EMERGING
INFECTIOUS DISEASES®

Population genetic approaches to assigning the source of human pathogens: host associated genetic import in Campylobacter jejuni.

Journal:	*Emerging Infectious Diseases*
Manuscript ID:	EID-06-0620.R1
Manuscript Type:	Research
Date Submitted by the Author:	n/a
Complete List of Authors:	McCarthy, Noel; Oxford University, Department of Zoology Colles, Frances; Oxford University, Department of Zoology Dingle, Kate; Oxford University, Nuffield Department of Clinical Laboratory Sciences Bagnall, Mary; Veterinary Laboratories Agency Manning, Georgina; Nottingham Trent University Maiden, Martin; The Peter Medawar Building, Department of Zoology Falush, Daniel; Oxford University, Department of Statistics
Keywords:	Bacterial typing techniques, Epidemiology, Population genetics, Population dynamics, Zoonoses, Campylobacter jejuni, Public Health, Disease reservoirs, Ecology, Communicable disease control

UNIVERSITY OF OXFORD

Department of Zoology and
The Peter Medawar Building for Pathogen Research
South Parks Road
Oxford OX1 3SY

Phone: +44 (0) 1865-281535
Fax: +44 (0) 1865-281535
E-mail: noel.mccarthy@zoo.ox.ac.uk

"Population genetic approaches to assigning the source of human pathogens: host associated genetic import in Campylobacter jejuni." MS # EID-06-0620

Dear Dr. Drotman,

Many thanks for considering this manuscript for publication in Emerging Infectious Diseases. The reviewers have read the manuscript very carefully and provided many constructive comments. We have focussed in particular on simplifying the text and making it clearer to an audience of microbiologists and epidemiologists. We chose EID specifically to reach this audience and are grateful for help in making the manuscript more appealing to these target groups. The revised manuscript is significantly shorter than our first submission and has an improved narrative flow.

The reviewers and editorial boards comments are considered in turn in the order raised.

Reviewer 1

1. the title is too general and suggests microbiological proof of the import of genetic material. Should be reformulated in a way that it is more covering the contents of the paper.
As the reviewer notes we don't have microbiological proof of the import of genetic material, but do present population genetic proof that it occurs in natural populations of this species. This is central to one aspect of the paper, which reviewer 2 identifies as "a very interesting subtheme that I think should be emphasized more." We are therefore keen to keep this second part of the title, "host associated genetic import in Campylobacter jejuni", as it is. For the first part of the title "Population genetic approaches to assigning the source of human pathogens", this could be made more specific by replacing the "Human pathogens" with either "freely recombining human pathogens" or with just "Campylobacter jejuni". We prefer to keep it more general than C. jejuni since the approach is of more broad relevance and this is worth highlighting. The second part of the title makes explicit that the example used is C. jejuni. If it was preferable in your judgement, we would be happy to add the words "freely recombining"to the title.

2. Abstract: the last sentence claims a wide-use of this approach. However, like the comparison between Salmonella and Campylobacter is totally different (as indicated by the authors), the general use of this approach is too ambitious.
We wanted to emphasize that it has wider application than in *C.jejuni* alone but agree that the description as a general approach may suggest universal applicability. We now specify

that the approach is applicable only "to pathogens that undergo frequent genetic recombination".

3. page 4, line 22: the comparison with Salmonella is made. Please compare the results of this Campylobacter study and the validity of the attribution with the Salmonella studies (compare predicted assigned accuracy) (in the discussion).
We are now quite explicit in the introduction about why methods that work for Salmonella do not work well for Campylobacter. This comparison should make the point of the paper significantly more accessible to bacterial epidemiologists. We do not try to compare results quantitatively in the discussion since this 1) is dependent on the availability of similar reference populations and 2) is not well known even for Salmonella where very good work has been done but accuracy has not been tested in an absolutely empirical fashion, let alone for C. jejuni.

4. page 6-7: the use of the training set should be described more clearly.
We have made several changes. Firstly we introduce the general approach at the end of the introduction. Secondly we have removed the (distracting) general description of what STRUCTURE software can do and only describe the approach that was employed here. Thirdly we have described training sets in a more concrete fashion. (second half of the first paragraph of "Population Assignment" under methods).

5. page 8; line 26: what is the "assignment accuracy" in relation to the 67%?
6. page 8, line 30: this means that 37% of the 67% gap is closed?
We now make it clear that the results presented below this point refer to "the proportion of the gap" between perfect prediction and the value expected by random guessing.

7. page 8, lines 52 and 54: I cannot relate these 58% and 16% to the data in Table 4. Please make clear to the reader!
8. pages 8-9, lines 56 resp 1-2: the, for the paper, essential prediction of 80% correct, does not follow clearly from the description you have given.
Table 4 shows results of analyses that assign individuals to one of three possible host sources (sheep, cow, chicken). These results relate to analyses involving assignment to one of two possible sources (either ruminant versus chicken or sheep versus cow) to further explore the results shown in Table 4 . The text has been reworked to make this distinction absolutely clear.

9. page 11, line 10: ...despite the lack of identifiable host specific markers... You may speculate that there are markers as there is "something chicken" in the poultry strains. Can the step be made from the MLST-data to the specific marker? Or do you expect that it is only the circulation of the genetic pool in the host without any biological relevance? (if your reply is that I missed the message of the paper, it may be because the way it is
written is a bit confusing...)
We have removed the phrase "despite the lack of identifiable host specific markers", which did not make anything clearer. We have also substantially reworked our description of host differences to make clear that there is a strong neutral component to the differences, as shown by the involvement of 6 out of 7 loci in the exchange, but also that selective differences are possible, particularly in parts of the genome that we have not surveyed. We have substantially reworked this section, as described more fully in the reply to the comments of reviewer 2.

10. page 11, line 29-30: analysing more strains may not solve this problem when the strains are rapidly exchanged between the ruminant species!
We agree with the reviewer that analyzing more strains (or indeed more genes per isolate) may not allow differentiation between cattle and sheep derived isolates. This possibility was presented in the preceding sentence. The order of these sentences is now reversed to ensure that the possible lack of differentiation between these two host species stands out more clearly.

11. page 12, line 57: will this estimation of the burden of human illness based upon the MLST data more reliable than the estimation of the BoI based upon (for example) case control studies?
This approach to estimating burden of disease is complementary to other approaches such as case-control studies. How accurate it will be is dependent on how good the reference populations representing different sources become. At present there are large and increasing food animal isolate collections although isolates from a range of environmental sources are limited. Our hope is that 1) the increasingly widespread application of this inherently standardized approach will ensure sampling of the full range of source populations and 2) further refinement of analysis will increase accuracy. In this setting the approach can add a lot to what is known by case-control methods and can indeed contribute to case control and other approaches (e.g. case-case) by allowing identification of meaningful subgroups of isolates which likely have different origins. Making this comparison seems premature in the current manuscript.

12. page 12-13, lines 57, resp 3: when there is exchange of strains what disturbs the host association, will the differences between chicken and ruminants the most obvious one as there may be much more exchange of strains between strains that live in a more open environment (ruminants and wild birds)? Please speculate about that.
Chicken and ruminant isolates are far more similar to each other than to those from other sources on currently available databases. In particular, we are involved with work led by collaborators which shows very little overlap between isolates from wild birds on an open free-range farm and those from either chickens or cattle on that farm. Since this work is unpublished, we prefer to avoid making explicit statements on this topic in the paper.

13. conclusion of the paper: I am not convinced that this a kind of generic approach (or the approach may be generic but the chance that the outcome is usefull strongly depends on the biology of the microbe). At least this restriction should be made!!
We fully accept that the generalisability is to organisms sharing similar genetic properties and now make this explicit in the conclusion.

14. Figure 1: the downloaded version does not have colours.
We have changed the figure to greyscale to avoid inadequate contrast when printed in black and white.

Reviewer 2

Since the method depends upon population statistics, it is not likely to be useful for single outbreaks of a particular strain, but should be useful to find the source of continued low level food borne infections, if most of the infections are caused by practices in a single industry, i.e. chicken farms.

The method can be in some cases used to assign single "unknown" isolates. As examples of this, among the fully sequenced C. jejuni genomes 2 are from known sources, one being isolated from poultry (isolate RM-1221) and the other from a human involved in a cow's milk outbreak (isolate 81-176). Assigning these based on the reference datasets we used in the paper predicted their origin as chicken (99% probability) and cattle/sheep (97% probability) respectively. We have now clarified that this is a potential use in the paper, but also indicating the limitations, and that in some cases uncertainty will persist.

However, it also has a very interesting subtheme that I think should be emphasized more. Particular clonal groups and particular MLSTs do not give good discrinination, but particular alleles do. This suggests that there are particular alleles favored when the bacteria grow in particular species. Thus there are host specific alleles. Since the MLST was done on only seven gene pieces, this proposal is very surprizing. Homologous recombination is providing the signal. What the signal is depends upon the piece size being recombined in.
In the manuscript, there is the suggestion that the piece size is a few hundred bases from a study of piece size for peices coming from other species.
This recombination fragments size estimate comes from a publication that uses a standard population-genetic approach within the *C. jejuni* species. A similar estimate using different methods and study populations, again within species *C. jejuni* (not published in a peer reviewed journal), further supports the relatively short recombination fragment size. This, along with 6 of the 7 MLST genes in our sample having alleles predictive of host, is strong evidence that the signal is not due to adjacent genes in linkage disequilibrium with one or more of the MLST loci.

If this is so, then the MLST genes that are critical for discrimination (The genes giving discrimination were never named or discussed and I would like to see this included in the paper) are the host specific genes. However, I would expect the observed piece size for homologous recombination to possible be much larger for within species homologous transfer, particularly for host specific genes which are advantageous in the current environment. This question should be discussed in more detail.
We have substantially reworked our discussion of these issues to make our findings more explicit and the logical flow clearer. In particular, we now state (1) that the genes chosen for MLST encode for core metabolic function and are unlikely to be selected for host specific import. (2) Neutral processes are on their own capable of generating such patterns. (3) Consistent with a neutral process, 6 of the 7 MLST fragments were involved in the host-specific import into ST21 complex. For these reasons we don't include separate discussion for each gene in the paper.

Also, I do not see how they arrived at the figure of 86 loci that will have host specific properties, particularly whan they do not know piece size.
This estimate is based on extrapolating this apparently approximately homogeneous recombination process to the entire genome. Since we estimate the number of genes involved, rather than the number of events, it does not matter that we do not know the average size of the fragment, although our estimate may be conservative insofar as we have not detected all the events affecting the gene. The calculation was based on the 185 non ST21 members of the ST21 complex excluding the ST300 isolates (with conflicting alleles). These isolates had 67 informative alleles (excluding the 4 giving completely erroneous predictions). This indicates an average of 0.36 informative alleles per isolate among the 7

studied loci, i.e. 5.2% of all genes. This was then extrapolated to the full genome by multiplying by 1654, an estimate of the total number of genes. What should and should not be excluded in this calculation and whether it should be based on ST or isolate is somewhat arbitrary. Different approaches to these exclusions etc give estimates ranging from 69 to 115 genes, with the main message not being a precise number but just that it suggests that there is a lot of host associated import. In this resubmission we have changed this estimate to be based on ST rather than alleles for consistency with Figure 1 results (and the arguments for using the ST approach there) and made it conservative (and easier to explain) by including only alleles predictive of a host where that prediction was correct. We have described the calculation explicitly in the text.

Particular comments:
Abstract: the statement "That this is in part because bacteria import those alleles present in other Campylobacter in the same host by homologous recombination" should be given more explination in the abstract. I assume this statement does not mean that alleles being picked up are selectively neutral because there would be no reason for these to present particularly in a particular host--particularly for a species like C. jejuni that shows high variability and high rates of homologous transfer. If particular clones are preferentially found in particular species of animals as would be required if the transferred loci are selectively neutral, then why why were STypes and clonal complexes such poor predictors of host source?
The abstract has been substantially reworked. We give a more full explanation of the processes that give rise to the host specific recombination that we have identified. We have not included explanation of why ST and clonal complex are poor predictors of host source in the abstract due to limits of space but have described the main problems (discrimination for ST and loss of the information provided by recombination when considering clonal complex) clearly in the discussion. The reason that the signal is there even in selectively neutral loci is the combination of very frequent recombination and a marked dominance of within host species transmission compared to between host species transmission which is now made explicit in the abstract.

The discussion at the end of page ten does not seem to correspond to Fig1B. Type 300 is two alleles away, mixed source, why is it predicted to be blue? But the text suggests all two step groups are correctly assigned? In mixed groups are the numbers from the different hosts different such that your ideas are supported? i.e. most 300 type are from ruminates? How does one treat these large groups like 369 and 642 which are from mixed sources and the allele is blue? This figure is important but needs to be more carefully explained and perhaps redrawn giving relative number in size of box. Maybe all the impressive data is from groups with a small number of isolates and if all groups had a large enough set of isolates, isolates would come from both sources.
The reviewer raises the interesting question of whether the results from Figure 1B would be very different if isolates were the unit of analysis rather than ST. We had started with isolates in the analysis for Fig. 1A but decided that this analysis could produce a false positive result if local expansions of a particular ST in one species contributed a significant proportion of the overall signal. Our results are robust to this possible confounding factor and therefore are preferable. In fact the results were almost identical and highly significant for either approach. We persisted with the ST approach for the Fig 1B on the same basis and to be consistent. Again however, considering isolates rather than ST does not alter the conclusion. Specifically, for the STs where the result was wrong (an allele predicted one species and the isolates came from another) only 1 isolate was involved for each of STs 8,

268, 369 and 615. Where an allele predicted one source and we observed that ST in more than 1 source then 1 isolate was predicted correctly and 1 incorrectly for ST 43 and 31 predicted correctly and 3 incorrectly for ST262. For ST 300 where we had conflicting alleles (1 predicting bovid and one predicting chicken origin) there were 2 bovid isolates and 1 chicken.

In describing these results in the text we have now included type 300, the lack of which seemed to cause confusion. It is the only type for which we had alleles that conflicted in their prediction of source. It was not referred to in the text before which referred to the 4 two types (519, 748, 302 and 722) where there were two alleles predicting the same source – in which case all isolates from these types came from the predicted source as stated. We have also amended the legend of Figure 1B to emphasize prediction first and origin of predicted ST second to communicate the idea of this figure more clearly.

Legend of table 3--"from" not "form"
Page 11, line 55-- remove "of"
Done.

EID Editorial Board Comments to the Authors:
Carefully consider all the review comments. We at EID consider reader-friendliness an important factor in our final acceptance decisions. Our reviewers find that the current version of this paper is difficult to interpret for clinical microbiologists and epidemiologists. This shortcoming in particular needs to be addressed to make the paper useful to the broad EID audience.
In addition to the clarification resulting from the constructive comments of the reviewers we have improved the clarity for a mixed audience by
1. Stripping out non-essential material (such as alternative population genetic approaches not actually applied here and some description of issues in past work that were prominent in the introduction)
2. Dealing with the epidemiological and biological parts of the discussion separately before combining them in the conclusion.
3. Minimising specifically population genetic terminology where alternatives are available.
Parts of the discussion remain which are technical, considering statistical limitations of the method as applied. We think that these are necessary to include and have tried to keep them discrete – but would be happy to make them an appendix if this is felt to improve readability.

Many thanks for your further consideration of this work.

Sincerely,

Noel McCarthy MPH MSc (Medical Statistics) MRCPI MFPH
Wellcome Trust Clinical Training Fellow
Honorary Consultant Epidemiologist, Health Protection Agency

Population genetic approaches to assigning the source of human pathogens: host associated genetic import in *Campylobacter jejuni*.

Noel D. McCarthy*, Frances M.Colles*, Kate E.Dingle†, Mary C. Bagnall‡, Georgina Manning‡, Martin C.J. Maiden*, Daniel Falush§.

*Department of Zoology, Peter Medawar Building, University of Oxford, Oxford, UK.

†Nuffield Department of Clinical Sciences, University of Oxford, Oxford, UK.

‡Veterinary Laboratories Agency, Weybridge, Surrey, UK.

§Department of Statistics, Peter Medawar Building, University of Oxford, Oxford, UK.

Corresponding author:

Noel D. McCarthy
Department of Zoology
University of Oxford
South Parks Road
Oxford
OX1 3SY
Tel and Fax +44-1865-281535
Email noel.mccarthy@zoo.ox.ac.uk

ABSTRACT 175 WORDS

TEXT 3072 WORDS

TABLES 5

FIGURES 1

KEYWORDS

Bacterial typing techniques; Epidemiology; Population genetics; Population dynamics; Zoonoses; Campylobacter jejuni; Public Health; Disease reservoirs; Ecology; Communicable disease control

SUMMARY OF CONCLUSION

Campylobacter jejuni genomes carry a host signature allowing attribution of isolates to animal sources and offering insights into the biology of multi-host microbial species.

Abstract

Many human infections are zoonotic with a broad host range being particularly typical of emerging diseases. Establishing the sources of human infection supports effective disease control measures. Host association of the foodborne zoonosis *Campylobacter jejuni* was evaluated by analysis of multilocus sequence typing data for 713 isolates from poultry and bovids (cattle and sheep). The commonly used summary measures of genotype : sequence type and clonal complex, performed poorly, while a method using the full allelic profile achieved 80% accuracy in distinguishing isolates from these two host groups. We explored the biological basis for the better performance of allelic profiles. We show that strains isolated from particular hosts have imported a substantial number of alleles while circulating in that host species. These results imply that (1) although *Campylobacter* do jump frequently between host, the bulk of transmission is within species and (2) lineages can acquire a host signature and potentially adapt to the host through recombination. Assignment using this signature allows improved prediction of source in pathogens that undergo frequent genetic recombination such as *Campylobacter*.

Many human pathogens inhabit several animal host and environmental reservoirs, with a broad host range being particularly characteristic of emerging diseases (1). Identification of the relative contributions of pathogen sources and transmission routes is necessary to underpin evidence based disease control programmes (2). One approach to address it, microbial source tracking, is the application of microbial typing to isolates from human cases and possible sources in the food chain to allow attribution of disease to food sources at individual case and population levels (3 ,4). Evidence-based control programs based on this information have worked well with Salmonella at a population level in Denmark (4).

Source tracking depends on accurate estimation of the frequency of different subtypes in each host reservoir. In Salmonella, particular serotypes and phage subtypes are stably found in the same host (3). The biology underlying this success is firstly that specific clones are well adapted to specific hosts and secondly that the serophage type is a stable and reliable indicator of a specific clone. For other organisms it can be much harder to find reliable host associated markers. One example, *Campylobacter jejuni*, is the main cause of bacterial gastroenteritis in the western world and the most common bacterial zoonosis. Phenotyping has not worked well. Genetic methods of discrimination show enormous diversity within the species, with studies typically reporting about half as many genotypes as there are strains in the study (5-12). Many common genotypes are broadly distributed, while for rare genotypes it is not possible to estimate the relative frequency of genotypes in different host reservoirs accurately. These difficulties have meant that although host associations have been identified for particular genotypes, no generally useable approach has been developed.

Here, we develop an approach using multilocus sequence typing (MLST) data to identifying the reservoir of origin of a strain. We develop and test the approach using isolates from known sources, namely cattle, sheep and poultry, allowing us to compare our prediction

with the true origin of each strain. The resulting method can provide reasonable accuracy even for rare or unique genotypes and for clones that are broadly distributed. This success is based on exploiting the frequent recombination in Campylobacter, which has served to limit the accuracy of approaches based on the Salmonella paradigm.

Methods

Data

MLST of *C. jejuni* is based on sequencing 7 loci of length 402-507 base pairs separated from each other by at least 15,000 base pairs in the type strain (10). We used MLST data in three different forms. Firstly, the Sequence Type (ST). Each ST is a unique combination of 7 alleles. STs index the full discrimination available within MLST. Secondly, clonal complex, which are groups of closely related STs, e.g. differing at no more than 2 of the 7 alleles. Clonal complexes are thought to represent a group of strains that have a single recent clonal origin, but for which genetic identity has been broken down by mutation and recombination (10 ,13 ,14). Lastly we used the 7 allele fragments assuming that they each provided independent information.

We included all *C. jejuni* isolates from cattle, sheep and chickens on the pubmlst database (www.pubmlst.org) with date-stamp before 1 August 2004 and which had been published in peer reviewed literature or for which permission was obtained from those who had submitted the data. All but 10 of the isolates on pubmlst were available for inclusion by these criteria. Additional typed isolates made available by researchers during the contact process to seek permission to include unpublished isolates from the pubmlst database (n=27) were also included. It has been shown that *C. jejuni* recombines with *C. coli* (15). Those isolates with at least 4/7 alleles typical of *C. jejuni* are included. In total 713 isolates were available by these criteria. Isolates were from animal feces, live animals and dead animal tissue. The distribution of the data by host type and by year and country of isolation is summarised in Tables 1 and 2.

Population assignment

Differences in genotype frequency between populations allow probabilistic assignment of isolates to populations, even if there is some sharing of genotypes between those populations. We used STRUCTURE, a model-based clustering method designed to infer population structure and assign individuals to populations using multilocus genotype data (16). The source of the isolates to be assigned was predicted based on a training set that consisted of other relevant isolates. In order to do this, we used the USEPOPINFO option, which allows the population of origin to be known for some strains (in this case the training set) while for other strains (the isolates to be assigned) it is assumed unknown.

STRUCTURE estimates the genotype frequencies in each host species based on all of the isolates, as well as estimating the population of origin of isolates of unknown origin, taking into account uncertainty due to sample size. To allow maximum use of the data, some analyses employed a leave-one-out strategy in which a single isolate was assigned using the remaining strains as the training dataset, with the procedure being repeated in turn for each isolate.

The parameters we used for all STRUCTURE simulations were a no admixture model with lambda = 1, and gene frequencies uncorrelated between populations. We ran 1000 burn-in cycles and 10000 further repetitions for each analysis. Empirical assignment accuracy was measured as the average probability pk^* with which each isolate was assigned to the correct host source k^*. Predicted assignment accuracy (Discussion) is estimated as the average of $\sum_{k=1..K} p_k^2$, where each individual is assigned to one of K different sources.

The permutation test (Figure 1A) was performed by randomly permuting the actual host species amongst the predictions obtained from STRUCTURE repeated 10,000 times.

Results

There were 330 MLST genotypes among 713 isolates**.** Two isolates (ST-284 and ST-327) had 4 alleles typical of *C. jejuni* and 3 typical of *C. coli*. All others had 5 or more typical *C. jejuni* alleles. Table 3 shows assignment accuracy when using the whole dataset and a leave-one-out strategy to assign strains to three host species; cow, sheep and chicken based on the seven alleles, the clonal complex, the ST and combinations thereof. Since random guessing would be correct one third of the time, comparison of how much improvement genotype information made above random assignment is more informative than the percentage correct, i.e. what proportion of the gap between 33% correct expected using random assignment and 100% correct with perfect prediction has been closed. Assignment using the 7 alleles closed 37% of this gap compared to 10% for sequence type and 13% for clonal complex. Prediction did not improve substantially when ST or clonal complex information was added to allele information. These overall results emphasize the limits to using a sequence type or clonal complex as a summary of multilocus sequencing typing when predicting host of origin. We therefore used alleles in all further analysis as well as exploring the basis for the better performance of this approach.

Prediction of host of origin to three host sources based on alleles is shown in more detail in Table 4. The method performed much better in distinguishing strains from chicken and those from cow or sheep than in distinguishing between strains from the two bovid species. When we performed further analysis restricted to cattle and sheep isolates we achieved an assignment accuracy of 58% between these two species compared to 50% expected by chance, and thus explained only 16% of remaining uncertainty showing little detectable host association between these two closely related host species. Further comparison of chicken isolates with a combined population from cattle and sheep gave improved resolution and

allowed correct prediction 80% of the time (60% of uncertainty removed) indicating substantial host association.

Given the nature of the dataset, we must consider possible confounding factors such as differences in time or location of sampling, which may lead either to completely spurious associations or to overestimates of their magnitude. Indeed, there was evidence for modest time and geographical effects within our dataset. For example, in a comparison of UK chicken isolates from 1997 or earlier and 1998 or later (Table 5) 66% were could be assigned to the population of the correct period based on allelic profile. Similarly, considering UK and Dutch chicken isolates, 69% were assigned to the correct country. We therefore performed additional analyses where host was negatively associated with time and/or space (Table 5). Late UK chicken isolates (1998-2003) were assigned using early UK chicken (1997 or earlier) and late UK bovid isolates (1998-2003) as training sets, giving 77% assignment to chicken. UK chicken isolates were assigned using non-UK chicken and UK bovid isolates as training sets, giving 64% assignment to chicken. These analyses showed that host effect is stronger than that of time or space and that our results are not simply the result of confounding due to these factors.

In order to explore the mechanism underlying the better performance observed for allele based assignment and to better understand the biological processes producing this host signature in the bacterial genome we investigated assignment within ST-21 complex. This clonal complex comprises a substantial proportion of isolates and is highly diverse (5 ,10 ,17 ,18). There were 252 ST-21 complex isolates in our sample. Of these 188 were not ST-21 but differed at between 1 and 3 alleles from the central genotype. We assigned these 188 isolates to chicken or bovid host based only on the alleles at which they differed from ST-21, using all non ST-21 complex isolates as the training set. 66% of isolates were assigned to the correct

host. This analysis suggests that ST-21 complex isolates are picking up alleles that are characteristic of the host population. In order to demonstrate that this deviation from 50% is not a sampling artefact or chance effect, we restricted analysis to the 88 unique ST-host combinations, which largely eliminates the possible effects of clonal expansion within host, and performed a permutation test to assess the possible role of chance. 67% of these combinations were correctly assigned, which was a higher proportion than observed in any of 10,000 iterations in a permutation test (Figure 1A).

The overall accuracy of host assignment based on imported alleles is hampered by the fact that many of these alleles are individually too rare for their frequency in particular host gene pools to be estimated accurately. Imported alleles that are frequently observed give more accurate host prediction. To illustrate this visually (Figure 1B) we use as predictors only those alleles that are both found in at least 10 different ST-host combinations in the non ST-21 complex isolates and are also substantially differentiated between the chicken and bovid populations (based on a 65% cut off). All 4 isolates having 2 alleles both of which are suggestive of either chicken or bovid origin are indeed from the predicted source. In one case two potentially informative alleles gave conflicting information, one suggesting bovid origin and one chicken. Isolates with this ST came from both sources. Of the 24 STs with only 1 informative allele available, 18 are correctly assigned and only 4 incorrectly with 2 STs isolated from both chicken and bovid sources.

Discussion

These analyses confirm the association of *C. jejuni* genotypes with host species, demonstrating a clear distinction between isolates obtained from chickens and those obtained from bovids, when alleles are considered independently in statistical analysis. This finding was robust to sampling differences in time and place, suggesting that host effects were stronger than geographic and temporal effects, an important consideration if these associations are to be employed in epidemiological investigations. Moreover, populations of *C. jejuni* in farm animals such as bovids and chickens may be rather similar compared to those from other hosts (5 ,9) so that the approach may be even more accurate when considering *C. jejuni* from a more diverse host range. The distinction between cow and sheep isolates is much weaker. Differentiation between these species might be demonstrable if substantially greater genetic information were available. However the minor differences observed may be a sampling artifact with these species sharing a common gene pool.

The allele based method that we have used goes a long way towards solving the problem of excess discrimination in *Campylobacter* typing. Many alleles show differences in frequency between hosts. These alleles provide useful information on source for STs that are too rare to allow estimates of their frequency in different hosts, for example because they are entirely absent from training sets.

There are some limitations to the implementation of our approach as presented in this paper, which must be considered in any more extensive application. The current estimate of 80% accuracy in distinguishing chicken isolates from bovid ones may be somewhat over optimistic if sampling effects are important. Sampling effects would include the nature of the sample (feces, meat etc.) as well as time and place. For example, the dominant *Campylobacter* types found on processed carcasses have been shown to differ from those found on the live

poultry entering the processing plant (19). Nonetheless, we have shown that the easily identifiable sampling effects are overwhelmed by the host effect. Moreover, the analysis within ST-21 complex (Figure 1) is robust to both identified and unidentifiable sampling effects and we do not therefore believe that this is a major problem.

A further limitation of our allele-based application of STRUCTURE is that it assumes allelic independence, which is clearly violated for the dataset analyzed. There are two different ways of estimating assignment accuracy. The first, which we have used throughout this paper, is a holdout procedure, whereby source of origin of strains for which the actual origin is known is predicted using the rest of the sample as a training set, providing an unbiased empirical measure of accuracy. For purposes of prediction of isolates where the source is unknown, this procedure is not possible so it would be desirable to use estimates of accuracy that the algorithm itself provides. Because STRUCTURE assumes each allele is independent, its estimate of the accuracy with which it estimates the frequency of a particular multilocus genotype frequency is often overconfident. For example, in differentiating chicken isolates from those originating in cattle and sheep, STRUCTURE predicts 91% accuracy for itself, but empirically it only achieves 80% on average. A better estimate of uncertainty would be necessary for predictive purposes. More sophisticated genetic models that reflect the dependence amongst the loci should achieve more accurate assignment as well as better estimates of statistical uncertainty.

Accepting these limitations this approach nonetheless demonstrates the ability to assign isolates probabilistically to populations. When broad reference populations from the full range of possible sources are available it will be possible to apportion groups of isolates, such as those affecting a human population over a period of time, to their sources, although precision in the attribution of *C. jejuni* may be less than with for example, Salmonella, where host

species appear to harbour more differentiated populations (3). Prediction is possible with individual isolates, in some cases to a single source, although in some cases prediction will suggest a range of populations rather than one. For example, two of the fully sequenced *C. jejuni* genomes are from known sources, one being isolated from poultry (isolate RM-1221) (20) and the other from a human involved in a cow's milk outbreak (isolate 81-176) (17 ,18). Assigning these based on the reference datasets we used in the paper predicted their origin as chicken (99% probability) and cattle/sheep (97% probability) respectively.

The broad host range of *C. jejuni* spanning a range of mammalian, avian and other species make it a good model to study features that may be informative of the ecology of multi-host pathogens. *C. jejuni* imports fragments from other members of the species, which have been estimated to be typically a few hundred base pairs in length (21). Our analysis within ST21 complex demonstrates that isolates in this complex have imported genetic material prevalent in the population of *Campylobacter* carried by their host species (Figure 1). This observation implies firstly that there is persistent differentiation in allele frequencies between different host species and secondly that many of the ST21 isolates represent lineages that have persisted within the same host species long enough to import a substantial number of alleles.

We have surveyed 7 loci and found on average 0.32 host-specific alleles in the 81 STs other than ST21 that were members of ST21 complex, i.e. just under 5% of the alleles in this analysis. The imported genes were approximately evenly distributed between them, involving 6 of the 7 loci. The MLST loci were chosen because they represent core metabolic functions of *C. jejuni* (10) and are not obvious candidates for host adaptation. Therefore, we are probably observing the neutral level of genetic import. Extrapolating linearly from these seven loci to the 1654 gene coding sequences in the *C. jejuni* genome (22) gives an estimate of 76 genes with alleles typical of a particular host species within each ST21 complex isolate. This is

obviously a rough estimate since it is based on fairly limited data and because recombination and selection at other genes may behave differently. However this approximate estimate demonstrates the potential for substantial adaptation to the most recent host by homologous recombination. Indeed, homologous recombination may be an important factor in allowing a single bacterial species to stably colonize a wide range of host species while adapting to some extent to each.

In conclusion, a population genetic approach has allowed host assignment in *C. jejuni* where host specific markers are unavailable but host species populations are differentiated by allele frequency at a range of loci. Host association appears stronger than temporal and geographical effects. Homologous recombination generates a host signature in the *C. jejuni* genome and analyses taking advantage of this signal have improved accuracy of host prediction. The inherent standardization and portability of sequence typing in combination with the availability of such improved assignment techniques support the application of this approach to clarify aspects of *C. jejuni* epidemiology on a global scale and application to the study of other suitable microbes.

Acknowledgements

Angus Buckling, Peter Donnelly, Ken Forbes, Gil McVean and Andrew Sewell provided useful comments on drafts of the paper. NDMcC, MCJM and DF are funded by the Wellcome Trust.

Biographical sketch

Noel McCarthy is a research fellow at Oxford University and honorary consultant epidemiologist at the Health Protection Agency. His research interests centre on the application of scientific methodology to public health problems in infectious disease control.

References

1. Cleaveland S, Laurenson MK, Taylor LH. Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. *Philos Trans R Soc Lond B Biol Sci* 2001;356(1411):991-9.
2. Batz MB, Doyle MP, Morris G, Jr., Painter J, Singh R, Tauxe RV, et al. Attributing illness to food. *Emerg Infect Dis* 2005;11(7):993-9.
3. Hald T, Vose D, Wegener HC, Koupeev T. A Bayesian approach to quantify the contribution of animal-food sources to human salmonellosis. *Risk Anal* 2004;24(1):255-69.
4. Wegener HC, Hald T, Lo Fo Wong D, Madsen M, Korsgaard H, Bager F, et al. Salmonella control programs in Denmark. *Emerg Infect Dis* 2003;9(7):774-80.
5. Dingle KE, Colles FM, Ure R, Wagenaar JA, Duim B, Bolton FJ, et al. Molecular characterization of Campylobacter jejuni clones: a basis for epidemiologic investigation. *Emerg Infect Dis* 2002;8(9):949-55.
6. Rosef O, Kapperud G, Lauwers S, Gondrosen B. Serotyping of Campylobacter jejuni, Campylobacter coli, and Campylobacter laridis from domestic and wild animals. *Appl Environ Microbiol* 1985;49(6):1507-10.
7. Schouls LM, Reulen S, Duim B, Wagenaar JA, Willems RJ, Dingle KE, et al. Comparative genotyping of Campylobacter jejuni by amplified fragment length polymorphism, multilocus sequence typing, and short repeat sequencing: strain diversity, host range, and recombination. *J Clin Microbiol* 2003;41(1):15-26.
8. Siemer BL, Harrington CS, Nielsen EM, Borck B, Nielsen NL, Engberg J, et al. Genetic relatedness among Campylobacter jejuni serotyped isolates of diverse origin as determined by numerical analysis of amplified fragment length polymorphism (AFLP) profiles. *J Appl Microbiol* 2004;96(4):795-802.
9. Manning G, Dowson CG, Bagnall MC, Ahmed IH, West M, Newell DG. Multilocus sequence typing for comparison of veterinary and human isolates of Campylobacter jejuni. *Appl Environ Microbiol* 2003;69(11):6370-9.
10. Dingle KE, Colles FM, Wareing DR, Ure R, Fox AJ, Bolton FE, et al. Multilocus sequence typing system for Campylobacter jejuni. *J Clin Microbiol* 2001;39(1):14-23.
11. French N, Barrigas M, Brown P, Ribiero P, Williams N, Leatherbarrow H, et al. Spatial epidemiology and natural population structure of Campylobacter jejuni colonizing a farmland ecosystem. *Environ Microbiol* 2005;7(8):1116-26.
12. Hopkins KL, Desai M, Frost JA, Stanley J, Logan JM. Fluorescent amplified fragment length polymorphism genotyping of Campylobacter jejuni and Campylobacter coli strains and its relationship with host specificity, serotyping, and phage typing. *J Clin Microbiol* 2004;42(1):229-35.
13. Smith JM, Smith NH, O'Rourke M, Spratt BG. How clonal are bacteria? *Proc Natl Acad Sci U S A* 1993;90(10):4384-8.
14. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* 1998;95(6):3140-5.
15. Dingle KE, Colles FM, Falush D, Maiden MC. Sequence Typing and Comparison of Population Biology of Campylobacter coli and Campylobacter jejuni. *J Clin Microbiol* 2005;43(1):340-7.

16. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000;155(2):945-59.
17. Korlath JA, Osterholm MT, Judy LA, Forfang JC, Robinson RA. A point-source outbreak of campylobacteriosis associated with consumption of raw milk. *J Infect Dis* 1985;152(3):592-6.
18. Hofreuter D, Tsai J, Watson RO, Novik V, Altman B, Benitez M, et al. Unique features of a highly pathogenic Campylobacter jejuni strain. *Infect Immun* 2006;74(8):4694-707.
19. Slader J, Domingue G, Jorgensen F, McAlpine K, Owen RJ, Bolton FJ, et al. Impact of transport crate reuse and of catching and processing on Campylobacter and Salmonella contamination of broiler chickens. *Appl Environ Microbiol* 2002;68(2):713-9.
20. Fouts DE, Mongodin EF, Mandrell RE, Miller WG, Rasko DA, Ravel J, et al. Major structural differences and novel potential virulence mechanisms from the genomes of multiple campylobacter species. *PLoS Biol* 2005;3(1):e15.
21. Fearnhead P, Smith NG, Barrigas M, Fox A, French N. Analysis of Recombination in Campylobacter jejuni from MLST Population Data. *J Mol Evol* 2005;61(3):333-40.
22. Parkhill J, Wren BW, Mungall K, Ketley JM, Churcher C, Basham D, et al. The genome sequence of the food-borne pathogen Campylobacter jejuni reveals hypervariable sequences. *Nature* 2000;403(6770):665-8.

Table 1. Isolates by year of isolation and host species.

Year	Chicken	Cattle	Sheep	Total
1981	0	4	0	4
1982	2	1	2	5
1983	0	3	0	3
1984	0	1	0	1
1986	0	2	0	2
1988	2	18	0	20
1989	0	1	0	1
1990	54	1	0	55
1991	30	6	0	36
1992	1	3	0	4
1993	8	6	1	15
1994	6	1	0	7
1995	12	1	0	13
1996	35	0	0	35
1997	2	0	0	2
1998	40	41	68	149
1999	10	38	38	86
2000	15	6	0	21
2001	45	83	5	133
2002	0	0	2	2
2003	13	0	0	13
Unspecified	34	29	43	106
Total	309	245	159	713

Table 2. Isolates by country and host species.

Country	Chicken	Cattle	Sheep	Total
Canada	0	5	0	5
Czech Republic	8	0	0	8
Denmark	6	1	0	7
Netherlands	53	4	0	57
New Zealand	5	1	0	6
Northern Ireland	1	2	0	3
UK	217	218	158	593
USA	17	13	1	31
Unknown	2	1	0	3
Total	309	245	159	713

Table 3. Capacity of alleles, overall sequence type and clonal complex information to predict host species for *C. jejuni* isolates from cattle, sheep and chickens.

Genotype information used	Percent correct	Percent of uncertainty removed*
Alleles	58	37
ST	40	10
Clonal complex (1) †	42	13
Clonal complex (2) †	42	13
Alleles plus ST	60	40
Alleles plus clonal complex†	58	37

* Random selection would be expected to predict correctly 33% of the time. The proportion of the remaining uncertainty (67%) that is resolved is given here.

† Clonal complex (1) substituted ST for clonal complex where no clonal complex is assigned and Clonal complex (2) substituted a missing value code. Clonal complex (1) was used in addition to alleles to assess "Alleles plus clonal complex".

Table 4. Predicted host compared to actual host among *C.jejuni* from cattle, sheep and chickens.

	Sample	Predicted host		
Host	size (n)	**Chicken**	**Cow**	**Sheep**
Chicken	309	**66%**	14%	19%
Cow	245	12%	**50%**	38%
Sheep	159	10%	36%	**54%**

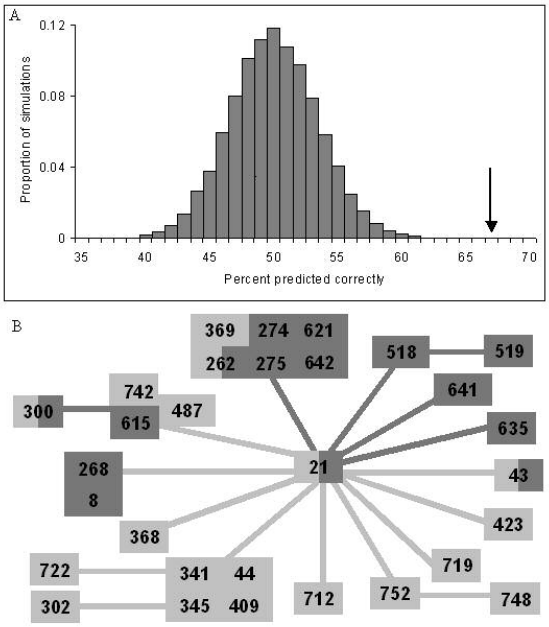
Table 5. Subpopulations for comparisons considering time and geography.

Description of source	Number
Early* UK chickens	114
Late† UK chickens	78
All UK chickens	217
Dutch chickens	53
Non UK chickens	92
Late† UK cattle and sheep	273

*1990 – 1997 = early

†1998 – 2003 = late

Figure 1. Prediction of source of origin within ST-21 complex.
A. Observed accuracy of prediction (arrow) compared with distribution of values obtained by permuting host labels so that the alleles varying from central genotype are not informative on host of origin.
B. Prediction of origin using only alleles for which substantial reference information is available. Light grey lines indicate presence of an allele different from ST-21 present mainly in chickens in the reference population (i.e. an allele that would predict chicken origin) and dark grey those present mainly in bovids (i.e. predicts bovid origin). Light grey boxes are STs found only in chickens and dark grey only in bovids. Mixed boxes indicate STs found in bovids and chicken.


A
0.12
0.08
0.04
0
Proportion of simulations
35
40
45
50
55
60
65
70
Percent predicted correctly
B
369
274
621
262
275
642
518
519
641
635
742
300
615
487
268
8
21
43
423
368
722
341
44
302
345
409
712
752
719
748

EMERGING
INFECTIOUS DISEASES
A Peer-Reviewed Journal Tracking and Analyzing Disease Trends

Centers for Disease Control and Prevention
1600 Clifton Road, N.E.
Mail stop D61
Atlanta, GA 30333
Phone: 404-371-5329
Fax: 404-371-5449

Checklist for Authors

First Author and Manuscript Title:

☒	This manuscript (or one with substantially similar content) has not been published and is not being considered for publication elsewhere.
☒	Corresponding author is the primary contact for proofing the manuscript and galleys.
☒	Financial support for this research is clearly disclosed in the manuscript.
☒	Any organization with a financial interest in the subject matter is disclosed in the manuscript.
☒	Authors have disclosed any conflict of interest related to this article.
☒	Research has been approved by appropriate human or animal subjects research review boards, which are named in the text of the manuscript. (NONE NEEDED)
☒	DNA and amino acid sequences have been submitted to a sequence database and accession numbers are used to refer to the sequences. AVAILABLE ON PUBMLST
☒	All persons who have made substantial contributions to this work but did not fulfill the authorship criteria are named in the Acknowledgments.
☐	Written permission has been obtained from all persons listed in the acknowledgments. VERBAL – WRITTEN COMING
☒	Written permission has been obtained from all persons listed as authors on this manuscript.
☒	Written permission has been obtained from the publishers of any figures or tables previously published or adapted from published figures or tables.
☒	Written permission has been obtained from persons identifiable in photographs, case descriptions, or pedigrees.
☒	Written permission has been obtained from persons named in personal communications (oral or written) stating that they agree to be named and that the information cited is accurate.
☒	All pages are double-spaced, numbered, and left justified (ragged right margin).
☒	All references are cited in the text, follow Uniform Requirements (http://www.icmje.org/index.html), and have been checked for accuracy and completeness.
☒	Legends for figures are at the end of the text.
☒	Each figure is in a separate file.
☐	The abstract meets the word count requirement for the type of manuscript (50 words for dispatches, 150 words for all others). 175 words
☒	All units of measure are expressed in SI units per Instructions to Authors.
☒	A short (2-3 sentence) biography is provided for the first author or both if two authors.

Additional notes or statements: