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 Wimaralathna[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[9],William G. Miller[9], Alison J. Cody[5], Keith A. Jolley[5], Martin M. J. Maiden[5. 10],Noel D. McCarthy[5,10. 11], Xavier Didelot[11],Craig T. Parker[9]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, Bath University, Claverton Down, Bath, BA2 7AY, UK; [3]Biostatistics Center, Kurume University, Kurume, Fukuoka 830-0011, 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, ON K1A 0K9, Canada; [7]Canadian Food Inspection Agency, 960 Carling Avenue, Ottawa, ON K1A 0Y9, Canada; [8]National Microbiology Laboratory at Lethbridge, Public Health Agency of Canada, PO Box 640, Township Rd. 9-1, Lethbridge, AB T1J 3Z4, Canada;[9]Produce Safety and Microbiology Research Unit, Agricultural Research Service, US Department of Agriculture, Albany, California, USA;[10]NIHR HPRU in Gastrointestinal Infections[11]University of Warwick, Coventry, CV4 7AL, UK; [11]Department of Infectious Disease Epidemiology, Imperial College London, London, W2 1PG, UK

#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

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 distances and genetic distances, 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*, source attribution based on genomic data has been successful in distinguishing the infected host species, but not geographical origin. We investigate biogeographical signals 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 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 origins. A further 383 UK clinical isolates were used to detect signals of foreign travel. Patient records indicated that 46 cases had travelled abroad less than two weeks prior to sampling, and 34 (74%) of those *Campylobacter* genomes, were deemed to be from a non-UK origin. Detection of signals of biogeographical differences in *Campylobacter* genomes will contribute to improved source attribution of clinical *Campylobacter* infection and inform intervention strategies to reduce campylobacteriosis.

Introduction

Geographical structuring 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 structuring has also been observed in the human stomach 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).

Tuberculosis and *H. pylori* are primarily human pathogens, but for *Campylobacter,* animals are the principal reservoir for human infection. International trade, particularly in agricultural animals such as chicken and poultry products, provides a vehicle forglobal 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, known as clonal complexes, that share four or more alleles at 7 multi-locus sequence typing (MLST) level (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 lineage transfers to a new animal host it may acquire DNA from the resident population by HGT. This has been shown in host generalist *Campylobacter jejuni* lineages isolated from chicken that sometimes contain alleles that originated in chicken-specialist genotypes (McCarthy et al., 2007, Wilson et al., 2008). Here we applied comparable approaches to investigate if HGT can lead to signatures of recombination that discriminate isolates from North America and the UK using genomic data. Using matched pairs of Canadian 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, augmented by 163 previously published genomes (Sheppard et al., 2014, Sheppard et al., 2013a, Sheppard et al., 2013b). Sequencing reads for all genomes studied are available from the NCBI short read archive associated with BioProject: PRJNA312235 (Individual SRA accession numbers can be found in Table S1).

Canadian isolates: Isolates were collected from chicken and bovine faecal samples between July 2004 and July 2006from farms at diverse locations in Alberta. Samples were placed on ice and processed within 6 h as described by (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°C for 24 h under microaerophilic conditions prior to plating 20 µl onto supplemented charcoal cefoperazone deoxycholate agar (CCDA). The plates were incubated for a further 48h at 42°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. A multiplex PCR assay was used to detect 16S rRNA gene sequences and *C. jejuni* and *C. coli* specific primers *map*A and *ceu*E, respectively

Commented [BP1]: Emma, Ed & Cathy to check details.

Commented [ELS2]: Yes, all chicken and cattle samples were collected in Alberta and dates entered.

(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). 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.

American 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 microtiter plate containing 6 ml 1× Anaerobe Basal Broth (Oxoid) amended with Preston supplement (when reconstituted consists of: amphotericin B (0.01 mg/ml), rifampicin (0.01mg/ml), trimethoprim lactate (0.01mg/ml), and polymixin B (5UI/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 cultureswere plated onto anaerobe basal agar (ABA) plates, amended with 5% laked horse blood and CAT supplement (cefoperazone (0.008mg/ml), amphotericin B (0.01 mg/ml), and teicoplanin (0.004 mg/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.

Commented [BP3]: Craig et al to check details.

DNA was extracted from a pure culture colony using Wizard Genomic DNA Purification Kit (Promega, Madison, WI). *Campylobacter* species was designated 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), 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 2 × 250 bp paired end v2 reagent kit on a MiSeq instrument (Illumina) at 13.5 pM, following manufacturer's protocols. Reads were obtained from SeqWright (Houston, TX).

UK isolates: Sequenced isolates from Canada and the USA were augmented by 163 previously published *Campylobacter* genomes collected between 1980 and 2012 from a range of sources, including cattle (54), chicken (80), pig (9), environmental (49), wild bird species (12) and human clinical cases (73) (Sheppard et al., 2014, Sheppard et al., 2013a, Sheppard et al., 2013b).

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

Population structure and selection of isolate pairs

Isolate genomes were archived in the web-accessible BIGSdb database that supports functionality for identifying 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 comprised 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 an approximation of the maximum likelihood algorithm (Tamura et al., 2013, Kumar et al., 2016).UK isolates from matching hosts were paired with their closest match from Canada. In total, 15 pairs of isolates were matched by source host and clonal complex (Figure 1).All paired isolates shared 1,378 genes giving rise to a core-genome alignment of 1,287,560 bp.

Commented [NM4]: Maybe put the criteria in here (< 1200 bp differences across the 1378 genes) rather than results.

Analysis of co-ancestry and inference of recombination hot regions

193 The predicted co-ancestry of the paired isolates was determined based on whole genome
194 sequences using *fine*STRUCTURE (Yahara et al., 2013) and visualized as a heat map (Figure
195 2). This algorithm infers the number of clusters (K) and partitions the strains into K
196 subgroups with indistinguishable genetic ancestry, based on the likelihood of co-ancestry
197 using a Bayesian MCMC (Markov chain Monte Carlo) approach (Lawson et al., 2012).
198 Previous estimates of recombination rate and generation time (Webb and Blaser, 2002,
199 Wilson et al., 2009, Morelli et al., 2010) were used to prepare a recombination map file
200 specifying the same recombination rate per-site per-generation of SNPs. The predicted
201 ancestry of co-inherited SNPs or chunks was calculated and SNPs with uncertain estimates of
202 their donor of more than 20 kb results were removed. The results were visualised in the
203 UCSC (The University of California Santa Cruz) browser (Kent et al., 2002) (Figure 2).
204 Following burn-in, Markov chain Monte Carlo (MCMC) iterations were run 100,000 times
205 in *fine*STRUCTURE (version 0.02) (Lawson et al., 2012) with a thinning interval of 100.
206 Population assignments runs were performed twice.
207
208 The time to the most recent common ancestor (TMRCA) in each pair was estimated using the
209 model described in Didelot et al., (2013) and summarised here briefly. Pairs of genomes
210 share a common ancestor $\boldsymbol{t}$ years ago and have been subject to mutation at a rate $\boldsymbol{\mu}$ and
211 recombination at rate $\boldsymbol{\rho}$. The mutation rate of 2.9×10^{-5} per site per year was used as reported
212 in Sheppard et al., (2010b), which is similar to the rates estimated in Wilson et al., (2008,
213 2009)) . The effect of recombination is to introduce a high density of polymorphism similar
214 to the ClonalFrame model (cite) but with the advantage that this density can vary between
215 recombination events to reflect differences in evolutionary distance between donors and

Commented [XD5]: I think this whole paragraph was copied by mistake and needs to be removed since here we do not use simulations

Commented [XD6]: Didelot X, Falush D (2007) Inference of bacterial microevolution using multilocus sequence data. Genetics 175:1251–66. doi: 10.1534/genetics.106.063305

Didelot X, Wilson DJ (2015) ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. PLOS Comput Biol 11:e1004041. doi: 10.1371/journal.pcbi.1004041

recipients(cite). In each pairwise comparison, the TMRCA and recombination rate parameters are estimated based on a core genome alignment, with 95% credibility intervals (Table 2).

Epidemiological markers of geographical clustering

Neighbour-joining phylogenetic trees were constructed for all genes that demonstrated pairwise diversity above 2% nucleotide diversity (Table S2). 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 (Table S1).

Attribution of clinical isolates to country based on 7 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.

Commented [XD7]: Morelli G, Didelot X, Kusecek B, et al (2010) Microevolution of Helicobacter pylori during prolonged infection of single hosts and within families. PLoS Genet 6:e1001036. doi: 10.1371/journal.pgen.1001036

Didelot X, Nell S, Yang I, et al (2013) Genomic evolution and transmission of Helicobacter pylori in two South African families. Proc Natl Acad Sci USA 110:13880–13885. doi: 10.1073/pnas.1304681110

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 sequence types ST-257, ST-283, ST-353, ST-354, ST-443, ST-573, ST-574 and ST-661 clonal complexes, cattle specialist ST-61 and ST-42 clonal complexes, and host generalist ST-21, ST-45, ST–206 and ST-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, < 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 predicted co-ancestry of the paired isolates was calculated based on core genome alignments using *fine*STRUCTURE (Yahara et al., 2013) (Figure 2). DNA sequence haplotype regions were coloured by predicted inheritance from donor isolates and the average frequency of co-ancestry of DNA 'chunks' from isolates within the same country (0.58) was significantly greater than that for isolates from different countries (0.32). The degree of inheritance for each gene was calculated and genes that have been predicted to inherit the most DNA from donor isolates of different countries was surmised (Table S2).

Commented [B8]: Koji - Pvalue and test used

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 2.9×10^{-5} per site per year reported in Sheppard et al. (2010b), which is consistent with estimates in Wilson et al. (2008, 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(cite) was estimated to be around 20-30 except in the two *C. coli* pairs with larger TMRCA, for which a much smaller value was estimated around r/m=4.

Commented [NM9]: The approx. tenfold longer TMRCA of C. coli vs C. jejuni pairs is so striking – and within pairs based on the 1200bp difference cut off that may operate differently across species, and a much less densely sampled C. coli population to generate sampling of close isolates. So many reasons may be contributing to this apparent difference even if true. To me this questions treating these coli and jejuni pairs as dealing with the same thing.

Commented [XD10]: Vos M, Didelot X (2009) A comparison of homologous recombination rates in bacteria and archaea. ISME J 3:199–208. doi: 10.1038/ismej.2008.93

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 2% 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 2% 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 (>2%) 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. 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 our collection of xxx isolates 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). The proportion of UK isolates correctly attributed to the UK reference population was 70%, while the proportion of North American isolatescorrectly attributed was slightly higher at 76%.

Attribution of clinical isolates to country based on sevenselected 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.

Commented [ET11]: Ben, did you ever have a chance to compare the performance of these 7 genes vs. the geographical attribution estimates obtained with genes with lesser geographical signal? It might be nice to have an idea of how much improvement in performance was obtained.

Commented [NM12]: This looks very good but looking at the figure I think that it is true to say that about 50% of those without a travel history would also have been assigned as non-UK. I think that it would be informative to give the figure for those without a travel history.

Discussion

In gut dwelling bacteria, isolation 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), secondly, as the 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 structuring 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, has 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) 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 of which indicative markers in the isolate genome, such as MLST locus alleles, can be placed within a source population. This would be 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

Commented [NM13]: Nothing particular to gut dwelling bacteria as opposed to e.g. maxillary sinus dwelling bacteria etc.

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 werematched 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 divergencewas <2% across the majority of the genome (Figure 2, Table S2).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 S2).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, that have been implicated in fluoroquinoline resistance, showed elevated recombination and phylogeographic variation (Table S2)(Luangtongkum et al., 2009, Ge et al., 2005). Clinical and agricultural prescription

Commented [NM14]: Here you cover some genes used for geography and others not. Good to make explicit. Only the supplementary material shows the 7 genes selected and this is reasonably central and would be better included even as a sentence in the paper. More generally I wonder about you considering the criteria for selecting genes along the lines of (1) first principles that substantial recombination needed to allow generation of a signal. (2) some biological processes e.g. Cipro resistance likely to support geographical patterning and others e.g. "initial interaction with host" not likely to be geographically informative etc. but could appear to be if e.g. more samples from a particular place were from a particular species, and that your selection was based on empirical association.

386 of broad-spectrum antibiotics such as quinolones varies worldwide. Since the late 1990's the
387 agricultural use of fluoroquinolones has declined following governmental intervention in
388 Europe and North America (Chang et al., 2015, Nelson et al., 2007). However, resistant
389 isolates remain common and the level of resistance can vary from country to country (Pham
390 et al., 2015). Higher levels of fluoroquinolone resistance has been observed among isolates
391 from patients returning from foreign travel (Gaudreau et al., 2014). The identification of
392 genes associated with efflux pumps, among those with high levels of inferred recombination,
393 suggests a role in the emergence of fluoroquinolones-resistance and provides a useful
394 indicator for geographic segregation of isolates.
395
396 Using signatures of local recombination in *Campylobacter* genomes, has the potential to
397 identify the country of origin and attribute the source of infection, among returning
398 travellers. In this study 74% of isolates from individuals, that had declared recent foreign
399 travel, were attributed to non-UK sources. However, in the absence of genetic elements that
400 segregate absolutely by geography, the model relies upon the availability of large reference
401 datasets from reservoir populations in different countries for frequency-dependent attribution.
402 Although this limits the applicability of the approach using currently available datathe
403 statistical genetics methodologies employed here provide a quantitative means for identifying
404 genomic signatures of allopatry. This potentially enables the evaluation of transmission
405 dynamics through global livestock trade networks. *Campylobacter* populations are highly
406 structured with some lineages having greater significance in human disease than others, either
407 because of enhanced capacity to survive through slaughter and food production [Ref] or
408 increased antimicrobial resistance (Wimalarathna et al., 2013, Cody et al., 2010). Monitoring

Commented [B15]: Yahara and Meric when published (In press, EM)

409 the spread of these strains may be useful for evidence-based interventions targeting strains
410 that are a significant global health burden.
411

Acknowledgements:

SKS is a WellcomeTrust 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 and USA sequencing for 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. We also acknowledge MRC CLIMB and HPC Wales for the use of HPC facilities.

Data Accessibility

Draft assembly genomes and short sequencing reads generated in this study have been deposited in NCBIGenBank database and/or the Short Read Archive associated with BioProject(s): PRJNA312235. Individual accession numbers can be found in table S1.

Author contributions

BP, GM, XD and SKS designed research; BP, GM, KY, HW, SM and XD performed research; BP, GM, KY, HW, SM, NM, XD, CTP and SKS analysed results; MDH, ELS, CDC, ENT, KKC, SH, WGM, 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.

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., LAWSON, D., DARLING, A. & FALUSH, D. 2010. Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics,* 186**,** 1435-49.

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.

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.

484 FRENCH, N., BARRIGAS, M., BROWN, P., RIBIERO, P., WILLIAMS, N., LEATHERBARROW, H., BIRTLES,
485 R., BOLTON, E., FEARNHEAD, P. & FOX, A. 2005. Spatial epidemiology and natural population
486 structure of Campylobacter jejuni colonizing a farmland ecosystem. *Environ Microbiol,* 7**,**
487 1116-26.
488 GAGNEUX, S. & SMALL, P. M. 2007. Global phylogeography of Mycobacterium tuberculosis and
489 implications for tuberculosis product development. *Lancet Infect Dis,* 7**,** 328-37.
490 GAUDREAU, C., BOUCHER, F., GILBERT, H. & BEKAL, S. 2014. Antimicrobial susceptibility of
491 Campylobacter jejuni and Campylobacter coli isolates obtained in Montreal, Quebec,
492 Canada, from 2002 to 2013. *J Clin Microbiol,* 52**,** 2644-6.
493 GE, B., MCDERMOTT, P. F., WHITE, D. G. & MENG, J. 2005. Role of efflux pumps and topoisomerase
494 mutations in fluoroquinolone resistance in Campylobacter jejuni and Campylobacter coli.
495 *Antimicrob Agents Chemother,* 49**,** 3347-54.
496 GRIEKSPOOR, P., COLLES, F. M., MCCARTHY, N. D., HANSBRO, P. M., ASHHURST-SMITH, C., OLSEN, B.,
497 HASSELQUIST, D., MAIDEN, M. C. & WALDENSTROM, J. 2013. Marked host specificity and
498 lack of phylogeographic population structure of Campylobacter jejuni in wild birds. *Mol Ecol,*
499 22**,** 1463-72.
500 GUERRY, P., EWING, C. P., SCHIRM, M., LORENZO, M., KELLY, J., PATTARINI, D., MAJAM, G.,
501 THIBAULT, P. & LOGAN, S. 2006. Changes in flagellin glycosylation affect Campylobacter
502 autoagglutination and virulence. *Mol Microbiol,* 60**,** 299-311.
503 GUNDOGDU, O., BENTLEY, S. D., HOLDEN, M. T., PARKHILL, J., DORRELL, N. & WREN, B. W. 2007. Re-
504 annotation and re-analysis of the Campylobacter jejuni NCTC11168 genome sequence. *BMC*
505 *Genomics,* 8**,** 162.
506 ISLAM, Z., VAN BELKUM, A., WAGENAAR, J. A., CODY, A. J., DE BOER, A. G., SARKER, S. K., JACOBS, B.
507 C., TALUKDER, K. A. & ENDTZ, H. P. 2014. Comparative population structure analysis of
508 Campylobacter jejuni from human and poultry origin in Bangladesh. *Eur J Clin Microbiol*
509 *Infect Dis,* 33**,** 2173-81.
510 JOKINEN, C. C., SCHREIER, H., MAURO, W., TABOADA, E., ISAAC-RENTON, J. L., TOPP, E., EDGE, T.,
511 THOMAS, J. E. & GANNON, V. P. 2010. The occurrence and sources of Campylobacter spp.,
512 Salmonella enterica and Escherichia coli O157:H7 in the Salmon River, British Columbia,
513 Canada. *J Water Health,* 8**,** 374-86.
514 JOLLEY, K. A. & MAIDEN, M. C. 2010. BIGSdb: Scalable analysis of bacterial genome variation at the
515 population level. *BMC Bioinformatics,* 11**,** 595.
516 KATOH, K., MISAWA, K., KUMA, K. & MIYATA, T. 2002. MAFFT: a novel method for rapid multiple
517 sequence alignment based on fast Fourier transform. *Nucleic Acids Res,* 30**,** 3059-66.
518 KENT, W. J., SUGNET, C. W., FUREY, T. S., ROSKIN, K. M., PRINGLE, T. H., ZAHLER, A. M. & HAUSSLER,
519 D. 2002. The human genome browser at UCSC. *Genome Res,* 12**,** 996-1006.
520 KIVISTO, R. I., KOVANEN, S., SKARP-DE HAAN, A., SCHOTT, T., RAHKIO, M., ROSSI, M. & HANNINEN,
521 M. L. 2014. Evolution and comparative genomics of Campylobacter jejuni ST-677 clonal
522 complex. *Genome Biol Evol,* 6**,** 2424-38.
523 KUMAR, S., STECHER, G. & TAMURA, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis
524 Version 7.0 for Bigger Datasets. *Mol Biol Evol,* 33**,** 1870-4.
525 LANE, D. J. 1991. 16S/23S rRNA sequencing. *In:* E., S. & M., G. (eds.) *Nucleic Acid Sequencing*
526 *Techniques in Bacterial Systematics.* New York: Wiley and Sons.
527 LAWSON, D. J., HELLENTHAL, G., MYERS, S. & FALUSH, D. 2012. Inference of population structure
528 using dense haplotype data. *PLoS Genet,* 8**,** e1002453.
529 LUANGTONGKUM, T., JEON, B., HAN, J., PLUMMER, P., LOGUE, C. M. & ZHANG, Q. 2009. Antibiotic
530 resistance in Campylobacter: emergence, transmission and persistence. *Future Microbiol,* 4**,**
531 189-200.

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.

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*.

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., DIDELOT, X., ANSARI, M. A., SHEPPARD, S. K. & FALUSH, D. 2014. Efficient inference of recombination hot regions in bacterial genomes. *Mol Biol Evol,* 31**,** 1593-605.

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.

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). Common clonal complexes (CC) are annotated 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 the number of chunks imported from a donor genome (column) to a recipient genome (row). Colour ranges from little import from the donor strain (yellow) to a large amount of imported DNA from the donor strain (blue). White indicates missing data. 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 summarising the co-ancestry matrix data. The average frequency of inferred recombination between donor to recipient grouped by import from isolates from the same country compared to isolates from different countries. There is significantly more import from donor strains of the same country compared to strains from different countries (*p-value, test*).

Commented [XD16]: To insert

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.

Commented [XD17]: I can only count six

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 algorithmStructure. 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 orderedby 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.

Commented [ELS18]: The key on the figure needs to be changed from yellow to 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). Theratio 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: 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 biogeographical epidemiological markers, including lists of (A) highly recombining genes as determined by per-nucleotide estimation of recombination intensity (recombination hot spots); (B) highly recombining genes as determined by pairwise analysis of nucleotide diversity (more than 2% diversity); and genes used to model biogeographical segregation in structure (orange). Genes with a role in fluoroquinolone resistance are highlighted in yellow.