Cross-species evolutionary rewiring in the enteric bacterium *Campylobacter*

Aidan J. Taylor[1]§, Guillaume Méric[2]†§, Koji Yahara[3], Ben Pascoe[2], Leonardos Mageiros[2,4], Evangelos Mourkas[2], Jessica K Calland[2], Santeri Puranen[5], Matthew D. Hitchings[6], Sion Bayliss[2], Keith A. Jolley[7], Carolin M. Kobras[1], Nicola J. Williams[8], Arnoud H. M. van Vliet[9], Julian Parkhill[10], Martin C. J. Maiden[7], Jukka Corander[5,11,12], Daniel Falush[13], Xavier Didelot[14], David J. Kelly[1*], Samuel K. Sheppard[2,7*]

[1]Department of Molecular Biology and Biotechnology, The University of Sheffield, Sheffield S10 2TN, United Kingdom;
[2]The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK;
[3]Antimicrobial Resistance Research Centre, National Institute of Infectious Diseases, Tokyo, Japan;
[4]Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Victoria 3004, Australia;
[5]Department of Mathematics and Statistics, Helsinki Institute for Information Technology, University of Helsinki, Helsinki, Finland;
[6]Swansea University Medical School, Institute of Life Science, Swansea SA2 8PP, UK;
[7]Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK;
[8]Department of Epidemiology and Population Health, Institute of Infection & Global Health, University of Liverpool, Leahurst Campus, Wirral, UK;
[9]School of Veterinary Medicine, University of Surrey, Guildford GU2 7AL, UK;
[10]Department of Veterinary Medicine, University of Cambridge, Madingley Rd, Cambridge CB3 0ES, UK;
[11]Department of Biostatistics, University of Oslo, Oslo, Norway;
[12]Parasites and Microbes, Wellcome Sanger Institute, Cambridge, UK;
[13]The Centre for Microbes, Development and Health, Institut Pasteur of Shanghai, Shanghai, China;
[14]School of Life Sciences and Department of Statistics, University of Warwick, Coventry CV4 7AL, UK.

*Corresponding author(s): David J. Kelly, d.kelly@sheffield.ac.uk; Samuel K. Sheppard, s.k.sheppard@bath.ac.uk;
†Present address: Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute, 75 Commercial Rd, Melbourne 3004, Victoria, Australia.

§ Contributed equally.

Keywords: comparative genomics, introgression, admixture, epistasis, evolution, *Campylobacter*, microbiology

Running title: Epistasis and recombination in *Campylobacter*

Abstract

The lateral transfer of genes among bacterial strains and species has opposing effects, conferring potentially beneficial adaptations whilst introducing disharmony in coadapted genomes. The prevailing outcome will depend upon the fitness cost of disrupting established epistatic interactions between genes. It is challenging to understand this in nature because it requires population-scale analysis of recombination and genomic coadaptation, and laboratory confirmation of the functional significance of genotype variation. By assigning the ancestry of DNA in the genomes of two species of the world's most common enteric bacterial pathogen, we show that up to 28% of the *Campylobacter coli* genome has been recently introgressed from *Campylobacter jejuni*. Then, by quantifying covariation across the genome we show that >83% of putative epistatic links are between introgressed *C. jejuni* genes in divergent genomic positions (>20kb apart), consistent with independent acquisition. Much of this covariation is between 16 genes, with just 5 genes accounting for 99% of epistatic SNP pairs. Laboratory mutagenesis and complementation cloning assays demonstrated functional links between these genes, specifically related to formate dehydrogenase (FDH) activity. These findings suggest that the genetic confederations that define genomic species may be transient. Even for complex traits such as central metabolism in the bacterial cell, conditions can arise where epistatic genes can be decoupled, transferred and reinstated in a new genetic background.

Introduction

Complex biological life is made possible by the interaction of genes. In bacteria, a vast array of genes, found together in almost infinite combinations, has allowed colonization of virtually every conceivable habitat on earth. While mutation remains the engine of genetic novelty, for most bacteria adaptation also involves the acquisition of genes from other strains and species through horizontal gene transfer (HGT)[1] potentially conferring new phenotypes to future generations. In some well documented cases, a single nucleotide substitution or acquisition of a small number of genes, can prompt new evolutionary trajectories with striking outcomes such as the emergence of virulent or antibiotic resistant strains[2]. With such dynamic genomic architecture, it may be tempting (and possibly useful[3]) to consider genes as independent units that 'plug and play' innovation into recipient genomes. This is clearly an oversimplification. In fact, genomes are highly interactive wherein the effect of one gene depends on another (epistasis). Therefore, it is likely that some introduced changes will disrupt gene networks and be costly to the original coadapted genetic background, particularly for complex phenotype innovation involving multiple genes.

Understanding how epistasis influences the evolution of phenotype diversity has preoccupied researchers since the origin of population genetics[4-10], with much emphasis placed upon the relative amounts of recombination and epistatic effect sizes[11,12]. In sexual populations, such as outbreeding metazoans, genetic variation is shuffled at each generation so genes can only rise in frequency if they have high mean fitness across genetic backgrounds. This means that it is unlikely that multiple distinct epistatic allele combinations will be maintained in the same population and barriers to gene flow, such as geographic isolation, may be required for marked phenotypic diversification[8]. In bacteria, however, rapid clonal reproduction allows multiple independent beneficial allele pairs to rise to high frequency in a single population. For example, in common enteric bacteria such as *Escherichia coli, Salmonella enterica* and *Campylobacter jejuni* the doubling time in the wild has been estimated at around 24 hours or less[13,14]. Therefore, though HGT occurs in these organisms[15], even in highly recombinogenic *C. jejuni*[13,16], there will likely be many millions of bacterial generations between recombination events at a given locus. This allows mutations that are beneficial only in specific genetic backgrounds to establish in a single population and linkage disequilibrium to form between different epistatic pairs[17].

In this coadapted genomic landscape, recombination presents something of a paradox. On the one hand, it promotes adaptation by conferring novel functionality on the recipient genome[18] and reduces competition between clones that carry different beneficial mutations (clonal interference), on the other it potentially introduces disharmonious genes or gene-combinations that will be discriminated against by selection[19]. Indeed, negative epistatic interactions between genes with different evolutionary histories have been proposed as a barrier to recombination[4-10,20], particularly between species. However, interspecies recombination is common in bacteria[18,21]. One of the most conspicuous examples of this can be seen in the common animal gut bacterium *Campylobacter*, which is among the most prolific causes of human bacterial gastroenteritis worldwide[22]. Specifically, introgression between the two most important pathogenic species, *C. jejuni* and *C. coli*, has led to the emergence of a globally distributed 'hybrid' *C. coli* lineage[23] that is responsible for almost all livestock and human infections with this species.

This evolutionary scenario provides an ideal context for addressing the paradox of recombination in coadapted genomes. First, because up to 23% of the core genome of

common *C. coli* has been recently introgressed from *C. jejuni*[24], potentially disrupting epistatic interactions. Second, because *C. jejuni* and *C. coli* have undergone an extended period of independent evolution (85% average nucleotide identity), therefore recombined sequence is conspicuous in the genome. Third, because the outcompetition of ancestral strains by introgressed *C. coli* is inconsistent with traditional assumptions of selection against hybrid lineages.

Here, we analyse the genomes of isolates representing known diversity within *C. jejuni* and *C. coli* and perform chromosome painting analysis to assign ancestry across the genomes of introgressed *C. coli*. We then perform a systematic screen for long-range covariation in the core genome to identify SNPs that are in strong linkage disequilibrium in independent genetic backgrounds and, therefore, are not explained by the clonal frame of the population. Finally, we investigate the function of covarying genes that account for most of the interactions and confirm epistasis among coadapted genes. By combining these analyses, we demonstrate how multiple interacting genes from one species can transfer to another, modifying the co-adapted genomic landscape of the recipient and promote a natural admixture-mediated genetic revolution[19,25].

Results

Campylobacter populations are highly structured with intermediate sequence clusters

Pan-genome analysis identified the presence and variation of every automatically annotated gene from every genome. This revealed a core genome of 631 gene orthologues in all *Campylobacter* isolates in this study. There were 1287 genes common to all *C. jejuni* isolates and 895, 1021 and 1272 common to *C. coli* clades 1, 2 and 3, respectively. Consistent with previous studies[24], neighbour-joining and ClonalFrameML trees based on genes within concatenated core genome alignments revealed population structure in which *C. jejuni* and *C. coli* clade 2 and 3 isolates each formed discrete clusters (Figure 1A, Supplementary Figure 1). Isolates designated as *C. coli* clade 1 were found in three clusters on the phylogeny: unintrogressed ancestral strains, and the ST-828 and ST-1150 clonal complexes which account for the great majority of strains found in agriculture and human disease[24]. Pan-genome analysis quantified the increase in unique gene discovery as the number of sampled genomes increased (Figure 1B). For all sequence clusters there was evidence of an open pan-genome with a trend towards continued rapid gene discovery within sequence clusters with fewer isolates. There was considerable accessory genome variation between species and clades, potentially associated with important adaptive traits (Supplementary Figure 2) and there was evidence that the average number of genes per genome was greater in *C. coli* ST-828 complex isolates than in *C. jejuni* (Figure 1C).

There is substantial introgression within the *C. coli* core genome

The large genetic distances among *C. coli* clade 1 isolates (Figure 1A), have been shown to be a consequence of the import of DNA from *C. jejuni* rather than accumulation of mutations during a prolonged period of separate evolution[23]. Using chromosome painting to infer the coancestry of core-genome haplotype data from CC-828 and CC-1150 isolates gave a detailed representation of the recombination-derived chunks from each *C. jejuni* donor group to each recipient individual (Supplementary Figure 3). The majority of introgressed SNPs were rare, occurring in fewer than 50 recipient genomes (Figure 2A). However, a large proportion of the introgressed *C. coli* clade 1 genomes contained DNA of *C. jejuni* ancestry in >98% of recipient isolates. Consistent with previous estimates[24], these regions where

introgressed DNA was largely fixed within the *C. coli* population, occurred across the genome and comprised up to 15% and 28% of the CC-828 and CC-1150 isolate genomes respectively (Figure 2A). When considering donor groups, the majority of introgressed DNA in *C. coli* involves genes that are present in multiple *C. jejuni* lineages (core genes) (Figure 2B).

Having identified *C. jejuni* ancestry within *C. coli* genomes, we investigated the sequence of events responsible for introgression. Most introgressed SNPs are found at low frequency in both clonal complexes. However, there was evidence of SNPs that are introgressed in both complexes as well as high frequency lineage specific introgression in both CC-828 and CC-1150 (Figure 2C). Specifically, 25% of the *C. jejuni* DNA found in >98% of CC-828 isolates was also found in CC-1150 (Figure 2D), implying that this genetic material was imported by the common ancestor(s) of both complexes. Subsequent to the divergence of these two complexes, introgression continued with nearly 75% of *C. jejuni* DNA present in one recipient clonal complex and not the other. This is consistent with an evolutionary history in which there was a period of progressive species and clade divergence reaching approximately 12% at the nucleotide level between *C. jejuni* and *C. coli* and around 4% between the three *C. coli* lineages. More recently, changes to the patterns of gene flow led one *C. coli* clade 1 lineage to import substantial quantities of *C. jejuni* DNA, and further lineage-specific introgression gave rise to two clonal complexes (CC-828 and CC-1150) that continued to accumulate *C. jejuni* DNA, independently creating the population structure observed today (Figure 2E).

The high magnitude introgression into *C. coli* clade 1 isolates has introduced thousands of nucleotide changes to the core genome. However, divergence in bacteria may be uneven across the genome. First, because recombination is more likely to occur in regions where donor and recipient genomes have high nucleotide similarity[21,26,27]. Second, because of 'fragmented speciation'[28], in which gene flow varies in different parts of the genome, such as regions responsible for adaptive divergence, leading to phylogenetic incongruence among genes. Consistent with previous estimates[24], we found that the three *C. coli* clades had similar high divergence with *C. jejuni* across the genome, ranging from 68% to 98% nucleotide identity for individual genes (Figure 2F), implying a period of divergence with low levels of gene flow. We found no evidence that high genetic differentiation between the species prevented recombination. While there was some evidence that more recombination occurred in regions of low nucleotide divergence (between unintrogressed *C. coli* clade 1 and *C. jejuni*), introgression occurred across the genome at sites with varying levels of nucleotide identity (Figure 2F). This level of recombination has greatly increased overall genetic diversity across the genome in *C. coli* clade 1 and introduced changes that have potential functional significance.

Much of the putative epistasis occurs between SNPs in introgressed genes

ClonalFrameML analysis revealed the importance of homologous recombination in generating sequence variation within the introgressed *C. coli*. Estimates of the relative frequency of recombination versus mutation ($R/\theta=0.43$), mean recombination event length ($\delta=152$bp) and average amount of polymorphism per site in recombined fragments ($\nu=0.07$), imply that recombination has had an effect (*r/m*) 4.57 times higher than *de novo* mutation during the diversification of CC-828. This is consistent with previous analysis and confirmed recombination as the major driver of molecular evolution in *C. coli*[13,24,29]. The continuous time Markov chain model for the joint evolution of pairs of biallelic sites on a phylogenetic tree (Supplementary Figure 4) was applied to investigate patterns of covariance for all pairs

of sites >20kb apart (Figure 3A). For most biallelic sites there were few branches on the tree where substitutions occurred, so that their evolution is compatible with separate evolution on the same clonal frame. However, 2874 covarying pairs evolved more frequently together than would be expected (*p*-value 10^{-8}) if they had evolved independently based on the tree, and hence indicated patterns of putative epistasis.

Among them, the location of 2618 putative epistatic pairs of sites was compared to the inferred ancestry (unintrogressed *C. coli* or *C. jejuni*) of sequence across the genome of CC-828 and CC-1150 *C. coli* strains (Figure 3B, Supplementary Data 1). For each epistatic pair, the major and minor haplotype were defined if there was haplotype polymorphism between *C. jejuni* and CC-828 and CC-1150 *C. coli.* This allowed quantification of the number of covarying sites that occurred between an ancestral *C. coli* (unintrogressed) and an introgressed *C. jejuni* allele, two introgressed alleles, and sites that do not segregate by species. Strikingly, the breakdown of the major and minor haplotype combinations among the 2618 epistatic pairs (Figure 3C, Source Data) shows the major haplotype for 83.5% of putative epistatic SNP pairs was *C. jejuni* indicating that both co-varying sites had *C. jejuni* ancestry, consistent with epistasis between introgressed ancestral *C. jejuni* sequence at divergent genomic positions. Investigation of the genes containing co-varying sites revealed that 2187 SNP pairs were in 16 genes with just five genes accounting for 99.1% of them (Figure 3D, Supplementary Data 2, Supplementary Figure 5).

Genomic context and physiological role of epistatically linked genes

The five genes accounting for the majority of epistatic interactions (*cj1167*, *cj1168c*, *cj1171c*/*ppi*, *cj1507c*/*modE* and *cj1508c*/*fdhD*) were investigated for their physiological role in *C. jejuni*. FdhD and ModE are proteins involved in the biogenesis of formate dehydrogenase (FDH). The FDH complex (FdhABC) oxidises formate to bicarbonate to generate electrons that fuel cellular respiration. Formate is an abundant electron donor produced by host microbiota and an important energy source for *Campylobacter in vivo*[30,31]. The remaining three genes, *cj1167* (annotated incorrectly as *ldh*, lactate dehydrogenase), *cj1168c* and *cj1171c* (*ppi*) are also grouped together on the genome, where *cj1167* and *cj1168c* are adjacent but with the open reading frames (ORFs) convergent and overlapping, while *ppi* is upstream, separated by two non-epistatically linked genes (*cj1169c* and *cj1170c*, Figure 4A). Considering the genomic arrangement, it is therefore clear that the putative epistatic links uncovered in this study essentially occur across two loci in the genome (*fdhD*/*modE* and *cj1167*/*cj1168c*/*ppi*), with each of the latter three genes linked with both *fdhD* and *modE* (Figure 4A). Given the known function of *fdhD* and *modE* in biogenesis of the FDH complex, we hypothesised that *cj1167*/*cj1168c*/*ppi* might also have some role in FDH biogenesis or activity in order to form a functional epistatic connection. We therefore constructed deletion mutants to investigate the possible role of these genes in FDH activity.

Initially, each of the mutants and their parental wildtype (*C. jejuni* NCTC11168) were grown in rich media (Muller-Hinton broth) and their formate dependent oxygen consumption rates determined (Figure 4B). *cj1167*, *cj1168c* and *ppi* mutants demonstrated wildtype levels of FDH activity, while activity in both *fdhD* and *modE* mutants was abolished. In order to confirm that the phenotype of the *fdhD* and *modE* mutants was not due to a polar effect on the surrounding *fdh* locus, these mutants were genetically complemented by reintroduction of a second copy of the wildtype gene into the rRNA locus, which restored near wildtype levels of FDH activity in both cases (Figure 4B).

As neither *cj1167*, *cj1168c* or *ppi* mutants showed altered FDH activity in cells grown in rich media, we considered that their function may be related to an FDH-specific nutrient requirement as would likely be found *in vivo*. Since the formate oxidising subunit of FDH, FdhA, specifically requires a molybdo- or tungsto-pterin (Mo/W) cofactor and a selenocysteine (SeC) residue for catalysis[32], Mo, W or Se supply presented possible targets. *cj1168c* encodes a DedA family integral membrane protein of unknown function. DedA proteins are solute transporters widespread in bacteria but are mostly uncharacterised[33]. However, a homologue of *cj1168c* in the heavy metal specialist beta-proteobacterium *Cupriavidus metallidurans* has been shown to be involved in selenite (SeO_3^{2-}) uptake[34]. We therefore speculated that Cj1168 could be a selenium oxyanion transporter that supplies Se for SeC biosynthesis. To test this, FDH activities were measured in *cj1168c* mutant and parental wildtype strains grown in minimal media with limiting concentrations of selenite or selenate (SeO_4^{2-}). The data in Figure 4C shows that the *cj1168c* mutant displayed significantly reduced FDH activity after growth with selenite in the low nM range, and this phenotype was partially restored by genetic complementation. We therefore designated *cj1168c* as *selF* (selenium transporter for formate dehydrogenase). However, although this phenotype does suggest that SelF is a selenium importer, another unrelated selenium transporter, FdhT (Cj1500), has previously been documented in *C. jejuni*[35], which is not epistatic with *fdhD* or *modE*. In contrast to this previous report we found considerable residual FDH activity still remained in an *fdhT* deletion mutant, which was fully restored to wildtype levels by complementation (Figure 4D).

Finally, we tested whether the residual FDH activity in our *fdhT* mutant was due to selenium uptake by SelF. An *fdhT selF* double mutant was generated and assayed for FDH activity after growth in minimal media containing limiting concentrations of selenite or selenate (Figure 4D). The *fdhT selF* double mutant demonstrated a significant additional reduction in FDH activity over the *fdhT* single mutant, a phenotype that was partially restored by complementing the double mutant with *selF*. Complementation of the double mutant with *fdhT* returned FDH activity to near wildtype levels (Figure 4D). Taken together, our data suggests that both FdhT and SelF facilitate selenium acquisition in *C. jejuni*, possibly representing low and high affinity transporters, respectively (Figure 4E).

Discussion

Hybridization between distantly related organisms can bring together new gene combinations and traits that potentially allow adaptation in a single evolutionary leap[36]. However, introducing disharmony among genes that have coevolved in epistasis can lead to reduced hybrid fitness, limiting the chance of this type of evolution by saltation[8,10]. Evidence for these contrasting paradigms comes largely from eukaryotes, but among prokaryotes, widespread HGT may seem to contradict assumptions about selection against hybrid lineages. For example, the transfer of mobile resistance genes between bacterial strains and species can confer a clear adaptive benefit, in a single step, allowing the recipient lineages to proliferate in the presence of antibiotics. Clearly, therefore, epistasis is not an absolute barrier to long-range recombination in bacteria.

There is ongoing debate about the extent to which the galaxy of accessory genes that are variously present or absent in many bacterial genomes constitute a cache of mobile genes from which innovation can be drawn and transferred between strains or species[37-40]. However, considering HGT of accessory genes does not fully address the extent to which recombination is constrained by epistatic fitness interactions between genes. First, because

genes or gene clusters introduced as autonomous elements, encoding specific novel traits, may cause minimal disruption to other essential cellular functions. Second, even when introduced genes result in a concomitant change elsewhere in the genome[41], the fitness cost may be outweighed by the benefit in a given niche. Third, and most importantly, because most recombination in bacteria is between homologous sequences[21], where a given gene is replaced by another version of the same gene.

Campylobacter is an ideal model organism for considering barriers to gene flow as interspecies recombination is well documented within the core genome. Consistent with previous studies[23,24,42], we found that a large proportion (15-28%) of the core genome of the two common *C. coli* lineages found in livestock and human disease (CC-828 and CC-1150) originated in *C. jejuni*. It is challenging to explain this level of genome-wide gene flow between species after a prolonged period of diversification (~15% nucleotide divergence) as it would likely disrupt coadapted gene networks. The evident success of hybrid *C. coli* in the agricultural niche suggests that the accumulation of *C. jejuni* DNA was not detrimental, but to specifically address if genome coadaptation was a barrier to recombination requires quantification of covariation among alleles.

Most bacteria, including *Campylobacter*, have highly genetically structured populations reflecting both neutral and adaptive evolutionary processes[18]. Therefore, a phylogenetically naive statistical model of coadaptation would afford equal significance to linkage disequilibrium resulting from selection and common ancestry. Even in recombinogenic bacteria such as *C. coli*, HGT is not sufficiently common to scramble the genome and abolish non-random SNP associations resulting from clonal population structure. Consistent with other models[43], the statistical test developed here accounts for the amount of covariation that would be expected based upon the clonal frame and identifies the same combinations of alleles in independent genetic backgrounds, thus providing evidence for coadaptation.

By combining the results from the chromosome painting and covariation models we were able to quantify the frequency of *C. coli* and *C. jejuni* SNPs in covarying allele pairs across the genome of introgressed *C. coli*. Recent interspecies admixture results in allele pairs, that correspond to *C. jejuni – C. coli* (and *vice versa*) and *C. jejuni – C. jejuni.* Comparing haplotype frequency provides the opportunity to contrast the disruptive and beneficial effects of homologous recombination in the bacterial genome. Under a neutral model, where recombination does not disrupt beneficial epistatic interactions, there would be more *C. jejuni – C. coli* than *C. jejuni – C. jejuni* covarying allele pairs in the recipient genome. Clearly, relatively low levels of covariation between ancestral *C. coli* and introgressed *C. jejuni* SNPs is expected, consistent with selection against disharmonious gene combinations in the coadapted recipient genome. However, the finding that *C. jejuni – C. jejuni* allele pairs constituted >83% of covarying introgressed haplotypes was striking. It is possible that in some cases both sites were introgressed in a single recombination event as bacteria can import very large pieces of DNA (>100kb)[44-46] but in *Campylobacter* LD for pairs of sites decreases with distance to approximately 20kbp and then remains at the same level for very distant sites[47]. Therefore, the divergent genome position of allele pairs (>20kb) implies that they were acquired independently. It follows, therefore, that the acquisition of the first introgression event was not fatal to the recipient genome, and was either mildly detrimental, neutral or beneficial. Acquisition of the second member of the pair then potentially enhanced the fitness restoring the integrated *C. jejuni – C. jejuni* coevolutionary unit.

Investigating covariation provides clues about genome plasticity and the potential benefit of introgression for niche adaptation. However, to directly address if genes are in epistasis it is necessary to test their function and confirm coadaptation. We demonstrated a functional link amongst the top scoring co-varying gene pairs with FDH, a key enzyme allowing the utilisation of formate as an electron donor *in vivo*[30,31]. In this study, FdhD and ModE were shown to be essential for FDH activity. While FdhD is a sulfur-transferase known to be required for the insertion of the pterin cofactor into FdhA[48] (Figure 4E), ModE is a transcriptional repressor that has been shown previously only to regulate the Mo/W uptake genes *mod* and *tup*[49,50]. However, the unexpected abolished FDH activity in a *modE* mutant indicates further functions for ModE in FDH biogenesis which warrant future investigation. Searching for functional links between *fdhD*/*modE* and *cj1167*/*selF*/*ppi*, revealed that a *selF* mutant strain had significantly reduced FDH activity under conditions of selenite limitation, a phenotype consistent with SelF being a Se oxyanion transporter and that functionally links SelF with FdhD/ModE. We suggest that *selF* rather than *fdhT* is epistatic because SelF confers an additional benefit for SeC biosynthesis (essential for FDH activity) under conditions of selenium limitation, for example as may be found in the host (Figure 4E). *cj1167* encodes a cytoplasmic NADPH dependent 2-oxoacid dehydrogenase but the current genome database annotation as lactate dehydrogenase (Ldh) is incorrect and its function is unknown[51]. There is no precedent for the involvement of such an enzyme in bacterial FDH or SeC biogenesis and we obtained no evidence for a functional connection between Cj1167 and FDH activity. However, the overlapping convergent gene arrangement of *cj1167* and *cj1168c* (*selF*) suggests a transcriptional architecture that might dictate these genes both form similar epistatic dependencies even if Cj1167 is not required for FDH activity. Finally, *cj1171c* (*ppi*) encodes a cytoplasmic peptidyl-prolyl *cis-trans* isomerase of the cyclophilin family. PPIases are general protein folding catalysts that often have pleiotropic and redundant functions[52] and we note that our *C. jejuni ppi* deletion mutant showed no growth defect as well as no reduction in FDH activity. It is possible that if Cj1171 does help promote the folding of e.g. FdhD or ModE, analysis of a simple deletion mutant may not reveal this if another PPIase can substitute in that genetic background.

Explaining the aberrant genome architecture among introgressed *C. coli* is challenging. Understanding the selective value of genes that promote proliferation depends on the overall genetic environment[19]. Our findings are consistent with an evolutionary scenario where an ancestral *C. coli* lineage underwent a niche transition and the surviving lineages (CC-828 and CC-1150) gained access to *C. jejuni* DNA (Figure 3E&F). As the adaptive landscape of the genome changed, potentially decoupling epistatic interactions that were previously selected, new gene combinations could be introduced by recombination and tested in the *C. coli* genetic background. Recent studies have shown that this type of genetic rewiring may be more common than previously thought[53-55] and when it leads to the proliferation of a hybrid organism it can be associated with the colonization of a new niche. Intensive livestock systems represent a possible explanation for the genetic revolution in *C. coli*. Host ecology can dramatically affect the evolution of gut-dwelling organisms, including *Campylobacter*[56] and colonization of this niche could promote the conditions necessary to promote introgressed *C. coli*. Whether host ecology is a factor or not, it is clear that conditions can arise where coadapted genes in a highly interactive bacterial genome can be transferred between species and reinstated as a single evolutionary unit in a new genome. This suggests that epistasis is not an absolute barrier to genome-wide recombination in structured bacterial populations.

Methods

Isolates, genome sequencing and assembly

A total of 973 isolates were used in this study, 827 from *C. coli* and a selection of 146 from a diversity of *C. jejuni* clonal complexes (Supplementary Data 3). Isolates were sampled mostly in the United Kingdom to maximise environmental and riparian reservoirs and thus the representation of genetic diversity in *C. coli*. Isolates were stored in a 20% (v/v) glycerol medium mix at -80°C and subcultured onto *Campylobacter* selective blood-free agar (mCCDA, CM0739, Oxoid). Plates were incubated at 42°C for 48 h under microaerobic conditions (5% CO2, 5% O2) generated using a CampyGen (CN0025, Oxoid) sachet in a sealed container. Subsequent phenotype assays were performed on Brucella agar (CM0271, Oxoid). Colonies were picked onto fresh plates and genomic DNA extraction was carried out using the QIAamp® DNA Mini Kit (QIAGEN; cat. number: 51306) according to the manufacturer's instructions. DNA was eluted in 100–200 µl of the supplied buffer and stored at −20°C. DNA was quantified using a Nanodrop spectrophotometer and high-throughput genome sequencing was performed on a MiSeq (Illumina, San Diego, CA, USA), using the Nextera XT Library Preparation Kit with standard protocols involving fragmentation of 2 µg genomic DNA by acoustic shearing to enrich for 600 bp fragments, A-tailing, adapter ligation and an overlap extension PCR using the Illumina 3 primer set to introduce specific tag sequences between the sequencing and flow cell binding sites of the Illumina adapter. DNA cleanup was carried out after each step to remove DNA < 150 bp using a 1:1 ratio of AMPure® paramagnetic beads (Beckman Coulter, Inc., USA). Short read paired-end data was assembled using the *de novo* assembly algorithm, SPAdes (version 3.10.0)[57]. All novel genome sequences (n=475) generated for use in this study are available on NCBI BioProjects PRJNA689604 and PRJEB11972. These were augmented with 498 previously published genomes and accession numbers for all genomes can be found in Supplementary Data 3[16,24,29,47,56,58-61].

Genome archiving, pan-genome content analyses and phylogenetic reconstruction

Contiguous genome sequence assemblies were individually archived on the web-based database platform BIGSDB[62] and sequence type (ST) and clonal complex (CC) designation were assigned based upon the *C. jejuni* and *C. coli* multi-locus sequence typing scheme[63]. To examine the full pan-genome content of the dataset, a reference pan-genome list was assembled as previously described[64]. Briefly, genome assemblies from all 973 genomes in this study were automatically annotated using the RAST/SEED platform[65], the BLAST algorithm was used to determine whether coding sequences from this list were allelic variants of one another or 'unique' genes, with two alleles of the same gene being defined as sharing >70% sequence identity on >10% of the sequence length. The prevalence of each gene in the collection of 973 genomes was determined using BLAST with a positive hit in a genome being defined as a local alignment of the reference sequence with the genomic sequence of >70% identity on >50% of the length, as previously described[66]. The resulting matrix was analysed for differentiating core and accessory genome variation. Genes present in all genomes were concatenated to produce a core-genome alignment, used for subsequent phylogenetic reconstructions. Phylogenetic trees were reconstructed using an approximation of maximum-likelihood phylogenetics in FastTree2[67]. This tree was used as an input for ClonalFrameML[68] to produce core genome phylogenies with branch lengths corrected for recombination.

Inference of introgression

All 973 genomes were aligned to a full reference sequence of *C. coli* strain CVM29710. We conducted imputation for polymorphic sites with missing frequency ≤10% using BEAGLE[69] as previously reported[70]. A total of 286,393 gapless SNPs (~17% of the average *C. coli* genome size) were used for recombination analyses. The coancestry of genome-wide haplotype data was inferred based on alignments using chromosome painting and FineStructure[71] as previously described[72]. Briefly, ChromoPainter was used to infer chunks of DNA donated from a list of 33 donor groups normalised for sample size to each of 677 ST-CC-828 and 12 CC-1150 recipient haplotypes. Results were summarised into a coancestry matrix containing the number of recombination-derived chunks from each donor to each recipient individual. FineStructure was then used for 100,000 iterations of both the burn-in and Markov chain Monte Carlo chain to cluster individuals based on the co-ancestry matrix. The results are visualized as a heat map with each cell indicating the proportion of DNA "chunks" (a series of SNPs with the same expected donor) a recipient receives from each donor.

Analysis of covariation in bacterial genomes

Non-random allele associations can result from selection and clonal population structure. To control for the latter, our approach identified SNP combinations in independent genetic backgrounds by accounting for the sequence variation associated with the inferred phylogeny. Based on the alignment of 677 genomes of *C. coli* CC-828, a first phylogenetic tree was created using PhyML[73]. ClonalFrameML[68] was then applied to correct the tree by accounting for the effect of recombination, and also to infer the ancestral sequence of each node. Covariance was assessed for pairs of biallelic sites across the genome using a Continuous Time Markov Chain (CTMC) model as follows. Briefly, let *A* and *a* denote the two alleles of the first site and *B* and *b* denote the two alleles of the second site, so that there are four states in total for the pair of sites (*ab*, *Ab*, *aB* and *AB*). The four substitution rates from *A* to *a*, from *a* to *A*, from *B* to *b* and from *b* to *B* are not assumed to be identical, to allow for differences in substitution rates in different parts of the genome and also to allow for non-equal rates of forward and backward substitution (for example as a result of recombination opportunities). Assuming no epistatic effect between the two sites ($\varepsilon=1$), the model M_0 has four free parameters (α_1, α_2, β_1 and β_2) representing independent substitutions at the two sites. We expand model M_0 with an additional fifth parameter $\varepsilon>1$ into model M_1 which is such that the state *AB* where the first site is allele *A* and the second site is allele *B* is favored relative to the other three sites *ab*, *aB* and *Ab*. Specifically, the state *AB* has a probability increased by a factor ε^2 in the stationary distribution of the CTMC of model M_1 compared to model M_0.

Both models M_0 (with 4 parameters) and M_1 (with 5 parameters) are fitted to the data using maximum likelihood techniques, where the likelihood is equal to the product for every branch of the tree of the state at the bottom of the branch given the state at the top. The two fitted models M_0 and M_1 are then compared using a likelihood-ratio test (LRT) as follows: since M_0 is nested with M_1, two times their difference in log-likelihood is expected to be distributed according to a chi-square distribution with number of degrees of freedom equal to the difference in their dimensionality, which is one. This LRT returns a *p*-value for the significance of a covariation effect, and a Bonferroni correction is applied to determine a conservative cutoff of significance that accounts for multiple testing. Furthermore, the test is applied only to pairs of sites separated by >20kb to reduce the chance that they were the result of a single recombination event, consistent with estimates of the length of recombined DNA sequence in quantitative bacterial transformation experiments[44,74] and evidence from *Campylobacter* genome analyses that show that LD for pairs of sites decreases with distance to approximately 20kbp and then remains at the same level for very distant sites[47]. It is still

possible of course that rare recombination events would stretch 20kbp[44-46], but for this to have an effect on the analysis of epistasis it would have to have happened several times for the same pairs of sites against different genomic background which becomes quite unlikely just by chance. This phylogenetically aware approach to testing for covariance presents the advantage to naturally account for both population structure and the effect of recombination[75]. The script implementing this coevolution test is available in R at: https://github.com/xavierdidelot/campy.

Quantifying covariation between recombined and unrecombined genomic regions

The results of the introgression and covariation analyses were combined so that for each pair of significantly covarying SNPs (*p*-value $<10^{-8}$), haplotype frequency was calculated among the 689 recipient introgressed *C. coli* clade-1 strains as well as among the donor *C. coli* (ancestral) and *C. jejuni* strains, respectively. If the most frequent haplotype of the pair is the same between the donor *C. coli* (ancestral) and *C. jejuni*, it was classified as 'no polymorphism'. Otherwise, if the most frequent haplotype accounted for >90% among the recipients, it was classified as either '*C. jejuni* (>90%)' or '*C. coli* (>90%)' if it was the same as that of donor *C. jejuni* or *C. coli* (ancestral) (inset in Figure 3C). If the most frequent haplotype accounted for ≤90% among the recipients, the top two most frequent haplotypes (written as major and minor haplotype in this manuscript) were indicated as either "*C. jejuni* / *C. coli*", "*C. jejuni* / other", "*C. coli* / *C. jejuni*", "*C. coli* / other", "other / *C. coli*", "other / *C. jejuni*", and "other / other", and the frequency of the major and minor haplotypes were calculated. For example, where the haplotype frequencies were as follows, AA=285, TA=192, TG=181, AG=27, A-=2, --=1, -A=1, AA is the major haplotype, frequency of which is 41.3%

Mutagenesis and complementation cloning

Genes *cj1167*, *cj1168c* (here designated *selF* for selenium transport for formate dehydrogenase), *cj1171c* (*ppi*), *cj1507c* (*modE*), *cj1508c* (*fdhD*) and *cj1500* (*fdhT*) were deleted by allelic exchange mutagenesis, with the majority of the open reading frame replaced by an antibiotic resistance cassette. Mutagenesis plasmids were generated by the isothermal assembly method using the HiFi system (NEB, UK). In brief, flanking regions of target genes were PCR amplified from genomic DNA using primers with adaptors homologous to either the backbone vector pGEM3ZF or the antibiotic resistance cassette (Supplementary Data 4). pGEM3ZF was linearised by digestion with HincII. The kanamycin and chloramphenicol resistance cassettes were PCR amplified from pJMK30 and pAV35, respectively[76]. Four fragments consisting of linearised pGEM3ZF, antibiotic resistance cassette and 2 flanking regions were combined in equimolar amounts and mixed with 2 x HiFi reagent (NEB, UK) and incubated at 50°C for 1 hour. The fragments combine such that the gene fragments flank the antibiotic resistance cassette, in the same transcriptional orientation, within the vector. Mutagenesis plasmids were transformed into *C. jejuni* NCTC 11168 by electroporation. Spontaneous double-crossover recombinants were selected for using the appropriate antibiotic and correct insertion into the target gene confirmed by PCR screening. For genetic complementation of mutants, genes *cj1168c* (*selF*), *cj1507c* (*modE*), *cj1508c* (*fdhD*) and *cj1500* (*fdhT*) were PCR amplified from genomic DNA, restriction digested with MfeI and XbaI, then ligated into similarly digested pRRA[77] (Supplementary Data 4). The orientation of insertion allowed the target gene to be expressed constitutively by a chloramphenicol resistance gene-derived promoter within the vector. Complementation plasmids were transformed into *C. jejuni* by electroporation. Spontaneous double-crossover recombinants were selected for using apramycin and correct insertion into the ribosomal locus confirmed by PCR screening.

Growth of *C. jejuni*

Microaerobic growth cabinets (Don Whitley, UK) were maintained at 42°C with an atmosphere of 10% O_2, 5% CO_2 and 85% N_2 (v/v). *C. jejuni* was grown on Columbia-base agar containing 5% v/v defibrinated horse blood. Selective antibiotics were added to plates as appropriate at the following concentrations: 50 µg ml^{-1} kanamycin, 20 µg ml^{-1} chloramphenicol, 60 µg ml^{-1} apramycin. Muller-Hinton (MH) broth supplemented with 20 mM L-serine was used as a rich medium. Minimal medium was prepared from a supplied MEM base (51200-38, Thermo Scientific, UK) with the following additions: 20 mM L-serine, 0.5 mM sodium pyruvate, 50 µM sodium metabisulfite, 4 mM L-cysteine. HCl, 2 mM L-methionine, 5 mM L-glutamine, 50 µM ferrous sulfate, 100 µM ascorbic acid, 1 µM vitamin B12, 5 µM sodium molybdate, 1 µM sodium tungstate. Selenium was then added as appropriate from stocks of sodium selenate or sodium selenite prepared in dH_2O. For assays, cells were washed and suspended in sterile phosphate-buffered saline (PBS, Sigma-Aldrich).

Respiration rates with formate

Cells were first grown in MH broth for 12 hours, then washed thoroughly in PBS before inoculating minimal media without an added selenium source. The appropriate concentrations were determined by serial dilution trials and it was subsequently found that *C. jejuni* has a strong preference for selenite over selenate, as equivalent FDH activity requires some 1000-fold greater concentration of selenate than selenite (Supplementary Figure 6). These cultures were grown for 8 hours before the cells were thoroughly washed again, then used to inoculate further minimal media, with a selenium source added as appropriate, and grown for 10 hours. This passaging was necessary to remove all traces of selenium from the inoculum, such that control cultures without selenium added had negligible (FDH) activity. Assay cultures were again thoroughly washed before the equivalent of 20 ml at an optical density of 0.8 at 600 nm was finally suspended in 1 ml of PBS. Formate-dependent oxygen consumption by whole cells was measured in a Clark-type electrode using 20 mM sodium formate as electron donor. The electrode was calibrated with air-saturated PBS assuming 220 nmol dissolved O_2 ml^{-1} at 37 °C. In the electrode, 200 µl of the dense cell suspension was added to 800 µl air-saturated PBS for a final volume of 1 ml. The chamber was sealed and the suspension allowed to equilibrate for 2 minutes. The assay was initiated by the addition of 20 µl of 1 M sodium formate (prepared in PBS) and the rate of oxygen consumption recorded for 90 s. The total protein concentration of the cell suspensions was determined by Lowry assay and the specific rate of formate-dependent oxygen consumption expressed as nmol oxygen consumed min^{-1} mg^{-1} total protein.

Data availability

Short-read sequence data for all isolates sequenced in this study are deposited in the sequence read archive (SRA) and can be found associated with NCBI BioProjects PRJNA689604 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA689604) and PRJEB11972 (https://www.ncbi.nlm.nih.gov/bioproject/PRJEB11972). These were augmented with 498 previously published genomes and assembled genomes are available on Figshare (doi.org/10.6084/m9.figshare.13521683). Accession numbers for all genomes are included in Supplementary Data 3. Source data are provided for this paper.

Acknowledgements

This work was supported by Wellcome Trust grants 088786/C/09/Z and Medical Research Council (MRC) grants MR/M501608/1 and MR/L015080/1 awarded to SKS. Bacterial sampling was funded by Food Standards Agency project FS101087. AJT was supported by a Biotechnology and Biological Sciences Research Council grant (BB/S014497/1) awarded to DJK. GM was supported by a NISCHR Health Research Fellowship (HF-14-13). JC was supported by the ERC grant no. 742158. Computational calculations were performed using computer servers from the MRC CLIMB project (UK), at National Institute of Genetics (Japan) and the Human Genome Center of the Institute of Medical Science (University of Tokyo, Japan) and at HPC Wales (UK).

Author Contributions

S.S., X.D. D.F. and K.Y. conceived and designed the study. S.S., N.W., A.V. and M.M. collected samples. G.M., A.T., L.M., M.H. and B.P. carried out Laboratory work. B.P., M.H., K.J., M.M., J.P. and S.S. supported data archiving. G.M., A.T., X.D., K.Y., L.M., S.P., S.S. and J.C. analysed the data. M.M., J.P., J.C., D.K. and D.F. contributed to data interpretation. S.S., B.P., G.M., C. K., A.T. and D.K. wrote the paper.

Competing Interests

The authors declare no competing interests.

Figure legends

Figure 1. Population genomics of *C. jejuni* and *C. coli.* (A) Phylogenetic tree reconstructed using neighbour-joining on a whole-genome alignment of 973 *C. jejuni* and *C. coli* isolates. Introgressed *C. coli* clades are represented with red (CC-828, n=677) and purple (CC-1150, n=12) circles, unintrogressed clade 1 (n=35) is shown in pink, clade 2 (n=45) in yellow and clade 3 (n=58) in green. A set of 146 *C. jejuni* genomes belonging to 30 clonal complexes (4 to 5 isolates per ST) are shown in blue. Recipient and donor populations, used to infer introgression in chromosome painting analysis, are indicated. The scale bar represents the number of substitutions per site. (B) Accumulation curves of the clade-specific pan-genome content of *Campylobacter* lineages, using the same colour code as panel A. Randomized genome sampling was used to obtain the average number of genes for each comparison (plain lines) and standard deviations (dotted lines). (C) The average number of genes/genome, identified by BLAST, is shown for *C. jejuni* and the different *C. coli* clades. Asterisks denote significant difference between distributions as inferred from a Dunn's multiple comparisons test after a Kruskal-Wallis test, as follows: ***: $p<0.001$; ****: $p<0.0001$.

Figure 2. Genome-wide introgression from *C. jejuni* to *C. coli*. (A) Summary of introgressed *C. jejuni* SNPs in *C. coli* CC-828 (n=677, red) and CC-1150 (n=12, purple) genomes using ChromoPainter; the number of introgressed core SNPs (coloured histograms; left y-axis) and core genes (white histograms; right y-axis) for a range of recipient strains proportions (at least 1, more than 50% and more than 98%) is shown. (B) The number of genes with different frequencies of maximum SNP introgression/gene in *C. coli* as a function of gene frequency in *C. jejuni*. Highly introgressed genes in CC-828 and CC-1150 tend to be core in *C. jejuni*. (C) Density plot (n=1000 bins) of specific and shared introgression events in CC-828 (x-axis) and CC-1150 (y-axis). The frequency of SNP introgression/gene is shown for both lineages. Close blue lines denote a high density of points. (D) Shared introgression between *C. coli* CC-828 and CC-1150. The number of SNPs being shared between the two lineages at various frequencies is shown in y-axis. (E) Diagram of *Campylobacter* species and clade (C1*,C2, C3) divergence with arrows indicating introgression from *C. jejuni* into *C. coli* (i) clade 1, (ii, iii) CC-828 and CC-1150, (iv, v) subsequent clonal expansion and ongoing introgression. (F) Pairwise nucleotide identity between *C. jejuni* and ancestral (unintrogressed) clade 1 *C. coli* core genes (black circles). Genes found to be introgressed in clade 1 CC-828 are highlighted in blue.

Figure 3. Covariation in introgressed *C. coli* genomes. (A) CC-828 and CC-1150 *C. coli* genomes were analysed using a continuous time Markov chain (CTMC) model and covariation was assessed for pairs of biallelic sites separated by at least 20kb along the genome while accounting for the effect of population structure and recombination. There are many biallelic sites that do not change often on the tree and few that do. Putative epistatic sites change more frequently than average with biallelic pairs found together on multiple branches. (B) Miami plot of each polymorphic site showing the maximum *p*-value for covarying biallelic pairs (>20kb apart) and the frequency of introgression in CC-828 and CC-1150. (C) The frequency of major and minor haplotype combinations (inset) among the 2578 pairs of covarying SNPs in the 689 *C. coli* clade-1 recipient genomes, revealing that the majority of long range covariation was between introgressed *C. jejuni* sites. (D) The position of putative epistatic sites mapped on the *C. coli* CVM29710 reference for covarying *C. jejuni*-*C. jejuni* SNPs (red) in 16 gene pairs (a to l), and other haplotype combinations (grey). (E and F) An evolutionary scenario for the observed patterns of covariation and introgression in natural *C. coli* populations: (i) *C. jejuni* (blue) and unintrogressed *C. coli* (red) co-exist

with genomes (internal circles) harbouring haplotype pairs (x-x) that segregate by species; (ii) Horizontal gene transfer, HGT, occurs (R1) disrupting covarying genetic elements and reducing the relative fitness of introgressed *C. coli* to varying degrees (grey arrow), few strains retain mixed *C. coli* - *C. jejuni* haplotypes; (iii) HGT continues (R2) and, where recombined mixed haplotypes survived, ancestral *C. jejuni* haplotype pairs are reinstated in introgressed *C. coli;* (iv,v) introgressed *C. coli* outcompete unintrogressed strains.

Figure 4. Genomic context and physiological roles of introgressed epistatically linked genes. (A) Genome organisation and percentage of co-varying SNP pairs (internal legend). (B-D) FDH activity of whole cells determined by oxygen consumption rates in a Clark-type electrode (nmol oxygen consumed per minute per mg of total protein) for (B) cells grown in rich media (excess selenium), (C) cells grown in minimal media with 0.5 nM sodium selenite, and (D) cells grown in minimal media with either 5 nM sodium selenite (left, open bars) or 5 µM sodium selenate (right, hashed bars). All data are means of at least 4 independent determinations and error bars are SD. *** denotes p value of <0.001 by students t-test. (E) Model for epistatically linked genes involving FDH biogenesis and activity. Host derived formate is converted to bicarbonate in the periplasm by the FDH complex to release electrons which are transferred from the iron sulfur (FeS) cluster of FdhA to the *b*-type hemes of FdhC and into the menaquinone (MK) pool where they can ultimately be used to reduce molecular oxygen via terminal oxidases. FdhA contains a selenocysteine residue and Mo/W-pterin cofactor (W/MoCo) at its active site, both of which are essential for catalysis. ModE is a DNA binding regulator which represses the Mo and W transporters Mod and Tup to regulate the cellular pool of Mo/W. W/MoCo is generated by the Moa pathway and inserted into apo-FdhA by the sulphur-transferase FdhD. Environmental selenite (the most abundant oxyanion) or selenate diffuses into the periplasm where it must be actively imported to the cytoplasm by FdhT and SelF. Putatively, FdhT is low affinity and functions efficiently when ample selenium is available. When selenium is limited, SelF can import sufficient selenium to maintain FDH production and activity. Cytoplasmic selenium is converted to selenophosphate by SelD, which is used by SelA to generate tRNA-SeC from tRNA-Ser. During translation of FdhA, tRNA-SeC is incorporated by the specific elongation factor SelB. Apo-FdhA with W/MoCo and SeC inserted is then transported to the periplasm by the TAT translocation system and incorporated into the FDH complex (See [31] for a review of FDH in *C. jejuni*).

References

1 Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299-304 (2000).

2 Koonin, E. V., Makarova, K. S. & Aravind, L. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.* **55**, 709-742 (2001).

3 Haldane, J. B. A defense of beanbag genetics. *Perspect Biol Med* **7**, 343-359 (1964).

4 Bateson, W. in *Darwin And Modern Science* 85–101 (Cambridge University Press, 1909).

5 Fisher, R. A. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb* **52**, 399–433 (1918).

6 Wright, S. Evolution in Mendelian populations. *Genetics* **16**, 97-159 (1931).

7 Dobzhansky, T. Studies on hybrid sterility. II. Localization of sterility factors in *Drosophila Pseudoobscura* hybrids. *Genetics* **21**, 113-135 (1936).

8 Dobzhansky, T. *Genetics And The Origin Of Species*. (Columbia University Press, 1937).

9 Muller, H. J. Bearing of the *Drosophila* work on systematics. *The new systematics*, 185-268 (1940).

10 Muller, H. Isolating mechanisms, evolution, and temperature. *Biol. Symp.* **6**, 71-125 (1942).

11 Kimura, M. Attainment of quasi linkage equilibrium when gene frequencies are changing by natural selection. *Genetics* **52**, 875-890 (1965).

12 Neher, R. A. & Shraiman, B. I. Competition between recombination and epistasis can cause a transition from allele to genotype selection. *Proc. Natl. Acad. Sci.* **106**, 6866-6871 (2009).

13 Wilson, D. J. *et al.* Rapid evolution and the importance of recombination to the gastroenteric pathogen *Campylobacter jejuni*. *Mol. Biol. Evol.* **26**, 385-397 (2009).

14 Gibson, B., Wilson, D. J., Feil, E. & Eyre-Walker, A. The distribution of bacterial doubling times in the wild. *Proc. R. Soc. B.* **285** (2018).

15 Vos, M. & Didelot, X. A comparison of homologous recombination rates in bacteria and archaea. *Isme J* **3**, 199-208 (2009).

16 Calland, J. K. *et al*. Quantifying bacterial evolution in the wild: a birthday problem for *Campylobacter* lineages. *Preprint at https://www.biorxiv.org/content/10.1101/2020.12.02.407999v1* (2020).

17 Arnold, B. J. *et al*. Weak epistasis may drive adaptation in recombining bacteria. *Genetics* **208**, 1247-1260 (2018).

18 Sheppard, S. K., Guttman, D. S. & Fitzgerald, J. R. Population genomics of bacterial host adaptation. *Nat. Rev. Genet.* **19**, 549-565 (2018).

19 Mayr, E. in *Evolution as a Process* 157–180 (Allen & Unwin, 1954).

20 Orr, H. A. Dobzhansky, Bateson, and the genetics of speciation. *Genetics* **144**, 1331-1335 (1996).

21 Fraser, C., Hanage, W. P. & Spratt, B. G. Recombination and the nature of bacterial speciation. *Science* **315**, 476-480 (2007).

22 Sheppard, S. K. *et al. Campylobacter* genotyping to determine the source of human infection. *Clin. Infect. Dis.* **48**, 1072-1078 (2009).

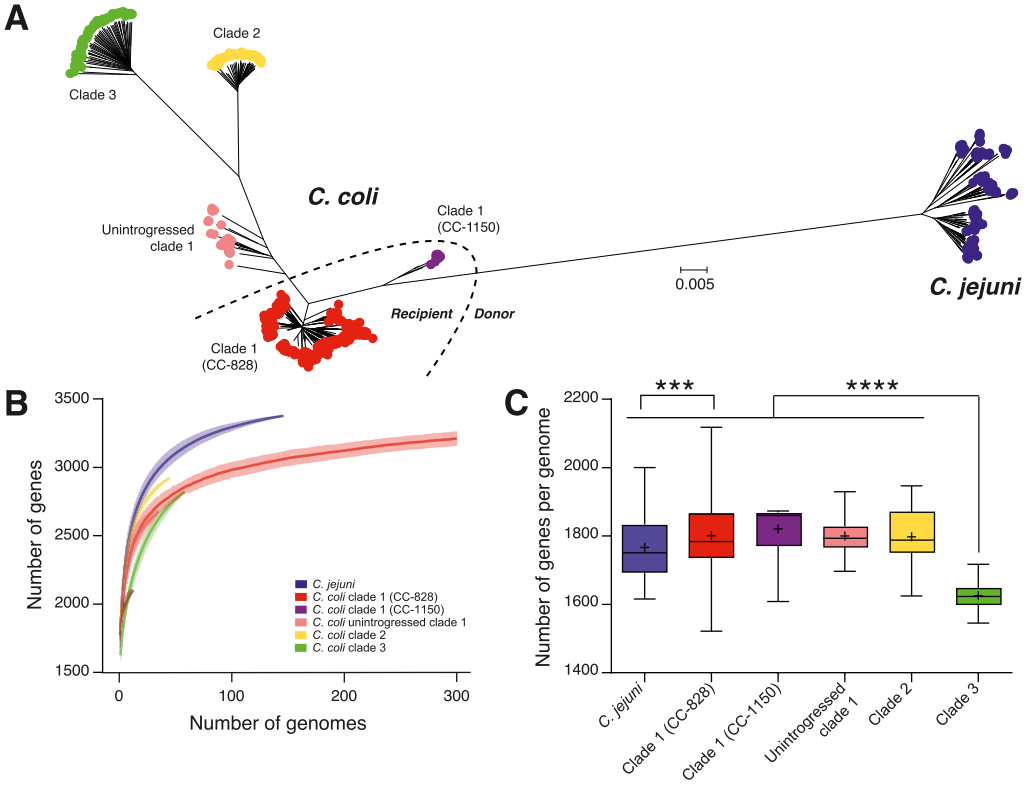
23 Sheppard, S. K., McCarthy, N. D., Falush, D. & Maiden, M. C. Convergence of *Campylobacter* species: implications for bacterial evolution. *Science* **320**, 237-239 (2008).

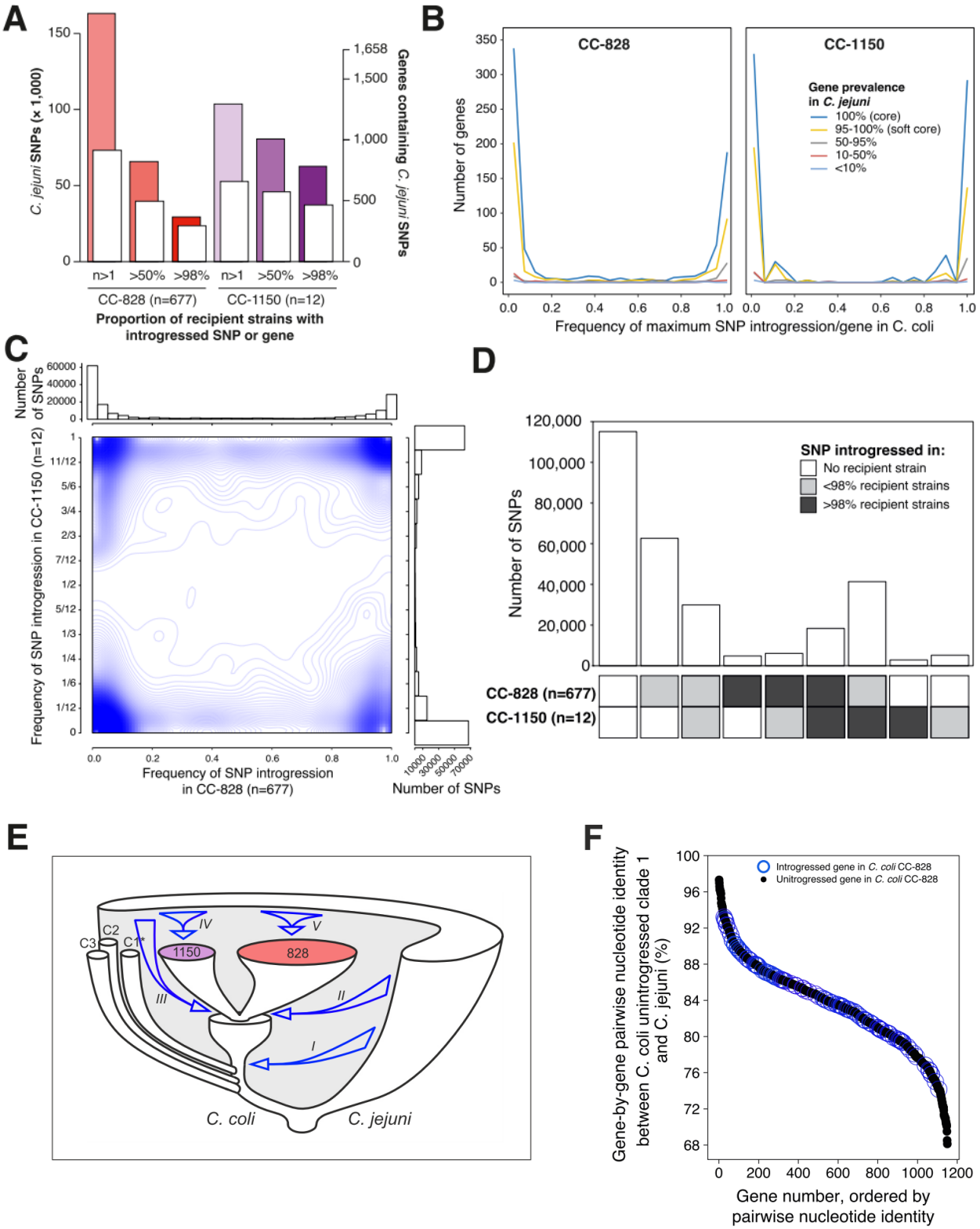
24 Sheppard, S. K. *et al.* Progressive genome-wide introgression in agricultural *Campylobacter coli*. *Mol. Ecol.* **22**, 1051-1064 (2013).

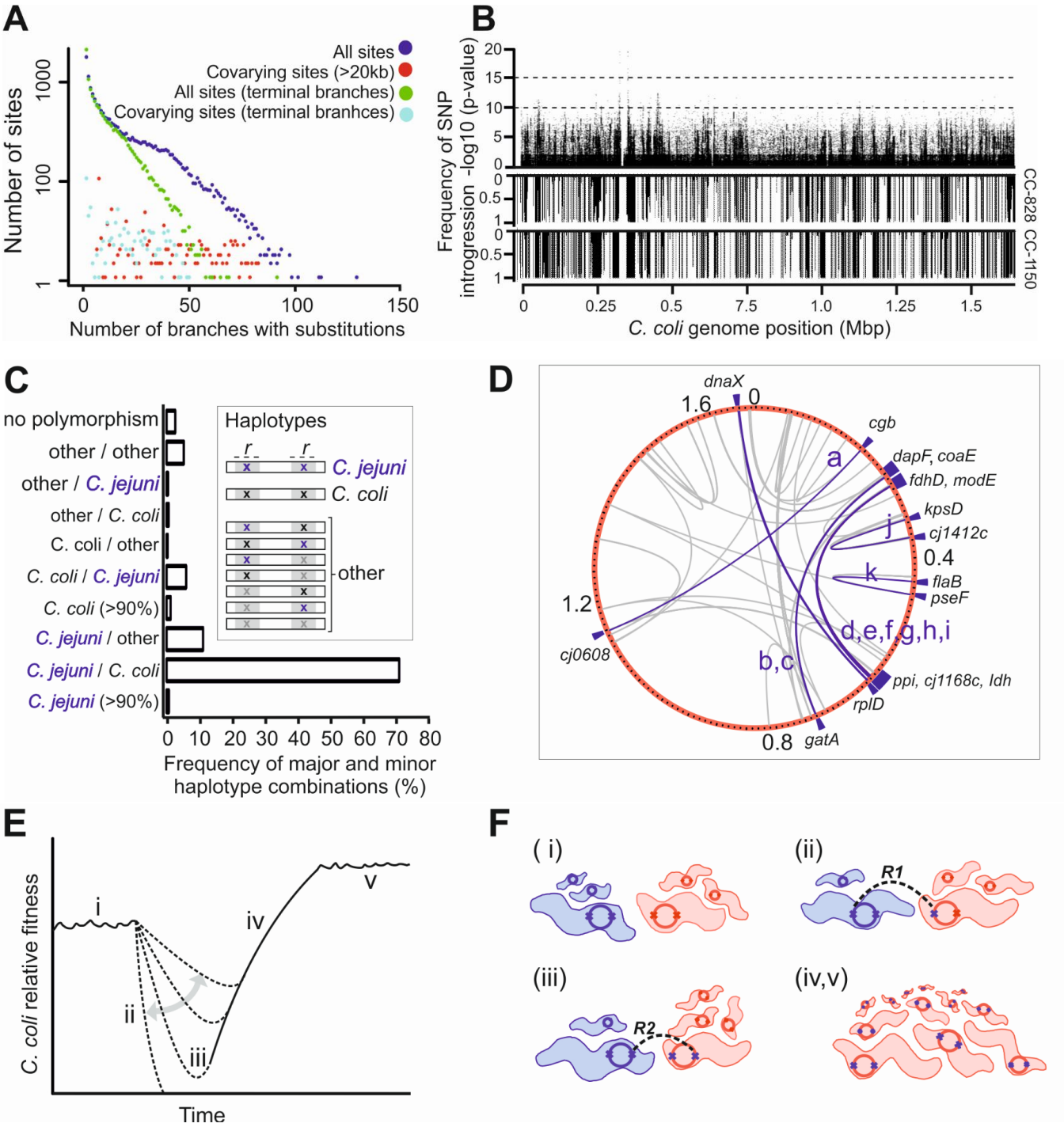
25 Provine, W. B. in *Genetics, Speciation And The Founder Principle* 43-76 (Oxford University Press, 1989).
26 Cohan, F. M. in *Genetics of Mate Choice: From Sexual Selection to Sexual Isolation* 359-370 (Springer, 2002).
27 Ansari, M. A. & Didelot, X. Inference of the properties of the recombination process from whole bacterial genomes. *Genetics* **196**, 253-265 (2014).
28 Retchless, A. C. & Lawrence, J. G. Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. *Proc. Natl. Acad. Sci.* **107**, 11453-11458 (2010).
29 Dearlove, B. L. *et al.* Rapid host switching in generalist *Campylobacter* strains erodes the signal for tracing human infections. *Isme J* **10**, 721-729 (2016).
30 Weerakoon, D. R., Borden, N. J., Goodson, C. M., Grimes, J. & Olson, J. W. The role of respiratory donor enzymes in *Campylobacter jejuni* host colonization and physiology. *Microb. Pathog.* **47**, 8-15 (2009).
31 Taylor, A. J. & Kelly, D. J. The function, biogenesis and regulation of the electron transport chains in *Campylobacter jejuni*: New insights into the bioenergetics of a major food-borne pathogen. *Adv. Microb. Physiol.* **74**, 239-329 (2019).
32 Smart, J. P., Cliff, M. J. & Kelly, D. J. A role for tungsten in the biology of *Campylobacter jejuni*: tungstate stimulates formate dehydrogenase activity and is transported via an ultra-high affinity ABC system distinct from the molybdate transporter. *Mol. Microbiol.* **74**, 742-757 (2009).
33 Doerrler, W. T., Sikdar, R., Kumar, S. & Boughner, L. A. New functions for the ancient DedA membrane protein family. *J. Bacteriol.* **195**, 3-11 (2013).
34 Ledgham, F., Quest, B., Vallaeys, T., Mergeay, M. & Covès, J. A probable link between the DedA protein and resistance to selenite. *Res. Microbiol.* **156**, 367-374 (2005).
35 Shaw, F. L. *et al.* Selenium-dependent biogenesis of formate dehydrogenase in *Campylobacter jejuni* is controlled by the *fdhTU* accessory genes. *J. Bacteriol.* **194**, 3814-3823 (2012).
36 Rieseberg, L. H. Hybrid origins of plant species. *Annu. Rev. Ecol. Syst.* **28**, 359-389 (1997).
37 Andreani, N. A., Hesse, E. & Vos, M. Prokaryote genome fluidity is dependent on effective population size. *Isme J* **11**, 1719-1721 (2017).
38 Vos, M. & Eyre-Walker, A. Are pangenomes adaptive or not? *Nature Microbiol.* **2**, 1576-1576 (2017).
39 Shapiro, B. J. The population genetics of pangenomes. *Nature Microbiol.* **2**, 1574-1574 (2017).
40 McInerney, J. O., McNally, A. & O'Connell, M. J. Why prokaryotes have pangenomes. *Nature Microbiol.* **2**, 17040 (2017).
41 Levin, B. R., Perrot, V. & Walker, N. Compensatory mutations, antibiotic resistance and the population genetics of adaptive evolution in bacteria. *Genetics* **154**, 985-997 (2000).
42 Sheppard, S. K., McCarthy, N. D., Jolley, K. A. & Maiden, M. C. J. Introgression in the genus *Campylobacter*: generation and spread of mosaic alleles. *Microbiology* **157**, 1066-1074 (2011).
43 Pensar, J. *et al.* Genome-wide epistasis and co-selection study using mutual information. *Nucleic Acids Res.* **47**, e112-e112 (2019).
44 Croucher, N. J., Harris, S. R., Barquist, L., Parkhill, J. & Bentley, S. D. A high-resolution view of genome-wide pneumococcal transformation. *PLoS Path.* **8**, e1002745 (2012).

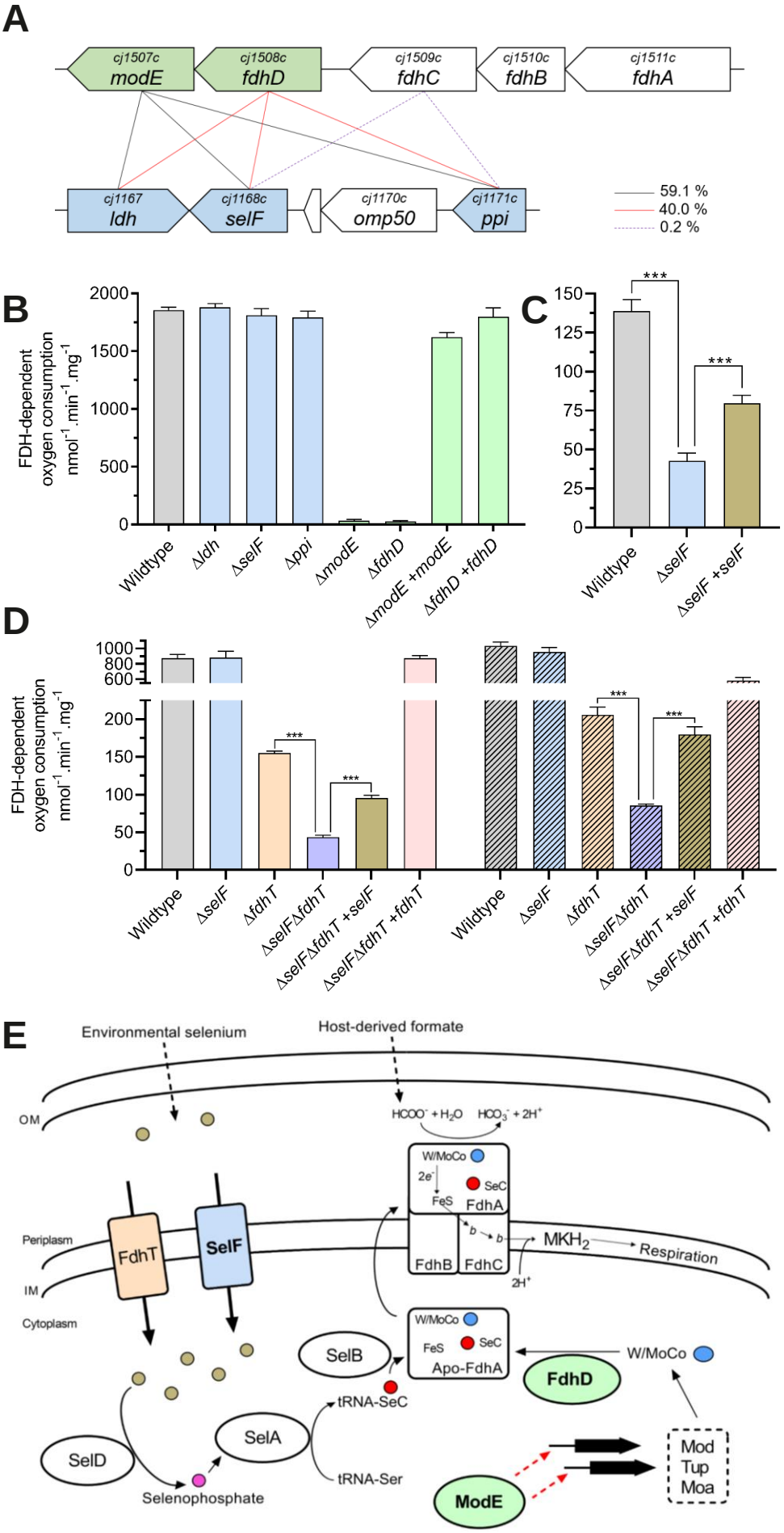
45 He, M. *et al.* Evolutionary dynamics of *Clostridium difficile* over short and long time scales. *Proc. Natl. Acad. Sci.* **107**, 7527-7532 (2010).
46 Power, P. M., Bentley, S. D., Parkhill, J., Moxon, E. R. & Hood, D. W. Investigations into genome diversity of *Haemophilus influenzae* using whole genome sequencing of clinical isolates and laboratory transformants. *BMC Microbiol.* **12**, 273 (2012).
47 Yahara, K. *et al.* Genome-wide association of functional traits linked with *Campylobacter jejuni* survival from farm to fork. *Environ. Microbiol.* **19**, 361-380 (2017).
48 Arnoux, P. *et al.* Sulphur shuttling across a chaperone during molybdenum cofactor maturation. *Nat. Commun.* **6**, 6148 (2015).
49 Taveirne, M. E., Sikes, M. L. & Olson, J. W. Molybdenum and tungsten in *Campylobacter jejuni*: their physiological role and identification of separate transporters regulated by a single ModE-like protein. *Mol. Microbiol.* **74**, 758-771 (2009).
50 Aguilar-Barajas, E., Díaz-Pérez, C., Ramírez-Díaz, M. I., Riveros-Rosas, H. & Cervantes, C. Bacterial transport of sulfate, molybdate, and related oxyanions. *BioMetals* **24**, 687-707 (2011).
51 Thomas, M. T. *et al.* Two respiratory enzyme systems in *Campylobacter jejuni* NCTC 11168 contribute to growth on l-lactate. *Environ. Microbiol.* **13**, 48-61 (2011).
52 Ünal, C. M. & Steinert, M. Microbial peptidyl-prolyl cis/trans isomerases (PPIases): virulence factors and potential alternative drug targets. *Microbiol. Mol. Biol. Rev.* **78**, 544-571 (2014).
53 Oren, Y. *et al.* Transfer of noncoding DNA drives regulatory rewiring in bacteria. *Proc. Natl. Acad. Sci.* **111**, 16112-16117 (2014).
54 Arnold, B. *et al.* Fine-scale haplotype structure reveals strong signatures of positive selection in a recombining bacterial pathogen. *Mol. Biol. Evol.* **37**, 417-428 (2019).
55 Wadsworth, C. B., Arnold, B. J., Sater, M. R. A. & Grad, Y. H. Azithromycin resistance through interspecific acquisition of an epistasis-dependent efflux pump component and transcriptional regulator in *Neisseria gonorrhoeae*. *mBio* **9**, e01419-01418 (2018).
56 Mourkas, E. *et al.* Agricultural intensification and the evolution of host specialism in the enteric pathogen *Campylobacter jejuni*. *Proc. Natl. Acad. Sci.* **117**, 11018-11028 (2020).
57 Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology* **19**, 455-477 (2012).
58 Sheppard, S. K. *et al.* Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. *Proc. Natl. Acad. Sci.* **110**, 11923-11927 (2013).
59 Sheppard, S. K. *et al.* Cryptic ecology among host generalist *Campylobacter jejuni* in domestic animals. *Mol. Ecol.* **23**, 2442-2451 (2014).
60 Pascoe, B. *et al.* Local genes for local bacteria: evidence of allopatry in the genomes of transatlantic *Campylobacter* populations. *Mol. Ecol.* **26**, 4497-4508 (2017).
61 Pascoe, B. *et al.* Genomic epidemiology of *Campylobacter jejuni* associated with asymptomatic pediatric infection in the Peruvian Amazon. *PLoS Negl. Trop. Dis.* **14**, e0008533 (2020).
62 Jolley, K. A. & Maiden, M. C. J. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11**, 595 (2010).
63 Dingle, K. E. *et al.* Multilocus sequence typing system for *Campylobacter jejuni*. *J. Clin. Microbiol.* **39**, 14-23 (2001).

64 Méric, G. *et al.* A reference pan-genome approach to comparative bacterial genomics: identification of novel epidemiological markers in pathogenic *Campylobacter*. *PLOS One* **9**, e92798 (2014).
65 Aziz, R. K. *et al.* The RAST server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).
66 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).
67 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLOS One* **5**, e9490 (2010).
68 Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comp. Biol.* **11**, e1004041 (2015).
69 Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210-223 (2009).
70 Yahara, K. *et al.* Chromosome painting *in silico* in a bacterial species reveals fine population structure. *Mol. Biol. Evol.* **30**, 1454-1464 (2013).
71 Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
72 Yahara, K., Didelot, X., Ansari, M. A., Sheppard, S. K. & Falush, D. Efficient inference of recombination hot regions in bacterial genomes. *Mol. Biol. Evol.* **31**, 1593-1605 (2014).
73 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307-321 (2010).
74 Mell, J. C., Shumilina, S., Hall, I. M. & Redfield, R. J. Transformation of natural genetic variation into *Haemophilus influenzae* genomes. *PLoS Path.* **7**, e1002151-e1002151 (2011).
75 Collins, C. & Didelot, X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS Comp. Biol.* **14**, e1005958 (2018).
76 van Vliet, A. H., Wooldridge, K. G. & Ketley, J. M. Iron-responsive gene regulation in a *Campylobacter jejuni fur* mutant. *J. Bacteriol.* **180**, 5291-5298 (1998).
77 Cameron, A. & Gaynor, E. C. Hygromycin B and apramycin antibiotic resistance cassettes for use in *Campylobacter jejuni*. *PLOS One* **9**, e95084 (2014).


A
Clade 2
Clade 3
C. coli
Unintrogressed clade 1
Clade 1 (CC-1150)
0.005
C. jejuni
Recipient
Donor
Clade 1 (CC-828)
B
Number of genes
Number of genomes
C. jejuni
C. coli clade 1 (CC-828)
C. coli clade 1 (CC-1150)
C. coli unintrogressed clade 1
C. coli clade 2
C. coli clade 3
C
Number of genes per genome
***
****
C. jejuni
Clade 1 (CC-828)
Clade 1 (CC-1150)
Unintrogressed clade 1
Clade 2
Clade 3


A
C. jejuni SNPs (× 1,000)
Genes containing C. jejuni SNPs
n>1
>50%
>98%
CC-828 (n=677)
CC-1150 (n=12)
Proportion of recipient strains with introgressed SNP or gene
B
CC-828
CC-1150
Number of genes
Gene prevalence in C. jejuni
100% (core)
95-100% (soft core)
50-95%
10-50%
<10%
Frequency of maximum SNP introgression/gene in C. coli
C
Number of SNPs
Frequency of SNP introgression in CC-1150 (n=12)
Frequency of SNP introgression in CC-828 (n=677)
Number of SNPs
D
SNP introgressed in:
No recipient strain
<98% recipient strains
>98% recipient strains
Number of SNPs
CC-828 (n=677)
CC-1150 (n=12)
E
C2
C3
C1*
IV
V
1150
828
III
II
I
C. coli
C. jejuni
F
Introgressed gene in C. coli CC-828
Unitrogressed gene in C. coli CC-828
Gene-by-gene pairwise nucleotide identity between C. coli unintrogressed clade 1 and C. jejuni (%)
Gene number, ordered by pairwise nucleotide identity


A
All sites
Covarying sites (>20kb)
All sites (terminal branches)
Covarying sites (terminal branhces)
Number of sites
Number of branches with substitutions
B
Frequency of SNP introgression -log10 (p-value)
CC-828
CC-1150
C. coli genome position (Mbp)
C
no polymorphism
other / other
other / C. jejuni
other / C. coli
C. coli / other
C. coli / C. jejuni
C. coli (>90%)
C. jejuni / other
C. jejuni / C. coli
C. jejuni (>90%)
Haplotypes
C. jejuni
C. coli
other
Frequency of major and minor haplotype combinations (%)
D
dnaX
cgb
dapF, coaE
fdhD, modE
kpsD
cj1412c
flaB
pseF
ppi, cj1168c, ldh
rplD
gatA
cj0608
a
j
k
d,e,f,g,h,i
b,c
E
C. coli relative fitness
Time
i
ii
iii
iv
v
F
(i)
(ii)
R1
(iii)
R2
(iv,v)


A
cj1507c modE
cj1508c fdhD
cj1509c fdhC
cj1510c fdhB
cj1511c fdhA
cj1167 ldh
cj1168c selF
cj1170c omp50
cj1171c ppi
59.1 %
40.0 %
0.2 %
B
FDH-dependent oxygen consumption nmol-1.min-1.mg-1
Wildtype
ΔldhΔselFΔppiΔmodEΔfdhDΔmodE +modEΔfdhD +fdhD
C
Wildtype
ΔselFΔselF +selF
D
Wildtype
ΔselFΔfdhTΔselFΔfdhT +selFΔselFΔfdhT +fdhT
E
Environmental selenium
Host-derived formate
OM
Periplasm
IM
Cytoplasm
FdhT
SelF
W/MoCo
SeC
FeS
FdhA
FdhB
FdhC
MKH2
Respiration
Apo-FdhA
SelB
tRNA-SeC
SelA
tRNA-Ser
SelD
Selenophosphate
FdhD
ModE
Mod
Tup
Moa