock-associated strains in this clade represent a specific clonal population of *C. jejuni* isolates from a common source, we believe that is unlikely. The chicken-associated strains used in the study were isolated over a 10-year period and by traditional typing approaches were highly diverse, comprising five different serotypes and 11 different phage types. They were isolated from different parts of the poultry food chain (from flocks to supermarkets) and different geographical locations throughout the United Kingdom. Furthermore, chicken flocks and chicken-ready meals in the United Kingdom are imported from worldwide sources. Similarly, the bovine and ovine strains were of diverse origin, including isolation from ox liver portions and isolation from different geographical locations. Given the diverse origin of these strains, the comparative phylogenomics results strongly suggest the presence of *C. jejuni* genes or genetic loci that are specific to adaptation in livestock.

Unexpectedly, 55.7% of clinical isolates were phylogenetically related to strains from nonlivestock sources, suggesting that environmental sources, or yet to be determined sources, contribute substantially to the burden of human *C. jejuni* infection. By contrast, comparative phylogenomics failed to identify relationships between strains associated with specific clinical outcomes, including GBS as reported by Leonard *et al.* (15). This observation may be a result of host immune factors influencing the outcome of *C. jejuni* clinical infection rather than genetic differences between strains. However, the possibility remains that clinical strains such as those causing septicaemia may possess additional virulence determinants contributing to this hyperinvasive phenotype that could not be detected by using the current *C. jejuni* DNA microarray based on strain NCTC11168.

In our analysis, the most prominent genetic marker indicative of the livestock-associated clade was a cluster of six genes within the O-linked flagellin glycosylation locus (*cj1321* to *cj1326*). This result contrasts with other glycoconjugate surface structures in *C. jejuni* such as the lipooligosaccharide, capsule and N-linked glycoproteins, and the rest of the O-linked glycosylation locus, none of which were specific for either clade. Given the similarity of the *cj1321* to *cj1326* genes to those involved in carbohydrate biosynthesis or sugar modifications, and the genes' location within the O-linked glycosylation locus, this result strongly suggests that these genes are involved in carbohydrate modification of the flagellum. Glycosylation of flagellin is increasingly being recognized in a number of Gram-negative pathogenic bacteria, including *Pseudomonas aeruginosa*, *Helicobacter pylori*, and *Aeromonas* spp (22–24). The modifications increase the hydrophilicity of flagellin and often influence the cells' immunogenicity and their interaction with eukaryotic cells (22, 23). The biological significance of the glycosylation island in *C. jejuni* remains to be determined, but different glycoforms appear to be expressed in different hosts or environments and may provide them with a specific survival advantage (25). Variant flagellin glycoforms have been shown to be expressed in different hosts or environments in the opportunistic pathogen *P. aeruginosa* that may provide the pathogen with a specific survival advantage (25). Recently, flagellin glycosylation in the plant pathogen *Pseudomonas syringae* has been shown to be involved in determining plant host specificity (26). We hypothesize that those genes *cj1321–cj1326* encode a distinct flagellin glycoform specifically maintained in livestock isolates, which could involve improved adhesion to the host cells, improved survival within livestock gut, and/or stimulation (or evasion) of the immune system. Thus, in *C. jejuni*, the potential for generating alternative surface-exposed glycoforms on the flagellin protein may suggest a mechanism for host specificity.

We have previously suggested a mechanism for the loss of genes in the *C. jejuni* flagellin glycosylation locus (including genes *cj1321–cj1326*) involving homologous recombination of highly similar genes in the motility accessory factor (*maf*) gene family (27). This observation may provide a genetic explanation for the apparent loss of these genes in the nonlivestock clade. Full characterization genes *cj1321–cj1326* and the entire 50-gene flagellin locus in NCTC11168 await further studies.

A probable secreted serine protease, encoded by CDS *cj1365* was identified as absent or divergent in all six of the "beach" isolates. Secreted serine proteases are well documented virulence factors in pathogens, serving multiple functions including the cleavage of human factor V that aggravates the hemorrhagic colitis characteristic of enterohemorrhagic infections (28). The cleavage of factor V is widespread among bacterial serine proteases secreted by pathogens that cause bloody diarrhea (28). The specific absence of the protease gene from the six environmental isolates has been confirmed by PCR analysis (Fig. 6) and by amplifying DNA flanking this gene and sequencing the resultant PCR products (A.V. Karlyshev and B.W.W., unpublished data). Given that the environmental isolates MLST types 177 and 179 are not found in human or chicken populations and are considered candidates as "nonpathogenic" *C. jejuni* strains, the protease gene is a good target for further investigation into the pathogenesis of *C. jejuni*. This observation is particularly useful because in contrast to other enteric pathogens such as *Salmonella enteritidis*, genetic virulence factors have been difficult to characterize in *C. jejuni*, as small animal models such as mice are inappropriate to model *C. jejuni*-associated disease. This

study demonstrates the value of comparative phylogenetics to studying isolates from their natural environment or disease state to uncovering potential virulence factors.

This typing method establishes an unequivocal association between *C. jejuni* strains and their source. The most promising typing method for *C. jejuni* to date is MLST, which has identified some clonal groups (21), but so far it has not been able to predict the source of a particular strain. Of the 111 strains included in this study only the six beach isolates had been previously MLST-typed and those strains belong to MLST clonal complexes 177 (three strains) and 179 (three strains) (21). A preliminary assessment of clades revealed a strong correlation between MLST and the Bayesian phylogeny (O.L.C., unpublished data). Thus a small-scale correlation between MLST and the phylogeny is observed. Strains from both MLST sequence type 177 and 179 formed separate clades within the nonlivestock clade of the phylogeny, each with unequivocal statistical support. The strong statistical support of the Bayesian-derived tree topology over small genetic distances within the livestock clade is particularly noteworthy. Moreover, the technique has the potential to provide a highly complementary approach to MLST. However, the power of this approach has been the elucidation of phylogenetic relationships of all strains in the study, including those clonal populations based on whole-genome comparisons, thus facilitating the identification of an unknown population structure. The other major advantage of the comparative phylogenomics approach described here is that once strains have been categorized it is easy to identify the genes/genetic loci specific to strain source (e.g., the putative glycosylation island and protease gene). A disadvantage of the comparative phylogenetics approach is that comparisons are often made to a single reference strain (i.e., the sequenced strain NCTC11168), therefore, additional sequences in the *C. jejuni* gene pool are excluded from subsequent phylogenomic analyses. The availability of additional *C. jejuni* gene sequences and the construction of a pan-*C. jejuni* species array should further increase the sensitivity of the comparative phylogenomics approach in distinguishing *C. jejuni* strains from diverse origins. The four outlying strains in the nonlivestock clade (one chicken, one bovine, and two ovine) may be explained as cross-contamination of the livestock source. Interestingly, the three ovine–bovine isolates in the nonlivestock clade show very little genetic divergence, particularly in comparison with the high levels of paraphyletic divergence of ovine–bovine isolates across the livestock clade. The ovine–bovine outliers are likely to represent a single outbreak, perhaps from an environmental source.

In this study combining the power of whole-genome comparisons using DNA microarrays with Bayesian-based algorithms to model phylogeny has proven to be far more informative than anticipated in uncovering a *C. jejuni* population structure that had been previously undetected with traditional typing systems. Surprisingly, the majority of human isolates appeared to be from nonlivestock sources. The likely sources are unknown, but water contamination is a commonly reported source for *C. jejuni* (4). It is noteworthy that the well characterized strain 81116, a human diarrhea strain traced back to contaminated water in Chelmsford, U.K. (29), is in this clade. Current strategies to reduce the burden of *C. jejuni* infection have focused on farm-to-fork interventions. In view of the data from this study, other potential sources, particularly within the environment, should be investigated as reservoirs for *C. jejuni* infection.

The approach described in this study provides a methodological prototype of robust phylogenomics that should be applicable for the study of other microbes. Recently, we applied the comparative phlyogenomics approach to other gastrointestinal bacterial pathogens, including *Yersinia enterocolitica* and *Clostridium difficile*, where distinct clades relating to strain source and biotype have been found (S.L. Howard, R. A. Stabler, M.W.G., and B.W.W., unpublished data). Although probably unsuitable for highly clonal species such as *Mycobacterium tuberculosis*, the comparative phylogenomics method has been found to be robust, reproducible, and a useful alternative for identifying potential genes associated with particular hosts and/or virulence.

We thank BμG@S (the Bacterial Microarray Group at St. George's, University of London) for advice and supply of the microarray; The Wellcome Trust for funding the multicollaborative microbial pathogen microarray facility under its Functional Genomics Resources Initiative; Lucy Brooks for assistance in depositing data in BμG@Sbase; Sarah Howard and Gemma Marsden for technical support; and Charles Penn and Stephen On for stimulating discussions. Clinical and animal strains included in the study were donated by Jennifer Frost (Health Protection Agency, London), Tom Humphrey (University of Bristol, Bristol, U.K.), Martin Maiden, (University of Oxford, Oxford), Diane Newell (Veterinary Laboratories Agency, Weybridge, U.K.), and Norman Gregson (Guy's Hospital, London). Environmental isolates were donated by Dr. David Wareing (Health Protection Agency, Preston, U.K.). This work was supported by a Medical Research Council studentship (to O.L.C.), the Food Standards Agency, and the Biotechnology and Biological Sciences Research Council.

1. Blaser, M. J., Berkowitz, I. D., LaForce, M., Cravens, J., Reller, B. & Wang, W.-L. L. (1979) *Ann. Int. Med.* **91,** 179–185.
2. Ang, C. W., De Klerk, M. A., Endtz, H. P., Jacobs, B. C., Laman, J. D., Van Der Meche, F. G. A. & van Doorn, P. A. (2001) *Infect. Immun.* **69,** 2462–2472.
3. Corry, J. E. & Atabay, H. I. (2001) *Symp. Ser. Soc. Appl. Microbiol.* 96S–114S.
4. Engberg, J., Gerner-Smidt, P., Scheutz, F., Moller Nielsen, E., On, S. L. & Molbak, K. (1998) *Clin. Microbiol. Infect.* **4,** 648–656.
5. Wood, R. C., MacDonald, K. L. & Osterholm, M. T. (1992) *J. Am. Med. Assoc.* **268,** 3228–3230.
6. Broman, T., Palmgren, H., Bergstrom, S., Sellin, M., Waldenstrom, J., Danielsson-Tham, M. L. & Olsen, B. (2002) *J. Clin. Microbiol.* **40,** 4594–4602.
7. Dziejman, M., Balon, E., Boyd, D., Fraser, C. M., Heidelberg, J. F. & Mekalanos, J. J. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 1556–1561.
8. Perrin, A., Bonacorsi, S., Carbonnelle, E., Talibi, D., Dessen, P., Nassif, X. & Tinsley, C. (2002) *Infect. Immun.* **70,** 7063–7072.
9. Hinchliffe, S. J., Isherwood, K. E., Stabler, R. A., Prentice, M. B., Rakin, A., Nichols, R. A., Oyston, P. C., Hinds, J., Titball, R. W. & Wren, B. W. (2003) *Genome Res.* **13,** 2018–2029.
10. Dorrell, N., Mangan, J. A., Laing, K. G., Hinds, J., Linton, D., Al-Ghusein, H., Barrell, B. G., Parkhill, J., Stoker, N. G., Karlyshev, A. V., *et al.* (2001) *Genome Res.* **11,** 1706–1715.
11. Leonard, E. E., 2nd, Takata, T., Blaser, M. J., Falkow, S., Tompkins, L. S. & Gaynor, E. C. (2003) *J. Infect. Dis.* **187,** 691–694.
12. Pearson, B. M., Pin, C., Wright, J., I'Anson, K., Humphrey, T. & Wells, J. M. (2003) *FEBS Lett.* **554,** 224–230.
13. Kim, C. C., Joyce, E. A., Chan, K. & Falkow, S. (2002) *Genome Biol.* **3,** 1–17.
14. Taboada, E. N., Acedillo, R. R., Carrillo, C. D., Findlay, W. A., Medeiros, D. T., Mykytczuk, O. L., Roberts, M. J., Valencia, C. A., Farber, J. M. & Nash, J. H. (2004) *J. Clin. Microbiol.* **42,** 4566–4576.
15. Leonard, E. E., 2nd, Tompkins, L. S., Falkow, S. & Nachamkin, I. (2004) *Infect. Immun.* **72,** 1199–1203.
16. Moen, B., Oust, A., Langsrud, O., Dorrell, N., Hinds, J., Marsden, G., Kohler, A., Wren, B. W. & Rudi, K. (2005) *Appl. Environ. Microbiol.* **71,** 2086–2094.
17. Rozen, S. & Skaletsky, H. (2000) *Methods Mol. Biol.* **132,** 365–386.
18. Ronquist, F. & Huelsenbeck, J. P. (2003) *Bioinformatics* **19,** 1572–1574.
19. Maddison, D. R. & Maddison, W. P.(2001) MACCLADE4: *Analysis of Phylogeny and Character Evolution* (Sinauer, Sunderland, MA).
20. The UK Intestinal Infectious Disease Study (2000) *A Report of the Study of Infectious Intestinal Diseases in England* (Her Majesty's Stationary Office, Norwich, U.K.).
21. Dingle, K. E., Colles, F. M., Wareing, D. R. A., Ure, R., Fox, A. J., Bolton, F. E., Bootsma, H. J., Willems, R. J. L., Urwin, R. & Maiden, M. C. J. (2001) *J. Clin. Microbiol.* **39,** 14–23.
22. Castric, P., Cassels, F. J. & Carlson, R. W. (2001) *J. Biol. Chem.* **276,** 26479–26485.
23. Schirm, M., Soo, E. C., Aubry, A. J., Austin, J., Thibault, P. & Logan, S. M. (2003) *Mol. Microbiol.* **48,** 1579–1592.
24. Gavin, R., Rabaan, A. A., Merino, S., Tomas, J. M., Gryllos, I. & Shaw, J. G. (2002) *Mol. Microbiol.* **43,** 383–397.
25. Arora, S. K., Wolfgang, M. C., Lory, S. & Ramphal, R. (2004) *J. Bacteriol.* **186,** 2115–2122.
26. Takeuchi, K., Taguchi, F., Inagaki, Y., Toyoda, K., Shiraishi, T. & Ichinose, Y. (2003) *J. Bacteriol.* **185,** 6658–6665.
27. Karlyshev, A. V., Linton, D., Gregson, N. A. & Wren, B. W. (2002) *Microbiology* **148,** 473–480.
28. Dutta, P. R., Cappello, R., Navarro-Garcia, F. & Nataro, J. P. (2002) *Infect. Immun.* **70,** 7105–7113.
29. Palmer, S. R., Gully, P. R., White, J. M., Pearson, A. D., Suckling, W. G., Jones, D. M., Rawes J. C. & Penner, J. L. (1983) *Lancet* **8319,** 287–290.