Quantifying bacterial evolution in the wild: a birthday problem for *Campylobacter* lineages

Jessica K. Calland[1], Ben Pascoe[1], Sion C. Bayliss[1], Evangelos Mourkas[1], Elvire Berthenet[2], Harry A. Thorpe[1], Matthew D. Hitchings[3], Edward J. Feil[1], Martin J. Blaser[4], Daniel Falush[5]# & Samuel K. Sheppard[1]#

[1]The Milner Centre for Evolution, University of Bath, Claverton down, Bath, UK; [2]French National Reference Center for Campylobacters and Helicobacters, University of Bordeaux, 33076, Bordeaux, France; [3]Institute of Life Sciences, Swansea University Medical School, Swansea University, Singleton Park, Swansea, UK; [4]Center for Advanced Biotechnology and Medicine, Rutgers University, New Brunswick, New Jersey, USA; [6]Centre for Microbes, Development and Health, Institute Pasteur Shanghai, China.

#Authors to which correspondence should be addressed: s.k.sheppard@bath.ac.uk & daniel.falush@ips.ac.cn

Short title: *Campylobacter* molecular clock

Keywords: Evolution / Pathogenesis / Molecular clock / Population structure / Population genomics.

Abstract

Measuring molecular evolution in bacteria typically requires estimation of the rate at which mutations accumulate in strains sampled at different times that share a common ancestor. This approach has been useful for dating ecological and evolutionary events that coincide with the emergence of important lineages, such as outbreak strains and obligate human pathogens. However, in multi-host (niche) transmission scenarios, where the pathogen is essentially an opportunistic environmental organism, sampling is often sporadic and rarely reflects the overall population, particularly when concentrated on clinical isolates. This means that approaches that assume recent common ancestry are not applicable. Here we present a new approach to estimate the molecular clock rate in *Campylobacter* that draws on the popular probability conundrum known as the 'birthday problem'. Using large genomic datasets and comparative genomic approaches, we identify isolate pairs where common ancestry is inferred within the sample time-frame – analogous to a shared birthday. Identifying synonymous and non-synonymous substitions, both within and outside of recombinant regions of the genome, we quantify clock-like diversification to estimate mutation rates for the common pathogenic species *Campylobacter coli* (2.4×10^{-6} s/s/y) and *Campylobacter jejuni* (3.4×10^{-6} s/s/y). Finally, using estimated mutation rates we assess the rate of turnover of lineages in our sample set over short evolutionary timescales. This provides a generalizable approach to calibrating mutation rates in populations of environmental bacteria and shows that multiple lineages are maintained, implying that large-scale clonal sweeps may take hundreds of years or more in these species.

Introduction

Theoretical models of a relatively constant rate of molecular change over time (Kimura 1968), the molecular clock, have become fundamental to explaining the evolution in bacteria (Kuo and Ochman 2009; Didelot et al. 2016). Spurred by the increasing availability of population-scale genome datasets, it is now common for comparative genomic studies to describe not only the relatedness of isolates but also how long ago they diverged (Kidgell et al. 2002; Mutreja et al. 2011; Mcadam et al. 2012; Cui et al. 2013; Mourkas et al. 2020). This can provide valuable information when combined with host, habitat or ecosystem data. For example, it is possible to investigate how events such as host transitions or global dissemination have influenced the emergence and spread of lineages that may display important phenotypes, including pathogenicity.

There are significant challenges when applying molecular clocks to date lineage diversification in natural bacterial populations. In particular, it is necessary to determine the rate at which the clock 'ticks' and the uniform accumulation of nucleotide substitutions over time. However, this is not simply a reflection of the background point mutation rate (associated with replication error) and the generation time of the bacterium (Weller and Wu 2015; Gibson et al. 2018), but is also influenced by horizontal gene transfer (HGT) that can introduce several mutations in a single event (Vos and Didelot 2009). Furthermore, the rate at which mutations accumulate in the population is influenced by the population size (Bromham 2009) and selection (positive and stabilizing) on different fitness effects (Eyre-Walker and Keightley 2007).

While debate continues about mutations that are effectively neutral, and hence provide accurate clock estimates (Gibson and Eyre-Walker 2019), there is clear utility for even approximations of the rate of genome change over time (Drummond et al. 2003; Biek et al. 2015). This has allowed the development of time-calibrated phylogenies explaining molecular evolution in numerous well-known pathogen species (Kidgell et al. 2002; Mutreja et al. 2011; Mcadam et al. 2012; Cui et al. 2013). However, even with large genome datasets and increasingly sophisticated models (Drummond and Rambaut 2007; Suchard et al. 2018), the accuracy of molecular evolution estimates is dependent upon the data from which they are derived, and two important considerations remain. First, the data should represent a longitudinal sample set

(Drummond et al. 2003; Arnold and Hanage 2017). Second, the data should be representative of the population as a whole.

It is conceptually simple to understand how a long time-frame between collection of the earliest and latest sample would increase the number of mutations recorded, and how sampling at consistent intervals could help to determine if accumulation was linear over time. Comparisons between modern samples and DNA from the stomach of a 5,300 year old frozen iceman 'Otzi' have been used to investigate the emergence of modern *Helicobacter pylori* lineages (Maixner et al. 2016). However, ancient pathogen samples are rarely available. More frequently, molecular clock rates are estimated using collections of contemporary isolates that often share a common ancestor older than the sample frame. Convincing estimations have been possible for medically important bacteria, through comparison of large numbers of closely related isolates (Didelot et al. 2012; Walker et al. 2013; Mathers et al. 2015; Menardo et al. 2019) but for many pathogens sampling of outbreaks may not provide an adequate representation of the bacterial population.

Most disease-causing bacteria are not obligate human pathogens. In this case, large reservoirs of isolates from which infection can arise may be infrequently sampled, despite their potential importance as emergent pathogenic strains. For example, *Campylobacter jejuni* and *C. coli* are among the most common causes of bacterial gastroenteritis worldwide but exist principally as commensal organisms in the gut of mammals and birds (Waldenström et al. 2002; Sheppard et al. 2011; Bronowski et al. 2014; Cody et al. 2015; Sheppard and Maiden 2015). Human infection results primarily via food contaminated with strains from wild and agricultural animals, especially chickens (Wilson et al. 2008; Sheppard et al. 2009; Strachan et al. 2009; Dearlove et al. 2016; Rosner et al. 2017; Thépault et al. 2017). In multi-host (niche) transmission scenarios such as this, where the pathogen is essentially an environmental organism, sampling is often sporadic and rarely reflects the overall population, particularly when concentrated on clinical isolates (Marin and Hedges 2018).

Overcoming the problem of sporadic or unrepresentative sampling for molecular clock estimation requires that sufficient numbers of isolates are collected to ensure that there are pairs that share a recent common ancestor (within the sampling period). However, with the enormous

effective population size of environmental bacteria populations, questions remain about how many isolates need to be sampled to achieve this. This is analogous to the well-known probability theory conundrum known as the birthday problem (Mathis 1991). This puzzle asks how many randomly chosen people need to be sampled so that a pair of them will share the same birthday. To be sure, requires a sample size of 366 (the number of possible birthdays), assuming that all birthdays are equally common, but a 99.9% probability is achieved with just 70 people and 50% with 23 people. This may seem counter intuitive but can be explained by considering that rather than comparing the birthday of a single individual to everyone else's, in fact comparisons are made between every pair of individuals, 23 x 22/2 = 253. The result is greater than half the number days in the year, hence the 50% probability. Clearly, there are challenges in relating this conceptual model to bacteria. First, it is not known how many possible lineages (here equivalent to birthdays) there are in natural bacterial populations. Second, how to define lineages or isolate pairs with recent common ancestry. Third, just as with birthdays, some lineages are far more common than others. For example, of >72,000 *C. jejuni* and *C. coli* isolates archived in the pubMLST database (Jolley et al. 2018), >50% belong to just 5 clonal complexes (out of 45).

Together, factors relating to isolate sampling and genome analysis conspire such that it may be difficult to distinguish nucleotide substitutions that reflect the passage of time (Didelot and Falush 2007; Biek et al. 2015). Here, we take a multi-layered approach to estimate the rate of molecular evolution of *C. jejuni* and *C. coli* using a large genome collection (2,425 genomes) representing isolates sampled over a 46-year period. We begin by identifying isolate pairs where: (i) common ancestry is inferred within the sample time-frame, and (ii) the most recently sampled isolate has accumulated SNPs over time. We then quantify synonymous and non-synonymous polymorphisms to take (some) account of selection, both within and outside of recombinant regions of the genome, and use synonymous polymorphisms to quantify clock-like diversification in *Campylobacter* (Kimura 1987; Gojobori et al. 1990). Finally, using estimated mutation rates we assess the rate of turnover of lineages in our sample sets over short evolutionary timescales. This provides a generalizable approach to calibrating mutation rates in populations of environmental bacteria and clues about lineage diversification in two important pathogenic bacteria.

Results

There is a weak temporal signal in C. coli and C. jejuni phylogenies

Core genome phylogenies revealed little evidence of clustering by collection date (**Figure 1**). Isolates belonging to common sequence types (STs) and clonal complexes were sampled over the 46-year period. These included poultry associated ST-353, ST-354 and ST-257 complexes, cattle associated ST-61 and ST-42 complexes, and host generalist ST-21, ST-45, ST-828 (*C. coli*) complexes (Sheppard et al. 2014) (**Figure 1 and Supplementary Table 1**). Linear regression of root-to-tip distances and sampling dates of *C. coli* and *C. jejuni* phylogenies **(Supplementary Figure 1)**, using TempEst software, provided very weak evidence of a temporal signal when the best-fitting root was estimated. The R^2 values were low for both *C. coli* (R^2 = 0.176) and *C. jejuni* (R^2 = 9.5 x 10^{-2}) phylogenies (**Supplementary Table 2**). Consistent with some other studies (Rieux and Balloux 2016), this poor branch-length to isolation date correlation suggests that estimation of the molecular clock rate from the entire dataset may be difficult. However, the accumulation of polymorphisms exhibited a positive correlation with sampling date in all datasets (**Supplementary Figure 1**) implying the maintenance of multiple STs and clonal complexes through time.

Sampling matched isolate pairs allows estimation of mutation rate

Estimation of molecular clock rates requires comparison of isolates from related, or preferably the same lineages, that have accumulated mutations over time. To achieve this there is a necessary balance between maximizing the time between sampling and accumulated SNPs whilst ensuring comparisons are made between related strains. Therefore, we plotted SNP difference against time difference to determine criteria for choosing comparable isolate pairs (**Figure 2**). The sample time difference was chosen to maximize the time between sampling and the number of comparable pairs belonging to the same lineage. Pair selection criteria were standardised for both species so that isolate pairs were excluded where the sampling time difference was <8 years or there were >5000 SNPs between them (**Figure 2, Supplementary Table 3, Table 1**). Based upon these criteria, 18 *C. coli* and 20 *C. jejuni* isolate pairs were considered for mutation rate calibration. These belonged to the ST-21, ST-22, ST-45, ST-1332, ST-828 clonal complexes and isolate pairs had a difference in sampling date of 8 to 11 years (*C. coli*) and 8 to 36 years (*C. jejuni*) (**Tables 1 and 2, Supplementary Table 3**).

Estimation of a molecular clock rate requires that SNPs accumulate over time, defined here as mutations per site per year (s/s/y). It is also possible that branch shortening can occur where there are fewer mutations in the more recent isolate of a pair resulting in a negative rate of molecular evolution (Duchêne et al. 2016). In this study, 13 out 18 *C. coli* and 11 out of 20 *C. jejuni* isolate pairs exhibited branch lengthening, i.e. more total mutations (within and outside recombined regions) were found in the more recent isolate (**Tables 3 and 4, Supplementary Table 4**). Only pairs having undergone measurable evolution (branch lengthening) were included in further analysis of the accumulation of mutations over time. For these isolate pairs, the total mutation rate was calculated as well as the mutation rate within and outside of recombined regions (**Table 5, Supplementary Table 5**). The mean mutation rate for non-recombined regions was 6.36 x 10^{-6} and 8.45 x 10^{-6} s/s/y for *C. coli* and *C. jejuni* respectively, or 11.46 and 13.53 mutations per genome per year (s/g/y) (**Table 5**).

Recombination drives molecular evolution in Campylobacter

Mutations in coding sequence based on gene definitions in reference isolate genomes (CVM N29710 (*C. coli*) and NCTC 11168 (*C. jejuni*)) introduced an average of 1569 and 242 SNPs in *C. coli* and *C. jejuni* paired genome datasets respectively. Of these, an average of only 222 (*C. coli*) and 106 (*C. jejuni*) were inferred to be the result of point mutation, with the remainder resulting from recombination (**Tables 3 and 4**). Recombination is therefore the major source of sequence variation in both species (**Figure 3, Tables 3 and 4**), introducing nearly six times as many polymorphisms in *C. coli* than in *C. jejuni* – consistent with previous estimates based upon MLST (Wilson et al. 2009). To assess the effect of mutations on amino acid sequences we quantified non-synonymous (N) and synonymous (S) mutations and determined the ratio per site (*dN/dS*) for all isolate pairs in recombined and non-recombined sequence (**Tables 3 and 4**). Point mutation accounted for an unequal amount of N and S polymorphism both within and between species (*C. coli,* N = 99, S = 123; *C. jejuni,* N = 63, S = 43). While recombination introduced many more polymorphisms than point mutation, in both species these were biased towards synonymous changes. Specifically, around six times as many S than N mutations were introduced by recombination in *C. coli* and approximately twice more in *C. jejuni* (*C. coli,* N = 546, S = 801; *C. jejuni,* N = 59, S = 77). Overall, average *dN/dS* ratios were consistent between species within recombined (*C. coli* 0.492, *C. jejuni* 0.490) and non-recombined (*C. coli* 0.594, *C. jejuni* 0.509) portions of the genome. However, because of the relative importance of

recombination (*r/m* = 37.240 (*C. coli*), *r/m* = 5.098 (*C. jejuni*)), on average N mutations were similar for *C. jejuni* from recombination and point mutation (59 and 63 respectively). However, recombination introduced 5.5 times more N mutations than point mutation in *C. coli* (**Tables 3 and 4**). Variation in *dN/dS* was observed between isolate pairs but was mostly indicative purifying selection (*dN/dS*<1). Evidence of positive selection (*dN/dS*>1) was only observed within recombined sequence in 6 isolate pairs (**Tables 3 and 4**).

Additional analysis of the distribution of recombination events revealed that an average of 13% (*C. coli*) and 2% (*C. jejuni*) of the genome has undergone recombination in at least one isolate pair since divergence from the common ancestor of each sub-tree (**Tables 6 and 7**). Recombination was distributed across the genome in both species but was elevated in certain regions of *C. coli* introducing more polymorphism at potential recombination hotspots (Yahara et al. 2014). However, recombination remained the main source of variation in both species (**Figure 3**).

Molecular clock estimates for C. coli and C. jejuni

Molecular clock estimates require that mutations accumulate at a consistent rate over time. We maximized the chance of identifying this signal in several ways. First, genomic variation within recombined regions was discounted as multiple SNPs can be introduced in a single evolutionary event – distorting clock estimates (Didelot and Falush 2007; Wilson et al. 2009; Croucher et al. 2011). Second, non-synonymous mutations were discounted as selection may be more likely to influence the frequency of variation at these sites. Third, only pairs in which the most recently sampled isolate contained more SNPs (branch lengthening) were used as they displayed measurable evolution over time. Based on these criteria, a similar average molecular clock rate was obtained for *C. coli,* 2.4 x 10^{-6} s/s/y (4.27 s/g/y), and *C. jejuni,* 3.4 x 10^{-6} s/s/y (5.42 s/g/y) (**Table 5**).

Coalescence and maintenance of lineages over time

Molecular clock estimates can vary within a population. Therefore the applicability of generalized clocks depend upon how much of the population has been sampled. To quantify this we estimated the average mutation rate (**μ**) (*C. coli* = 77.292 s/g/y, *C. jejuni* = 14.101 s/g/y), including all polymorphisms within and outside recombined sequence. These mutation rates

were used to determine the number of coalescences in the population at a given time point (here referred to as '*effective lineages'*) within the dataset. The maximum timeframe for comparison was 37 years for *C. coli* and 46 years for *C. jejuni* (short in evolutionary terms). This provided information about the number of ancestral strains and the rate of turnover of lineages within the dataset. The total number of potential pairs without accounting for genetic similarity (***Y***), was equal to the square of the total number of isolates (n^2) divided by two (to avoid double counting of isolate pairs), 180,600 and 1,663,488 for *C. coli* and *C. jejuni* respectively.

Having determined the mutation rate, we were able to predict the expected number of mutations over a given period of time. For example, 14 in 1 year for *C. jejuni*. We then subsampled all isolate pairs (***Y***) to determine how many isolate pairs had $\leq$14 SNPs between them – 76 isolate pairs. This is the possible number of isolate pairs that have arisen in 1 year. This process was repeated for each time cut-off, up to a maximum of 37 and 46 years for *C. coli* and *C. jejuni* respectively (**Tables 8 and 9**), to give the number of possible pairs for every time cut-off (***X***) (**Figure 2B and 2D**). Dividing ***Y/X*** resulted in the number of coalescences (*effective lineages*) at a given time interval in the past (***Z***) (**Tables 8 and 9**). For example, if the total mutation rate was 14 s/g/y and we were interested in the number of birthdays within 5 years of our dataset, we would multiply the mutation rate by 5 to result in 70 SNPs of evolution over 5 years. The number of *potential pairs* (***Y*** = 1,663,488) / *possible pairs* (***X*** = 174) = ~9,560 coalescences (ancestors) within this time period (**Supplementary Figure 2**, **Tables 8 and 9**).

The number of effective lineages at a given time-point can also be interpreted as the number of lineages that gave rise to those that are seen today. This provides valuable information about how the population is maintained over time and the extent to which it has diversified. For example, 1,263 *C. coli* lineages 37 years ago gave rise to an estimated 22,575 one year ago and 4,726 *C. jejuni* lineages 46 years ago gave rise to 21,888 lineages one year ago. This equates to an average increase in the number of effective lineages of 576 and 373 per year for *C. coli* and *C. jejuni* respectively. For *C. jejuni* it is clear that a considerable proportion (22%) of all lineages have been maintained throughout the 46 year sampling period and probably much longer (**Figure 4**). In contrast, only 6% of all effective lineages were present in the *C. coli* population 37 years ago. Perhaps the most striking finding is that the *C. coli* population has

rapidly diversified in recent years. For example, there has been an 800% increase in the number of effective lineages in the last 10 years, over 3 times the rate of increase observed in *C. jejuni.* (**Figure 4, Supplementary Figure 2**).

Discussion

The increasing availability of large genome datasets has great potential for improving molecular clock estimates in bacteria. However, significant challenges remain. While it is clear that the frequency of substitutions can vary between different species and strains (von Mering et al. 2007; Mcadam et al. 2012; Cui et al. 2013; Li et al. 2015; Duchêne et al. 2016; Gibson et al. 2018; Menardo et al. 2019), the extent to which nucleotide variation represents an intrinsic molecular clock is often less apparent. Biological factors such as generation time, population size and recombination rate, and ecological factors including cellular responses to habitat variation or stress and the strength of natural selection, influence the rate at which substitutions accumulate in populations (Denamur and Matic 2006). Therefore, obtaining a robust molecular clock estimate from natural bacterial populations requires an appropriate sample frame and careful consideration of the nature of observed sequence variation.

In cases where there is a clear temporal signal among isolates, it may be possible to obtain a robust molecular clock estimate by applying models to large genome datasets (Menardo et al. 2019). However, analysing all *C. jejuni* and *C. coli* genomes together gave a weak temporal signal. This is likely related to the population structure and biology of these organisms that is in stark contrast to many obligate human pathogens (Menardo et al. 2019). Consistent with many other zoonotic or environmental bacteria, *Campylobacter* is a diverse genus with multiple lineages (STs and clonal complexes) inhabiting multiple hosts/niches. This required a more targeted approach to microevolutionary analysis consistent with that used to investigate transmission in similarly variable organisms (Didelot et al. 2012).

Sub-sampling within the isolate collection, sampled over 46 years, identified closely related pairs of isolates with divergent sampling dates. Clearly, calibration of the molecular clock requires that mutations accumulate over time. This was not the case in all isolate pairs. In some cases, the most recently sampled isolate had accumulated fewer substitutions than the comparator strain leading to a negative mutation rate as observed in some other bacterial species

(Duchêne et al. 2016; Menardo et al. 2019). This could indicate time-dependency of molecular evolution (Ho et al. 2007; Ho et al. 2011), where deleterious mutations in the older isolate have been purged leading to differences in long and short term molecular clock estimates (Rocha et al. 2006; Duchêne et al. 2014). However, in organisms with complex ecology such as *Campylobacter*, it is also possible that closely related isolates occupy different sub-niches and experience different selection pressures even when sampled from the same host.

Returning to the birthday problem analogy, considering the number of isolate pairs (equivalent to people with the same birthday) obtained from the original genome dataset can provide clues about the extent of lineage diversity in the natural population. Using total mutation rates, we were able to assess the nature of coalescence across the sample time frame for each species. The coalescence we refer to here is equivalent to the number of ancestral strains at a particular time point (effective lineages) in the natural environment from which contemporary strains emerged. Effective population size (Ne) is commonly used to reflect the number of individuals in a population that contribute to subsequent generations (Kirchberger et al. 2020). This has been used to investigate bacteria but contrasting approaches can provide different estimates depending on the method used (Cui et al. 2015; von Mering et al 2007). The idea of effective lineages, described in this study, is related to Ne but is more specific for clonal organisms. Rather than typical Ne estimates for sexual populations, where the mating of two individuals is largely independent of what happened in previous generations, the number of effective lineages in a bacterial population reflects the number of distinct lineages that will survive and therefore contribute to future generations. This provides information on the genetic inertia of the population.

These analyses highlighted the importance of appropriate sampling when calibrating mutation rates and can help in determining the extent to which samples represent the population as a whole. Specifically, by considering the number of coalescences in a random population, we can look back through the sample time frame to estimate the number of effective lineages across a randomly sampled dataset. For example, suppose we would like to know if our contemporary isolates have a common ancestor in 1980. We know that a proportion of these ancestors gave rise to the diversity we see today but many lineages would go extinct and therefore not contribute (Louca et al. 2018). Based on an average mutation rate of 14 s/g/y for *C. jejuni*,

there would be 560 SNPs over 40 years total evolution between a strain pair. So, one can then ask how many pairs are close enough genetically for that to be the case. This gives an estimate of the effective number of ancestors in 1980 that gave rise to the contemporary dataset - equivalent to the number of birthdays.

For *Campylobacter*, it is clear that multiple lineages have persisted over a long period of time. This indicates that although the population is large, the strains are not turning over at a particularly fast rate and are maintained over time. The absence of lineage replacement is inconsistent with some models of bacterial evolution that predict periodic population bottlenecks (Koeppel et al. 2007) but this can be explained in several ways. First, it is possible that the 37/46 year sampling period in this study is not sufficient time to out-compete a rival strain. Second, bacteria occupy different niches that are sustained so strains are not in direct competition. Third, the fitness differences among strains are not great enough for one lineage to out-compete another.

As well as the maintenance of multiple lineages, there is also evidence for variation in the number of effective lineages that contributed to successive generations between the two major pathogenic *Campylobacter* species. While this was consistently higher for *C. jejuni* throughout much of the sample frame there was a rapid increase in the number of *C. coli* lineages that began around 8 years ago (**Figure 4**). The reason for this is unclear. The average mutation rate estimates were similar for *C. jejuni* and *C. coli*, 3.4 x 10^{-6} and 2.4 x 10^{-6} s/s/y respectively, equating to approximately 5.4 (*C. jejuni*) and 4.3 (*C. coli*) mutations per genome per year. This is somewhat lower that previous estimates for *C. jejuni* calculated from 7-locus MLST (2.79 x 10^{-5} s/s/y) (Wilson et al. 2009) but is within the range of molecular clock estimates calculated from genomic variation for *Enterococcus faecium* (9.35 x 10^{-6} s/s/y) *Y. pestis* (1.57 x 10^{-8} s/s/y) (Duchêne et al. 2016).

While the average mutation rate was consistent for *C. coli* and *C. jejuni*, the relative number of polymorphisms introduced by homologous recombination and mutation (*r/m*) differed markedly, with on average 37-fold (*C. coli*), compared to 5-fold (*C. jejuni*), greater impact on sequence variation. HGT is known to be an important driver of genome evolution in *Campylobacter* (Wilson et al. 2009; Sheppard et al. 2010) but these estimates are considerably

higher than previous ones using 7-locus MLST (Vos and Didelot 2009). Recombination introduced nearly twice as many synonymous than non-synonymous mutations, but even taking this into account, recombined sequence accounted for around 79% of all non-synonymous variation. This highlights the importance of HGT in rapidly evolving *Campylobacter* genomes and provides evidence that recombination may have been an important factor in the recent diversification of *C. coli* (Sheppard et al. 2008; Sheppard et al. 2011; Sheppard et al. 2013), potentially associated with an adaptive radiation (Rainey and Travisano 1998; Flohr et al. 2013) linked to the colonization of agricultural niches (Thakur et al. 2006). However, this should be balanced against the evidence of purifying selection within recombined sequence (*dN/dS* = 0.492 for *C. coli* and 0.49 for *C. jejuni)* and the removal of non-synonymous mutations through negative selection (Rocha et al. 2006).

Finally, throughout this study we have emphasized the importance of sampling so that measures of molecular evolution are obtained by comparing recent samples with a true ancestor. The uneven distribution of lineages within the population and the possibility that they differ in key evolutionary measures (*r/m* and *dN/dS*), means that our molecular clock estimate may not be applicable to all *Campylobacter* lineages (Didelot et al. 2012; Croucher et al. 2013; Didelot et al. 2013; Everitt et al. 2014). Perhaps this is best illustrated by considering two host-specialist *C. jejuni* lineages, one associated with chickens and the other with cattle (Sheppard et al. 2011; Mourkas et al. 2020). There are 19 billion chickens on earth compared to 1.3 billion cattle (Bar-On et al. 2018) and *C. jejuni* colonizes up to 80% of chickens (Dhillon et al. 2006) with much lower rates in cattle. As the efficiency by which natural selection acts on sequence variation is related to effective population size (Gojobori et al. 1990), the rate of fixation and removal of mutations will be much faster in *C. jejuni* in chickens. Furthermore, chickens have a higher body temperature than cattle therefore the *C. jejuni* will grow faster, have a shorter generation time, and accumulate mutations at a higher rate (Weller and Wu 2015). From this simple example, which ignores many important factors (eg. subniche structure, host transition bottle necking, resident microbiome) it is clear molecular evolution can be influenced by population-scale forces down to the physiology of the individual cell. The approach employed in this study goes some way towards mitigating effects that confound generalized molecular clock estimates. Focussing on well-defined closely related isolate pairs inevitably reduces the number of comparisons from which the mean molecular clock rate is estimated. However, consideration

of the distribution of effective lineages within the population is essential for identifying robust molecular clock estimates in environmental bacteria with complex multi-host ecology and massive effective population sizes.

Materials and Methods

Isolate sampling, genome sequencing and assembly

The accuracy of molecular clock estimates are improved by sampling strains over long time periods. To achieve this, an isolate collection was assembled comprising 53 isolates sampled between 1978 and 1985 (12 *C. coli*, 41 *C. jejuni*) and derived from multiple sources (human, duck, cattle, dog, turkey, wild bird and pig (**Supplementary Table 1**)). Samples were streaked onto mCCDA (PO0119A Oxoid Ltd, Basingstoke, UK) with CCDA Selective Supplement (SR0155E Oxoid Ltd, Basingstoke, UK) and incubated at 37˚C for 48h in a microaerobic atmosphere (85% N_2, 10% CO_2, and 5% O_2) using CampyGen Compact sachets (Thermo Fisher Scientific Oxoid Ltd, Basingstoke UK). Single colonies from each plate was then sub-cultured onto Mueller Hinton (MH) (CM0337 Oxoid Ltd, Basingstoke, UK) agar and grown for an additional 48h at 37˚C and stored in 20% glycerol stocks at -80˚C.

DNA was extracted using the QIAamp DNA Mini Kit (QIAGEN, Crawley, UK), according to manufacturer's instructions. DNA was quantified using a Nanodrop spectrophotometer before sequencing on an Illumina MiSeq sequencer using the Nextera XT library preparation kits with standard protocols. Paired end libraries were sequenced using 2 × 300 bp 3rd generation reagent kits (Illumina). Short read data was assembled using the *de novo* assembly algorithm, SPAdes (version 3.10.0 35) (Bankevich et al. 2012) generating an average of 49 contigs (range: 2 -115) for a total average assembled genome size of 1.69 Mbp (range: 1.62-1.80). The average N50 was 189,430 bp (range: 81,283-974,529). These isolate genomes were augmented with 1,783 *C. jejuni* and 589 *C. coli* genomes archived in BIGSdb (Jolley and Maiden 2010) representing isolates sampled from multiple sources (human, cattle, chicken, cat, dog, duck, environmental waters, farm environments, geese, lamb, rabbit, sand, seal, wild birds, turkey, pig) between 1970 and 2016 (**Supplementary Table 1).** The total isolate collection comprised 2,425 *Campylobacter* genomes, including *C. jejuni* belonging to 286 STs and 36 clonal complexes, and *C. coli* to 125 STs and 1 clonal complex. All assembled genomes and raw reads have been deposited in the NCBI repository associated with BioProject: PRJNA524315. Individual

accession numbers can be found in **Supplementary Table 1**. Assembled genomes of all isolates used in the study are available in FigShare DOI: 10.6084/m9.figshare.7886810.

C. coli and C. jejuni phylogenies and assessing temporal signal and 'clock-likeness'

Phylogenies were constructed for 601 *C. coli* and 1,824 *C. jejuni* isolates (**Supplementary Table 1, Figure 1**). The genomes were aligned against a reference (*C. coli* CVM N29710, accession number: NC_022347.1 and *C. jejuni* NCTC 11168, accession number: NC_002163.1) using MAFFT with default parameters of minimum nucleotide identity of 70% over >50% of the gene and BLAST-n word size 20. Core genes, shared by all isolates within a species (2,014 for *C. coli* and 1,668 for *C. jejuni*) were concatenated and used to construct Maximum likelihood (ML) trees using FastTree version 2.1.8 and the Generalised time-reversible (*GTR*) model of nucleotide evolution (Price et al. 2010). Isolates were analysed to test for a temporal signal of the accumulation of genetic variation over time (**Supplementary Figure 1**). This was carried out prior to mutation rate analysis using a phylogeny of genetic distances and sampling dates, and root to tip regression implemented in the software TempEst v1.5.1 (Rambaut et al. 2016). Core genome phylogenies contained dated-tip isolates sampled between 1970 and 2016 for *C. coli* and *C. jejuni*.

Selection of closely related isolate pairs

An ideal dataset for mutation rate analysis would include isolate pairs with divergent sampling dates, sufficient to measure mutation rate over time, while remaining close enough (clustering on the tree) to share reliable recent common ancestry. Furthermore we required as many pairs as possible for confidence in average mutation rates. In order to achieve this, pairwise nucleotide identity and year between isolation date matrices were constructed separately for 601 (*C. coli*) and 1,824 (*C. jejuni*) isolates. Using a bespoke R script (https://github.com/SionBayliss/CallandMolClock), the distribution of nucleotide identity was determined for isolate pairs within sequential isolation date categories of 1 year or more (1-37 for *C. coli*, 1-46 for *C. jejuni*) by comparing every isolate to all other isolates (**Figure 2**). In each analysis, isolates were used only once as the ancestral or derived strain.

Recombination and mutation inference (quantifying nucleotide change)

The raw reads of genomes (**Supplementary Table 1**) of isolate pairs (**Table 1**) were mapped to the complete reference genomes: *C. coli* YH501 (accession: NZ_CP015528.1) and *C. jejuni* NCTC 11168 using the BWA-MEM algorithm (Li 2013). Variants were called using Freebayes v1.1.0-dirty (Garrison and Marth 2012) and SNP effects predicted and annotated using SnpEff version 4.3 (Cingolani et al. 2012) (**Supplementary Table 4**). These tools were included in the haploid variant calling pipeline, 'snippy' v3.0 (https://github.com/tseemann/snippy). Core genome sub-tree alignments were constructed using snippy-core. Mutations introduced by point mutation and recombination were inferred on the alignments using Gubbins v2.4.1 (default settings) (Croucher et al. 2015) for each isolate pair (**Supplementary Table 4**). The snippy pipeline was used to identify synonymous and non-synonymous polymorphism within and outside of inferred recombinant regions (Croucher et al., 2015). *dN/dS* ratios were calculated for sites across the core genome using the synonymous/non-synonymous analysis program (SNAP) v2.1.1 based on the Nei and Gojobori 1986 method (Korber 2000) (www.hiv.lanl.gov). By quantifying point mutation and recombination and synonymous and nonsynonymous polymorphism, we were able to infer different molecular evolution rate estimates. These included (i) the total mutation rate, used to calculate the number of effective lineages and (ii) the rate of accumulation of synonymous mutations occurring outside of recombinant regions, used to estimate the molecular clock. Hotspots of recombination occurring across multiple isolate pairs were observed.

Estimating the number of coalescences at yearly intervals (Birthday problem)

To consider the extent to which a given sample set represented genetic diversity within the population we developed a pipeline that calculated the number of coalescences (*effective lineages*, **Z**) at yearly time intervals (Z1, Z2, Z3….Zn) within the datasets. This is described by the equation **Z** = **Y/X,** Where: **Y** = all potential isolate pairs ($n^2/2$); **X** = the number of possible pairs for each time interval (*t1, t2,t3….tn*) that is less than the predicted number of mutations that have occurred over a given time interval ($\mu(t(1-n))$; $\boldsymbol{\mu}$ = mutation rate; ***t*** = time interval between sampling dates, 1-46 and 1-37 years for *C. jejuni* and *C. coli* respectively. The resultant **Z** value for each time period is the estimated number of effective lineages (Birthdays) at each time cut-off, equivalent to the number of lineages sharing a common ancestor at a particular time interval (**Supplementary Figure 2**).

Acknowledgements

SKS lab funded by the Medical Research Council (MR/L015080/1, MR/S009264/1, MR/T030062/1). JKC is supported by a BBSRC-CASE studentship (BB/P504750/1). DF was supported by an MRC senior research fellowship (MR/M501608/1). EM is supported by a University of Bath Faculty of Science URSA studentship.

Contributors

JKC, SKS and DF designed the study and wrote the paper with BP. JKC, BP performed genomic analysis with input from HT, SCB and EJF. EB and MB cultured isolates for sequencing. EM, MDH and BP sequenced and assembled genomes. All authors contributed and approved the final manuscript.

Conflict of Interest

The authors declare no conflict of interest.

References

Arnold BJ, Hanage WP. 2017. Longitudinal samples of bacterial genomes potentially bias evolutionary analyses. Unpublished data bioRxiv https://www.biorxiv.org/content/10.1101/103465v1, last accessed October 18, 2020

Bar-On YM, Phillips R, Milo R. 2018. The biomass distribution on Earth. Proc Natl Acad Sci. 115(25):6506–6511.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, et al. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 19(5):455-477.

Biek R, Pybus OG, Lloyd-Smith JO, Didelot X. 2015. Measurably evolving pathogens in the genomic era. Trends Ecol Evol. 30(6):306–313.

Bromham L. 2009. Why do species vary in their rate of molecular evolution? Biol Lett. 5:401-404.

Bronowski C, James CE, Winstanley C. 2014. Role of environmental survival in transmission of *Campylobacter jejuni*. FEMS Microbiol Lett. 356(1):8–19.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Luan Wang SJL, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 6(2):80–92.

Cody AJ, McCarthy ND, Bray JE, Wimalarathna HML, Colles FM, Jansen van Rensburg MJ, Dingle KE, Waldenstrom J, Maiden MCJ. 2015. Wild bird-associated Campylobacter jejuni isolates are a consistent source of human disease, in Oxfordshire, United Kingdom. Env Micro Reports 7(5)782-788.

Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, Van Der Linden M, Mcgee L, Von Gottberg A, Song JH, Ko KS, et al. 2011. Rapid Pneumococcal Evolution in Response to

Clinical Interventions. Science 331(6016):430–434.

Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, Bentley SD, Hanage WP, Lipsitch M. 2013. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat Genet. 45:656–663.

Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 43(3):e15.

Cui Y, Yu C, Yan Y, Li D, Li Y, Jombart T, Weinert LA, Wang Z, Guo Z, Xu L, et al. 2013. Historical variations in mutation rate in an epidemic pathogen, Yersinia pestis. Proc Natl Acad Sci U S A. 110(2):577–582.

Cui Y, Yang X, Didelot X, Guo C, Li D, Yan Y, Zhang Y, Yuan Y, Yang H, Wang J et al. 2015. Epidemic clones, oceanic gene pools, and eco-LD in the free-living marine pathogen Vibrio parahaemolyticus. Mol Biol Evol. 32(6):1396-410.

Dearlove BL, Cody AJ, Pascoe B, Méric G, Wilson DJ, Sheppard SK. 2016. Rapid host switching in generalist Campylobacter strains erodes the signal for tracing human infections. ISME J. 10(3):721–729.

Denamur E, Matic I. 2006. Evolution of mutation rates in bacteria. Mol Microbiol. 60(4).

Dhillon AS, Shivaprasad HL, Schaberg D, Wier F. 2006. Campylobacter jejuni in broiler chickens. Avian Diseases. 50(1):55-58.

Didelot X, Falush D. 2007. Inference of Bacterial Microevolution Using Multilocus Sequence Data. Genetics. 175(3):1251–1266.

Didelot X, Eyre DW, Cule M, Ip CLC, Ansari MA, Griffiths D, Vaughan A, O'Connor L, Golubchik T, Batty EM, et al. 2012. Microevolutionary analysis of Clostridium difficile

genomes to investigate transmission. Genome Biol. 13(12):R118.

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(34):13880–13885.

Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. 2016. Within-host evolution of bacterial pathogens. Nat Rev Microbiol. 14:150-162.

Dingle KE, Colles FM, Wareing DRA, Ure R, Fox AJ, Bolton FE, Bootsman HJ, Willems RJL, Urwin R, Maiden MCJ. 2001. Multilocus sequence typing system for Campylobacter jejuni. Journal of Clinical Microbiology 39(1):14-23.

Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. 2003. Measurably evolving populations. Trends Ecol Evol. 18:481–488.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BioMed Cent. 7(214):1–8.

Duchêne S, Holmes EC, Ho SYW. 2014. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. Proceedings Biol Sci. 281(1786).

Duchêne S, Holt KE, Weill F-X, Le Hello S, Hawkey J, Edwards DJ, Fourment M, Holmes EC. 2016. Genome-scale rates of evolutionary change in bacteria. Microb Genomic. 2(11).

Everitt RG, Didelot X, Batty EM, Miller RR, Knox K, Young BC, Bowden R, Auton A, Votintseva A, Larner-Svensson H, et al. 2014. Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus. Nat Commun. 5(3956).

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. Nat Rev Gen. 8:610-618.

Flohr RCE, Blom CJ, Rainey PB, Beaumont HJE. 2013. Founder niche constrains evolutionary adaptive radiation. PNAS. 110(51):20663-20668.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. Unpublished data, arXiv preprint arXiv:1207.3907 [q-bio.GN]. Last accessed on November 24 2020.

Gibson B, Wilson D, Feil E, Eyre-Walker A. 2018. The Distribution of Bacterial Doubling Times in the Wild. Proc Biol Sci. 285(1880).

Gibson B, Eyre-Walker A. 2019. Investigating evolutionary rate variation in bacteria. J Mol Evol. 87:317-326.

Gojobori T, Moriyama EN, Kimura M. 1990. Molecular clock of viral evolution, and the neutral theory. Proc Natl Acad Sci U S A. 87(24):10015-10018.

Ho SYW, Shapiro B, Phillips MJ, Cooper A, Drummond AJ, Sullivan J. 2007. Evidence for Time Dependency of Molecular Rate Estimates. Sullivan J, editor. Syst Biol. 56(3):515–522.

Ho SYW, Lanfear R, Bromham L, Phillips MJ, Soubrier J, Rodrigo AG, Cooper A. 2011. Time-dependent rates of molecular evolution. Mol Ecol. 20(15):3087–3101.

Jolley KA, Maiden MC. 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics. 11:595.

Jolley KA, Bray JE, Maiden MCJ. 2018. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res. 24(3):124.

Kidgell C, Reichard U, Wain J, Linz B, Torpdahl M, Dougan G, Achtman M. 2002. Salmonella typhi, the causative agent of typhoid fever, is approximately 50,000 years old. Infect Genet Evol. 2(1):39–45.

Kimura M. 1968. Evolutiotionary rate at the molecular level. Nature. 217(5129):624-626.

Kimura M. 1987. Molecular evolutionary clock and the neutral theory. J Mol Evol. 26(1–2):24–33.

Kirchberger PC, Schmidt ML, Ochman H. 2020. The ingenuity of bacterial genomes. Annu Rev Microbiol. 8;74:815-834.

Koeppel A, Perry EB, Sikorski J, Krizanc D, Warner A, Ward DM, Rooney AP, Brambilla E, Connor N, Ratcliff RM, et al. 2008. Identifying the fubdamental units of bacterial diversity: A paradigm shift to incorporate ecology into bacterial systematics. PNAS 105(7):2504-2509.

Korber B. 2000. HIV Signature and Sequence Variation Anaylsis. Computational Analysis of HIV Molecular Sequences, Allen G. Rodrigo and Gerald H. Learn, eds. Dordrecht, Netherlands: Kluwer Academic Publishers, Chapter 4, pages 55-72.

Kuo CH, Ochman H. 2009. Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria. Biol Direct. 4:35.

Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Unpublished data, arXiv:1303.3997v2 [q-bio.GN]. Last accessed on October 18 2020

Li S-J, Hua Z-S, Huang L-N, Li J, Shi S-H, Chen L-X, Kuang J-L, Liu J, Hu M, Shu W-S. 2015. Microbial communities evolve faster in extreme environments. Sci Rep. 4(1):6205.

Louca S, Shih PM, Pennell MW, Fischer WW, Parfrey LW, Doebeli M. 2018. Bacterial diversification through geological time. Nat Ecol Evol. 2, 1458-1467.

Maixner F, Krause-Kyora B, Turaev D, Herbig A, Hoopmann MR, Hallows JL, Kusebauch U, Egarter Vigl E, Malfertheiner P, Megraud F, et al. 2016. The 5,300-year-old Helicobacter pylori genome of the Iceman HHS Public Access. Science 351(6269):162–165.

Marin J, Hedges SB. 2018. Undersampling genomes has biased time and rate estimates throughout the tree of life. Mol Biol Evol. 35(10):2595.

Mathers AJ, Stoesser N, Sheppard AE, Pankhurst L, Giess A, Yeh AJ, Didelot X, Turner SD, Sebra R, Kasarskis A, et al. 2015. Klebsiella pneumoniae Carbapenemase (KPC)-Producing K. pneumoniae at a Single Institution: Insights into Endemicity from Whole-Genome Sequencing. Antimicrob Agents Chemother. 59:1656–1663.

Mathis, FH. 1991. A generalized birthday problem. SIAM review. 33(2):265-270.

Mcadam PR, Templeton KE, Edwards GF, Holden MTG, Feil EJ, Aanensen DM, Bargawi HJA, Spratt BG, Bentley SD, Parkhill J, et al. 2012. Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant Staphylococcus aureus. Proc Natl Acad Sci U S A. 109(23):9107–9112.

Menardo F, Duchêne S, Brites D, Gagneux S. 2019. The molecular clock of Mycobacterium tuberculosis. PLoS Pathog. 15(9):e1008067.

von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N, Bork P. 2007. Quantitative Phylogenetic Assessment of Microbial Communities in Diverse Environments. Science 315(5815):1126–1130.

Mourkas E, Taylor AJ, Meric G, Bayliss SC, Pascoe B, Mageiros L, Calland JK, Hitchings MD, Ridley A, Vidal A, et al. 2020. Agricultural intensification and the evolution of host specialism in the enteric pathogen Campylobacter jejuni. PNAS 117(20):11018-11028.

Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, Croucher NJ, Choi SY, Harris SR, Lebens M, et al. 2011. Evidence for several waves of global transmission in the seventh cholera pandemic. Nature 477(7365):462–465.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3(5):418-426.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. Poon AFY, editor. PLoS One. 5(3):e9490.

Rainey PB, Travisano M. 1998. Adaptive radiation in a heterogeneous environment. Nature. 394(6688):69-72.

Rambaut A, Lam TT, Carvalho LM, Pybus OG. 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2(1).

Rieux A, Balloux F. 2016. Inferences from tip-calibrated phylogenies: a review and a practical guide. Mol Ecol. 25(9):1911–1924.

Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, Feil EJ. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol. 239:226–235.

Rosner BM, Schielke A, Didelot X, Kops F, Breidenbach J, Willrich N, Gölz G, Alter T, Stingl K, Josenhans C, et al. 2017. A combined case-control and molecular source attribution study of human Campylobacter infections in Germany, 2011–2014. Sci Rep. 7(1):5139.

Sheppard SK, McCarthy ND, Falush D, Maiden MCJ. 2008. Convergence of Campylobacter species: implications for bacterial evolution. Science 320(5873):237-239.

Sheppard SK, Dallas JF, Strachan NJC, MacRae M, McCarthy ND, Wilson DJ, Gormley FJ, Falush D, Ogden ID, Maiden MCJ, et al. 2009. Campylobacter genotyping to determine the source of human infection. Clin Infect Dis. 48(8):1072–1078.

Sheppard SK, Dallas JF, Wilson DJ, Strachan NJC, McCarthy ND, Jolley KA, Colles FM, Rotariu O, Ogden ID, Forbes KJ, et al. 2010. Evolution of an Agriculture-Associated Disease Causing Campylobacter coli Clade: Evidence from National Surveillance Data in Scotland. Hartskeerl RA, editor. PLoS One. 5(12):e15708.

Sheppard SK, Colles FM, Mccarthy ND, Strachan NJC, Ogden ID, Forbes KJ, Dallas JF, Maiden MCJ. 2011. Niche segregation and genetic structure of Campylobacter jejuni populations from wild and agricultural host species. Mol Ecol. 20(16):3484–3490.

Sheppard SK, Didelot X, Jolley KA, Darling AE, Pascoe B, Meric G, Kelly DJ, Cody A, Colles FM, Strachan NJC, et al. 2013. Progressive genome-wide introgression in agricultural Campylobacter coli. Mol Ecol. 22(4)1051-1064.

Sheppard SK, Cheng L, Méric G, De Haan CPA, Llarena AK, Marttinen P, Vidal A, Ridley A, Clifton-Hadley F, Connor TR, et al. 2014. Cryptic ecology among host generalist Campylobacter jejuni in domestic animals. Mol Ecol. 23(10):2442–2451.

Sheppard SK, Maiden MCJ. 2015. The evolution of Campylobacter jejuni and Campylobacter coli. Cold Spring Harb Perspec Biol 7(8):a018119.

Strachan NJC, Gormley FJ, Rotariu O, Ogden ID, Miller G, Dunn GM, Sheppard SK, Dallas JF, Reid TMS, Howie H, et al. 2009. Attribution of *Campylobacter* Infections in Northeast Scotland to Specific Sources by Use of Multilocus Sequence Typing. J Infect Dis. 199(8):1205–1208.

Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol 8:4(1):vey016.

Thakur S, Morrow WEM, Funk JA, Bahnson PB, Gebreyes WA. 2006. Molecular epidemiologic investigation of Campylobacter coli in swine production systems, using multilocus sequence typing. Appl Environ Microbiol. 72(8):5666-5669.

Thépault A, Méric G, Rivoal K, Pascoe B, Mageiros L, Touzain F, Rose V, Béven V, Chemaly M, Sheppard SK. 2017. Genome-Wide Identification of Host- Segregating Epidemiological Markers for Source Attribution in Campylobacter jejuni. Appl Environ Microbiol. 83(7).

Vos M, Didelot X. 2009. A comparison of homologous recombination rates in bacteria and archaea. ISME J. 3(2):199–208.

Waldenström J, Broman T, Carlsson I, Hasselquist D, Achterberg RP, Wagenaar JA, Olsen B. 2002. Prevalence of Campylobacter jejuni, Campylobacter lari, and Campylobacter coli in different ecological guilds and taxa of migrating birds. Appl Environ Microbiol. 68(12):5911–5917.

Walker TM, C Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW, et al. 2013. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis. 13:137–146.

Weller C, Wu M. 2015. A generation-time effect on the rate of molecular evolution in bacteria. Evolution (N Y). 69(3):643–652.

Wilson DJ, Gabriel E, Leatherbarrow AJH, Cheesbrough J, Gee S, Bolton E, Fox A, Fearnhead P, Hart CA, Diggle PJ. 2008. Tracing the Source of Campylobacteriosis. PLoS Genet. 4(9):e1000203.

Wilson DJ, Gabriel E, Leatherbarrow AJH, Cheesbrough J, Gee S, Bolton E, Fox A, Hart CA, Diggle PJ, Fearnhead P. 2009. Rapid evolution and the importance of recombination to the gastroenteric pathogen Campylobacter jejuni. Mol Biol Evol. 26(2):385–397.

Yahara K, Didelot X, Ansari MA, Sheppard SK, Falush D. 2014. Efficient inference of recombination hot regions in bacterial genomes. Molecular Biology and Evolution 31(6): 1593-1605.

Figure legends

Figure 1. Little evidence of clustering of isolate sampling dates in *Campylobacter* phylogenies. Maximum likelihood (ML) core genome phylogenetic trees of *C. coli* (**A**) (n = 601) and *C. jejuni* (**B**) (n = 1824) constructed using FastTree version 2.1.8 (Price et al. 2010) and the *GTR* model of nucleotide evolution. Both phylogenies show the distribution of the sample time frame used in this study with major *Campylobacter* clonal complexes identified and terminal nodes coloured according to isolation decade (orange = 1970s, yellow = 1980s, white = 1990s, green = 2000s, blue = 2010s). Scale bars represent the estimated number of mutations per site. Terminal nodes sampled from different decades can be seen scattered throughout both trees with little evidence of clustering by decade. Isolates sampled from the 2000s and 2010s are most abundant within each dataset.

Figure 2. Pair selection criteria curves for inclusion in rate estimates. Visual representation of possible pairs of isolates at all time cut-offs across the sample time frame for *C. jejuni* (**B**) and *C. coli* (**D**). As time difference between pairs increases, distinguishing between individual curves becomes distorted. Therefore, a selection of years were plotted (**A** and **C**) (black = all pairs >1 year difference, pink = >2 years, blue = >4 years, purple = >6 years, orange = >8 years, red = >10 years, green = >15 years). All isolates were paired with the nearest isolate (genetic distance), matched according to difference in year of isolation (coloured lines) for both *C. jejuni* (**A**) and *C. coli* (**C**) (orange line). Dashed boxes (**A** and **C**) show magnified images of the closest pairs from all curves. Grey scale bars (**B** and **D**) indicate the time difference cut-off of each curve for every time point in the sample date frame.

Figure 3. Mutation and recombination in *C. coli* and *C. jejuni*. Average genome-wide SNP positions (red dots = synonymous polymorphisms, blue dots = nonsynonymous polymorphisms) per isolate pair in relation to inferred recombined regions (grey blocks). Each plot represents one pair of isolates considered in rate calibration for *C. coli* (**A**) and *C. jejuni* (**B**) and are ordered according to **Table 1**. *y axis* = number of substitutions in relation to particular bp position of the reference genome (*C. coli* = YH501, *C. jejuni* = NCTC11168) and varies between pairs. *x axis* = position of reference genome in bins of 10,000 bp. The cladogram shows the relatedness of isolate pairs based on nucleotide identity, scale bar indicates

polymorphisms per site. It is evident from both **A** and **B** that recombination is the main source of variation in *C. coli* and *C. jejuni*.

Figure 4. Lineage expansion in *C. jejuni* and *C. coli*. (A) Number of effective lineages (y axis) at each time point within the sample time frame (x axis) for *C. coli* (grey) and *C. jejuni* (black). (**B**) Diagrammatic representation of lineage expansion in *C. coli* and *C. jejuni* showing contrasting lineage diversification scenarios.

Supplementary Figure 1. Root-to-tip linear regression of *C. coli* and *C. jejuni* implemented in the software, TempEst. Root-to-tip genetic distance (y axis) is correlated against sampling times (x axis) for phylogenies of 601 *C. coli* (**A**) and 1,824 *C.jejuni* (**B**). Although both *C. coli* and *C. jejuni* datasets show a weak temporal signal, positive correlations can be seen for both species.

Supplementary Figure 2. Methods for calculating the number of effective lineages within the population. **(A)** The total number of *C. jejuni* and *C. coli* isolates in the population and all potential pairwise comparisons between putative ancestral (black) and contemporary (white) strains to give the total number of potential isolate pairs, **Y. (B)** Isolate pair selection based on divergent sampling date (>8 years) and a nucleotide identity threshold <5000 SNPs. **(C)** Total mutation rate ($\boldsymbol{\mu}$) calculated for all chosen pairs. The rate of accumulation of all synonymous (Sd), nonsynonymous (Sn) substitutions, within (rec) and outside (mut) of recombined regions, was estimated since the most recent common ancestor (MRCA, red circle). The difference in substitutions between each pair was divided by the difference in isolation years to give $\boldsymbol{\mu}$. **(D)** The mutation rate was used to estimate the number of SNPs that were to accumulate over a time period and the number of possible isolate pairs at given time intervals (t*1, t2,t3….tn*) for each species.

Table titles

Table 1: Isolate pair information for *C. coli*

Table 2: Isolate pair information for *C. jejuni*

Table 3: Estimates of evolutionary potential of nucleotide change across all *C. coli* genomes in each pair of isolates

Table 4: Estimates of evolutionary potential of nucleotide change across all *C. jejuni* genomes in each pair of isolates

Table 5: Average rate calibrations in *C. jejuni* and *C. coli*

Table 6: Recombination information for each isolate in each *C. coli* pair considered for rate calibration

Table 7: Recombination information for each isolate in each *C. jejuni* pair considered for rate calibration

Table 8: *C. coli* "birthday problem" data and estimates of coalescence across sample time frame

Table 9: *C. jejuni* "birthday problem" data and estimates of coalescence across sample time frame

Supplementary Table 1: Isolate list information

Supplementary Table 2: TempEst root-tip regression analysis estimates

Supplementary Table 3: List of possible pairs of isolates

Supplementary Table 4: Additional SNP annotations

Supplementary Table 5: Individual pair rates

Figure 1.

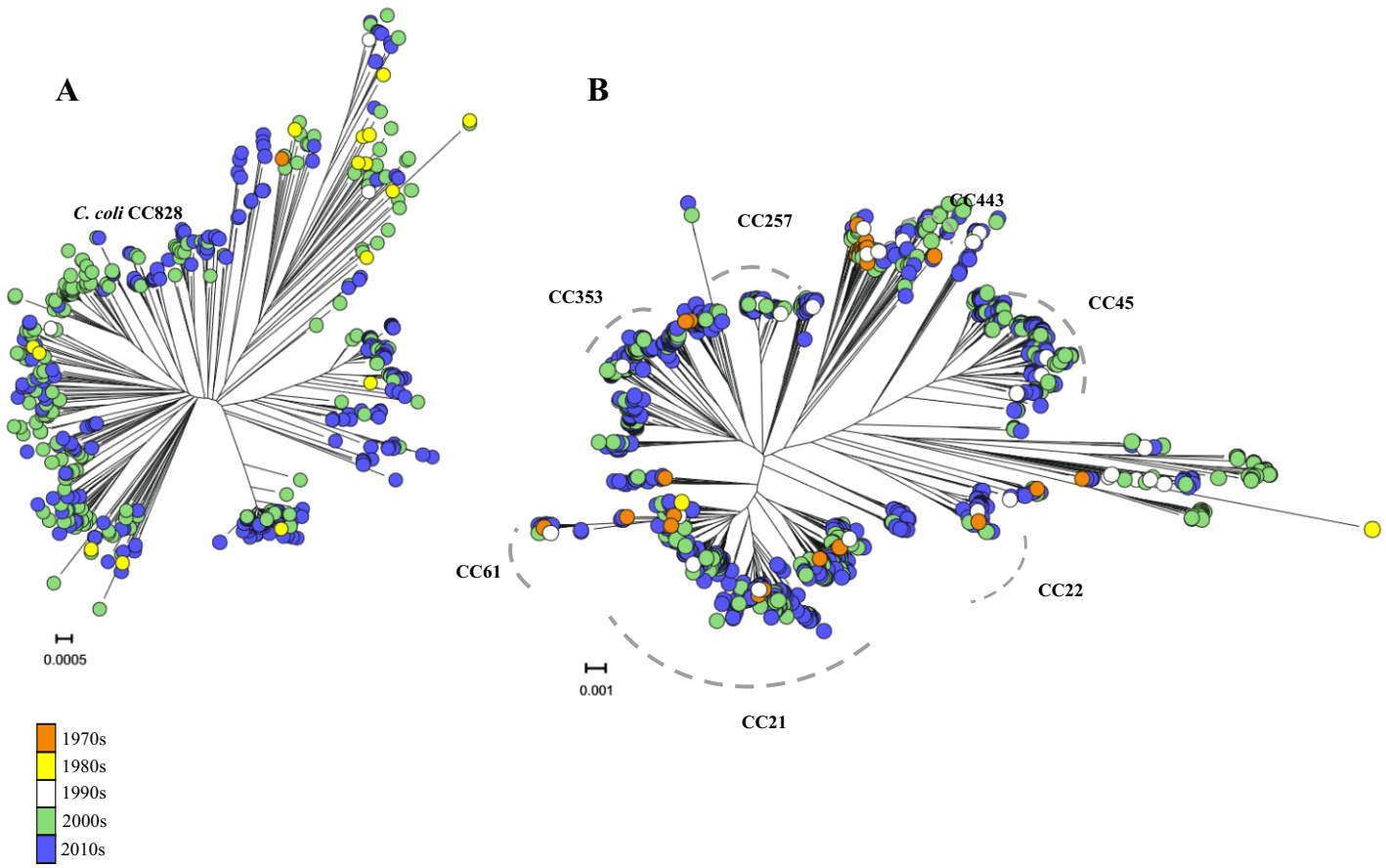

A
C. coli CC828
0.0005
B
CC257
CC443
CC353
CC45
CC61
CC22
CC21
0.001
1970s
1980s
1990s
2000s
2010s

Figure 2.

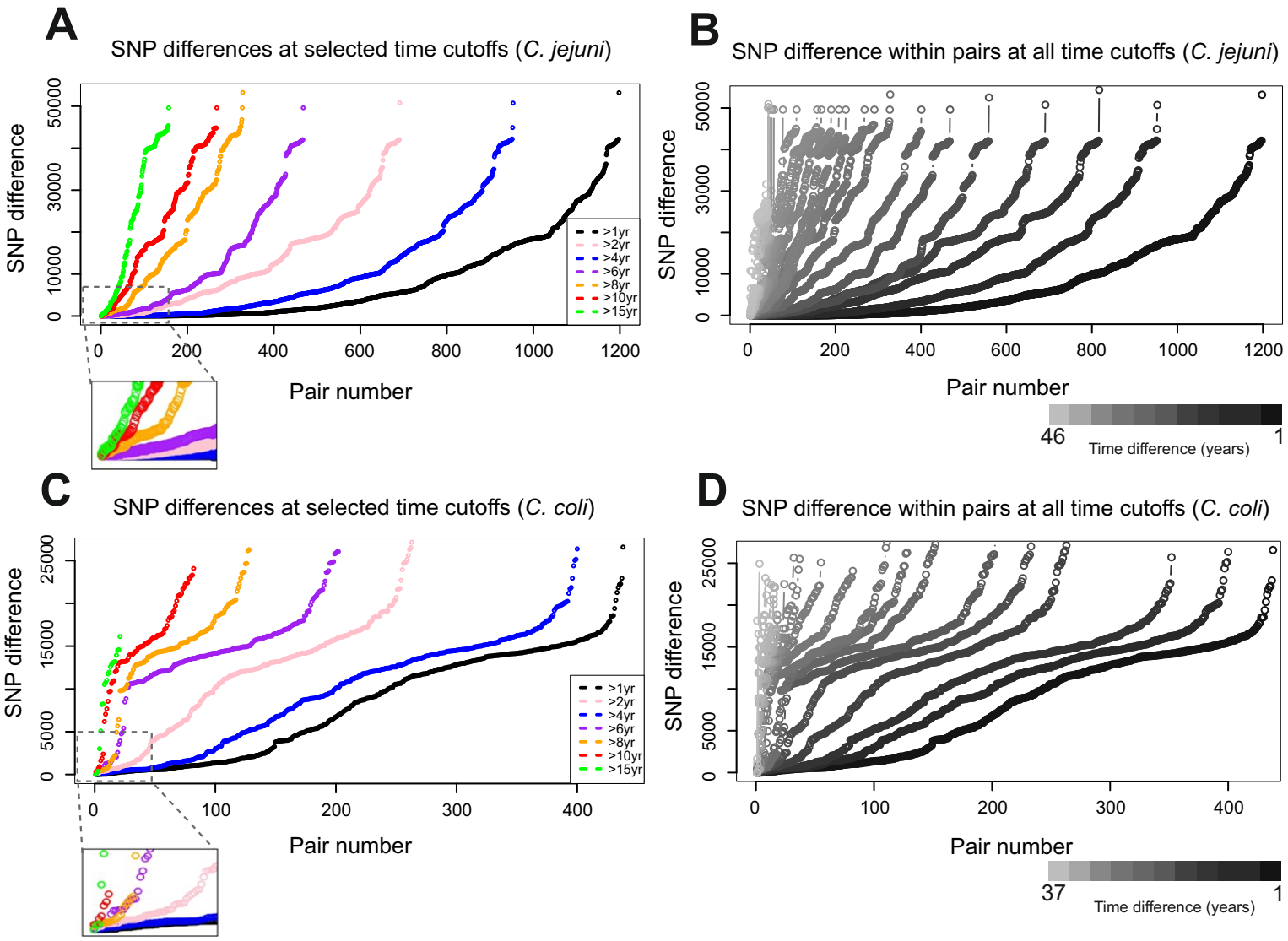

A
SNP differences at selected time cutoffs (C. jejuni)
SNP difference
Pair number
>1yr
>2yr
>4yr
>6yr
>8yr
>10yr
>15yr
B
SNP difference within pairs at all time cutoffs (C. jejuni)
SNP difference
Pair number
46
1
Time difference (years)
C
SNP differences at selected time cutoffs (C. coli)
SNP difference
Pair number
>1yr
>2yr
>4yr
>6yr
>8yr
>10yr
>15yr
D
SNP difference within pairs at all time cutoffs (C. coli)
SNP difference
Pair number
37
1
Time difference (years)

Figure 3.

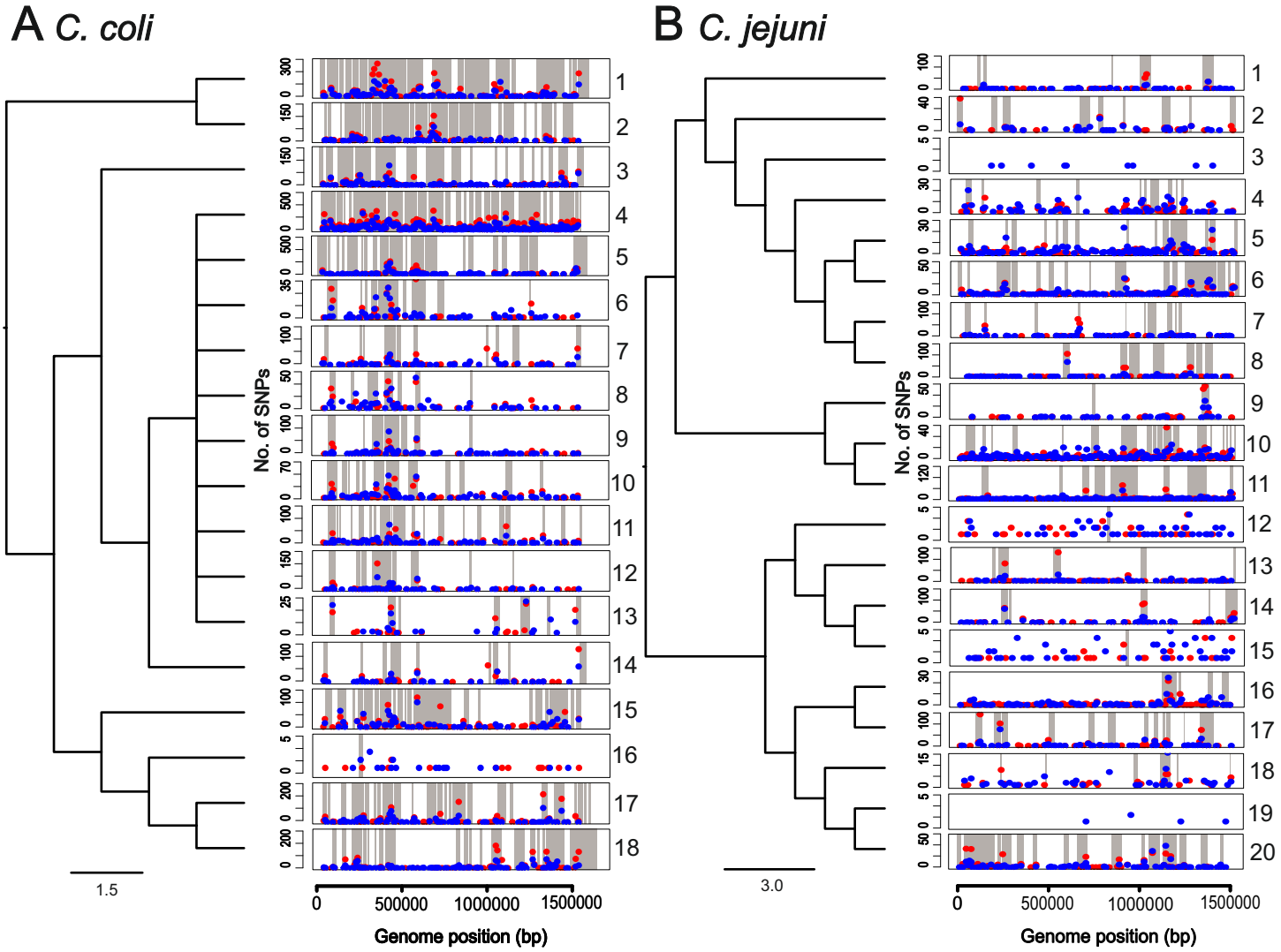
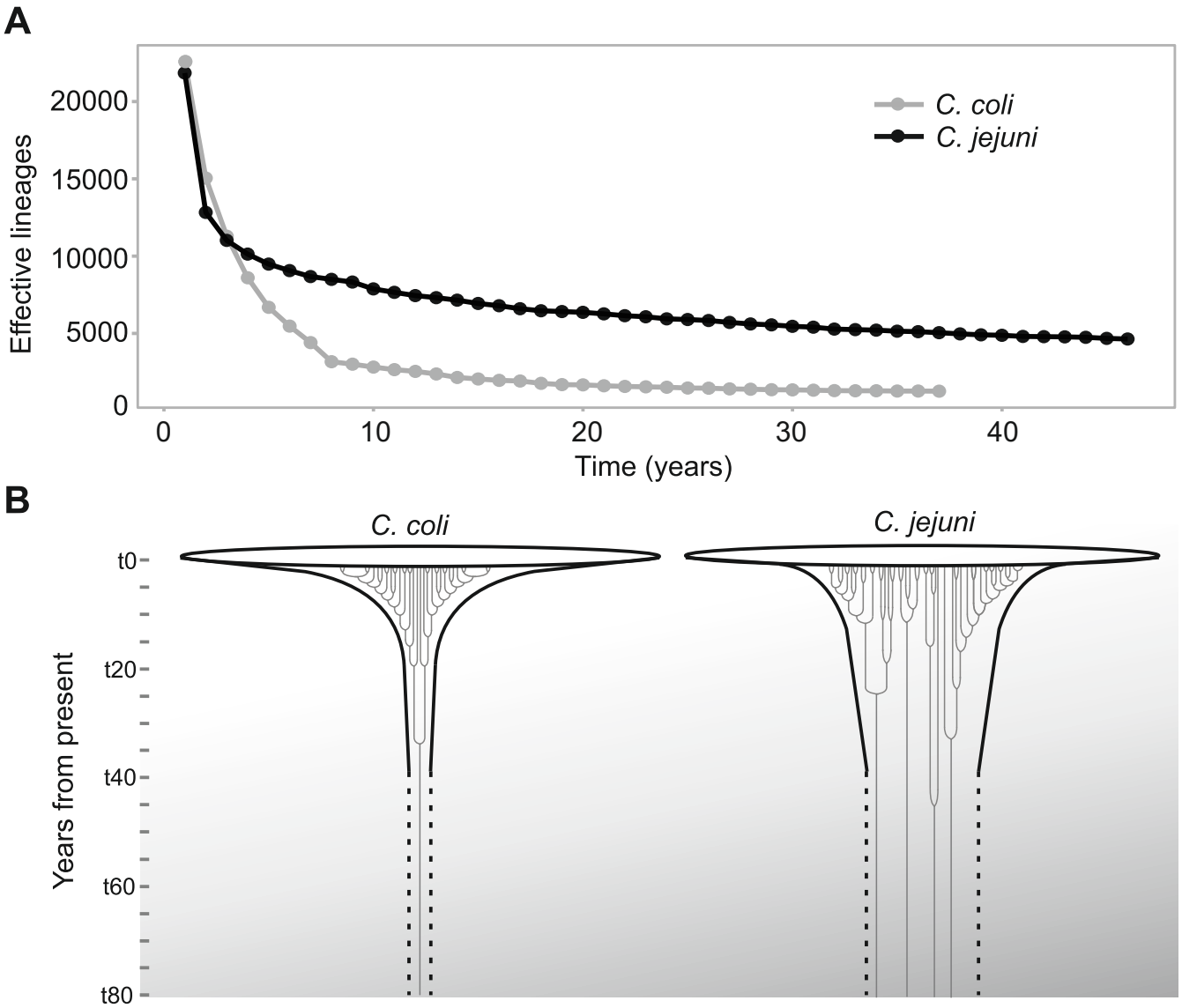
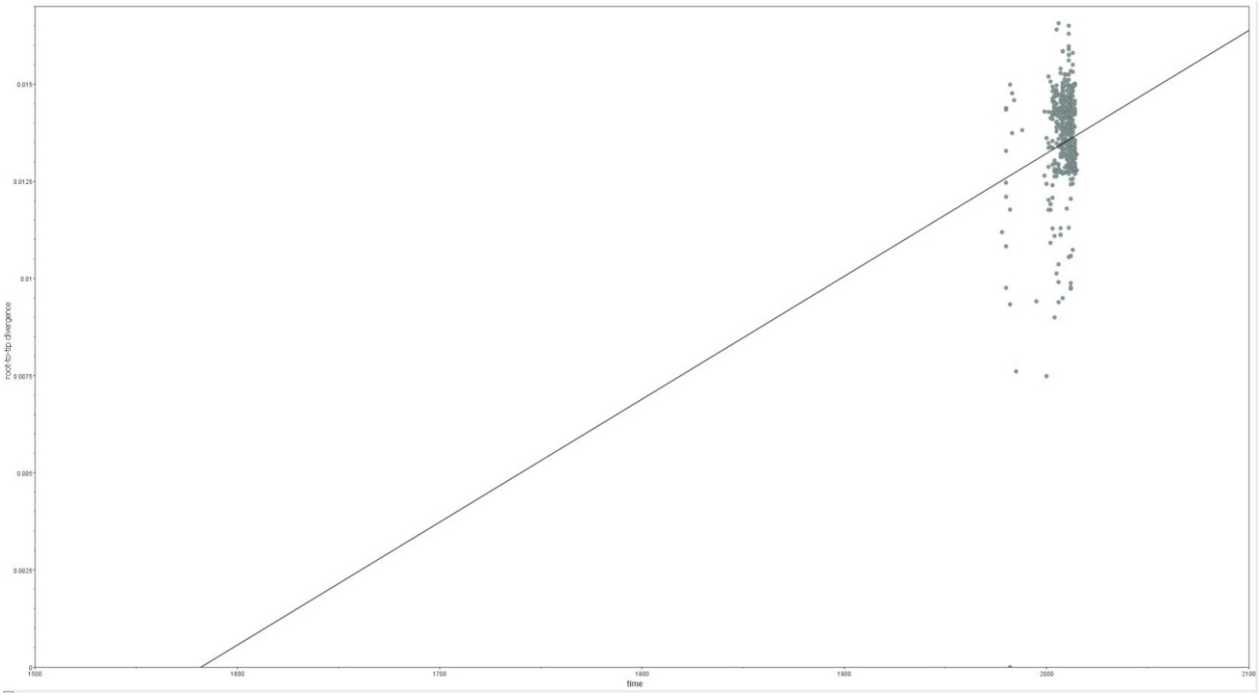

A C. coli
B C. jejuni
No. of SNPs
Genome position (bp)
0
500000
1000000
1500000
1.5
3.0

Figure 4.

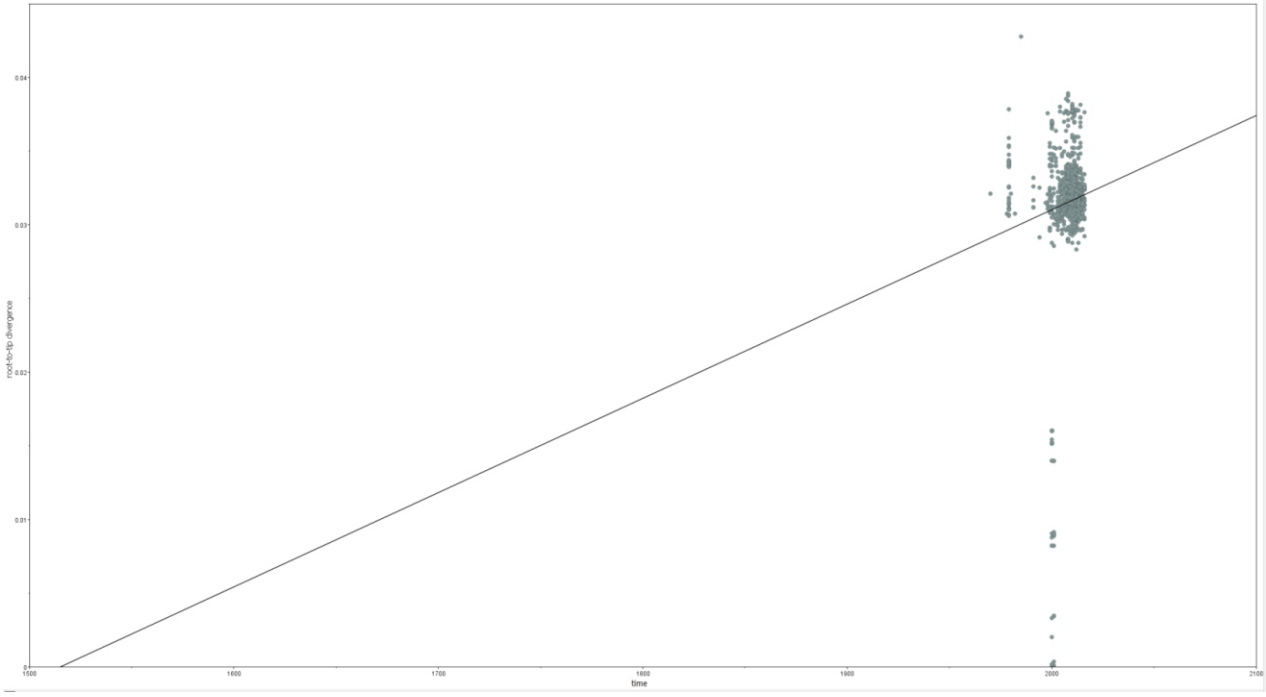

A
Effective lineages
20000
15000
10000
5000
0
0
10
20
30
40
Time (years)
C. coli
C. jejuni
B
C. coli
C. jejuni
Years from present
t0
t20
t40
t60
t80

Supplementary Figure 1.

A) *C. coli*



B) *C. jejuni*



Supplementary Figure 2.

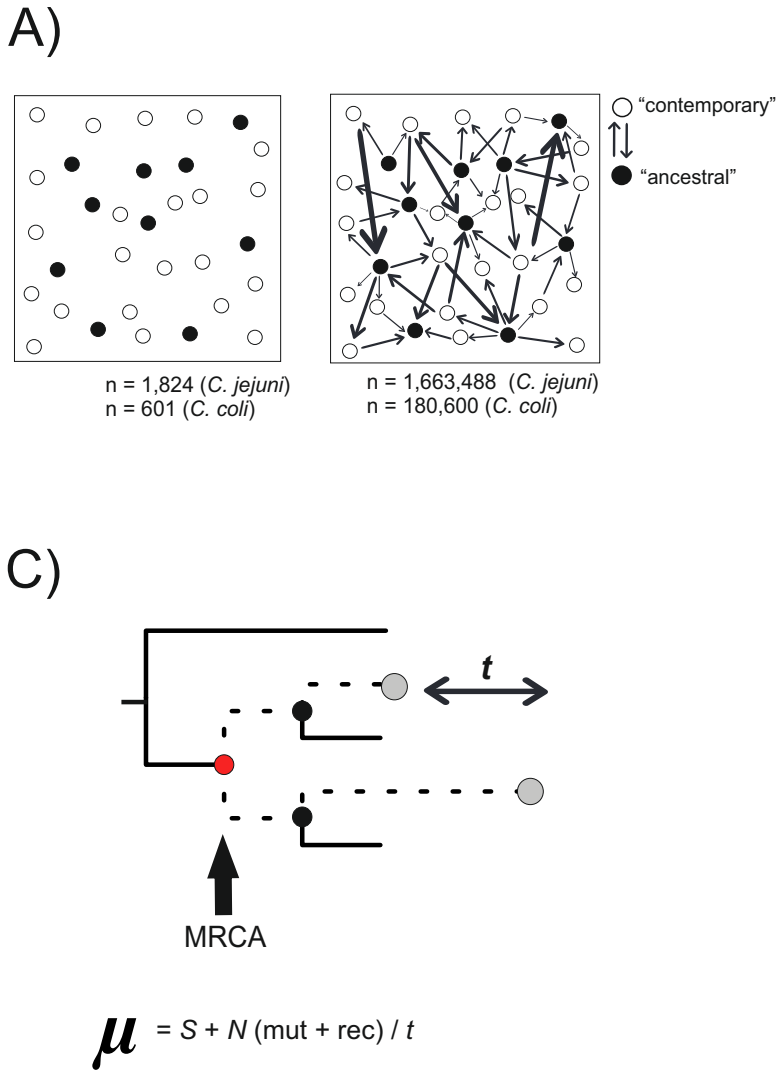


Table 1: Isolate pair information for *Campylobacter coli*

Pair no.	Isolate	Year of isolation	ST*	CC**	Source	Country	Difference in years
1	4316.LDI12946	2006	827	ST-828CC	Environmental	UK	8
	3158.LDI6744	2014	827	ST-828CC	Environmental	UK	
2	4281.LDI12911	2007	827	ST-828CC	Environmental	UK	8
	3804.CCN182coli	2015	827	ST-828CC	Duck	UK	
3	1783.SS_328	2006	827	ST-828CC	Cattle	UK	8
	3166.LDI6752	2014	827	ST-828CC	Environmental	UK	
4	454.SS_047	2005	827	ST-828CC	Chicken	UK	8
	3150.LDI6736	2013	827	ST-828CC	Environmental	UK	
5	3899.H042120298	2004	962	ST-828CC	Environmental	UK	8
	2701.OXC6817	2012	962	ST-828CC	Human	UK	
6	1770.SS_018	2006	827	ST-828CC	Chicken	UK	8
	3160.LDI6746	2014	827	ST-828CC	Environmental	UK	
7	439.SS_031	2005	827	ST-828CC	Chicken	UK	8
	3152.LDI6738	2013	827	ST-828CC	Environmental	UK	
8	1789.SS_335	2006	827	ST-828CC	Wild bird	UK	8
	3174.LDI6760	2014	827	ST-828CC	Environmental	UK	
9	1798.SS_344	2001	827	ST-828CC	Chicken	UK	10
	1753.SS_273	2011	827	ST-828CC	Human	UK	
10	3766.H065100499	2006	827	ST-828CC	Environmental	UK	8
	3176.LDI6762	2014	827	ST-828CC	Environmental	UK	
11	3922.H054900335	2005	825	ST-828CC	Environmental	UK	8
	3161.LDI6747	2013	825	ST-828CC	Environmental	UK	
12	4348.LDI12978	2005	827	ST-828CC	Environmental	UK	8
	3154.LDI6740	2013	827	ST-828CC	Environmental	UK	
13	3925.H043900429	2004	827	ST-828CC	Environmental	UK	8
	2690.OXC6785	2012	827	ST-828CC	Human	UK	
14	3877.H054000445	2005	825	ST-828CC	Environmental	UK	8
	3169.LDI6755	2013	825	ST-828CC	Environmental	UK	
15	1797.SS_343	2005	825	ST-828CC	Chicken	UK	8
	3162.LDI6748	2013	825	ST-828CC	Environmental	UK	
16	3800.UNOR5482c	2003	2195	ST-828CC	Environmental	UK	9
	1855.SS_614	2012	2195	ST-828CC	Human	UK	
17	3803.UNOR13691b	2003	827	ST-828CC	Environmental	UK	8
	1792.SS_338	2011	827	ST-828CC	Chicken	UK	
18	3167.LDI6753	2001	1541	ST-828CC	Environmental	UK	11
	1852.SS_595	2012	6795	ST-828CC	Human	UK	

Table 2: Isolate pair information for *Campylobacter jejuni*

Pair no.	Isolate	Year of isolation	ST*	CC**	Source	Country	Difference in years
1	5932.Manchester	2003	43	ST-21CC	Human	UK	13
	29.NC_002163	2016	43	ST-21CC	Human	UK	
2	275.13264	1999	257	ST-257CC	Human	UK	12
	2140.OXC5779	2011	257	ST-257CC	Human	UK	
3	267.13256	1991	42	ST-42CC	Human	UK	12
	85.cow3583	2003	3583	ST-42CC	Cattle	UK	
4	5921.Aberdeen	2002	43	ST-21CC	Human	UK	14
	5855.LITTER.B.E	2016	43	ST-21CC	Chicken	UK	
5	269.13258	1998	48	ST-48CC	Lamb	UK	9
	1806.SS_381	2007	48	ST-48CC	Chicken	UK	
6	1776.SS_029	2005	45	ST-45CC	Chicken	UK	10
	4809.CTA244	2015	45	ST-45CC	Dog	France	
7	5735.79.23	1979	1457	ST-443CC	Duck	USA	32
	2397.Seal73	2011	1457	ST-443CC	Seal	UK	
8	255.Hn30	2003	50	ST-21CC	Human	UK	12
	4801.CTA093	2015	50	ST-21CC	Dog	France	
9	299.SS_202	2008	257	ST-257CC	Chicken	UK	8
	5814.A4.G	2016	257	ST-257CC	Chicken	UK	
10	303.SS_214	2008	48	ST-48CC	Chicken	UK	8
	5986.LE.72	2016	48	ST-48CC	Chicken	UK	
11	268.13257	1999	45	ST-45CC	Human	UK	12
	2533.OXC6314	2011	45	ST-45CC	Human	UK	
12	5592.RM1285	1997	50	ST-21CC	Chicken	USA	16
	5560.MTVDSCj07	2013	50	ST-21CC	Chicken	USA	
13	2928.SS_0890	1999	1326	ST-45CC	Wild Bird	Sweden	12
	2451.Seal186	2011	1326	ST-45CC	Seal	UK	
14	5038.CjRM3147	2001	22	ST-22CC	Human	Mexico	15
	5902.Camp_108c	2016	22	ST-22CC	Chicken	UK	
15	1800.SS_375	2007	45	ST-45CC	Chicken	UK	8
	4811.CTA277	2015	45	ST-45CC	Dog	France	
16	463.SS_058	2005	45	ST-45CC	Chicken	UK	10
	5006.SS_2784	2015	1701	ST-45CC	Chicken	UK	
17	62.cowa21	2006	21	ST-21CC	Cattle	UK	9
	4834.CTA710	2015	21	ST-21CC	Dog	France	
18	5738.79.248	1979	50	ST-21CC	Turkey	USA	30
	5595.WP2202	2009	50	ST-21CC	Chicken	USA	
19	272.13261	1998	61	ST-61CC	Cattle	UK	18
	191.Cj2008.872	2016	61	ST-61CC	Human	France	
20	5743.79.315	1979	50	ST-21CC	Human	USA	36
	4832.CTA693	2015	50	ST-21CC	Dog	France	

Table 3: Estimates of evolutionary potential of nucleotide change across all *C. coli* genomes in each pair of isolates

	SNPs in rec		SNPs out rec						
Pair no.	**S**	**N**	**S**	**N**	**rdN/dS**	**mdN/dS**	**SNPs/year***	**mol. clock rates****	**Total SNPs rate (m + r)/yr*****
1	202	249	61	105	0.731	0.541	-17.375	-6.000	-78.375
	0	0	13	20	0.000	0.575			
2	0	0	0	0	0.000	0.595	6.500	1.875	32.000
	104	95	15	34	0.473	0.597			
3	167	170	8	14	0.474	0.564	6.000	2.000	45.500
	276	349	24	31	0.362	0.558			
4	0	0	0	0	0.000	0.540	23.125	6.250	65.000
	171	157	50	113	0.540	0.619			
5	0	0	0	0	0.000	0.595	20.000	10.250	171.750
	609	569	82	64	0.564	0.509			
6	110	138	17	24	0.578	0.653	-2.125	-1.375	28.375
	199	284	6	17	0.376	0.571			
7	74	96	8	10	0.744	0.542	5.125	1.375	54.875
	350	210	19	35	0.692	0.579			
8	283	295	14	16	0.938	0.611	2.875	0.375	3.125
	247	314	17	23	0.288	0.543			
9	11717	6313	2467	1520	0.598	0.528	-256.400	-151.200	-1228.200
	4905	3639	955	644	1.144	1.115			
10	40	44	1	0	0.482	0.554	4.000	1.125	69.750
	282	316	10	24	0.679	0.611			
11	7	7	11	10	1.177	0.566	-1.875	-1.250	-3.625
	2	1	1	4	0.000	0.585			
12	0	0	0	0	0.000	0.620	10.250	4.125	62.250
	257	140	33	45	0.527	0.575			
13	49	75	36	74	0.889	0.622	10.875	5.250	177.250
	702	645	78	112	0.575	0.562			
14	419	262	19	44	0.689	0.750	10.000	3.125	107.125
	899	524	44	90	0.720	0.548			
15	947	622	79	39	0.458	0.613	-8.875	-8.125	-174.250
	191	107	14	37	0.354	0.500			
16	549	370	58	64	0.493	0.581	-15.333	-6.444	-126.111
	0	0	0	0	0.464	0.588			
17	681	510	60	129	0.611	0.618	-3.125	1.625	110.500
	1102	932	73	96	0.592	0.580			
18	0	0	0	0	0.000	0.592	27.273	13.818	545.909
	3304	2220	152	127	0.515	0.474			
SD	2101.48	1204.22	431.68	266.48	0.311	0.101	62.649	36.435	339.754
Mean	801	546	123	99	0.492	0.594	11.457	4.266	113.339

*synonymous and nonsynonymous substitution outside of recombination; **synonymous substitutions outside of recombination; ***all substitutions within and without of recombination

S = observed synonymous mutations, N = observed nonsynonymous mutations, SD = standard deviation

Table 4: Estimates of evolutionary potential of nucleotide change across all *C. jejuni* genomes in each pair of isolates

	SNPs in rec		SNPs out rec						
Pair no.	**S**	**N**	**S**	**N**	**rdN/dS**	**mdN/dS**	**SNPs/year***	**mol. clock rates****	**Total SNPs rate (m + r)/yr*****
1	0	0	0	3	0.000	0.000	0.000	0.000	0.077
	0	1	0	1	0.000	0.000			
2	30	50	23	40	0.488	0.505	0.167	0.083	-6.750
	0	0	24	47	0.000	0.580			
3	0	0	0	0	0.000	0.521	9.667	3.250	10.667
	4	3	39	67	0.342	0.517			
4	0	0	0	1	0.000	0.000	0.571	0.000	0.571
	0	0	0	8	0.000	0.000			
5	314	157	33	48	0.462	0.565	-8.000	-2.556	-57.444
	18	20	10	14	2.236	0.524			
6	136	205	90	174	0.730	0.545	58.300	26.800	32.100
	20	62	358	432	0.649	0.511			
7	0	0	0	0	0.000	0.524	7.219	3.000	10.844
	52	61	96	118	0.667	0.534			
8	16	47	32	32	1.131	0.544	27.583	10.917	39.583
	67	128	163	212	0.684	0.569			
9	107	63	20	36	0.000	0.563	-5.500	-1.875	-27.250
	1	2	5	6	1.225	0.503			
10	0	0	1	4	0.000	0.500	7.125	1.875	27.125
	97	62	16	40	0.459	0.534			
11	0	0	0	0	0.000	0.534	9.583	2.750	10.250
	0	7	33	77	0.786	0.542			
12	242	114	50	61	0.635	0.547	-6.313	-2.875	-28.000
	3	8	4	11	0.605	0.598			
13	259	138	27	42	0.311	0.560	-5.500	-2.083	-39.667
	0	0	2	8	0.000	0.538			
14	31	29	2	11	0.406	0.515	4.333	2.067	13.067
	128	61	33	40	0.639	0.571			
15	370	211	64	88	0.508	0.513	45.125	23.500	-20.000
	16	50	252	288	0.322	0.577			
16	172	69	79	101	0.399	0.531	-19.600	-7.500	-44.400
	0	0	4	9	0.000	0.500			
17	37	62	117	210	0.550	0.565	-39.222	-13.000	-51.000
	0	0	0	0	0.000	0.521			
18	10	26	0	0	1.265	0.578	2.300	0.367	10.267
	174	88	11	47	0.632	0.542			
19	177	128	22	44	0.840	0.511	-0.833	0.056	0.556
	202	119	23	33	0.441	0.520			
20	270	201	24	39	0.639	0.482	3.972	1.278	-0.167
	121	204	70	125	0.576	0.566			
SD	102.78	67.13	72.07	89.59	0.461	0.165	20.772	9.085	27.767
Mean	77	59	43	63	0.490	0.509	13.534	5.424	14.101

*synonymous and nonsynonymous substitution outside of recombination; **synonymous substitutions outside of recombination; ***all substitutions within and without of recombination

S = observed synonymous mutations, N = observed nonsynonymous mutations, SD = standard deviation

Table 5: Average rate calibrations in *C. coli* and *C. jejuni*

		C. coli			*C. jejuni*		
	Units**	**Min**	**Mean**	**Max**	**Min**	**Mean**	**Max**
Total substitution rate*	s/s/y	1.7 x 10-6	6.3 x 10-5	3.0 x 10-4	4.8 x 10-8	8.8 x 10-6	2.3 x 10-5
Total substitution rate absent of recombining sequences	s/s/y	1.6 x 10-6	6.4 x 10-6	1.5 x 10-6	1.0 x 10-7	8.5 x 10-6	3.6 x 10-5
Synonymous substitution rate in recombining sequences	s/s/y	4.9 x 10-6	3.1 x 10-4	1.8 x 10-3	2.1 x 10-7	1.9 x 10-6	7.6 x 10-6
Synonymous mutation rate absent of recombining sequences (molecular clock)	s/s/y	2.1 x 10-7	2.4 x 10-6	7.7 x 10-6	3.8 x 10-8	3.4 x 10-6	1.7 x 10-5
Nonsynonymous substitution rate in recombining sequences	s/s/y	4.4 x 10-8	2.4 x 10-5	1.1 x 10-4	5.0 x 10-8	1.4 x 10-6	4.8 x 10-6
Nonsynonymous mutation rate absent of recombining sequences	s/s/y	4.8 x 10-7	3.2 x 10-6	7.8 x 10-6	3.1 x 10-7	4.8 x 10-6	1.6 x 10-5

***all substitutions from within and outside recombination**

****substitutions per site per year (*C. jejuni* = 1.6 Mbp, *C. coli* = 1.8 Mbp)**

Table 6: Recombination information for each isolate in each *C. coli* pair considered for rate calibration

Pair no.	Rec blocks from root of subtree	Rec blocks from last common ancestor	r/m	Genome length (bp)	Bases in clonal frame (bp)	% of recombined genome
1	16	9	8.407	1,627,573	1,588,753	2.99
	7	0	0.978	1,609,340	1,606,957	0.15
2	0	0	0.000	1,666,864	1,666,864	0.00
	7	7	3.923	1,665,656	1,659,185	0.39
3	167	15	35.708	1,617,708	1,220,734	24.90
	162	10	38.985	1,603,282	1,201,817	25.84
4	6	0	10.145	1,616,057	1,593,221	1.41
	15	5	16.141	1,507,717	1,468,305	3.14
5	7	0	5.457	1,651,605	1,614,910	2.22
	52	45	20.475	1,659,211	1,595,168	7.09
6	162	7	46.548	1,621,038	1,225,532	24.71
	161	6	38.660	1,613,935	1,209,402	25.67
7	163	6	44.503	1,613,236	1,215,532	25.05
	171	14	54.284	1,618,186	1,220,026	24.94
8	180	28	39.419	1,621,543	1,223,908	24.84
	160	8	37.994	1,615,211	1,210,318	25.65
9	191	191	6.901	1,579,494	1,287,491	28.63
	172	172	37.839	1,616,244	1,203,711	25.92
10	192	5	143.956	1,616,455	1,209,888	25.48
	197	10	152.767	1,598,501	1,190,972	26.38
11	33	1	48.051	1,648,672	1,514,755	8.12
	32	0	47.809	1,649,085	1,515,390	8.11
12	2	0	4.321	1,653,346	1,647,288	0.37
	12	10	12.071	1,652,455	1,617,482	2.14
13	11	10	9.962	1,582,966	1,569,538	0.85
	36	35	19.288	1,633,910	1,600,539	9.07
14	16	16	31.305	1,645,474	1,588,868	3.44
	32	32	47.809	1,642,338	1,508,643	8.14
15	70	40	63.266	1,655,982	1,494,469	11.56
	35	5	46.657	1,647,567	1,545,999	6.53
16	23	23	7.188	1,651,444	1,521,539	7.87
	23	23	5.289	1,658,746	1,600,962	4.35
17	25	15	50.581	1,617,864	1,541,049	5.10
	72	60	59.602	1,618,755	1,514,712	6.58
18	218	0	42.182	1,597,971	1,119,919	31.12
	306	88	102.167	1,618,051	1,093,701	40.07
SD	85	43	35.448	30,559	189,708	11.74
Mean	87	25	37.240	1,625,375	1,427,987	13.30

SD = standard deviation

Table 7: Recombination information for each isolate in each *C. jejuni* pair considered for rate calibration

Pair no.	Rec blocks from root of subtree	Rec blocks from last common ancestor	r/m	Genome length (bp)	Bases in clonal frame (bp)	% of recombined genome
1	1	0	6.389	1,639,229	1,627,395	0.72
	1	0	6.389	1,634,719	1,622,885	0.72
2	6	6	3.042	1,618,681	1,606,663	0.74
	0	0	0.000	1,618,074	1,618,074	0.00
3	4	0	3.378	1,632,540	1,623,184	0.63
	5	1	0.058	1,626,877	1,618,438	0.66
4	1	0	6.389	1,635,771	1,623,937	0.72
	1	0	6.389	1,597,308	1,585,474	0.74
5	14	14	9.837	1,637,250	1,630,577	2.00
	2	2	2.247	1,639,143	1,637,492	0.10
6	23	23	2.992	1,622,009	1,610,144	1.73
	5	5	0.191	1,613,353	1,602,784	0.66
7	0	0	0.000	1,565,731	1,565,731	0.00
	4	4	0.502	1,605,399	1,576,627	1.79
8	6	6	1.295	1,605,876	1,587,041	1.17
	16	16	1.765	1,602,144	1,595,123	1.62
9	0	0	0.000	1,624,255	1,618,823	1.66
	2	2	0.594	1,635,479	1,630,084	0.33
10	0	0	0.000	1,638,978	1,638,978	0.00
	9	9	2.581	1,617,064	1,597,375	1.34
11	0	0	0.000	1,637,286	1,637,286	0.00
	1	1	0.070	1,634,549	1,633,954	0.04
12	14	5	3.102	1,631,249	1,569,740	3.77
	11	2	1.059	1,635,689	1,595,743	2.45
13	5	5	5.282	1,624,931	1,600,666	1.49
	0	0	0.000	1,635,860	1,635,860	0.00
14	11	3	7.825	1,531,436	1,515,728	1.03
	11	3	8.203	1,479,611	1,464,823	1.01
15	37	12	11.074	1,577,196	1,503,167	5.23
	27	2	4.535	1,578,646	1,527,929	3.21
16	13	5	3.493	1,629,646	1,594,397	3.68
	8	0	2.246	1,585,000	1,559,014	2.55
17	11	11	4.135	1,622,264	1,614,837	0.46
	0	0	0.000	1,608,872	1,608,872	0.00
18	3	3	0.442	1,589,659	1,585,982	0.23
	7	7	6.710	1,626,126	1,625,934	1.05
19	22	11	13.058	1,633,924	1,575,472	3.66
	28	17	10.982	1,571,410	1,495,992	5.30
20	8	8	32.016	1,529,278	1,290,989	18.64
	49	19	35.641	1,600,590	1,336,645	17.51
SD	11	6	7.585	35,024	74,747	3.94
Mean	9	5	5.098	1,609,328	1,579,746	2.22

SD = standard deviation

Table 8: *C. coli* "birthday problem" data and estimates of coalescence across sample time frame

Time cut-off (*t*)	No. of SNPs in time (*st*)	No. of pairs < st (X)	Effective lineages (Z)
1	77	8	22,575
2	155	12	15,050
3	232	16	11,288
4	309	21	8,600
5	386	27	6,689
6	464	33	5,473
7	541	41	4,405
8	618	57	3,168
9	696	60	3,010
10	773	64	2,822
11	850	68	2,656
12	928	71	2,544
13	1,005	76	2,376
14	1,082	84	2,150
15	1,159	88	2,052
16	1,237	92	1,963
17	1,314	94	1,921
18	1,391	102	1,771
19	1,469	107	1,688
20	1,546	108	1,672
21	1,623	112	1,613
22	1,700	115	1,570
23	1,778	117	1,544
24	1,855	119	1,518
25	1,932	123	1,468
26	2,010	124	1,456
27	2,087	128	1,411
28	2,164	130	1,389
29	2,241	133	1,358
30	2,319	134	1,348
31	2,396	136	1,328
32	2,473	139	1,299
33	2,551	140	1,290
34	2,628	140	1,290
35	2,705	142	1,272
36	2,783	142	1,272
37	2,860	143	1,263
			mut + rec SNPs/yr rate for C. coli = 77.3
			Y (all potential pairs) = 180,600

X = possible pairs; Y = potential pairs; Z = Coaelescences ("birthdays")

Table 9: *C. jejuni* "birthday problem" data and estimates of coalescence across sample time frame

Time cut-off (*t*)	No. of SNPs in time (*st*)	No. of pairs < st (X)	Effective lineages (Z)
1	14	76	21,888
2	28	129	12,895
3	42	150	11,090
4	56	163	10,205
5	71	174	9,560
6	85	182	9,140
7	99	190	8,755
8	113	194	8,575
9	127	198	8,401
10	141	209	7,959
11	155	215	7,737
12	169	221	7,527
13	183	225	7,393
14	197	230	7,233
15	212	237	7,019
16	226	242	6,874
17	240	249	6,681
18	254	254	6,549
19	268	256	6,498
20	282	258	6,448
21	296	262	6,349
22	310	267	6,230
23	324	270	6,161
24	338	276	6,027
25	353	278	5,984
26	367	281	5,920
27	381	287	5,796
28	395	292	5,697
29	409	295	5,639
30	423	300	5,545
31	437	303	5,490
32	451	310	5,366
33	465	312	5,332
34	479	314	5,298
35	494	318	5,231
36	508	320	5,198
37	522	324	5,134
38	536	329	5,056
39	550	333	4,995
40	564	335	4,966
41	578	340	4,893
42	592	341	4,878
43	606	342	4,864
44	620	344	4,836
45	635	349	4,766
46	649	352	4,726
			mut + rec SNPs/yr rate for *C. jejuni* = 14.1
			Y (all potential pairs) = 1,663,488

X = possible pairs; Y = potential pairs; Z = Coaelescences ("birthdays")