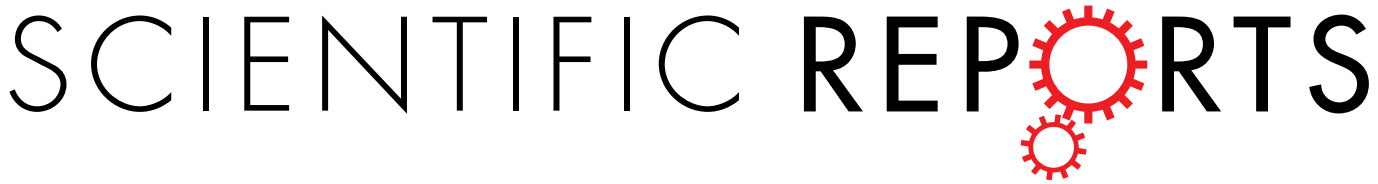

OPEN

Genomic plasticity and rapid host switching can promote the evolution of generalism: a case study in the zoonotic pathogen *Campylobacter*

Received: 12 December 2016
Accepted: 25 July 2017
Published online: 29 August 2017

Dan J. Woodcock[1], Peter Krusche[1], Norval J. C. Strachan[2], Ken J. Forbes[3], Frederick M. Cohan[4], Guillaume Méric[5] & Samuel K. Sheppard[5,6]

Horizontal gene transfer accelerates bacterial adaptation to novel environments, allowing selection to act on genes that have evolved in multiple genetic backgrounds. This can lead to ecological specialization. However, little is known about how zoonotic bacteria maintain the ability to colonize multiple hosts whilst competing with specialists in the same niche. Here we develop a stochastic evolutionary model and show how genetic transfer of host segregating alleles, distributed as predicted for niche specifying genes, and the opportunity for host transition could interact to promote the emergence of host generalist lineages of the zoonotic bacterium *Campylobacter*. Using a modelling approach we show that increasing levels of homologous recombination enhance the efficiency with which selection can fix combinations of beneficial alleles, speeding adaptation. We then show how these predictions change in a multi-host system, with low levels of recombination, consistent with real *r/m* estimates, increasing the standing variation in the population, allowing a more effective response to changes in the selective landscape. Our analysis explains how observed gradients of host specialism and generalism can evolve in a multihost system through the transfer of ecologically important loci among coexisting strains.

Adaptation is typically thought to lead to gradual ecological specialization, in which populations progress towards an optimal phenotype. This can occur among competing organisms in sympatry, particularly when resources diversify or if the cost of maintaining homeostasis in different environmental conditions is high[1]. However, resource or host generalism are also widely observed in nature[2–4] and it is generally accepted that natural environmental heterogeneity can promote the maintenance of phenotypic variation where it confers an ecological advantage[1,4].

Host generalism is exhibited by some bacteria that infect multiple hosts resulting in important implications for the spread of disease from animals to humans. In some zoonotic bacteria, such as *Staphylococcus aureus*, livestock-associated lineages are largely host-restricted[5], allowing the direction and time-scale of host transfer to be estimated by comparison of genotype information[6]. In contrast, in *Escherichia coli* it is difficult to link ecological niche with genotype as isolates from all major phylogroups are represented in multiple isolate sources[7]. Other organisms including *Campylobacter jejuni* and *Salmonella enterica*[8] represent an intermediate between these. For example, comparison of *C. jejuni* isolates from various sources, by multilocus sequence typing (MLST)[9] and

[1]Warwick Systems Biology Centre, Coventry House, University of Warwick, Coventry, CV47AL, UK. [2]School of Biological Sciences, University of Aberdeen, Cruickshank Building. St Machar Drive, Aberdeen, AB24 3UU, UK. [3]School of Medicine and Dentistry, The University of Aberdeen, Foresterhill, Aberdeen, AB25 2ZD, UK. [4]Department of Biology, Wesleyan University, Middletown, CT, 06459-0170, USA. [5]The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath, BA2 7AY, UK. [6]Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS, UK. Correspondence and requests for materials should be addressed to S.K.S. (email: s.k.sheppard@bath.ac.uk)

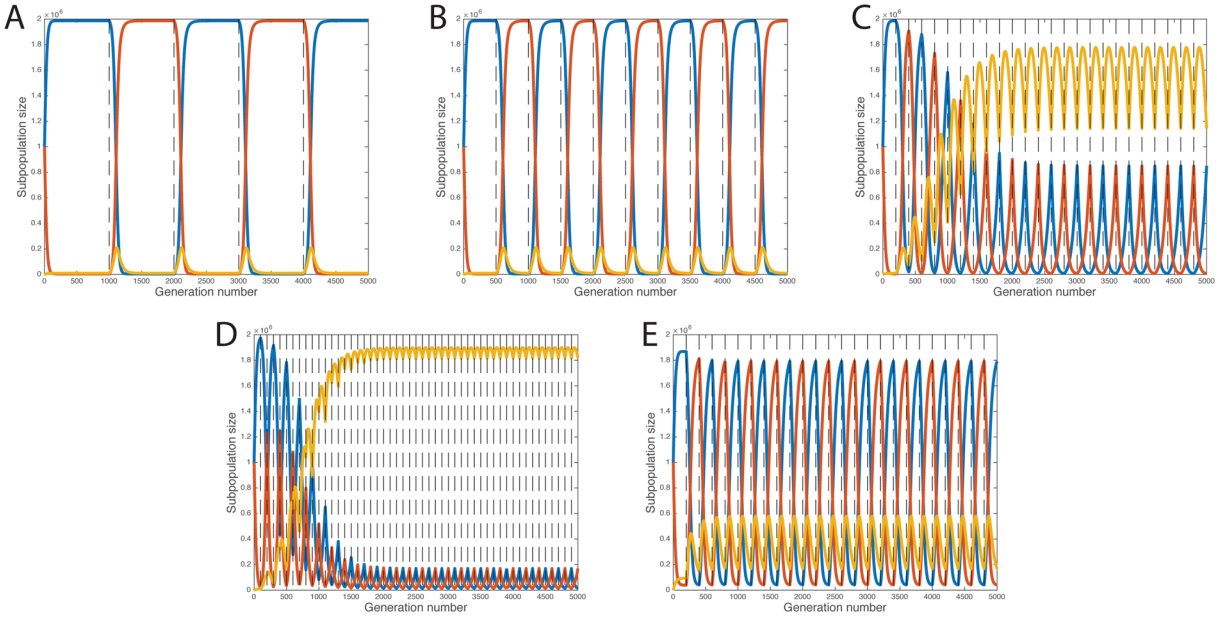

Figure 1. Effect of host switching frequency in the three phenotype model. The abundance of three phenotypes, the Host 1 specialist (blue), the Host 2 specialist (red) and generalist (yellow) were modelled as a set of ordinary differential equations. Simulations were run for 5000 generations where host-swtiching occurred every (**A**) 1000 generations (**B**) 500 generations (**C**) 200 generations (**D**) 100 generations, where the probability of switching between hosts, \phi, was set at 0.0001. Another simulation was run where the probability of switching between phenotypes increased to 0.001, with host switching every 200 generations (**E**).

whole genome sequencing[10,11], has shown evidence for host restricted lineages, found predominantly in one host species, as well as lineages commonly isolated from hosts as diverse as chickens, cattle and wild birds[12].

Out-competition by host specialists, might be expected to select against generalist *Campylobacter* lineages. However, generalists remain among the most common lineages in agricultural animals and are a major cause of human disease[13]. There are several factors that may be involved in the emergence of generalism as a successful strategy. First, the development, industrialisation and globalisation of livestock farming have created a vast open niche in which *Campylobacter* has expanded from pre-agriculture wild animal hosts. Second, factors associated with livestock husbandry and habitations promote close contact between different livestock species providing opportunities for *Campylobacter* transmission from one host to another. Third, *Campylobacter* is a highly recombinogenic organism[14] and lineages regularly acquire genetic elements through horizontal gene transfer (HGT). This genomic plasticity has the potential to introduce DNA segments, or whole genes, into the recipient genome, potentially conferring novel function.

The coexistence of *Campylobacter* lineages with generalist and specialist ecology remains poorly understood and little is known about the factors that promote the emergence and maintenance of lineages with distinct ecology (ecotypes). Here, using information on natural genomic variation in 15 genes from 130 *C. jejuni* genomes from chicken and cattle[15], as well as a computational model of bacterial evolution, we find that the maintenance of genetic variance of putative niche-specifying genes in the population can promote the emergence of generalism as seen in nature. By quantifying how resource competition, rapid host switching, and horizontal gene transfer interact to affect the variance in the population, we provide a generalised framework for considering the emergence of generalist ecotypes.

Results

Emergence of generalism from a background of specialists is dependent on frequency of host transitions in a three phenotype model. Evidence that frequent host migration can lead to the emergence of ecological generalist genotypes is well established in work on the evolution of viruses[16–21]. This principal has not been well studied among the bacteria, so to demonstrate that ecological generalism can be influenced by the frequency of host transitions we first considered a very simple model of a bacterial ecosystem under finite resources (model description in methods). In this model, the bacteria adopt one of three phenotypes/strategies: specialist in host 1; specialist in host 2; generalist. Bacteria can switch between these strategies at a set rate, ϕ. Simulations were performed with this model for 5000 generations with four different transition durations, τ (Fig. 1). At $\tau = 1000$, the dominant phenotypes are host specialists, as generalists briefly proliferate shortly after a host transition but are outcompeted and return to a basal level (Fig. 1A). This pattern is also seen at $\tau = 500$ (Fig. 1B). However, at $\tau = 200$, we observe that the generalists emerge as the dominant phenotype (Fig. 1C), which is even more pronounced when $\tau = 100$ (Fig. 1D). As such, we conclude that under this simple model, generalists can theoretically emerge from a background of specialists, consistent with evolutionary theory among multi-host viruses[16]. Furthermore, the fitness of generalists is intrinsically linked to the frequency of host transitions. An

explanation for this is that when the host transitions are sufficiently frequent that the specialists cannot grow rapidly enough to reach the carrying capacity before the next transition, generalists gain a foothold in the overall population, which increases with each transition. This also has implications for the rate of adaptation because if adaptation increases, then the rate at which the specialists (as the dominant phenotype) can adapt to reach the carrying capacity. Indeed, we find that if we increase the probability of phenotype switching by an order of magnitude to $\phi = 0.001$, then the specialists are able to adapt sufficiently quickly and the generalists are no longer the dominant phenotype when $\tau = 200$ (Fig. 1E).

Long term adaptation promotes specialism but recombination enhances colonization in a subsequent host transition. Expanding upon the principles established in the three phenotype model, we developed a fully stochastic model of a bacterial population, which we refer to as the Genome Evolution by Recombination and Mutation (GERM) model. This model incorporated several distinct gene loci that contributed varying amounts to the overall genotype fitness depending on their constituent alleles and the current host. Alleles were derived from gene-by-gene analysis of *C. jejuni* genomes from natural populations. A total of 5 host segregating genes (proxy for niche-specifying genes) per host (chicken and cattle) and 5 MLST genes were chosen to give a total of 15 genes for input into the full model (See methods). As such, there were five alleles corresponding directly to specialism in either host, with no pre-defined alleles corresponding to generalists.

The effect of homologous recombination was characterized in 200 independent population GERM simulations, with a transition from the composite niche to Host 1 after 200 generations, followed by a transition to Host 2 after another 1000 generations. Simulations were performed at five r/m ratios: 0; 0.1; 1; 10; 100. In all simulations, the mean number of cells decreased sharply from the initial condition and then recovered to approach an equilibrium level between birth and death just before the transition to Host 1 (Fig. 2). In the composite niche, level of proliferation was proportional to the rate of homologous recombination with concomitant increase in mean fitness and population genetic variance (Fig. 2C and D). This is because higher recombination rates result in greater genetic variance, and so by Fisher's fundamental theorem of natural selection[22], the rate of increase in the mean fitness will be greater and the population will thrive. Following the first host transition (Fig. 2A, numeral I), there was a brief increase in the population due to increased resource availability, but in all cases the mean number of cells quickly returned to the equilibrium state where it continued until the next host transition (numeral II), with the mean number of cells ordered largely as before, with the number of cells increasing with recombination rate (Fig. 2C). After the transition to Host 2, the populations were decimated in all cases, as the alleles which would have conveyed enhanced fitness in Host 2 have been purged from the population, meaning that the bacteria are ill-equipped to survive in the new host and so they will die. Almost all of the populations died out in the low recombining groups ($r/m = 0$ and 0.1), a large proportion died out in the intermediate group ($r/m = 1$), with the greatest number of surviving populations in the highly recombining groups.

Intermediate recombination rates enhance population mean fitness after multiple rapid host transitions. As in the single host transition model, the mean number of cells in the composite niche reached equilibrium levels ordered by their homologous recombination rate (Fig. 3A) and consistent with their mean fitness levels (Fig. 3B). At the end of growth in the composite niche, the intermediate and high recombination rates ($r/m = 1$, 10 and 100) have a similar ability to survive, with the highest recombination level displaying high variance (Fig. 3C). After the composite niche a number of host transitions were simulated where the mean number of cells shifted between two equilibrium states depending on the host species. The mean number of cells was always higher in Host 1 because some alleles conferring increased fitness in Host 2 will inevitably be purged from the populations in the Host 1 niche. Reversing the species order had the same effect for Host 2. At the end of the last Host 2 growth cycle all the recombining populations ($r/m > 0$) had a similar mean number of cells, but the variance differed, with the smallest variance at $r/m = 1$ and $r/m = 10$. Similarly, in the final Host 1 niche, we can see in Fig. 3A that the intermediate recombination rate ($r/m = 1$) is associated with the highest mean number of cells. This shows that in contrast to the single host transition model, recombination at a high level is not advantageous under repeated host switching.

Emergence of host generalist strategy as a consequence of frequent host switches. We used a Dirichlet process clustering algorithm[23] on all simulations to identify characteristic profiles of population dynamics for the different homologous recombination rates (Fig. 4). Three broad population dynamic profiles were observed: (i) populations that were primarily adapted to Host 1 (Clusters 1–5); (ii) populations that were primarily adapted to Host 2 (Clusters 7–9); (iii) populations that were adapted to both niches (Cluster 6). This is consistent with a classification as a specialist for either species, or as a generalist. The membership of these 3 adaptation profile types relates to recombination rate (Table 1). Simulations with no recombination were predominantly found in Cluster 2, with a substantial amount found in other clusters. This is to be expected as the outcome was driven entirely by stochasticity acting on the population and so genes were purged almost at random in the composite niche, yielding a set of outcomes which were maintained during the host transitions as more alleles were lost. In contrast, it can be seen that simulations of all of the recombining populations are predominantly found in Cluster 4, which is a Host 1 specialist cluster, albeit with a relatively high equilibrium population number during the Host 2 niche compared to the other Host 1 specialist clusters. In Cluster 4, it can be seen that an $r/m = 1$ gives the greatest occupancy, at 91.8%, explaining the high numbers of cells seen in Fig. 3. The membership of the generalist cluster, Cluster 6, is also represented across all recombination rates, with the highest percentage coming from a relatively low recombination rate ($r/m = 0.1$).

The gradient of host generalism is mirrored in natural *Campylobacter* populations. The degree of host specialism and generalism in *in silco*, resulting from model simulations, was compared to data from

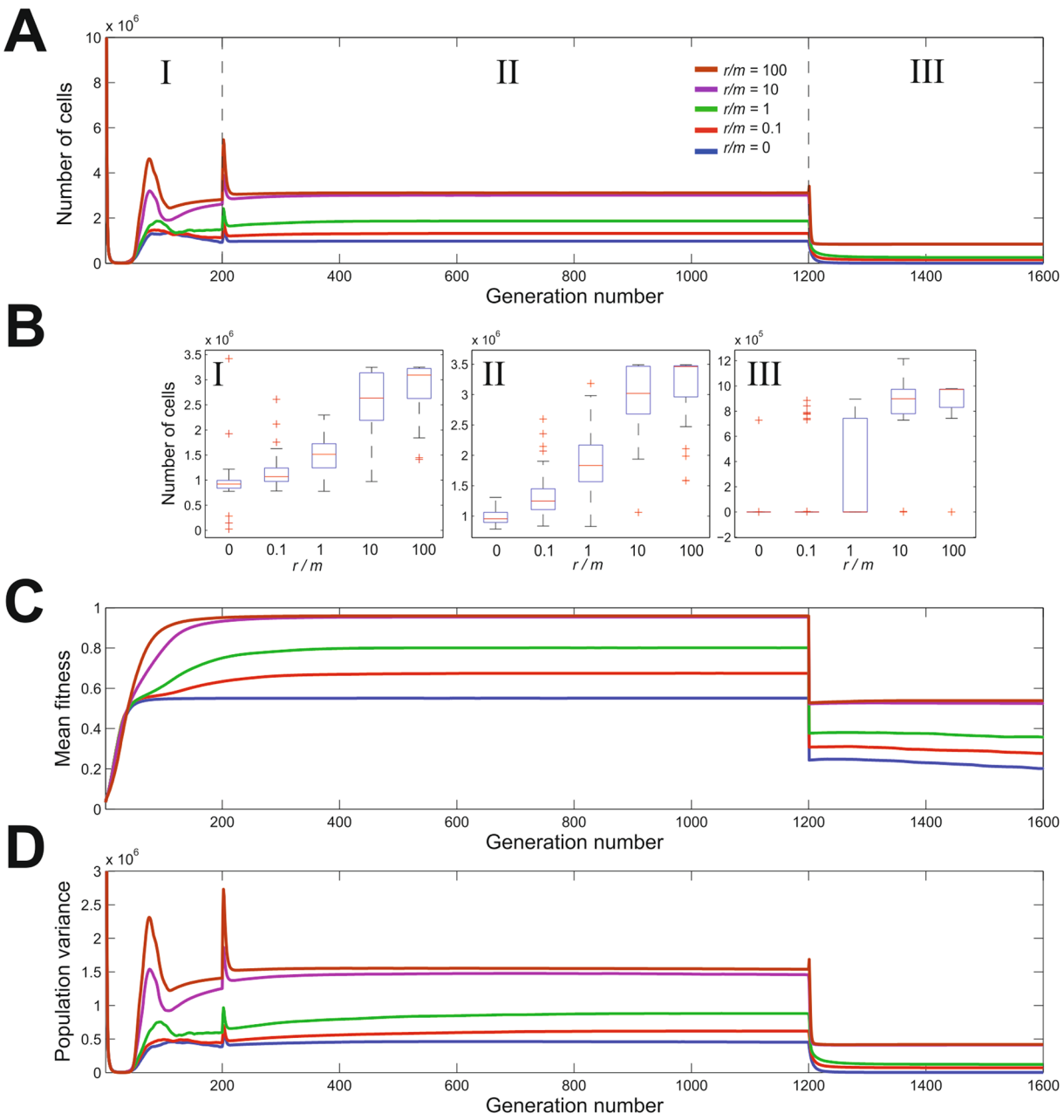

Figure 2. Long term host adaptation followed by a host switch with different recombination rates. The number of cells (**A**), population mean fitness (**C**) and population variance (**D**) were monitored in 200 independent simulations performed at recombination to mutation ratios of: $r/m = 0$ (blue), $r/m = 0.1$ (red), $r/m = 1$ (green), $r/m = 10$ (magenta) and $r/m = 100$ (brown). Model parameters were different in phases I, II and III. Phase I corresponds to a composite niche no fitness-related selection. The transition from I to II corresponds to the addition of selective pressure favoring genes specifying adaptation to host 1, and the transition from phase II to III, corresponds to a single host transition, with a change to selective pressures favoring genes specifying adaptation to host 2. Panel B shows the number of cells at the end of every phase for populations with different recombination rates.

natural *Campylobacter* populations. Mapping isolation source data of 30239 *Campylobacter* STs within the PubMLST database[24] to a core genome ClonalFrame ML tree of the 74 common *C. jejuni* and *C. coli* STs revealed that few STs demonstrate absolute generalism or specialism (Fig. 5a). Rather there is a gradient ranging from STs predominantly found in one host to those frequently isolated from multiple host sources (Fig. 5b). This could be influenced by sampling heterogeneity, as there is some correlation between the number of sampled isolates for a particular ST and its host range size (Figure S1). Therefore, the clustering of isolate pairs, with shared host source richness, was estimated by correlating nucleotide divergence in the core genome with the number of hosts (Fig. 5c) by a corrected rarefaction analysis (by randomising sample size and resampling isolates to prevent over-sampling of particular dominant STs). STs on the phylogeny were located close to STs with similar host species

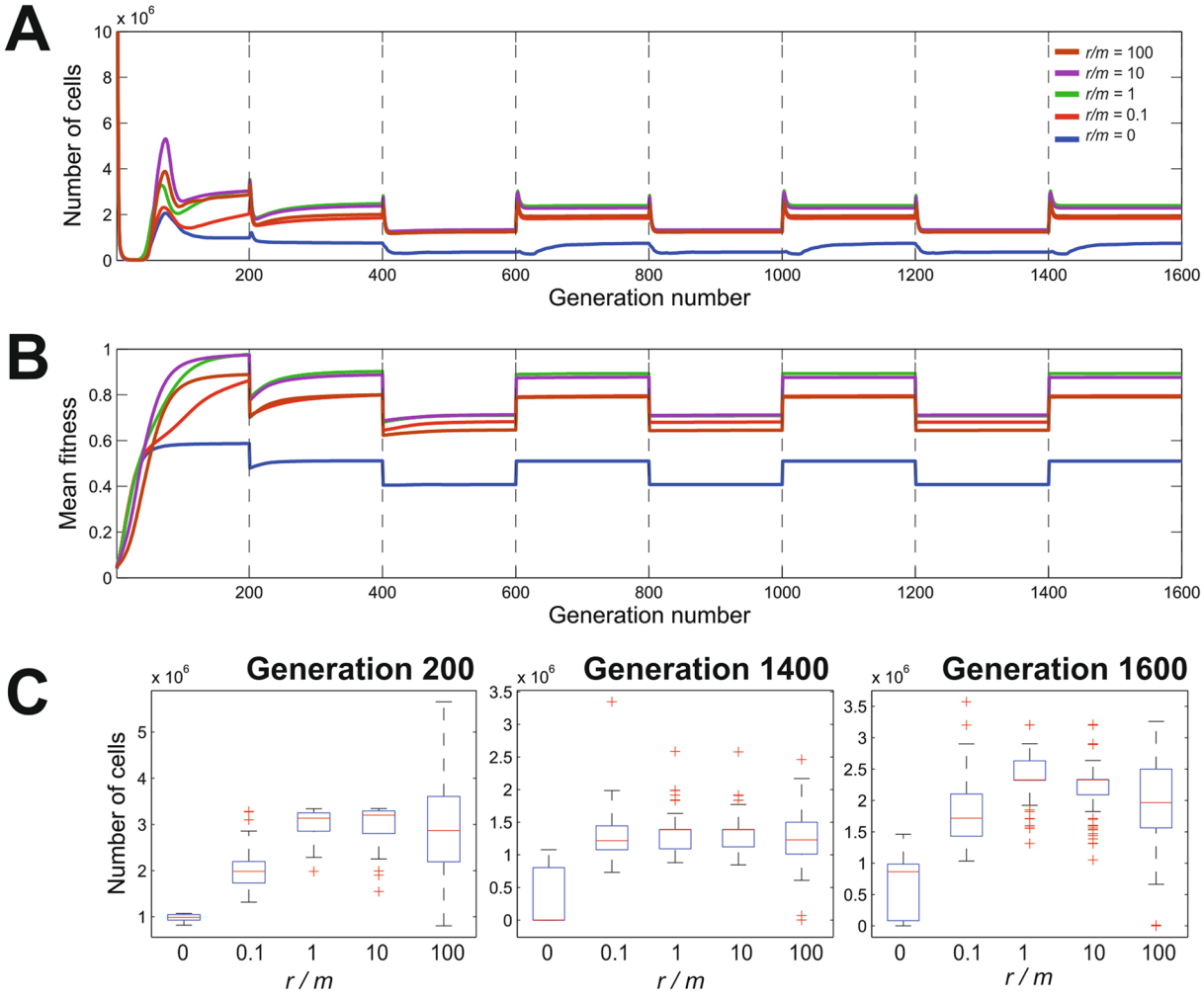

Figure 3. Rapid multiple host transitions with different recombination rates. Number of cells (**A**) and mean fitness (**B**) for recombination to mutation ratios of $r/m = 0$ (blue), $r/m = 0.1$ (red), $r/m = 1$ (green), $r/m = 10$ (magenta) and $r/m = 100$ (brown) as they progress through several niche transfers (broken lines), for which selective pressures are alternatively imposed to favor genes specifying adaptation to one or the other host. (**C**) Population growth distribution at various recombination rates at generations 200 (transition from composite niche to the first host), 1400 (last transition growth phase in host 2) and 1600 (end of simulation after 6 host switches).

richness suggesting that there is some evolutionary signal which determines the likelihood of an ST's degree of specialism, but there was no evidence from the tree that specialist and generalist lineages had different degrees of genetic diversification.

Discussion

The GERM model provides a context for considering how genome plasticity may influence the proliferation of *Campylobacter* in a multihost environment. In simple simulations, rapid acquisition of niche-specifying genes promoted better colonization in a new host. This is consistent with the Fisher–Muller evolutionary model[25,26] where recombination functions to bring together fit alleles, which would otherwise compete for fixation in the population, into a single lineage speeding the overall increase in population mean fitness[27]. In line with classical population genetic theory for sex[28–30], the efficiency of selection on bacteria is enhanced by this shuffling of alleles. Therefore, the population with the highest recombination rate will expand to fill the niche more rapidly after a genetic bottleneck. This demonstrates a clear short-term adaptive benefit to rapid recombination, but does not explain why most bacteria recombine at low rates in nature.

Where survival is predominantly influenced by a few genetic determinants, for example the acquisition of essential antibiotic resistance genes[31], high recombination rates would be favoured. However, this is an unusually simple evolutionary scenario and bacterial habitats comprise numerous interacting selective pressures. Because increased genetic variation leads to faster adaptation[25], the potential for the population to survive future genetic bottlenecks is related to the fitness variance. In populations that recombine at a low rate, a relatively high fitness variance is often maintained. Therefore, if the species is likely to encounter frequent environmental changes,

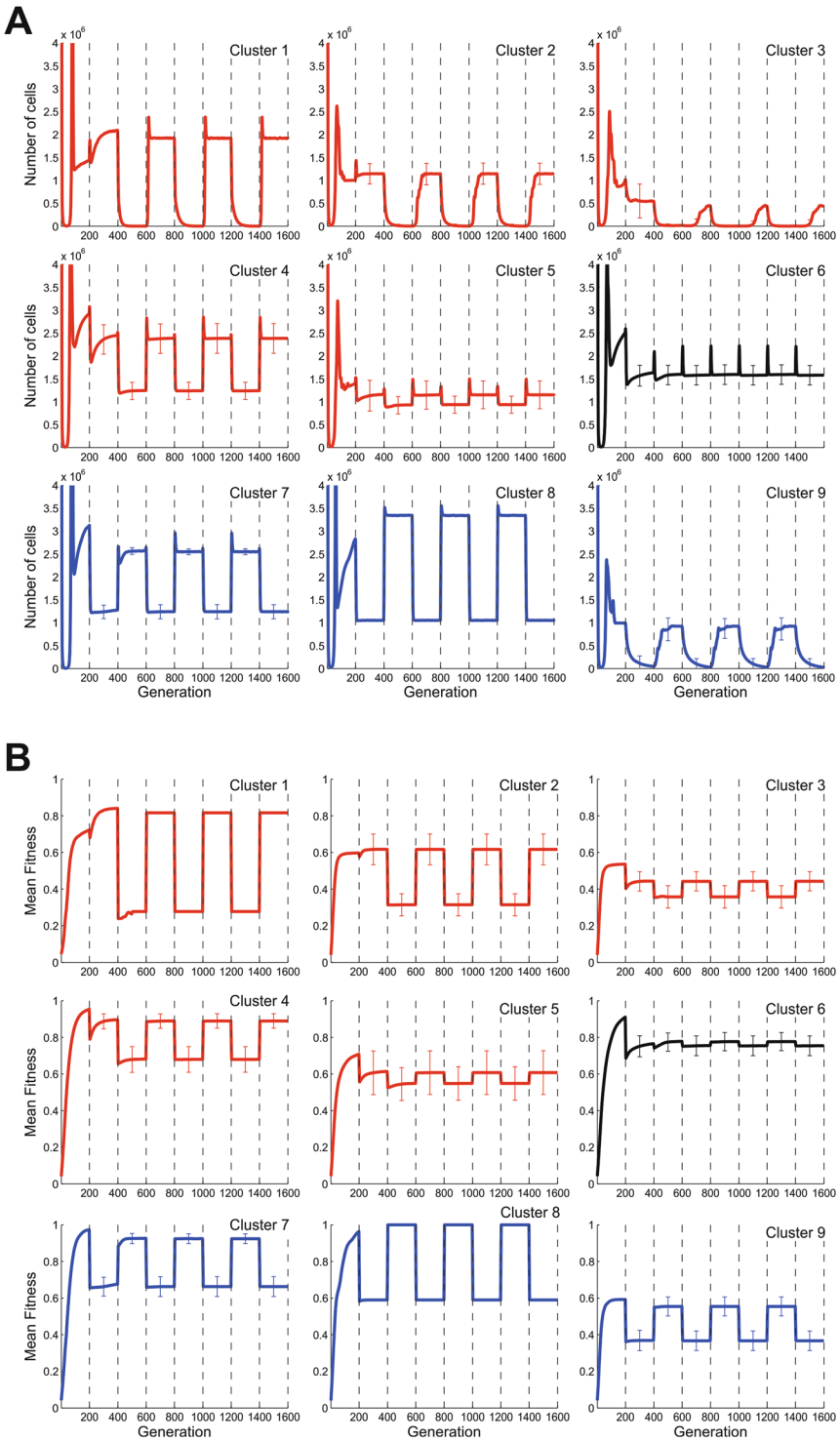

Figure 4. Population dynamic profile clusters. Population profiles for all simulates at tested recombination rates. Nine distinct profile clusters were identified and the mean number of cells (**A**) and the mean fitness (**B**) is shown for example simulations for profile cluster. Black broken lines indicate a host transition where selective pressures switch in favour of the other host. Error bars are given at the midpoint of each niche and correspond to one standard deviation. Ecological groups were inferred from the profiles, with specialist groups for Host 1 (red line) and Host 2 (blue line) as well as a generalist group (black line).

Cluster	Inferred ecological group	Number of simulations (%)				
		$r/m = 0$	$r/m = 0.1$	$r/m = 1$	$r/m = 10$	$r/m = 100$
1	Specialist Host 1	0	0	0	0	0.6
2	Specialist Host 1	41.1	0	0	0	1.2
3	Specialist Host 1	17.3	0	0	0	0.6
4	Specialist Host 1	0	48.2	91.8	86.4	56.7
5	Specialist Host 1	21.8	18.3	0	1	18.7
6	Generalist	0	33	7.7	11.5	19.9
7	Specialist Host 2	0	0	0.5	1	0.6
8	Specialist Host 2	0	0.5	0	0	0
9	Specialist Host 2	19.8	0	0	0	1.8

Table 1. Proportion of the representative clusters of population profiles at various recombination to mutation ratios. For each ratio (r/m), the percentage of simulations that followed particular representative patterns (clusters 1–9 from Fig. 3) from a total of 200 simulations were indicated. Ecological groups were inferred from Fig. 3.

such as host switches, it may be beneficial to have a lower recombination rate than would be optimal in the Fisher-Muller model.

In nature, each niche will have an associated carrying capacity. As a species expands to approach this carrying capacity, competition will inevitably increase, meaning that the influence of external factors will be of increased importance to the fate of the organism. To avoid this competition, an organism could adapt to a new less competitive niche, consistent with the 'tangled bank hypothesis'[32–34] for the evolution of sexual reproduction. In our competitive model, fitness variance is higher in populations with low recombination rates, providing a competitive advantage under the tangled bank hypothesis that would facilitate a transition to a new niche. This provides an explanation for the low recombination rates observed in some natural bacterial populations, and explains the presence of multiple lineages as infrequent recombination will allow the uptake of adaptive genes but may be too infrequent to prevent adaptive divergence between lineages[35].

Based on model simulations, the most favorable recombination to mutation ratio for promoting *Campylobacter* survival in the new niche whilst maintaining fitness variance within the population was $r/m = 0.1–1$ which is comparable to that calculated in natural *Campylobacter* populations ($r/m = 0.44$) (13). In multiple niche transition simulations, at intermediate recombination rates ($r/m = 1$), many populations did not completely die out, but resisted the introduction to a novel host recovering after a few passages, as seen by population size and mean fitness increases over time. This provides a context for considering the balance between rapid adaptation, mediated by recombination, and maintenance of genetic variance, allowing each population to survive in both host environments over time (Figs 2 and 3). Absolute host specialists, which went extinct in the second host, were uncommon, and most populations demonstrated some capacity to survive in both hosts.

In most cases, simulated genotypes were more successful in one or other niche. This is consistent with evidence from natural populations where lineages such as the ST-257 and ST-61 clonal complexes are predominantly associated with chicken and cattle respectively but are also isolated from both niches[10]. However, in some multi-host simulations, populations emerged that were affected very little by the host switches (Fig. 4). These populations can be considered true ecological generalists, comparable to the ST-21 and ST-45 clonal complexes that are regularly isolated from chickens, cattle and other hosts[10].

Ecological specialism and generalism have been well described in animals. For example, the Giant Panda, *Ailuropoda melanoleuca*, is a paradigm of specialism, confined to six isolated mountain ranges in south-central China, where bamboo comprises 99% of its diet[36], while the American Black Bear, *Ursus americanus*, is a generalist, opportunist omnivore with a broad range including temperate and boreal forests in northern Canada and subtropical areas of Mexico[37]. Specialization is a potentially precarious strategy as change to the environment can cause extinction if organisms are unable to move between niches or hosts. Consistent with this, generalist lineages would be expected to be older, preceding specialists which cluster closer to the tips of the phylogenetic tree[38]. In *Campylobacter*, there was no correlation between tree tip length and number of hosts suggesting similar evolutionary timescales for specialist and generalist STs. This contrast with studies of metazoans may, in part, be explained by the ability of *Campylobacter* to rapidly acquire niche-specifying elements leading to rapid adaptation of multiple lineages.

In addition to genomic plasticity, the scale of environmental variation can act on the type of ecological strategy observed among the inhabitants[4,39]. For example, in a highly stratified environment, adaptation may occur early with traits becoming fixed, whereas in a graduated environment individuals may be more likely to show a reversible phenotype response[4]. Specific niches of different bacterial lineages may be less well defined than for some animal species, but isolation source data from the PubMLST database indicates that *Campylobacter* STs show a gradient of host generalism. This is influenced by the opportunity to colonize the new niche and the capacity to survive the transition. Consistent with the findings in this study, rapid host transitions provide an opportunity for the proliferation of organisms with varying levels of host specialization including generalists that are capable of proliferating in more than one host.

In this study, by combining data from natural bacterial populations and a selection driven computer modeling, we simulated and predicted the evolution of host association strategies in the zoonotic pathogen *Campylobacter*. In practice, bacteria in natural populations usually exist in complex changeable ecosystems with numerous

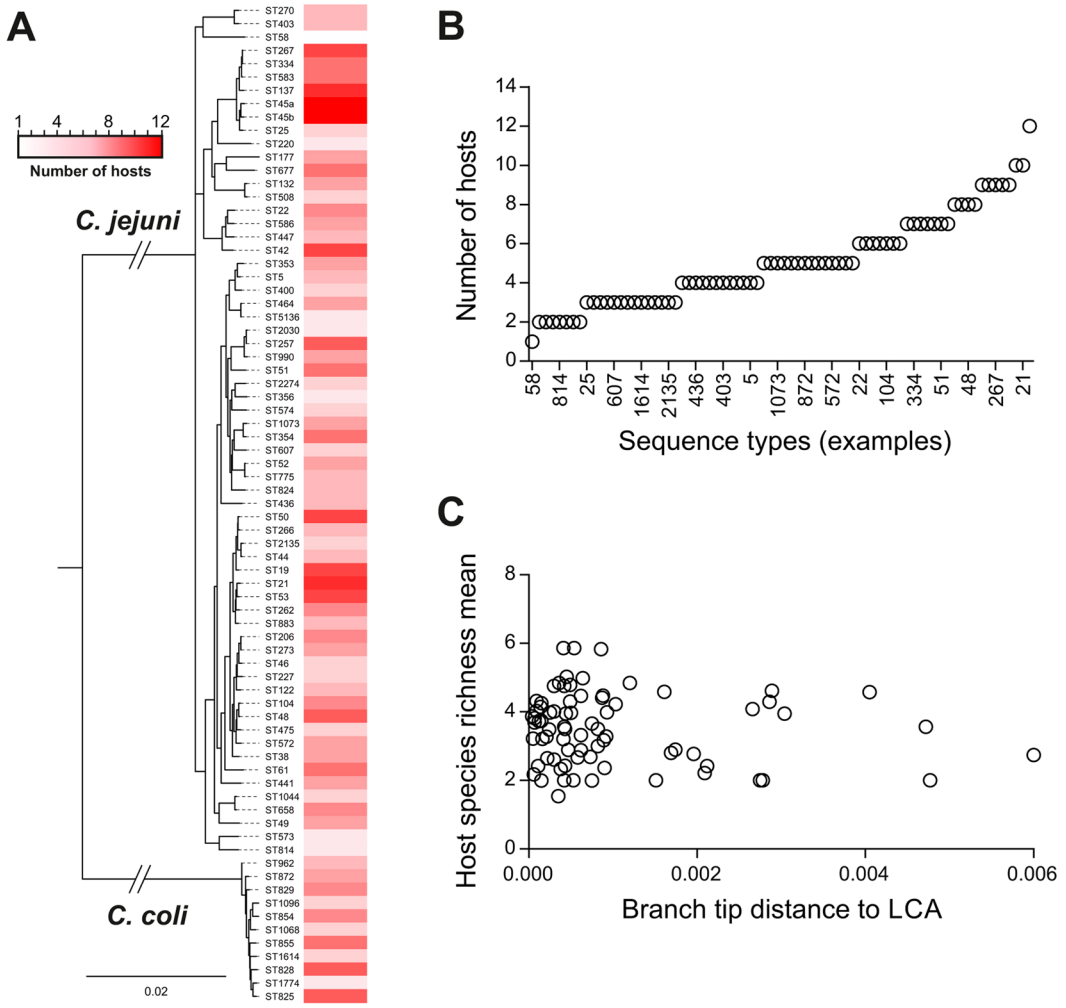

Figure 5. Phylogeny of generalist and specialist *Campylobacter* lineages. (**A**) Phylogenetic tree and isolation source of 74 common C. jejuni and C. coli sequence types (STs). The tree was constructed from a concatenated gene-by-gene alignment of 595 core genes, using ClonalFrameML with the standard model where recombination parameters are shared by all branches using the transition/transversion parameter estimated by PhyML. The heatmap represents the number of hosts in which particular STs were isolated, based on analysis of 30239 pubMLST isolate records. The scale bar represents the number of substitutions per site; (**B**) Quantification of the variation in number of hosts from which each ST has been sampled for all lineages shown in the tree, highlighting a gradient between host specialism and generalism; (**C**) Comparison of clustering on the tree, calculated as the estimated number of SNP corresponding to the average branch tip distance to the last common ancestor (LCA) with the number of hosts of isolation of each examined ST. There is no correlation between the two datasets (linear regression; $r^2 = 0.037$).

selection pressures. Here we show that recombination allows a more rapid response after a genetic bottleneck, as in a host transition, by increasing the efficiency with which selection can fix combinations of beneficial alleles. Furthermore, in a dynamic setting of host switching, recombination rate was observed to be a key factor in the colonization and maintenance in multiple niches. Livestock in modern intensive agricultural systems are different to ancestral host populations in numerous ways associated with diet and stocking density. The implications of this are potentially significant as, under conditions favoring rapid host switching, the emergence of host generalist zoonotic pathogens can be simulated. Our model therefore provides a context for considering how recombining bacteria, such as *Campylobacter*, could evolve to meet the challenges of anthropogenic environmental change. This could promote the emergence of multi-host pathogens and increase their capacity to overcome deliberate human interventions.

Methods

Three-phenotype model. We first consider a very simple abstract model of a bacterial ecosystem under finite resources. In this model, the bacteria adopt one of three phenotypes (or strategies), namely: Specialist in Host 1 (S_1), Specialist in Host 2 (S_2), and Generalist (G). Bacteria of each phenotype grow at an exponential rate proportional to their relative fitness f, which is dependent on their current host, and can switch between

strategies with an equal probability, ϕ. Death is proportional to the ratio of the total population at any given time $T(t) = S_1(t) + S_2(t) + G(t)$, and the carrying capacity k, and occurs at rate δ. The differential equations governing this system can be written as

$$\begin{aligned} dS_1(t) &= f_{S_1} S_1(t) + \phi(S_2(t) + G(t) - S_1(t)) - \delta S_1(t)T(t)/k dt, \\ dS_2(t) &= f_{S_2} S_2(t) + \phi(S_1(t) + G(t) - S_2(t)) - \delta S_2(t)T(t)/k, \\ dG(t) &= f_G G(t) + \phi(S_1(t) + S_2(t) - G(t)) - \delta G(t)T(t)/k. \end{aligned}$$

The fitness values were scaled by a growth rate of log(2) so that the time can be interpreted as generational time (i.e. in an unconstrained scenario, the population would double for each integer value of t). During the simulation, the entire population was subjected to a host transition at periodic intervals of τ generations.

In the simulations, bacteria start in Host 1 with initial values of $S_1 = S_2 = 1{,}000{,}000$, which was half of the carrying capacity, $k = 2{,}000{,}000$. The initial number of generalists was set at zero. The absolute fitness values for S_1 in Host 1 and S_2 in Host 2 was set at 1, and for S_2 in Host 1 and S_1 in Host 2 was set at 0.95. The absolute fitness for generalists in both hosts was set at 0.98. Relative fitness for each phenotype was calculated by dividing the absolute fitness by the population mean fitness. We also set $\phi = 0.0001$ and $\delta = 1$.

Extending the model to a more realistic scenario. A three phenotype switching model is a greatly simplified scenario. Bacteria in natural populations are influenced by complex interactions between the genotypes, expressed genes and their environment. It is not currently possible to incorporate every potential interaction in a bacterial ecosystem into a computational model, but it is possible to extend our simple three-phenotype model to represent a more realistic scenario for real bacterial populations.

In the full model we incorporate the cumulative effect of several genes, which together determine the fitness of each bacterium in a given host. This was informed using real population genomic data comprising of genomes found in both chicken and cow hosts. Based upon differential allele frequency, 10 niche segregating genes were chosen and augmented with 5 MLST genes where alleles do not segregate by host. This simulates the expected distribution of genetic variation in niche specifying and neutral genes. It is important to note that we are not assigning any particular functional role in host adaptation the host segregating genes, but are using them as a proxy for niche-specifying genes based upon the allele distribution. The aim was to create a simplified conceptual model of how generalists may co-exist with specialists, given a similar population structure to that observed in nature. Another significant change to the three-phenotype model is that individual genes can be exchanged between bacteria corresponding to homologous recombination in natural populations. This allows alleles from one bacterium to replace alleles at the same locus in the genome of another. This means that adapting between specialists is a more protracted process, as each host-specialist gene will need to be acquired before the fittest genotype is achieved. As such, the rate of recombination needs to be considered as this will likely affect the outcome, much like the switching rate in the three-phenotype model.

Bacterial genomes. For simulations using data from natural *Campylobacter* populations, 130 *C. jejuni* and *C. coli* isolates were used, including 87 strains sampled from chicken and cattle (Table S1). This provides the best available basis for obtaining a representative sample collection for generalizable findings but it should be noted that the results in this study are based on data where isolates from clinical and agricultural host sources are over represented. The genomes were previously published and isolates were described as belonging to clonal complexes on the basis of sharing 4 or more identical alleles at seven multilocus sequence typing loci[15,40]. In total for this study, 30 genomes from the ST-21 clonal complex, 28 from the ST-45 clonal complex, 7 from the ST-353 clonal complex, 6 from the ST-206 clonal complex, 6 from the ST-61 clonal complex, and 5 from the ST-48 clonal complex were used. ST-353 complex isolates are known to be chicken-associated and ST-61 complex isolates to be cattle-associated, whereas ST-21, ST-45, ST-206 and ST-48 are host generalists[10]. Genomes were archived on a web-based platform system based on BIGSdb, which also implements analytic and sequence exporting tools[24]. An additional 75 genomes representing 74 different STs from *C. jejuni* and *C. coli* were used. These genomes were sequenced as part of other studies (Food Standards Scotland i-CaMPS-3 Contract S14054, DEFRA project OZ0625), and from the PubMLST Database[24].

Model input data. Using a gene-by-gene approach[41,42], loci in the 130 genomes were identified by BLAST comparison to the *C. jejuni* strain NCTC11168 reference genome (Genbank accession number: AL111168) with a $>70\%$ nucleotide sequence identity on $\geq 50\%$ of sequence considered sufficient to call a locus match, as in other studies[43–45]. A whole-genome MLST[41,42] matrix was produced summarizing the presence/absence and allelic diversity at each locus in each genome, based upon these BLAST parameters. From this matrix, 1080 core genes were found to be shared by all cattle and chicken isolates from our dataset. The proportion of each allele at each locus was calculated in both cattle and chicken and then subtracted to identify alleles that were common in one group, but rare in the other. These were then summed these values at each locus to get the discriminative capacity of each locus for each host species. Loci in which the alleles segregated by host were considered as a proxy for niche-specifying genes in the model. While this was done in preference to simulating data, no inference is made based on the function or potential adaptive advantage conferred. A total of 5 putative niche-specifying genes per host (chicken and cattle) and 5 MLST genes picked at random (15 genes in total) were used as the input genotype for the model (Table S2). The inclusion of MLST loci (*aspA*, *uncA*, *pgm*, *glnA*, *gltA*) provides a reference for comparison.

The Genome Evolution by Recombination and Mutation (GERM) model. The GERM model is a simplified representation of a bacterial population and associated processes, which allows us to simulate bacterial

evolution *in silico* by tracking individual bacteria of variable genotypes as they are exposed to various selective environmental pressures. Furthermore, by simulating with a stochastic sampling algorithm, we can also incorporate some degree of the randomness inherent in natural populations, and hence investigate the importance of stochastic effects by performing simulations with the same initial population. Similar models have been proposed previously[46], and our approach extends and builds on these models in a number of ways. In the model in this study, each individual bacterial cell is represented as a 15-locus genotype as described above. In a population of N cells, each cell C is described entirely by the alleles a at locus j which compose the genome, denoted a_j, $j = 1, 2, \ldots, 15$. Each allele at each locus is represented by an integer ranging from 0 to ∞. Basing our algorithm on real data, where there are approximately 20 possible alleles at each locus ($20^{15} \approx 3.27 \times 10^{19}$ different combinations) presents computational challenges if we model the population using proportions of genotypes as in previous models[46]. To account for this we store the entire population at any one time and perform operations at the individual level. Working with the population directly, instead of adjusting proportions of STs, allows the investigation of the population dynamics at different population sizes, which is particularly pertinent after selective sweeps when the number in the population itself will drop as the population adapts to the new environment. The model incorporates six basic processes: mutation, homologous recombination, resource consumption, cell death, cell division and host migration. Each process occurs once per generation and is stochastic, therefore occurring with a probability defined for each cell. These can be interpreted as rates per generation.

Cell division, mutation and recombination. Mutation and homologous recombination occur at the level of the individual locus and cell death and cell division occur at the individual bacterial cell level. With cell division, an identical copy of the cell that divides is added to the population, this occurs with probability b. Mutation occurs with probability m and, unlike existing models[46] any allele that mutates is deemed to offer no selective advantage (the fitness of that allele is 0). Similarly, a recombination event occurs at probability r and an allele that recombines is assigned the value of another allele randomly chosen from those at the same locus within the current population. In natural systems, these processes are typically considered rare with upper rate estimates for homologous recombination of 10^{-6} per gene per generation[35]. However, these rate estimates can be affected by a number of factors[47,48] and so typically in studies of bacterial evolution, the preferred measure of the magnitude of recombination is the relative frequency of recombination compared to mutation, the r/m ratio[49–53]. To quantify the effects of varying levels of homologous recombination on niche adaptation in the GERM model, we used r/m ratio as the ratio of rates at which alleles are substituted as a result of recombination and mutation. Partly for reasons of computational tractability, the GERM model simulates a simplification of the size and complexity of a natural system, and imposes enhanced selection against maladapted sequences types. Because of this, r and m rate estimates from natural populations are adjusted so that sequence types that arise in the population have a similar opportunity on average to proliferate in both the model and the natural environment and are not removed from the population by chance alone. Consistent with existing estimates from multiple bacterial species including *Campylobacter*[48], we run simulations at r/m ratios ranging from 0 to 100 corresponding to a mutation rate of 0.01, with recombination rates of 0, 0.001, 0.01, 0.1 and 1 to facilitate comparisons with natural populations.

Fitness. In this study, host-specific alleles at niche specifying genes are considered to confer a fitness advantage to the cell in one or other host. The fitness of allele a, in a given host h is defined as $f^{\{h\}}(a)$ and this reflects the fitness conferred by that allele to its environment, with 1 corresponding to a perfectly adapted allele conferring maximal fitness, ranging to an allele that provides no benefit to the survival of the cell and has 0 fitness. We then follow Levin[46] and calculate the fitness of an individual cell as the sum of the allelic fitness values assuming that each allele contributes equally. The fitness function is therefore:

$F(C) = \frac{1}{n}\sum_{j=1}^{15} f(a_j)$. As a consequence, a different fitness landscape, defined by another host, affects the fitness of a cell through the fitnesses of its constituent alleles. It is by changing the fitness landscapes that host transition is simulated. Linkage disequilibrium is not factored into the model, although genes were selected from divergent genomic positions to limit linkage effects, and the model was not designed to test the specific function of individual genes. In nature, complex interactions between genes and the environment are likely to correspond to fitness function involving operations somewhere between additive and multiplicative[54] but less information on this is available for bacteria than in the more well studied diploid scenario.

Resource. The fate of a bacterium in the GERM model is not only dependent on fitness, but also on the availability of resources. This introduces a dynamic relationship between the fate of cells, their fitness, and the population size, and confers a soft carrying capacity for a niche dependence on the interplay between these aspects. Resource is modeled as a generic entity for which no distinction is made for the type of resource in a given niche, and the affinity for the consumption of a resource by a bacterium is independent of genotype. For each cell, the chance of using a resource occurs with probability u^+, multiplied by the amount of available resource. If a cell is already using a resource, it finishes with probability u^- in which case the resource is then consumed and is not returned to the environment. A resource is generated with constant probability g and is added to the pool of available resource. As such, for any given death in the population, there is a probability that it is caused either by a lack of fitness, or the inability to find and utilize resource. For a given fitness, the chance of death due to fitness or resources changes as the number of cells rises. For very small populations, the cause of death is predominantly fitness-related as resources are abundant. As the population increases, and the chance of finding a resource diminishes, resource-related death quickly starts to influence the fate of the cell but then plateaus as the population approaches carrying capacity. This nonlinear relationship is because, when unconstrained, the population will double in size during the course of a generation, where the resources will only increase by a fixed amount each time, regardless of population size. Conversely, as fitness increases, the proportion of deaths caused by fitness decreases to the point that resource related death becomes dominant. This is important as the mechanism of

death is different in each case: in fitness-related death, the probability of death is inversely proportional to the fitness. However, for resource-related death, the chance of death is independent of fitness and occurs entirely at random. As such, when resources become scarce, the fitness of the members of a population becomes largely less important to their fate. Therefore, resource availability is fundamental to the population dynamics, as well as suppressing the speed of convergence to the optimal genotype and maintaining genetic variance.

Cell death. Cell death occurs in two stages, fitness-related death and resource-related death. Fitness-related death is dependent on a fitness function, as discussed above and detailed further the supplementary material, which provides a probability of dying per generation for a given genotype. Resource-related death can happen only when a cell is not using a resource and occurs with probability d regardless of fitness. In both cases, the cell is removed from the population. The algorithm also incorporates a host transition or a selective sweep by switching between fitness functions. In this case, the resources can also be reset (although those already using a resource will continue as before). Further details of the model and stochastic simulation are included in the supplementary material.

Simulation. To account for the randomness inherent in the process of evolution, we simulated the system with a variant of a stochastic sampling algorithm (SSA)[55], involving Monte Carlo decision processes, was used for simulating the system. The number of cells required to represent a natural population is extremely large, therefore some constraints were imposed to ensure computational tractability. We sampled from Poisson distributions and binomial distributions when the quantities in question were unbounded and bounded respectively. For simplicity, we set the probability of a cell division event b to be equal to 1 so that every iteration can be interpreted as a generation. For mutation and recombination we imposed a constraint that each of these events can occur only once every time step. There were ten time steps per generation. The algorithm iterated through the five processes in order of resource allocation, mutation, recombination, cell death and cell division. We use three fitness functions, one each from chicken and cow, which we subsequently refer to as "Host 1" and "Host 2" respectively, and one derived using data from both hosts referred to as the composite fitness function. The composite fitness function is used to initialize the simulation and allow a burn-in period where the genetic composition to arrange themselves in a way that favors neither host. The algorithms start in using the composite fitness function to allow the algorithm to burn-in and deplete unfit allele combinations without purging alleles from one of the hosts. When the population has recovered and reached equilibrium with resource generation, we switch the fitness function to one of the single hosts. This occurred after 200 generations. We subsequently alternated the fitness function between chicken and cattle with the frequency defined by the user. This corresponds to the movement of the entire population to the new host, a simplification of the more realistic process in which bacteria would move to one of several distinct and concurrently evolving ecosystems. This layer of abstraction was partially chosen for computational tractability, but also to aid in the interpretation of results as the number of stochastic events would increase exponentially as the model complexity increases which would potentially dominate any patterns in our results. As such, this constraint is suitable for the scope of this study, particularly as our model is predicated on *Campylobacter* populations which are often transmitted *en masse* through stool. To initialize the algorithm, we generated a population of 50 million cells consisting of alleles found in the data set, randomly assigned throughout the population in a uniform manner rather than weighted by their abundance in the data set. The same initial population was used for every simulation run. This way, the populations and processes can be kept identical, and so that the only differences between runs with the same parameters were stochastic effects.

Two experiments were carried out at a range of homologous recombination to mutation ratios (0–100).

1. Long term adaptation followed by host switch – after the first composite burn in, a switch to host 2 was performed. This was run for 1000 generations to simulate long term colonization and adaptation and was then followed by a switch to host 1.
2. Rapid host switching – after the composite burn in – rapid host switching (every 200 generations) was performed.

Host generalism and phylogeny. We used 75 genomes from isolates representing 74 *C. jejuni* and *C. coli* STs, from 30239 pubMLST isolate records (http://pubmlst.org/campylobacter/), to investigate the degree of host generalism among different lineages. From these, concatenated gene-by-gene alignments of 595 core genes[15], constructed as previously published[43,44], were used to infer phylogenetic relationships. A core genome maximum likelihood tree was produced using ClonalFrameML with the standard model where recombination parameters are shared by all branches using the transition/transversion parameter estimated by PhyML, with the HKY85 substitution model (ratio of 8.517) and a gamma shaped parameter of 0.141. Point estimates after CFML analysis where: R/theta = 0.203573; 1/delta = 0.0062502; nu (mean divergence of imported DNA) = 0.0470359. Each ST was labelled with the number of distinct hosts from which it has been isolated, based on data submitted to the pubMLST database. The 74 STs were reported to have been isolated from 20 distinct animal hosts, including humans. The maximum number of hosts associated with a single ST was 12 (ST-45) and the minimum was 1 (ST-58). A tree and the heatmap representing the number of hosts, was prepared and visualized in Evolview[56]. The mean host species richness score was correlated with the genetic distance (derived from the number of SNPs) from the tip of the tree to the first branching point for each of the isolates. Rarefaction analysis incorporated the MLST data at the level of ST and structured by host species. To correct for variations in sample size a sample of 50 isolates was used for each ST. Resampling (n = 1000) the dataset was performed using the Monte Carlo Excel add-in @RISK version 6.2.1 (Palisade Ivybridge, United Kingdom) and the shuffle option in PopTools (www.poptools.org accessed: 16/11/2016) add-in to Microsoft Excel).

References

1. Van Tienderen, P. H. Evolution of Generalists and Specialist in Spatially Heterogeneous Environments. *Evolution* **45**, 1317–1331, doi:10.2307/2409882 (1991).
2. Woolhouse, M. E., Taylor, L. H. & Haydon, D. T. Population biology of multihost pathogens. *Science* **292**, 1109–1112 (2001).
3. Fried, G., Petit, S. & Reboud, X. A specialist-generalist classification of the arable flora and its response to changes in agricultural practices. *BMC ecology* **10**, 20, doi:10.1186/1472-6785-10-20 (2010).
4. Kassen, R. The experimental evolution of specialists, generalists, and the maintenance of diversity. *Journal of Evolutionary Biology* **15**, 173–190, doi:10.1046/j.1420-9101.2002.00377.x (2002).
5. Fitzgerald, J. R. Livestock-associated Staphylococcus aureus: origin, evolution and public health threat. *Trends Microbiol* **20**, 192–198, doi:10.1016/j.tim.2012.01.006 (2012).
6. Lowder, B. V. *et al*. Recent human-to-poultry host jump, adaptation, and pandemic spread of Staphylococcus aureus. *Proc Natl Acad Sci USA* **106**, 19545–19550, doi:10.1073/pnas.0909285106 (2009).
7. Meric, G., Kemsley, E. K., Falush, D., Saggers, E. J. & Lucchini, S. Phylogenetic distribution of traits associated with plant colonization in Escherichia coli. *Environ Microbiol* **15**, 487–501, doi:10.1111/j.1462-2920.2012.02852.x (2013).
8. Baumler, A. & Fang, F. C. Host specificity of bacterial pathogens. *Cold Spring Harbor perspectives in medicine* **3**, a010041, doi:10.1101/cshperspect.a010041 (2013).
9. Sheppard, S. K. *et al*. Niche segregation and genetic structure of Campylobacter jejuni populations from wild and agricultural host species. *Molecular ecology* **20**, 3484–3490, doi:10.1111/j.1365-294X.2011.05179.x (2011).
10. Sheppard, S. K. *et al*. Cryptic ecology among host generalist Campylobacter jejuni in domestic animals. *Mol Ecol* **23**, 2442–2451, doi:10.1111/mec.12742 (2014).
11. Dearlove, B. L. *et al*. Rapid host switching in generalist Campylobacter strains erodes the signal for tracing human infections. *ISME J*. doi:10.1038/ismej.2015.149 (2015).
12. Gripp, E. *et al*. Closely related Campylobacter jejuni strains from different sources reveal a generalist rather than a specialist lifestyle. *Bmc Genomics* **12**, 584, doi:10.1186/1471-2164-12-584 (2011).
13. Sheppard, S. K. *et al*. Campylobacter genotyping to determine the source of human infection. *Clin Infect Dis* **48**, 1072–1078, doi:10.1086/597402 (2009).
14. Wilson, D. J. *et al*. Rapid evolution and the importance of recombination to the gastroenteric pathogen Campylobacter jejuni. *Mol Biol Evol* **26**, 385–397, doi:10.1093/molbev/msn264 (2009).
15. Sheppard, S. K. *et al*. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. *PNAS* **110**, 11923–11927, doi:10.1073/pnas.1305559110 (2013).
16. Elena, S. F., Agudelo-Romero, P. & Lalic, J. The evolution of viruses in multi-host fitness landscapes. *The open virology journal* **3**, 1–6, doi:10.2174/1874357900903010001 (2009).
17. Geoghegan, J. L., Senior, A. M., Di Giallonardo, F. & Holmes, E. C. Virological factors that increase the transmissibility of emerging human viruses. *Proc Natl Acad Sci USA* **113**, 4170–4175, doi:10.1073/pnas.1521582113 (2016).
18. Doron, S. *et al*. Transcriptome dynamics of a broad host-range cyanophage and its hosts. *ISME J* **10**, 1437–1455, doi:10.1038/ismej.2015.210 (2016).
19. Kubinak, J. L., Cornwall, D. H., Hasenkrug, K. J., Adler, F. R. & Potts, W. K. Serial infection of diverse host (Mus) genotypes rapidly impedes pathogen fitness and virulence. *Proc Biol Sci* **282**, 20141568, doi:10.1098/rspb.2014.1568 (2015).
20. McGee, L. W. *et al*. Payoffs, not tradeoffs, in the adaptation of a virus to ostensibly conflicting selective pressures. *Plos Genet* **10**, e1004611, doi:10.1371/journal.pgen.1004611 (2014).
21. Nikolin, V. M. *et al*. Antagonistic pleiotropy and fitness trade-offs reveal specialist and generalist traits in strains of canine distemper virus. *PLoS One* **7**, e50955, doi:10.1371/journal.pone.0050955 (2012).
22. Fischer, R. The Genetical Theory of Natural Selection. (The Clarendon Press, 1930).
23. Kurihara, K., Welling, M. & Vlassis, N. Accelerated variational Dirichlet process mixtures. *Advances in Neural Information Processing Systems* **19**, 761–768 (2007).
24. Jolley, K. A. & Maiden, M. C. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC bioinformatics* **11**, 595, doi:10.1186/1471-2105-11-595 (2010).
25. Fisher, R. A. *The genetical theory of natural selection*. 1 edn, Vol. 1 (Clarendon Press, 1930).
26. Muller, H. J. Some genetic aspects of sex. *The American Naturalist* **66**, 118–138 (1932).
27. Gerrish, P. J. & Lenski, R. E. The fate of competing beneficial mutations in an asexual population. *Genetica* **102–103**, 127–144 (1998).
28. Weismann, A. *The Evolutionary Theory*. 1 edn, (Edward Arnold, 1904).
29. Felsenstein, J. The evolutionary advantage of recombination. *Genetics* **78**, 737–756 (1974).
30. Barton, N. H. & Charlesworth, B. Why Sex and Recombination? *Science* **281**, 1986–1990 (1998).
31. Spratt, B. G. *et al*. Recruitment of a penicillin-binding protien gene from *Neisseria flavescens* during the emergence of penicillin resistance in *Neisseria meningitidis*. *Proceedings of the National Academy of Sciences USA* **86**, 8988–8992 (1989).
32. Bell, G. The Masterpiece of Nature: the Evolution and Genetics of Sexuality. (Croom Helm, 1982).
33. Koella, J. C. The Tangled Bank: The maintenance of sexual reproduction through competitive interactions. *Journal of Evolutionary Biology* **1**, 95–116 (1988).
34. Doebeli, M. Quantitative Genetics and Papulation Dynamics. *Evolution* **50**, 532–546 (1996).
35. Wiedenbeck, J. & Cohan, F. M. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol Rev* **35**, 957–976, doi:10.1111/j.1574-6976.2011.00292.x (2011).
36. Lü, Z., Wang, D., Garshelis, D. L. & Group, I. S. B. S. *Ailuropoda melanoleuca*. *The IUCN Red List of Threatened Species*. http://dx.doi.org/10.2305/IUCN.UK.2008.RLTS.T712A13069561.en (2008).
37. Garshelis, D. L., Crider, D., van Manen, F. & Group, I. S. B. S. Ursus americanus. The IUCN Red List of Threatened Species. (2008).
38. Stireman, J. O. 3rd The evolution of generalization? Parasitoid flies and the perils of inferring host range evolution from phylogenies. *J Evol Biol* **18**, 325–336, doi:10.1111/j.1420-9101.2004.00850.x (2005).
39. Futuyma, D. J. & Moreno, G. The Evolution of Ecological Specialization. *Annual Review of Ecology and Systematics* **19**, 207–233 (1988).
40. Sheppard, S. K. *et al*. Progressive genome-wide introgression in agricultural Campylobacter coli. *Mol Ecol* **22**, 1051–1064, doi:10.1111/mec.12162 (2013).
41. Maiden, M. C. *et al*. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat Rev Microbiol* **11**, 728–736, doi:10.1038/nrmicro3093 (2013).
42. Sheppard, S. K., Jolley, K. A. & Maiden, M. C. J. A Gene-By-Gene Approach to Bacterial Population Genomics: Whole Genome MLST of Campylobacter. *Genes* **3**, 261–277 (2012).
43. Meric, G. *et al*. Ecological Overlap and Horizontal Gene Transfer in Staphylococcus aureus and Staphylococcus epidermidis. *Genome Biol Evol* **7**, 1313–1328, doi:10.1093/gbe/evv066 (2015).
44. Meric, G. *et al*. A reference pan-genome approach to comparative bacterial genomics: identification of novel epidemiological markers in pathogenic Campylobacter. *PLoS One* **9**, e92798, doi:10.1371/journal.pone.0092798 (2014).
45. Pascoe, B. *et al*. Enhanced biofilm formation and multi-host transmission evolve from divergent genetic backgrounds in Campylobacter jejuni. *Environ Microbiol*. doi:10.1111/1462-2920.13051 (2015).

46. Levin, B. R. & Cornejo, O. E. The population and evolutionary dynamics of homologous gene recombination in bacterial populations. *Plos Genet* **5**, e1000601, doi:10.1371/journal.pgen.1000601 (2009).
47. Barrick, J. E. & Lenski, R. E. Genome dynamics during experimental evolution. *Nat Rev Genet* **14**, 827–839, doi:10.1038/nrg3564 (2013).
48. Vos, M. & Didelot, X. A comparison of homologous recombination rates in bacteria and archaea. *ISME J* **3**, 199–208, doi:10.1038/ismej.2008.93 (2009).
49. Milkman, R. & Bridges, M. M. Molecular evolution of the Escherichia coli chromosome. III. *Clonal frames. Genetics* **126**, 505–517 (1990).
50. Falush, D. *et al*. Recombination and mutation during long-term gastric colonization by Helicobacter pylori: estimates of clock rates, recombination size, and minimal age. *Proc Natl Acad Sci USA* **98**, 15056–15061, doi:10.1073/pnas.251396098 (2001).
51. Fearnhead, P., Smith, N. G., Barrigas, M., Fox, A. & French, N. Analysis of recombination in Campylobacter jejuni from MLST population data. *J Mol Evol* **61**, 333–340, doi:10.1007/s00239-004-0316-0 (2005).
52. Feil, E. J. Small change: keeping pace with microevolution. *Nat Rev Microbiol* **2**, 483–495, doi:10.1038/nrmicro904 (2004).
53. Fraser, C., Hanage, W. P. & Spratt, B. G. Neutral microepidemic evolution of bacterial pathogens. *Proc Natl Acad Sci USA* **102**, 1968–1973, doi:10.1073/pnas.0406993102 (2005).
54. Phillips, P. C. Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* **9**, 855–867, doi:10.1038/nrg2452 (2008).
55. Gillespie, D. T. Exact Stochastic Simulation of Coupled Chemical Reactions. *The Journal of Physical Chemistry* **81**, 2340–2361 (1977).
56. Zhang, H., Gao, S., Lercher, M. J., Hu, S. & Chen, W. H. EvolView, an online tool for visualizing, annotating and managing phylogenetic trees. *Nucleic Acids Res* **40**, W569–572, doi:10.1093/nar/gks576 (2012).

Acknowledgements

This work was supported by the Biotechnology and Biological Sciences Research Council (BBSRC) grant BB/I02464X/1, the Medical Research Council (MRC) grants MR/M501608/1 and MR/L015080/1, and the Wellcome Trust grant 088786/C/09/Z. GM was supported by a NISCHR Health Research Fellowship (HF-14–13).

Author Contributions

D.J.W. and S.K.S. designed the study. D.J.W., P.K. and N.J.C.S. performed experiments. D.J.W., P.K., N.J.C.S., K.J.F., F.M.C., G.M. and S.K.S. analysed and interpreted results. D.J.W., G.M. and S.K.S. wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-09483-9

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.