PeerJ Preprints NOT PEER-REVIEWED

Local genes for local bacteria: evidence of allopatry in the genomes of transatlantic *Campylobacter* populations

Ben Pascoe[1,2], Guillaume Méric[1], Koji Yahara[3,4], Helen Wimalarathna[5], Susan Murray[4], Matthew D. Hitchings[4], Emma L. Sproston[6], Catherine D. Carrillo[7], Eduardo N. Taboada[8], Kerry K. Cooper[9], Steven Huynh[10], Alison J. Cody[5], Keith A. Jolley[5], Martin C. J. Maiden[5,11], Noel D. McCarthy[5,11,12], Xavier Didelot[13], Craig T. Parker[10] and Samuel K. Sheppard[1,2,5]#

[1]The Milner Centre for Evolution, Department of Biology and Biochemistry, Bath University, Claverton Down, Bath, BA2 7AY, UK; [2]MRC CLIMB Consortium, UK; [3]Department of Bacteriology II, National Institute of Infectious Diseases, Musashimurayama, Tokyo, 208-0011, Japan; [4]Swansea University Medical School, Swansea University, Singleton Park, Swansea, SA2 8PP; [5]Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS, UK; [6]Bureau of Microbial Hazards, Health Canada, 251 Sir Frederick Banting Driveway, Ottawa, K1A 0K9, Canada; [7]Canadian Food Inspection Agency, 960 Carling Avenue, Ottawa, K1A 0Y9, Canada; [8]National Microbiology Laboratory at Lethbridge, Public Health Agency of Canada, PO Box 640, Township Rd. 9-1, Lethbridge, Alberta, T1J 3Z4, Canada; [9]Department of Biology, California State University, Northridge, Northridge, California, USA; [10]Produce Safety and Microbiology Research Unit, Agricultural Research Service, US Department of Agriculture, Albany, California, USA; [11]NIHR Health Protection Research Unit in Gastrointestinal Infections, UK; [12]University of Warwick, Coventry, CV4 7AL, UK; [13]Department of Infectious Disease Epidemiology, Imperial College London, London, W2 1PG, UK

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2638v2 | CC BY 4.0 Open Access | rec: 19 May 2017, publ: 19 May 2017

PeerJ Preprints NOT PEER-REVIEWED

#Address correspondence to: Professor Samuel K. Sheppard, The Milner Centre for Evolution, Department of Biology and Biochemistry, Bath University, Claverton Down, Bath, BA2 7AY, UK. Telephone: +44(0)1225 385046; Fax: +44(0)1225 386779; Email: S.K.Sheppard@bath.ac.uk

Running title: *Campylobacter* biogeography

Keywords: Allopatry; *Campylobacter;* Genomics; Source attribution; Recombination; Phylogeny

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2638v2 | CC BY 4.0 Open Access | rec: 19 May 2017, publ: 19 May 2017

Abstract

The genetic structure of bacterial populations can be related to geographical locations of isolation. In some species, there is a strong correlation between geographical distance and genetic distance, which can be caused by different evolutionary mechanisms. Patterns of ancient admixture in *Helicobacter pylori* can be reconstructed in concordance with past human migration, whereas in *Mycobacterium tuberculosis* it is the lack of recombination that causes allopatric clusters. In *Campylobacter*, analyses of genomic data and molecular typing have been successful in determining the reservoir host species, but not geographical origin. We investigated biogeographical variation in highly recombining genes to determine the extent of clustering between genomes from geographically distinct *Campylobacter* populations. Whole genome sequences from 294 *Campylobacter* isolates from North America and the UK were analysed. Isolates from within the same country shared more recently recombined DNA than isolates from different countries. Using 15 UK/American closely matched pairs of isolates that shared ancestors, we identify regions that have frequently and recently recombined to test their correlation with geographical origin. The seven genes that demonstrated the greatest clustering by geography were used in an attribution model to infer geographical origin which was tested using a further 383 UK clinical isolates to detect signatures of recent foreign travel. Patient records indicated that in 46 cases travel abroad had occurred less than two weeks prior to sampling and genomic analysis identified that 34 (74%) of these isolates were of a non-UK origin. Identification of biogeographical markers in *Campylobacter* genomes will contribute to improved source attribution of clinical *Campylobacter* infection and inform intervention strategies to reduce campylobacteriosis.

PeerJ Preprints NOT PEER-REVIEWED

Introduction

Geographical structure is well documented in bacteria and analysing genetic variation among isolates can provide information about the global spread of important pathogens. For example, after spreading with Neolithic human hosts (Comas et al., 2013), lineages of *Mycobacterium tuberculosis* populations can be classified into geographical groups based upon local genetic diversification of DNA sequences (Achtman, 2008, Gagneux and Small, 2007). Phylogeographic structure has also been observed in the human gastric bacterium *Helicobacter pylori*, where a rapidly evolving genome with high levels of horizontal gene transfer (HGT) allows the reconstruction of recent human migrations to the extent that genetic admixture among the bacteria reflects interactions among human populations (Falush et al., 2003, Moodley et al., 2009).

M. tuberculosis and *H. pylori* are primarily human pathogens. However, in the foodborne pathogen *Campylobacter*, animals are the principal reservoir for human infection. International trade, particularly in agricultural animals including chicken and poultry products, provides a vehicle for global spread. In this case, local phylogeographic signals can be weakened not only by the rapid movement of lineages around the world, but also by genomic changes that occur within the reservoir host. This may make it difficult to attribute the country of origin based on the *Campylobacter* isolate genome alone. Sequence-based analyses have shown that populations of the main human disease-causing *Campylobacter* species, *C. jejuni* and *C. coli*, are highly structured into clusters of related lineages, which can be identified by MLST as clonal complexes (CC’s). Members of CC’s share four or more MLST alleles with a pre-defined central genotype, which gives the CC its name, for example

ST-21 defines CC-21 (Dingle et al., 2005, Sheppard et al., 2010b). In *C. jejuni*, host-associated clonal complexes can be identified based upon the frequency with which particular genotypes are isolated from different hosts (Sheppard et al., 2011, Sheppard et al., 2014). Many of these lineages are globally distributed (Sheppard et al., 2010a) but despite this strong host signal, there is evidence for phylogeographic structuring and the proliferation of distinct lineages in different countries (McTavish et al., 2008, Asakura et al., 2012).

Horizontal gene transfer in recombining bacteria, such as *Campylobacter* (Sheppard et al., 2008, Wilson et al., 2008, Sheppard et al., 2013a), can provide information about ecological differences between lineages. For example, when a *Campylobacter* lineage transfers to a new animal host it may acquire DNA from the resident population by HGT. This has been shown in host generalist *C. jejuni* lineages isolated from chicken that sometimes contain alleles that originated in chicken-specialist genotypes (McCarthy et al., 2007, Wilson et al., 2008). In this study, we applied comparable approaches to investigate if HGT can lead to signatures of recombination that discriminate between isolates from North America and the UK using genomic data. Using matched pairs of North American and UK isolates, we identify genes that are prone to recombination, and will therefore pick up a local DNA more rapidly, and hypothesise that these genes may acquire a biogeographical signal.

Materials and Methods

Bacterial isolates and genome sequencing

A total of 294 sequenced isolates were analysed, of which 131 genomes were generated in this study and augmented by 163 previously published genomes (Sheppard et al., 2014, Sheppard et al., 2013a, Sheppard et al., 2013b). Sequencing reads for all genomes sequenced in this study are available from the NCBI short read archive associated with BioProject: PRJNA312235. All assembled genomes used in this study can also be downloaded from FigShare (doi: 10.6084/m9.figshare.4906634).

Canadian isolates: Isolates were collected from chicken and bovine faecal samples between July 2004 and July 2006 from farms at diverse locations in Alberta. Samples were placed on ice and processed within 6 h as previously described (Jokinen et al., 2010). Approximately 5 g of faecal matter was mixed with 5 ml of phosphate buffered saline (PBS) to form uniform slurry. One-millilitre aliquots of the PBS-faecal samples were added to 20 ml of Bolton broth containing 5% (v/v) lysed horse blood and selective supplement (Diergaardt et al., 2004) and incubated at 42^{o}C for 24 h under microaerobic conditions prior to plating 20 µl onto supplemented charcoal cefoperazone deoxycholate agar (CCDA). The plates were incubated for a further 48 h at 42^{o}C. Human samples were acquired from clinical laboratories in three Canadian provinces. These were re-plated from frozen glycerol stocks and the DNA extracted as described below.

Presumptive *Campylobacter* colonies were cultured onto blood agar plates and tested using biochemical oxidase and catalase tests. *Campylobacter* species identification was performed using a multiplex PCR assay that included 16S rRNA gene primers and *C. jejuni* (*mapA*) and *C. coli*-specific (*ceuE*) primers (Denis et al., 1999). Positive *Campylobacter* isolates were sub-cultured on Mueller-Hinton agar and genomic DNA was extracted using the Wizard Genomic DNA Purification Kit as per manufacturer's instructions (Promega, Madison, WI). DNA integrity was checked on an agarose gel and purity and concentration determined by optical density. Purified genomic DNA was sent to Canada's Michael Smith Genome Sciences Centre (Vancouver, Canada) and sequenced using the Illumina HiSeq 2000 platform. Sequence reads were assembled into contigs using the SPAdes assembler (v3.0) (Bankevich et al., 2012).

US isolates: Isolates were collected from cattle faecal samples between December 2008 and June 2010 from diverse locations within the Salinas Valley watershed, California. Samples were placed on ice and processed within 12 h. Cattle faeces were inoculated into a six-well microtitre plate containing 6 ml 1× Anaerobe Basal Broth (Oxoid) amended with Preston supplement (when reconstituted consists of: amphotericin B (10 µg/ml), rifampicin (10 µg/ml), trimethoprim lactate (10 µg/ml), and polymyxin B (5 UI/ml) (Oxoid), using a sterile cotton swab. Microtiter plates were placed inside plastic ZipLoc bags filled with a microaerobic gas mixture (1.5% O_2, 10% H_2, 10% CO_2, and 78.5% N_2) and incubated for 24 h at 37°C, while shaking at 40 rpm. Subsequently, 10-µl of these enrichment cultures were plated onto anaerobe basal agar (ABA, Oxoid) plates, amended with 5% laked horse blood and CAT supplement (cefoperazone (8 µg/ml), amphotericin B (10 µg/ml), and teicoplanin (4 µg/ml) (Oxoid)). All plates were then incubated under microaerobic conditions at 37°C for 24 h. Bacterial cultures were passed through 0.2 µm mixed cellulose ester filters onto ABA plates and incubated at 37°C under microaerobic conditions. After 24 h, single colonies were streaked onto fresh ABA plates and incubated 24–48 h at 37°C for purification.

DNA was extracted from a pure culture colony using the Wizard Genomic DNA Purification Kit (Promega, Madison, WI). *Campylobacter* species was identified by 16S rDNA sequencing, using the primer pairs as described by Lane (1991). Genome sequencing was performed on an Illumina MiSeq sequencer using the KAPA Low-Throughput Library Preparation Kit with Standard PCR Amplification Module (Kapa Biosystems, Wilmington, MA), following manufacturer's instructions except for the following changes; 750 ng DNA was sheared at 30 psi for 40 s and size selected to 700–770 bp following Illumina protocols. Standard desalted TruSeq LT and PCR primers were ordered from Integrated DNA Technologies (Coralville, IA) and used at 0.375 and 0.5 µM final concentrations, respectively. PCR was reduced to 3–5 cycles. Libraries were quantified using the KAPA Library Quantification Kit (Kapa), except with 10 µl volume and 90-s annealing/extension PCR, then pooled and normalized to 4 nM. Pooled libraries were re-quantified by ddPCR on a QX200 system (Bio-Rad, Hercules, CA), using the Illumina TruSeq ddPCR Library Quantification Kit and following manufacturer's protocols, except with an extended 2-min annealing/extension time. Libraries were sequenced using a 2 × 250 bp paired end v2 reagent kit on a MiSeq instrument (Illumina, San Diego, CA) at 13.5 pM, following manufacturer's protocols. Genomes were assembled using the Roche Newbler assembler (version 2.3).

Published isolates: We augmented our collection of isolates sequenced in this study with 163 previously published *Campylobacter* isolate genomes from Canada, the USA and the UK collected between 1980-2012 from a range of sources, including cattle (54), chicken (80), pig

(9), environmental (49), wild bird species (12) and human clinical cases (73) (Figure S1)(Sheppard et al., 2014, Sheppard et al., 2013a, Sheppard et al., 2013b).

UK clinical test isolates: In addition to this collection of sequenced and publicly available *Campylobacter* genomes, we used a further 383 clinical samples collected from the John Radcliffe Hospital in Oxford between June and October 2011 as a test dataset to attribute source according to geography (Table S2)(Cody et al., 2013). These genomes were downloaded from http://pubmlst.org/campylobacter/.

Population structure

Isolate genomes were archived on an open-source BIGSdb database which identifies gene presence and allelic variation by comparison to a reference locus list (Jolley and Maiden, 2010, Sheppard et al., 2012, Meric et al., 2014). This list comprises 1,623 locus designations from the annotated genome of *C. jejuni* strain NCTC11168 (Genbank accession number: NC_002163.1) (Gundogdu et al., 2007, Parkhill et al., 2000). Reference loci were identified in each of the 294 isolate genomes using BLAST. Loci were recorded as present if the sequence had ≥70% nucleotide identity over ≥50% of the gene length. Each gene was aligned individually using MAFFT (Katoh et al., 2002), and concatenated into a single multi-FASTA alignment file for each isolate for a total alignment of 1,585,605 bp. Phylogenetic trees were constructed from a whole-genome alignment of *C. jejuni* (n=229) and *C. coli* (n=55) isolates based on 103,878 and 806,657 variable sites, respectively, using FastTree (version 2) and an approximation of the maximum likelihood algorithm (Tamura et al., 2013, Kumar et al., 2016).

Selection of isolate pairs

To minimise the effect of host adaptation and maximize the opportunity of identifying genetic signatures of geographic separation, a subset of 15 isolate pairs were chosen based upon their phylogenetic clustering. In each case, isolate pairs contained one Canadian and one UK isolate of the same clonal complex sampled from the same host species. Paired isolates shared 1,378 genes resulting in a core-genome alignment of 1,287,560 bp.

Analysis of co-ancestry and inference of recombination hot regions

The co-ancestry of the paired isolates was inferred based on whole genome sequences using chromosome painting and fineSTRUCTURE (Lawson et al., 2012), as previously described (Yahara et al., 2013). ChromoPainter (version 0.02) was used to infer the number of DNA 'chunks' donated from a donor to a recipient for each recipient haplotype, and the results summarized in a co-ancestry matrix indicating average isolate similarity across the entire genome. fineSTRUCTURE was then used for 100,000 iterations of both the burn-in and Markov chain Monte Carlo (MCMC) chain to cluster individuals based on the co-ancestry matrix. The results are visualized as a heat map with each cell indicating the proportion of DNA 'chunks' a recipient receives from each donor.

The time to the most recent common ancestor (TMRCA) of each pair was estimated using the model described in Didelot et al. (2013) and summarised here briefly. Pairs of genomes share a common ancestor $\boldsymbol{t}$ years ago and have been subject to mutation at a rate $\boldsymbol{\mu}$ and recombination at rate $\boldsymbol{\rho}$. The mutation rate of $2.9\text{x}10^{-5}$ per site per year was used as reported in (Sheppard et al., 2010b), which is similar to the rates estimated in Wilson et al., (2008, 2009). The effect of recombination is to introduce a high density of polymorphism similar to the ClonalFrame model (Didelot and Falush, 2007, Didelot and Wilson, 2015) but with the advantage that this density can vary between recombination events to reflect differences in evolutionary distance between donors and recipients (Morelli et al., 2010, Didelot et al., 2013). In each pairwise comparison, the TMRCA and recombination rate parameters are estimated based on a core genome alignment, with 95% credibility intervals.

Epidemiological markers of geographical clustering

Neighbour-joining phylogenetic trees were constructed for all genes that demonstrated an average of above 1% pairwise nucleotide diversity across all 15 pairs of isolates. Individual gene phylogenies were constructed in MEGA for all 57 genes. Isolates were assigned to a putative source population based on the seven highly recombining genes that showed the greatest level of clustering by geography. Probabilistic assignment of geographical source is based on the allele frequencies in the reference population data sets for each of the seven loci. This analysis was performed using Structure, a Bayesian model-based clustering method designed to infer population structure and assign individuals to populations using multilocus genotype data (Sheppard et al., 2010a, Pritchard et al., 2000). Canadian and USA isolates were combined as a North American population for comparison with UK isolates.

Attribution of clinical isolates to country based on seven geographically segregating genes

The source attribution model was tested with isolates of a known source. Self-assignment of a random subset of the comparison dataset was conducted by removing a third of the isolates from each candidate population (n = 73). The remainder were used as the reference set (78 North American isolates to compare with 68 UK isolates). Structure was run for 100,000 iterations following a burn-in period of 10,000 iterations using the no admixture model to assign individuals to putative populations. The assignment probability for each source was calculated for each isolate individually and were attributed to origin populations when the attribution probability was greater than 0.50.

PeerJ Preprints NOT PEER-REVIEWED

Results

Core genomes of isolates from North America and the UK were compared, and there was no observable clustering by country or continent on a neighbour-joining tree (Figure 1). STs sampled in both *Campylobacter* populations belonged to clonal complexes that can be classified as specialist and host generalist based upon the frequency at which they have been isolated from different hosts. These included chicken specialist clonal complexes CC-257, CC-283, CC-353, CC-354, CC-443, CC-573, CC-574 and CC-661, cattle specialist CC-61 and CC-42, and host generalist CC-21, CC-45, CC–206 and CC-48 complexes (Figure 1 and Table S1).

Matched isolates share more common ancestry with isolates from the same country

To minimise the effect of host adaptation and maximize the opportunity of identifying genetic signatures of geographic separation, a subset of 15 isolate pairs were chosen based upon their phylogenetic clustering with less than 1,200 bp difference in 1,378 core genome loci. In each case, isolate pairs contained one Canadian and one UK isolate of the same clonal complex sampled from the same host species (Table 1). The co-ancestry of the paired isolates was inferred based on core genome alignments using chromosome painting and fineSTRUCTURE (Lawson et al., 2012)(Yahara et al., 2013)(Figure 2). The total proportion of DNA ‘chunks’ in a recipient from isolates within the same country (median 0.59) was significantly higher than that from isolates from different countries (median 0.33)(p<10^{-9}, Wilcoxon’ rank sum test).

Matched isolates share recent common ancestors but have since experienced significant recombination

The estimated time since the most recent common ancestor (TMRCA) was calculated for each UK/American pair of genomes as previously described (Didelot et al., 2013), using the mutation rate of 2.9x10^{-5} per site per year reported in Sheppard et al. (2010b), which is consistent with estimates in Wilson et al. (2009). In each pairwise comparison, the level of divergence along the genome (Figure 3) was used to estimate the TMRCA and recombination rate, with 95% credibility intervals around these parameters (Table 2). All pairs were estimated to have shared ancestors between one and five years ago, with two exceptions, namely the two *C. coli* pairs, for which the TMRCA was around 25 years ago. The ratio r/m of rates at which recombination and mutation introduce polymorphism was estimated to be around 20-30 except in the two *C. coli* pairs with larger TMRCA, for which a smaller value was estimated around r/m=4. Most existing r/m estimates have been calculated using 7 MLST housekeeping genes (Vos and Didelot, 2009, Wilson et al., 2008, Wilson et al., 2009). Other estimates have been derived through comparison of relatively small numbers of *Campylobacter* genomes (Llarena et al., 2016). R/m estimates can vary considerably depending on the isolate collection and the genes used in the analysis. For example, ranging from 0-100 among *Helicobacter* isolates within a human population from within a single settlement in South Africa (Didelot et al., 2013). Given the potential for sample-dependent variation, the r/m estimates in this study are consistent with previous estimates. Further, variation in TMRCA estimates and r/m between *C. coli* compared to *C. jejuni* pairs in this study may reflect differences between the species, but more sampling of the *C. coli* population is necessary to investigate this further.

Highly recombining genes as markers of geographical attribution

A pairwise comparison of the matched pairs was used to quantify the level of divergence in each gene within the core genome (1,147 genes) of the paired isolates. Most genes showed low diversity, indicative of closely related pairs. Polymorphism in genes with less than 1% divergence between pairs (white and red in Figure 3) are likely to be the result of mutation or recombination with a tract of DNA with high nucleotide identity, so that only one or two substitutions are visible. Genes with greater than 1% divergence between pairs are likely to have recombined as numerous substitutions have been introduced (blue in Figure 3). Fifty-seven genes (e.g. *Cj0034c* and *Cj0635*) had a high level (>1%) of nucleotide divergence and high probability of recombination in all 15 pairs. This result did not arise just by chance: overall recombination was inferred in around 25% of the genes in each pair and so if recombination was random, the probability that all 15 pairs had recombined for a given gene would be extremely small ($0.25^{15}=9.3x10^{-10}$).

Individual gene trees were generated for these 57 genes from which the most recombination could be identified (Figure S3). The seven genes that gave the clearest geographic clustering were used for further analysis of geographical attribution using Structure as previously described (Sheppard et al., 2010a, Pritchard et al., 2000). A self-test was performed on a subset of our isolate collection and in 76.7% of cases the source continent was correctly attributed. The percentages of correctly attributed isolates by population were not significantly different, at 76.9% for North America and 76.5% for the UK. Where an isolate was incorrectly attributed to a population there was a higher average reported attribution probability (0.85) in the case of UK isolates compared with North American isolates (0.67). When applied to the remainder of our isolate collection, the proportion of UK isolates correctly attributed to the UK reference population was 70%, while the proportion of North

American isolates correctly attributed was slightly higher at 76%. This was not improved when using data from all 57 highly recombining genes as input for the attribution model in Structure (43% of UK and 72% of North American isolates correctly attributed).

Attribution of clinical isolates to country based on seven selected genes

The same geographical attribution model was applied to 383 clinical *C. jejuni* isolates from the Oxfordshire *Campylobacter* Surveillance Study in the UK, accessed via pubMLST.org/campylobacter, and for which details of recent foreign travel were provided (Cody et al., 2013). The model correctly assigned 34 of the 46 (73.9%) isolates where recent foreign travel had previously been declared, to a non-UK source of origin (Figure 4). In total, approximately half (47%) of the collected clinical isolates could be attributed to the UK.

Discussion

Isolation of bacteria in different host species and barriers to recombination between populations overtime, can lead to population differentiation reflected in the genome. In *C. jejuni*, this can be seen at different levels. The proliferation of certain lineages to a particular host species that are abundant in one host and rare or absent in others (Sheppard et al., 2011, Griekspoor et al., 2013, Sheppard et al., 2010a). Increased frequency of host associated nucleotide substitutions in multiple lineages (that reflect adaptation to the host) drift in physically isolated populations (Sheppard et al., 2013b). This host-associated genetic structure can be informative for understanding the evolution of *C. jejuni* (Dearlove et al., 2016), but can also be used in a more practical way to identify the source of isolates causing human infection by identifying genomic signatures (resulting from adaptation or drift) in the infecting isolate that are associated with populations in particular reservoir hosts (Sheppard et al., 2009, Wilson et al., 2008). Quantitative source attribution models, based upon the probability that a particular clinical isolate originated in different reservoirs, have been widely used to estimate the risk of human infection from different food production animals and other sources (Colles et al., 2008, French et al., 2005, Mullner et al., 2009, Sheppard et al., 2009, Roux et al., 2013, Griekspoor et al., 2013, Viswanathan et al., 2016, Thepault et al., 2017) and have informed intervention strategies and public health policy (Cody et al., 2013, Cody et al., 2012).

The accuracy of probabilistic source attribution models is influenced by the degree to which indicative markers in the isolate genome, such as MLST locus alleles, can be placed within a source population. This is relatively straightforward for markers that segregate absolutely by source, but in *C. jejuni* and *C. coli* it is common that alleles are present in more than one population, but at different frequencies. In simple attribution models using MLST data, *C. jejuni* and *C. coli* isolates from chickens in the Netherlands, Senegal and the USA have been more closely related to UK chicken isolate populations rather than to populations from other host species in the same country (Sheppard et al., 2010a). While genomic signatures of host association can transcend geographic structuring within *C. jejuni* and *C. coli* populations, there can be differences in the genotypes that are isolated from different countries (Mohan et al., 2013, Asakura et al., 2012, Kivisto et al., 2014, Islam et al., 2014, Prachantasena et al., 2016). This presents challenges, not only for attributing the source of infections among travellers returning from foreign locations (Mughini-Gras et al., 2014), but also for understanding disease epidemiology in the context of a global food industry.

Following the occupation of a new niche *C. jejuni* and *C. coli* can acquire DNA signatures through recombination (Wilson et al., 2009, Sheppard et al., 2013a, Sheppard et al., 2008) and local DNA signatures via HGT, from resident strains. To quantify the extent to which isolates from the same country share DNA sequence, we compared 15 isolate pairs from different countries, that to minimise the effect of clonal inheritance and host-associated variation were matched by both clonal complex and source. The predicted ancestry of co-inherited SNPs was nearly twice as high among isolates from same country compared to those from different countries. While this represents a relatively weak signal of geographic association, compared to host association, there was a quantifiable local (national) signal that can be used to investigate geographical clustering.

Since recombination introduces more nucleotide substitutions than during mutation in *C. jejuni* and *C. coli* (Webb and Blaser, 2002, Wilson et al., 2009, Morelli et al., 2010), genes with evidence of elevated recombination rates, that share a gene pool, will more rapidly acquire local signals of sequence variation than genes with lower recombination rates. These genes represent potential targets for use as biogeographic epidemiological markers. Pairwise isolate comparison revealed that nucleotide divergence was <1% across the majority of the genome (Table S3); however, some genes consistently had more sequence variation in multiple isolate pairs, potentially indicating enhanced recombination at these loci.

Several of these genes have been annotated with functions associated with DNA processing, transcription, repair and maintenance. This may reflect the mechanisms of recombination and horizontal gene transfer. Other genes with evidence of elevated recombination included those associated with surface exposed proteins with roles in glycosylation, motility and secretion which would form part of an initial interaction with the host/environment (Table S3). Variation in recombination rate could be influenced by differential selection pressure. The *C. jejuni* N-acetyltransferase PseH (Cj1313) plays a key role in O-linked glycosylation, which contributes to flagellar formation, motility and pseudoaminic acid biosyntheseis (Song et al., 2015, McNally et al., 2006) and is important in host colonisation (Guerry et al., 2006). The variable outer membrane protein gene *porA*, which has been used as part of extended MLST schemes (Dingle et al., 2008, Cody et al., 2009) was also among those genes with evidence of elevated recombination. This may explain why weak allopatric signals have been associated with sequence variation in the *porA* gene in addition to source attribution signals (Sheppard et al., 2010a, Smid et al., 2013, Mughini-Gras et al., 2014).

Three efflux pump genes *Cj0034c*, *Cj0619* and *Cj1174* genes, SNPs in which have been implicated in fluoroquinoline resistance, showed elevated recombination and phylogeographic variation (Table S3)(Luangtongkum et al., 2009, Ge et al., 2005). Clinical and agricultural prescription of broad-spectrum antibiotics such as quinolones varies worldwide. Since the late 1990's the agricultural use of fluoroquinolones has declined following governmental intervention in Europe and North America (Chang et al., 2015, Nelson et al., 2007); however, resistant isolates remain common and the level of resistance can vary from country to country (Pham et al., 2015). Higher levels of fluoroquinolone resistance have been observed among isolates from infected individuals who have recently returned from foreign travel (Gaudreau et al., 2014). This is consistent with the higher levels of use in other parts of the world (Zhong et al., 2017). The identification of efflux pump genes among those with high levels of inferred recombination suggests that fluoroquinolones-resistance provides a useful indicator for geographic segregation of isolates.

MLST-based attribution models have been successful in assigning genomes to host reservoirs, using large test datasets (10s of thousands of isolates) to train the model. With additional isolates from other countries and appropriate source information, signatures of local recombination in *Campylobacter* genomes have the potential to identify the country of origin and attribute the source of infection among returning travellers. In this study, 74% of isolates from individuals that had declared recent foreign travel were attributed to non-UK sources; however, in the absence of genetic elements that segregate absolutely by geography, the model relies upon the availability of large reference datasets from reservoir populations in different countries for frequency-dependent attribution. Although this limits the applicability of the approach using currently available data the statistical genetics methodologies employed here provide a quantitative means for identifying genomic signatures of allopatry. This potentially enables the evaluation of transmission dynamics through global livestock

PeerJ Preprints NOT PEER-REVIEWED

trade networks. *Campylobacter* populations are highly structured with some lineages having greater significance in human disease than others, either because of enhanced capacity to survive through slaughter and food production (Yahara et al., 2016) or increased antimicrobial resistance (Wimalarathna et al., 2013, Cody et al., 2010). Monitoring the spread of these strains may be useful for evidence-based interventions targeting strains that are a significant global health burden.

PeerJ Preprints NOT PEER-REVIEWED

Acknowledgements:

SKS is a Wellcome Trust Fellow (088786/C/09/Z) and research in his laboratory is funded by grants from the Medical Research Council (MR/L015080/1), the Food Standards Agency (FS246004) and the Biotechnology and Biological Sciences Research Council (BB/I02464X/1). We acknowledge Canada's Michael Smith Genome Sciences Centre, Vancouver, Canada for whole genome sequencing. GM is supported by a postdoctoral research fellowship from Health and care research, Wales (HF-14-13). KY was supported by a JSPS Research Fellowships for Young Scientists. This publication made use of the PubMLST website (http://pubmlst.org/) developed by Keith Jolley and Martin Maiden (Jolley and Maiden, 2010) and sited at the University of Oxford. The development of that website was funded by the Wellcome Trust. We also acknowledge MRC CLIMB and HPC Wales for the use of HPC facilities and thank William G. Miller for taking the time to read the manuscript prior to publication and provide helpful feedback.

Data Accessibility

Draft assembly genomes and short sequencing reads for all genomes sequenced in this study are available from the NCBI short read archive associated with BioProject: PRJNA312235 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA312235). All assembled genomes used in this study can also be downloaded together from FigShare (doi: 10.6084/m9.figshare.4906634). Individual accession numbers can be found in supplementary table S1.

Author contributions

PeerJ Preprints NOT PEER-REVIEWED

BP, GM, XD and SKS designed research; BP, GM, KY, HW, SM and XD performed research; BP, GM, KY, HW, SM, XD, CTP and SKS analysed results; MDH, ELS, CDC, ENT, KKC, SH, AJC, KAJ, MCJM, NM and SKS provided isolates, genomes or software and BP, GM, CTP and SKS wrote the manuscript.

Conflict of Interest Statement

The authors declare no competing interests.

References

ACHTMAN, M. 2008. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annu Rev Microbiol,* 62**,** 53-70.

ASAKURA, H., BRUGGEMANN, H., SHEPPARD, S. K., EKAWA, T., MEYER, T. F., YAMAMOTO, S. & IGIMI, S. 2012. Molecular evidence for the thriving of Campylobacter jejuni ST-4526 in Japan. *PLoS One,* 7**,** e48394.

BANKEVICH, A., NURK, S., ANTIPOV, D., GUREVICH, A. A., DVORKIN, M., KULIKOV, A. S., LESIN, V. M., NIKOLENKO, S. I., PHAM, S., PRJIBELSKI, A. D., PYSHKIN, A. V., SIROTKIN, A. V., VYAHHI, N., TESLER, G., ALEKSEYEV, M. A. & PEVZNER, P. A. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol,* 19**,** 455-77.

CHANG, Q., WANG, W., REGEV-YOCHAY, G., LIPSITCH, M. & HANAGE, W. P. 2015. Antibiotics in agriculture and the risk to human health: how worried should we be? *Evol Appl,* 8**,** 240-7.

CODY, A. J., CLARKE, L., BOWLER, I. C. & DINGLE, K. E. 2010. Ciprofloxacin-resistant campylobacteriosis in the UK. *Lancet,* 376**,** 1987.

CODY, A. J., MAIDEN, M. J. & DINGLE, K. E. 2009. Genetic diversity and stability of the porA allele as a genetic marker in human Campylobacter infection. *Microbiology,* 155**,** 4145-54.

CODY, A. J., MCCARTHY, N. D., JANSEN VAN RENSBURG, M., ISINKAYE, T., BENTLEY, S. D., PARKHILL, J., DINGLE, K. E., BOWLER, I. C., JOLLEY, K. A. & MAIDEN, M. C. 2013. Real-time genomic epidemiological evaluation of human Campylobacter isolates by use of whole-genome multilocus sequence typing. *J Clin Microbiol,* 51**,** 2526-34.

CODY, A. J., MCCARTHY, N. M., WIMALARATHNA, H. L., COLLES, F. M., CLARK, L., BOWLER, I. C., MAIDEN, M. C. & DINGLE, K. E. 2012. A longitudinal 6-year study of the molecular epidemiology of clinical campylobacter isolates in Oxfordshire, United kingdom. *J Clin Microbiol,* 50**,** 3193-201.

COLLES, F. M., JONES, T. A., MCCARTHY, N. D., SHEPPARD, S. K., CODY, A. J., DINGLE, K. E., DAWKINS, M. S. & MAIDEN, M. C. 2008. Campylobacter infection of broiler chickens in a free-range environment. *Environ Microbiol,* 10**,** 2042-50.

COMAS, I., COSCOLLA, M., LUO, T., BORRELL, S., HOLT, K. E., KATO-MAEDA, M., PARKHILL, J., MALLA, B., BERG, S. & THWAITES, G. 2013. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nature genetics,* 45**,** 1176-1182.

DEARLOVE, B. L., CODY, A. J., PASCOE, B., MERIC, G., WILSON, D. J. & SHEPPARD, S. K. 2016. Rapid host switching in generalist Campylobacter strains erodes the signal for tracing human infections. *ISME J,* 10**,** 721-9.

DENIS, M., SOUMET, C., RIVOAL, K., ERMEL, G., BLIVET, D., SALVAT, G. & COLIN, P. 1999. Development of a m-PCR assay for simultaneous identification of Campylobacter jejuni and C. coli. *Lett Appl Microbiol,* 29**,** 406-10.

DIDELOT, X. & FALUSH, D. 2007. Inference of bacterial microevolution using multilocus sequence data. *Genetics,* 175**,** 1251-66.

DIDELOT, X., NELL, S., YANG, I., WOLTEMATE, S., VAN DER MERWE, S. & SUERBAUM, S. 2013. Genomic evolution and transmission of Helicobacter pylori in two South African families. *Proc Natl Acad Sci U S A,* 110**,** 13880-5.

DIDELOT, X. & WILSON, D. J. 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol,* 11**,** e1004041.

DIERGAARDT, S. M., VENTER, S. N., SPREETH, A., THERON, J. & BROZEL, V. S. 2004. The occurrence of campylobacters in water sources in South Africa. *Water Res,* 38**,** 2589-95.

DINGLE, K. E., COLLES, F. M., FALUSH, D. & MAIDEN, M. C. 2005. Sequence typing and comparison of population biology of Campylobacter coli and Campylobacter jejuni. *J Clin Microbiol,* 43**,** 340-7.

DINGLE, K. E., MCCARTHY, N. D., CODY, A. J., PETO, T. E. & MAIDEN, M. C. 2008. Extended sequence typing of Campylobacter spp., United Kingdom. *Emerg Infect Dis,* 14, 1620-2.

FALUSH, D., WIRTH, T., LINZ, B., PRITCHARD, J. K., STEPHENS, M., KIDD, M., BLASER, M. J., GRAHAM, D. Y., VACHER, S., PEREZ-PEREZ, G. I., YAMAOKA, Y., MEGRAUD, F., OTTO, K., REICHARD, U., KATZOWITSCH, E., WANG, X., ACHTMAN, M. & SUERBAUM, S. 2003. Traces of human migrations in Helicobacter pylori populations. *Science,* 299, 1582-5.

FRENCH, N., BARRIGAS, M., BROWN, P., RIBIERO, P., WILLIAMS, N., LEATHERBARROW, H., BIRTLES, R., BOLTON, E., FEARNHEAD, P. & FOX, A. 2005. Spatial epidemiology and natural population structure of Campylobacter jejuni colonizing a farmland ecosystem. *Environ Microbiol,* 7, 1116-26.

GAGNEUX, S. & SMALL, P. M. 2007. Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. *Lancet Infect Dis,* 7, 328-37.

GAUDREAU, C., BOUCHER, F., GILBERT, H. & BEKAL, S. 2014. Antimicrobial susceptibility of Campylobacter jejuni and Campylobacter coli isolates obtained in Montreal, Quebec, Canada, from 2002 to 2013. *J Clin Microbiol,* 52, 2644-6.

GE, B., MCDERMOTT, P. F., WHITE, D. G. & MENG, J. 2005. Role of efflux pumps and topoisomerase mutations in fluoroquinolone resistance in Campylobacter jejuni and Campylobacter coli. *Antimicrob Agents Chemother,* 49, 3347-54.

GRIEKSPOOR, P., COLLES, F. M., MCCARTHY, N. D., HANSBRO, P. M., ASHHURST-SMITH, C., OLSEN, B., HASSELQUIST, D., MAIDEN, M. C. & WALDENSTROM, J. 2013. Marked host specificity and lack of phylogeographic population structure of Campylobacter jejuni in wild birds. *Mol Ecol,* 22, 1463-72.

GUERRY, P., EWING, C. P., SCHIRM, M., LORENZO, M., KELLY, J., PATTARINI, D., MAJAM, G., THIBAULT, P. & LOGAN, S. 2006. Changes in flagellin glycosylation affect Campylobacter autoagglutination and virulence. *Mol Microbiol,* 60, 299-311.

GUNDOGDU, O., BENTLEY, S. D., HOLDEN, M. T., PARKHILL, J., DORRELL, N. & WREN, B. W. 2007. Re-annotation and re-analysis of the Campylobacter jejuni NCTC11168 genome sequence. *BMC Genomics,* 8, 162.

ISLAM, Z., VAN BELKUM, A., WAGENAAR, J. A., CODY, A. J., DE BOER, A. G., SARKER, S. K., JACOBS, B. C., TALUKDER, K. A. & ENDTZ, H. P. 2014. Comparative population structure analysis of Campylobacter jejuni from human and poultry origin in Bangladesh. *Eur J Clin Microbiol Infect Dis,* 33, 2173-81.

JOKINEN, C. C., SCHREIER, H., MAURO, W., TABOADA, E., ISAAC-RENTON, J. L., TOPP, E., EDGE, T., THOMAS, J. E. & GANNON, V. P. 2010. The occurrence and sources of Campylobacter spp., Salmonella enterica and Escherichia coli O157:H7 in the Salmon River, British Columbia, Canada. *J Water Health,* 8, 374-86.

JOLLEY, K. A. & MAIDEN, M. C. 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics,* 11, 595.

KATOH, K., MISAWA, K., KUMA, K. & MIYATA, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res,* 30, 3059-66.

KIVISTO, R. I., KOVANEN, S., SKARP-DE HAAN, A., SCHOTT, T., RAHKIO, M., ROSSI, M. & HANNINEN, M. L. 2014. Evolution and comparative genomics of Campylobacter jejuni ST-677 clonal complex. *Genome Biol Evol,* 6, 2424-38.

KUMAR, S., STECHER, G. & TAMURA, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol,* 33, 1870-4.

LANE, D. J. 1991. 16S/23S rRNA sequencing. *In:* E., S. & M., G. (eds.) *Nucleic Acid Sequencing Techniques in Bacterial Systematics.* New York: Wiley and Sons.

LAWSON, D. J., HELLENTHAL, G., MYERS, S. & FALUSH, D. 2012. Inference of population structure using dense haplotype data. *PLoS Genet,* 8, e1002453.

LLARENA, A. K., ZHANG, J., VEHKALA, M., VALIMAKI, N., HAKKINEN, M., HANNINEN, M. L., ROASTO, M., MAESAAR, M., TABOADA, E., BARKER, D., GAROFOLO, G., CAMMA, C., DI GIANNATALE, E., CORANDER, J. & ROSSI, M. 2016. Monomorphic genotypes within a generalist lineage of Campylobacter jejuni show signs of global dispersion. *Microb Genom,* 2**,** e000088.

LUANGTONGKUM, T., JEON, B., HAN, J., PLUMMER, P., LOGUE, C. M. & ZHANG, Q. 2009. Antibiotic resistance in Campylobacter: emergence, transmission and persistence. *Future Microbiol,* 4**,** 189-200.

MCCARTHY, N. D., COLLES, F. M., DINGLE, K. E., BAGNALL, M. C., MANNING, G., MAIDEN, M. C. & FALUSH, D. 2007. Host-associated genetic import in Campylobacter jejuni. *Emerg Infect Dis,* 13**,** 267-72.

MCNALLY, D. J., HUI, J. P., AUBRY, A. J., MUI, K. K., GUERRY, P., BRISSON, J. R., LOGAN, S. M. & SOO, E. C. 2006. Functional characterization of the flagellar glycosylation locus in Campylobacter jejuni 81-176 using a focused metabolomics approach. *J Biol Chem,* 281**,** 18489-98.

MCTAVISH, S. M., POPE, C. E., NICOL, C., SEXTON, K., FRENCH, N. & CARTER, P. E. 2008. Wide geographical distribution of internationally rare Campylobacter clones within New Zealand. *Epidemiol Infect,* 136**,** 1244-52.

MERIC, G., YAHARA, K., MAGEIROS, L., PASCOE, B., MAIDEN, M. C., JOLLEY, K. A. & SHEPPARD, S. K. 2014. A reference pan-genome approach to comparative bacterial genomics: identification of novel epidemiological markers in pathogenic campylobacter. *PLoS One,* 9**,** e92798.

MOHAN, V., STEVENSON, M., MARSHALL, J., FEARNHEAD, P., HOLLAND, B. R., HOTTER, G. & FRENCH, N. P. 2013. Campylobacter jejuni colonization and population structure in urban populations of ducks and starlings in New Zealand. *Microbiologyopen,* 2**,** 659-73.

MOODLEY, Y., LINZ, B., YAMAOKA, Y., WINDSOR, H. M., BREUREC, S., WU, J. Y., MAADY, A., BERNHOFT, S., THIBERGE, J. M., PHUANUKOONNON, S., JOBB, G., SIBA, P., GRAHAM, D. Y., MARSHALL, B. J. & ACHTMAN, M. 2009. The peopling of the Pacific from a bacterial perspective. *Science,* 323**,** 527-30.

MORELLI, G., DIDELOT, X., KUSECEK, B., SCHWARZ, S., BAHLAWANE, C., FALUSH, D., SUERBAUM, S. & ACHTMAN, M. 2010. Microevolution of Helicobacter pylori during prolonged infection of single hosts and within families. *PLoS Genet,* 6**,** e1001036.

MUGHINI-GRAS, L., SMID, J. H., WAGENAAR, J. A., A, D. E. B., HAVELAAR, A. H., FRIESEMA, I. H., FRENCH, N. P., GRAZIANI, C., BUSANI, L. & VAN PELT, W. 2014. Campylobacteriosis in returning travellers and potential secondary transmission of exotic strains. *Epidemiol Infect,* 142**,** 1277-88.

MULLNER, P., SPENCER, S. E., WILSON, D. J., JONES, G., NOBLE, A. D., MIDWINTER, A. C., COLLINS-EMERSON, J. M., CARTER, P., HATHAWAY, S. & FRENCH, N. P. 2009. Assigning the source of human campylobacteriosis in New Zealand: a comparative genetic and epidemiological approach. *Infect Genet Evol,* 9**,** 1311-9.

NELSON, J. M., CHILLER, T. M., POWERS, J. H. & ANGULO, F. J. 2007. Fluoroquinolone-resistant Campylobacter species and the withdrawal of fluoroquinolones from use in poultry: a public health success story. *Clin Infect Dis,* 44**,** 977-80.

PARKHILL, J., WREN, B. W., MUNGALL, K., KETLEY, J. M., CHURCHER, C., BASHAM, D., CHILLINGWORTH, T., DAVIES, R. M., FELTWELL, T., HOLROYD, S., JAGELS, K., KARLYSHEV, A. V., MOULE, S., PALLEN, M. J., PENN, C. W., QUAIL, M. A., RAJANDREAM, M. A., RUTHERFORD, K. M., VAN VLIET, A. H., WHITEHEAD, S. & BARRELL, B. G. 2000. The genome sequence of the food-borne pathogen Campylobacter jejuni reveals hypervariable sequences. *Nature,* 403**,** 665-8.

PHAM, N. T., THONGPRACHUM, A., TRAN, D. N., NISHIMURA, S., SHIMIZU-ONDA, Y., TRINH, Q. D., KHAMRIN, P., UKARAPOL, N., KONGSRICHAROERN, T., KOMINE-AIZAWA, S., OKITSU, S., MANEEKARN, N., HAYAKAWA, S. & USHIJIMA, H. 2015. Antibiotic Resistance of

Campylobacter jejuni and C. coli Isolated from Children with Diarrhea in Thailand and Japan. *Jpn J Infect Dis*.

PRACHANTASENA, S., CHARUNUNTAKORN, P., MUANGNOICHAROEN, S., HANKLA, L., TECHAWAL, N., CHAVEERACH, P., TUITEMWONG, P., CHOKESAJJAWATEE, N., WILLIAMS, N., HUMPHREY, T. & LUANGTONGKUM, T. 2016. Distribution and Genetic Profiles of Campylobacter in Commercial Broiler Production from Breeder to Slaughter in Thailand. *PLoS One,* 11**,** e0149585.

PRITCHARD, J. K., STEPHENS, M. & DONNELLY, P. 2000. Inference of population structure using multilocus genotype data. *Genetics,* 155**,** 945-59.

ROUX, F., SPROSTON, E., ROTARIU, O., MACRAE, M., SHEPPARD, S. K., BESSELL, P., SMITH-PALMER, A., COWDEN, J., MAIDEN, M. C., FORBES, K. J. & STRACHAN, N. J. 2013. Elucidating the aetiology of human Campylobacter coli infections. *PLoS One,* 8**,** e64504.

SHEPPARD, S. K., CHENG, L., MERIC, G., DE HAAN, C. P., LLARENA, A. K., MARTTINEN, P., VIDAL, A., RIDLEY, A., CLIFTON-HADLEY, F., CONNOR, T. R., STRACHAN, N. J., FORBES, K., COLLES, F. M., JOLLEY, K. A., BENTLEY, S. D., MAIDEN, M. C., HANNINEN, M. L., PARKHILL, J., HANAGE, W. P. & CORANDER, J. 2014. Cryptic ecology among host generalist Campylobacter jejuni in domestic animals. *Mol Ecol,* 23**,** 2442-51.

SHEPPARD, S. K., COLLES, F., RICHARDSON, J., CODY, A. J., ELSON, R., LAWSON, A., BRICK, G., MELDRUM, R., LITTLE, C. L., OWEN, R. J., MAIDEN, M. C. & MCCARTHY, N. D. 2010a. Host association of Campylobacter genotypes transcends geographic variation. *Appl Environ Microbiol,* 76**,** 5269-77.

SHEPPARD, S. K., COLLES, F. M., MCCARTHY, N. D., STRACHAN, N. J., OGDEN, I. D., FORBES, K. J., DALLAS, J. F. & MAIDEN, M. C. 2011. Niche segregation and genetic structure of Campylobacter jejuni populations from wild and agricultural host species. *Mol Ecol,* 20**,** 3484-90.

SHEPPARD, S. K., DALLAS, J. F., STRACHAN, N. J., MACRAE, M., MCCARTHY, N. D., WILSON, D. J., GORMLEY, F. J., FALUSH, D., OGDEN, I. D., MAIDEN, M. C. & FORBES, K. J. 2009. Campylobacter genotyping to determine the source of human infection. *Clin Infect Dis,* 48**,** 1072-8.

SHEPPARD, S. K., DALLAS, J. F., WILSON, D. J., STRACHAN, N. J., MCCARTHY, N. D., JOLLEY, K. A., COLLES, F. M., ROTARIU, O., OGDEN, I. D., FORBES, K. J. & MAIDEN, M. C. 2010b. Evolution of an agriculture-associated disease causing Campylobacter coli clade: evidence from national surveillance data in Scotland. *PLoS One,* 5**,** e15708.

SHEPPARD, S. K., DIDELOT, X., JOLLEY, K. A., DARLING, A. E., PASCOE, B., MERIC, G., KELLY, D. J., CODY, A., COLLES, F. M., STRACHAN, N. J., OGDEN, I. D., FORBES, K., FRENCH, N. P., CARTER, P., MILLER, W. G., MCCARTHY, N. D., OWEN, R., LITRUP, E., EGHOLM, M., AFFOURTIT, J. P., BENTLEY, S. D., PARKHILL, J., MAIDEN, M. C. & FALUSH, D. 2013a. Progressive genome-wide introgression in agricultural Campylobacter coli. *Mol Ecol,* 22**,** 1051-64.

SHEPPARD, S. K., DIDELOT, X., MERIC, G., TORRALBO, A., JOLLEY, K. A., KELLY, D. J., BENTLEY, S. D., MAIDEN, M. C., PARKHILL, J. & FALUSH, D. 2013b. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. *Proc Natl Acad Sci U S A,* 110**,** 11923-7.

SHEPPARD, S. K., JOLLEY, K. A. & MAIDEN, M. C. 2012. A Gene-By-Gene Approach to Bacterial Population Genomics: Whole Genome MLST of Campylobacter. *Genes (Basel),* 3**,** 261-77.

SHEPPARD, S. K., MCCARTHY, N. D., FALUSH, D. & MAIDEN, M. C. 2008. Convergence of Campylobacter species: implications for bacterial evolution. *Science,* 320**,** 237-9.

SMID, J. H., MUGHINI GRAS, L., DE BOER, A. G., FRENCH, N. P., HAVELAAR, A. H., WAGENAAR, J. A. & VAN PELT, W. 2013. Practicalities of using non-local or non-recent multilocus sequence typing data for source attribution in space and time of human campylobacteriosis. *PLoS One,* 8**,** e55029.

SONG, W. S., NAM, M. S., NAMGUNG, B. & YOON, S. I. 2015. Structural analysis of PseH, the Campylobacter jejuni N-acetyltransferase involved in bacterial O-linked glycosylation. *Biochem Biophys Res Commun,* 458**,** 843-8.

TAMURA, K., STECHER, G., PETERSON, D., FILIPSKI, A. & KUMAR, S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol,* 30**,** 2725-9.

THEPAULT, A., MERIC, G., RIVOAL, K., PASCOE, B., MAGEIROS, L., TOUZAIN, F., ROSE, V., BEVEN, V., CHEMALY, M. & SHEPPARD, S. K. 2017. Genome-Wide Identification of Host-Segregating Epidemiological Markers for Source Attribution in Campylobacter jejuni. *Appl Environ Microbiol,* 83.

VISWANATHAN, M., PEARL, D. L., TABOADA, E. N., PARMLEY, E. J., MUTSCHALL, S. & JARDINE, C. M. 2016. Molecular and Statistical Analysis of Campylobacter spp. and Antimicrobial-Resistant Campylobacter Carriage in Wildlife and Livestock from Ontario Farms. *Zoonoses Public Health*.

VOS, M. & DIDELOT, X. 2009. A comparison of homologous recombination rates in bacteria and archaea. *ISME J,* 3**,** 199-208.

WEBB, G. F. & BLASER, M. J. 2002. Dynamics of bacterial phenotype selection in a colonized host. *Proc Natl Acad Sci U S A,* 99**,** 3135-40.

WILSON, D. J., GABRIEL, E., LEATHERBARROW, A. J., CHEESBROUGH, J., GEE, S., BOLTON, E., FOX, A., FEARNHEAD, P., HART, C. A. & DIGGLE, P. J. 2008. Tracing the source of campylobacteriosis. *PLoS Genet,* 4**,** e1000203.

WILSON, D. J., GABRIEL, E., LEATHERBARROW, A. J., CHEESBROUGH, J., GEE, S., BOLTON, E., FOX, A., HART, C. A., DIGGLE, P. J. & FEARNHEAD, P. 2009. Rapid evolution and the importance of recombination to the gastroenteric pathogen Campylobacter jejuni. *Mol Biol Evol,* 26**,** 385-97.

WIMALARATHNA, H. M., RICHARDSON, J. F., LAWSON, A. J., ELSON, R., MELDRUM, R., LITTLE, C. L., MAIDEN, M. C., MCCARTHY, N. D. & SHEPPARD, S. K. 2013. Widespread acquisition of antimicrobial resistance among Campylobacter isolates from UK retail poultry and evidence for clonal expansion of resistant lineages. *BMC Microbiol,* 13**,** 160.

YAHARA, K., FURUTA, Y., OSHIMA, K., YOSHIDA, M., AZUMA, T., HATTORI, M., UCHIYAMA, I. & KOBAYASHI, I. 2013. Chromosome painting in silico in a bacterial species reveals fine population structure. *Molecular Biology and Evolution,* 30**,** 1454-64.

YAHARA, K., MERIC, G., TAYLOR, A. J., DE VRIES, S. P., MURRAY, S., PASCOE, B., MAGEIROS, L., TORRALBO, A., VIDAL, A., RIDLEY, A., KOMUKAI, S., WIMALARATHNA, H., CODY, A. J., COLLES, F. M., MCCARTHY, N., HARRIS, D., BRAY, J. E., JOLLEY, K. A., MAIDEN, M. C., BENTLEY, S. D., PARKHILL, J., BAYLISS, C. D., GRANT, A., MASKELL, D., DIDELOT, X., KELLY, D. J. & SHEPPARD, S. K. 2016. Genome-wide association of functional traits linked with Campylobacter jejuni survival from farm to fork. *Environ Microbiol*.

ZHONG, L. L., STOESSER, N., DOI, Y., SHEN, C., HUANG, X. & TIAN, G. B. 2017. Carriage of beta-lactamase-producing Enterobacteriaceae by Chinese travellers. *Lancet Infect Dis,* 17**,** 138-139.

PeerJ Preprints NOT PEER-REVIEWED

Tables and Figures

Figure 1: Population structure of *Campylobacter* isolates used in this study. Phylogenetic trees were constructed from a whole-genome alignment of **(A)** *C. jejuni* (n=229) and **(B)** *C. coli* (n=55) isolates based on 103,878 and 806,657 variable sites, respectively, using an approximation of the maximum likelihood algorithm (Tamura et al., 2013, Kumar et al., 2016). Leaves on the tree are coloured by source country, UK (green circles), Canada (red) and USA (blue). Ancestral *C. coli* clades (1, 2 and 3)(Sheppard et al., 2010b) are annotated and common clonal complexes (CC) based on four or more shared alleles in seven MLST house-keeping genes (Dingle et al., 2005).

Figure 2: Co-ancestry matrix with population structure and genetic flux. (A) The colour of each cell of the matrix indicates proportion of DNA chunks in a recipient genome (row) from a donor genome (column). The colour ranges from little (yellow) to a large amount of DNA from the donor strain (blue). Diagonal white cells indicate chunks of DNA that are shared between the pairs of isolates and masked in the comparison in Figure 2B. The trees above and to the left show clustering of the paired isolates with leaves coloured by source country (UK in green, Canada in red). (**B)** Box plot comparing total proportion of chunks of DNA inherited by a recipient from donors either within or between countries. The total proportion is significantly higher for chunks of DNA from donor strains of the same country compared to those from different countries (p<10^{-9}, Wilcoxon rank sum test).

Figure 3: Pairwise comparison of nucleotide diversity in the core genome. Above: Estimated values of the per-nucleotide statistic reflecting relative intensity of recombination at each site plotted along the NCTC11168 reference genome. **Left:** Core genome phylogeny of selected paired isolates (matched by CC and source host), with clonal complex indicated. **Centre:** Matrix of gene-by-gene pairwise comparison along the NCTC11168 reference genome of our selected pairs. Each row represents a pairwise comparison of selected paired of isolates. Each column is a gene from the NCTC11168 reference genome. Panels of the matrix are coloured based on nucleotide divergence for that gene in each pair: from no nucleotide diversity (0%, white), through some nucleotide diversity (~1%, red) to high levels of nucleotide diversity (up to 2%, blue). The per-nucleotide scan of relative intensity of recombination is aligned with our gene-by-gene pairwise comparison of nucleotide diversity and the location of seven putative epidemiological markers for geographical segregation are indicated.

Figure 4: Assignment of human clinical cases of campylobacteriosis to origin country, including patients with history of recent foreign travel. (A) Assignment of human clinical cases of campylobacteriosis to origin country using epidemiological markers of biogeography and the Bayesian clustering algorithm Structure. Each isolate is represented by a vertical bar, showing the estimated probability that it comes from each of the putative source countries, including the UK (green), USA (blue) and Canada (red). Isolates are ordered by attributed source. (B) Boxplots of predicted attribution probabilities for the three locations. (C) Isolates from Oxford clinical dataset with declared history of recent foreign travel. The model correctly assigned 34 of 46 (73.9%) isolates to a non-UK origin. (D) Attribution of Oxford clinical isolates between UK, USA and Canada source populations. Isolates with declared recent foreign travel are shown in blue.

Table 1: Isolate pairs matched by clonal complex and host.

Table 2: Shared ancestry analysis and estimation of pairwise recombination rates. The time to the most recent common ancestor (TMRCA) for each selected pair was estimated with 95% confidence intervals (TMRCA-CI). The ratio of rates at which recombination and mutation introduce polymorphism (r/m) was also calculated with 95% confidence intervals (r/m-CI). In addition, the number of definitely recombined genes (probability > 95%) is also shown. The two *C. coli* pairs are coloured in red.

Supplementary material

Figure S1: Neighbour-joining trees of all 57 genes showing greater than 1% diversity between pairs. Genes used in attribution model are labelled in red.

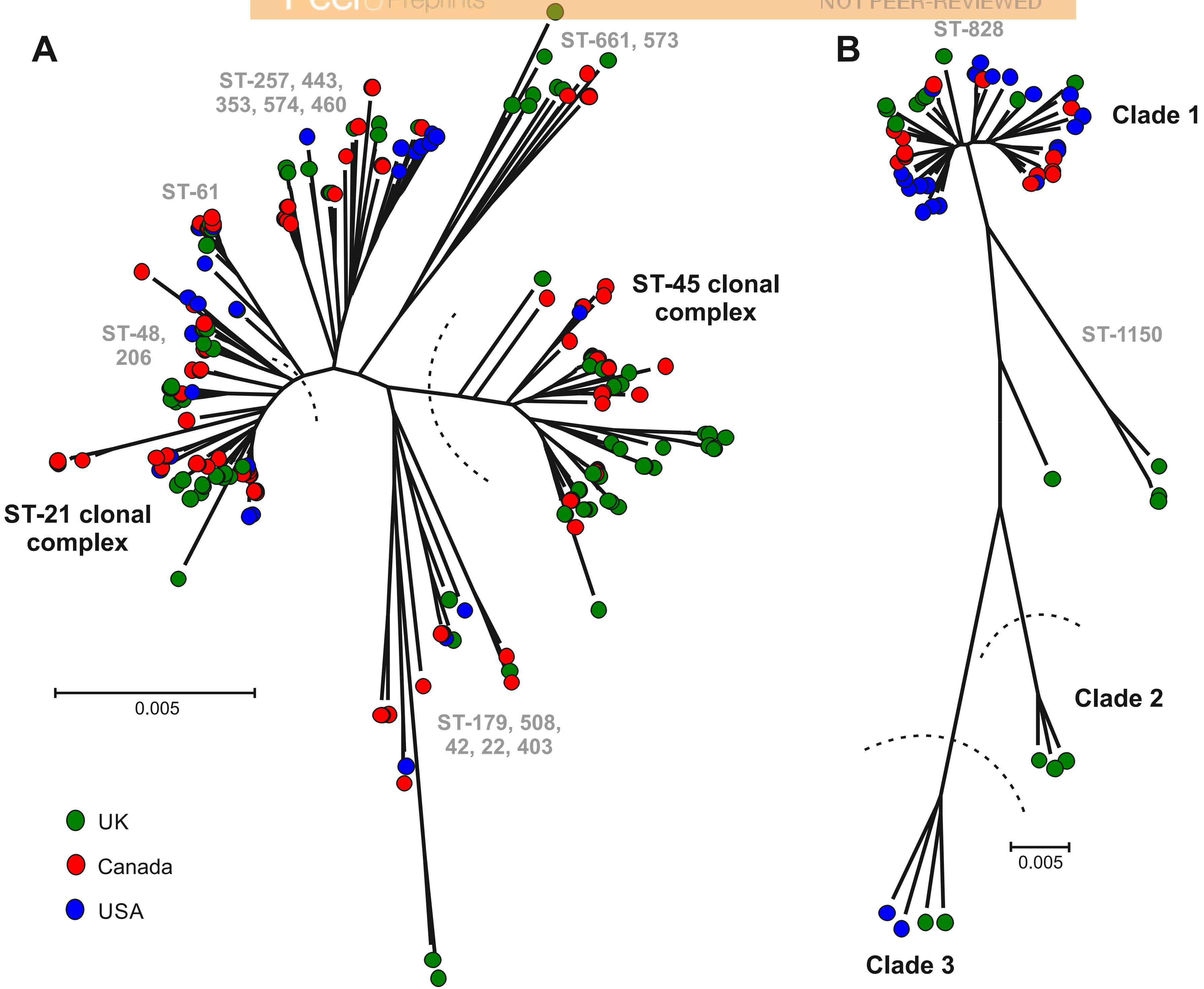
Figure S2: Phylogeny of 7 highly recombining epidemiological markers used to attribute biogeography using structure.

Table S1: List of isolates used, including details of genome accession numbers.

Table S2: List of Oxford clinical isolates used to test our biogeography attribution model. Isolate genomes and metadata downloaded from pubMLST.org/Campylobacter

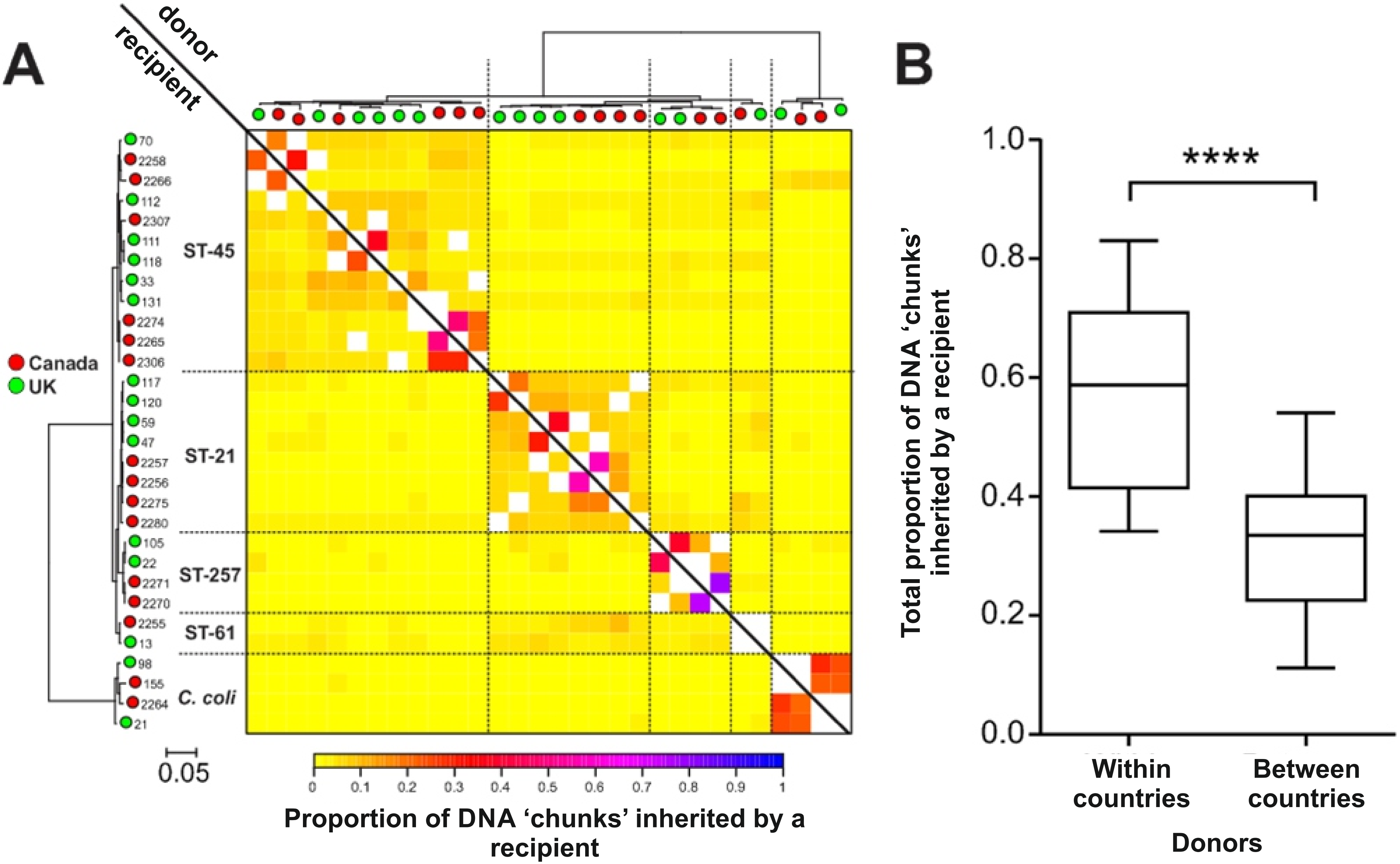
Table S3: List of biogeographical epidemiological markers, including lists of highly recombining genes as determined by pairwise analysis of nucleotide diversity (more than 1% diversity); and genes used to model biogeographical segregation in structure.

PeerJ Preprints NOT PEER-REVIEWED

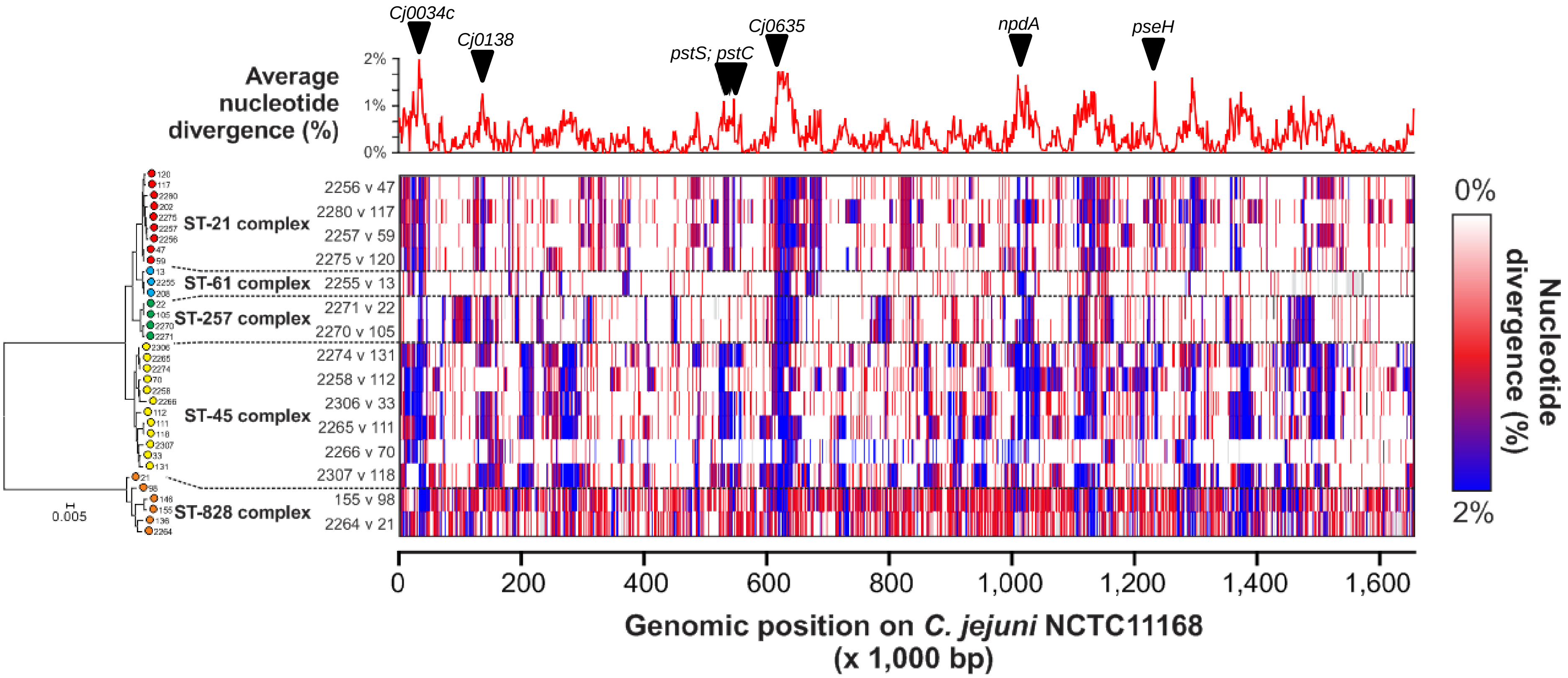


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2638v2 | CC BY 4.0 Open Access | rec: 19 May 2017, publ: 19 May 2017

PeerJ Preprints NOT PEER-REVIEWED

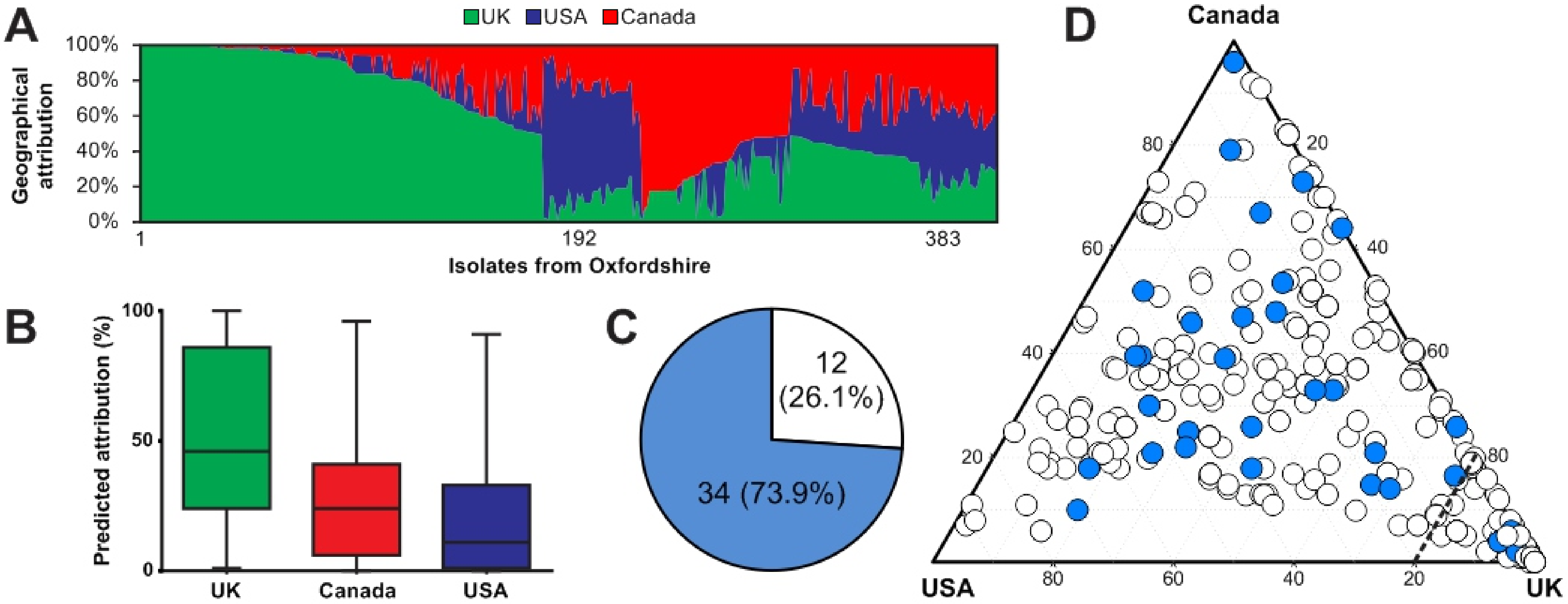


PeerJ Preprints NOT PEER-REVIEWED



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2638v2 | CC BY 4.0 Open Access | rec: 19 May 2017, publ: 19 May 2017

PeerJ Preprints NOT PEER-REVIEWED



PeerJ Preprints NOT PEER-REVIEWED

				MLST genes							
Pair	Isolate	Origin	Host	aspA	glnA	gltA	glyA	pgm	tkt	uncA	Clonal Complex
1	2256	Canada	cattle	2	1	1	3	2	1	5	ST-21
	47	UK	cattle	2	1	1	3	2	1	5	ST-21
2	2280	Canada	human	2	1	1	3	2	1	5	ST-21
	117	UK	human	2	1	1	3	2	1	5	ST-21
3	2271	Canada	chicken	9	2	4	62	4	5	17	ST-257
	22	UK	chicken	9	2	4	62	4	5	6	ST-257
4	2274	Canada	duck	4	7	10	4	1	7	1	ST-45
	131	UK	duck	4	7	10	4	1	7	1	ST-45
5	2258	Canada	chicken	4	7	10	4	1	7	1	ST-45
	112	UK	chicken	4	7	10	4	1	7	1	ST-45
6	2306	Canada	human	4	7	10	4	1	7	1	ST-45
	33	UK	human	4	7	10	4	1	7	1	ST-45
7	2255	Canada	cattle	1	4	2	2	6	3	17	ST-61
	13	UK	cattle	1	4	2	2	6	3	17	ST-61
8	2264	Canada	chicken	33	39	30	203	113	47	17	ST-828
	21	UK	chicken	33	39	30	82	104	43	17	ST-828
9	2257	Canada	cattle	2	1	1	3	2	1	5	ST-21
	59	UK	cattle	2	1	1	3	2	1	5	ST-21
10	2275	Canada	human	2	1	1	3	2	1	5	ST-21
	120	UK	human	2	1	1	3	2	1	5	ST-21
11	2270	Canada	chicken	9	2	4	62	4	5	17	ST-257
	105	UK	chicken	9	2	4	62	4	5	6	ST-257
12	2265	Canada	chicken	4	7	10	4	1	7	1	ST-45
	111	UK	chicken	4	7	10	4	1	7	1	ST-45
13	2266	Canada	chicken	4	7	10	4	1	7	1	ST-45
	70	UK	chicken	4	7	10	4	1	7	1	ST-45
14	2307	Canada	human	4	7	10	4	1	7	1	ST-45
	118	UK	human	4	7	10	4	1	7	1	ST-45
15	155	Canada	cattle	33	39	30	82	104	85	68	ST-828
	98	UK	cattle	33	39	30	82	104	56	17	ST-828

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2638v2 | CC BY 4.0 Open Access | rec: 19 May 2017, publ: 19 May 2017

PeerJ Preprints NOT PEER-REVIEWED

Isolate pair	TMRCA	TMRCA-CI	r/m	r/m-CI	Definitely recombined genes (probability>0.95)
2256 vs 47	2.8	[2.5;3.2]	23.1	[20.2;26.3]	210
2280 vs 117	3.9	[3.2;4.5]	23.1	[19.0;28.3]	273
2271 vs 22	1.9	[1.6;2.3]	34.5	[28.8;39.6]	194
2274 vs 131	3.3	[2.9;3.9]	38.8	[32.0;43.6]	385
2258 vs 112	3.4	[3.0;3.8]	32.1	[28.4;37.0]	336
2306 vs 33	3.7	[3.2;4.2]	24.5	[21.0;27.9]	280
2255 vs 13	1.2	[1.0;1.5]	25.2	[20.1;30.2]	99
2264 vs 21	**22.7**	**[20.7;24.8]**	**3.9**	**[3.2;4.9]**	**187**
2257 vs 59	3	[2.5;3.5]	23.5	[19.3;27.8]	219
2275 vs 120	2.7	[2.3;3.1]	24.1	[20.4;27.9]	194
2270 vs 105	2.2	[1.9;2.5]	30.5	[26.6;34.8]	224
2265 vs 111	3.7	[3.3;4.2]	32.8	[28.6;36.7]	372
2266 vs 70	1.3	[1.1;1.5]	38	[33.4;41.4]	147
2307 vs 118	3.9	[3.4;4.6]	31.9	[26.2;37.4]	379
155 vs 98	**27.1**	**[25.0;29.3]**	**3.6**	**[3.0;4.3]**	**236**

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2638v2 | CC BY 4.0 Open Access | rec: 19 May 2017, publ: 19 May 2017