Genome-wide association of functional traits linked with *Campylobacter jejuni* survival from farm to fork

Koji Yahara[1]#, Guillaume Méric[2]#, Aidan J. Taylor[3], Stefan P. W. de Vries[4], Susan Murray[5], Ben Pascoe[2,6], Leonardos Mageiros[5], Alicia Torralbo[5]†, Ana Vidal[7], Anne Ridley[7], Sho Komukai[1], Helen Wimalarathna[8], Alison J. Cody[8], Frances M. Colles[8], Noel McCarthy[8,9]§, David Harris[10], James E. Bray[8], Keith A. Jolley[8], Martin C. J. Maiden[8,9], Stephen D. Bentley[10], Julian Parkhill[10], Christopher D. Bayliss[11], Andrew Grant[4], Duncan Maskell[4], Xavier Didelot[12], David J. Kelly[3]*, Samuel K. Sheppard[2,6,8]*

[1]Department of Bacteriology II, National Institute of Infectious Diseases, Tokyo, Japan; [2]The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath, UK; [3]Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield, UK; [4]Department of Veterinary Medicine, University of Cambridge, Madingley Road, Cambridge, UK; [5]Swansea University Medical School, Institute of Life Science, Swansea University, Swansea, UK; [6]MRC CLIMB Consortium, UK; [7]Animal and Plant Health Agency (APHA), Addlestone, UK; [8]Department of Zoology, Oxford University, Oxford, UK; [9]NIHR Health Protections Research Unit in Gastrointestinal Infections, University of Oxford, UK; [10]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK; [11]Department of Genetics, University of Leicester, Leicester, UK; [12]Department of Infectious Disease Epidemiology, Imperial College, London, UK.

# These authors contributed equally; § Current affiliation: Warwick Medical School, University of Warwick, Coventry, United Kingdom; † Current affiliation: Department of Animal Health, Campus de Excelencia Internacional Agroalimentario ceiA3, University of Cordoba, Cordoba, Spain; * Corresponding authors: Prof. Samuel K. Sheppard; s.k.sheppard@bath.ac.uk ; phone: +44(0)1225385046; Prof. David J. Kelly; d.kelly@sheffield.ac.uk; phone: +44(0)1142224414.

Keywords: GWAS, population genomics, poultry processing chain, *Campylobacter*, zoonosis, industrial food safety, survival

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as an 'Accepted Article', doi: 10.1111/1462-2920.13628

Abstract

Campylobacter jejuni is a major cause of bacterial gastroenteritis worldwide, primarily associated with the consumption of contaminated poultry. *C. jejuni* lineages vary in host range and prevalence in human infection, suggesting differences in survival throughout the poultry processing chain. From 7,343 MLST-characterised isolates, we sequenced 600 *C. jejuni* and *C. coli* isolates from various stages of poultry processing and clinical cases. A genome-wide association study (GWAS) in *C. jejuni* ST-21 and ST-45 complexes identified genetic elements over-represented in clinical isolates that increased in frequency throughout the poultry processing chain. Disease-associated SNPs were distinct in these complexes, sometimes organised in haplotype blocks. The function of genes containing associated elements was investigated, demonstrating roles for *cj1377c* in formate metabolism, *nuoK* in aerobic survival and oxidative respiration, and *cj1368-70* in nucleotide salvage. This work demonstrates the utility of GWAS for investigating transmission in natural zoonotic pathogen populations and provides evidence that major *C. jejuni* lineages have distinct genotypes associated with survival, within the host specific niche, from farm to fork.

Importance

Bacteria that live in animal guts have enhanced potential for transmission to humans if they are able to survive outside of the host. This becomes a major problem if these organisms are pathogenic to humans, especially if they are common in animal reservoirs. *Campylobacter jejuni* is such an organism; responsible for more than 280,000 annual cases of food poisoning in the UK alone, often associated with the consumption of poultry meat contaminated at slaughter. This bacterium is well suited to growth in the micro-aerophilic environment of the animal gut but is not well adapted to proliferation outside of the host. Therefore, questions remain about how *C. jejuni* is able to survive through the poultry processing chain to infect humans. In this study we examined a large number of isolates genotyped at 7 loci, and showed that some lineages increased, as a proportion of the population, through processing. This implies different capacities to survive outside of the host. Using a genome-wide association study approach, we were able to identify novel genetic elements that were associated with survival, and by testing the function of these genes using deletion mutants, we were able to identify functional differences that promote transmission. This combined comparative genomic-phenotyping approach provided evidence for the role of formate metabolism, oxidative respiration, and nucleotide salvage in survival from farm to fork, opening new possibilities for targeted interventions.

Introduction

More than 60% of human infections are caused by pathogens that infect both humans and animals, annually accounting for around a billion cases of illness and death worldwide (1, 2). One of the major transmission routes for zoonoses is contaminated food, and rising demand imposes ever-increasing pressure on safe sustainable production (2, 3). *Campylobacter* is the most common cause of foodborne bacterial gastroenteritis in industrialized countries (4). The two main disease-causing species, *C. jejuni* and *C. coli,* cause approximately 2.4 million cases of food poisoning each year in the US (almost 1% of the population) with a significant associated economic cost (5). A large proportion of human infections, 40-80%, result from eating poultry meat contaminated through processing (6-8) and the slaughter house is known to be an important location for the spread of *Campylobacter* to the surface of retail meat (9-11).

Genotyping studies, including multi-locus sequence typing (MLST), have been instructive in showing that *Campylobacter* is not a genetically monomorphic organism but comprises highly diverse assemblages with numerous phenotypes (8, 12-14). Within this complexity there is sufficient genetic structuring to identify genotypes associated with particular animal and bird hosts (13), where *Campylobacter* can be a common component of the gut microbiota. It is, therefore, possible to identify isolates from chickens as a major source of human infection by comparison of clinical isolates with those from reservoir hosts (8, 15-18). Characterization of isolates from different stages in poultry processing has shown that some genotypes increase as a proportion of the population on chicken after slaughter compared to the levels in live chickens (19-22). This is in spite of the evidence that *C. jejuni* appears to be more susceptible than many bacterial pathogens to changes in temperature, oxidation, hydrostatic pressure and acidity (23, 24). In the laboratory, poor growth of these bacteria at atmospheric oxygen concentrations or at a temperature below 30° C implies that proliferation outside of the host or on food is unlikely. However, the high incidence indicates a capacity for survival (25). Genomic analyses have revealed that many common bacterial stress response genes such as *rpoS*, *soxRS*, *oxyR*, *rpoH* and *cspA* may be absent in *C. jejuni* (26), which suggests that there could be unknown survival mechanisms in *Campylobacter* that may coordinate the response to environmental stress and promote the proliferation of strains that are adapted to survival outside the host.

Genome-wide association studies (GWAS), using methods derived from human genetics, are increasingly being used in microbial genetics to identify genetic elements that are associated with particular phenotypes (27-31). Although association studies have been successful in identifying genetic variants that contribute to complex traits in humans (32), these methods have been challenging in bacteria. The main reason for this is the strong population structure resulting from clonal reproduction (33). For example, in *C. jejuni* and *C. coli* populations, lineages cluster into clonal complexes that share genetic elements correlated with a particular phenotype of interest, but some of these elements are unrelated to the phenotype and are simply inherited through clonal descent. This impedes the use of simple association mapping approaches.

In this study, we use a GWAS approach to compare genome sequences from *Campylobacter* isolates sampled throughout the poultry processing chain and from human campylobacteriosis cases. Using 600 whole-genome sequences, we explore whether specific alleles or sequences are significantly associated with human disease by first analysing the two major *C. jejuni* host generalist clonal complexes (ST-21 and ST-45) separately, and then exploring signals across other *C. jejuni* and *C. coli* clonal complexes. By investigating the function of genes with clinically associated genetic variation, it is possible to identify candidate survival determinants that may influence transmission to humans through the major infection route.

Results

Differential poultry processing chain survival and disease-associated genetic variations

Substantial variations were observed in the prevalence of STs from major clonal complexes at different stages in the poultry processing chain (Figure 1). For example, ST-21 increased in prevalence from farm to clinical samples. This is consistent with variation in the ability of *C. jejuni* lineages to survive different poultry processing chain conditions. Using GWAS methodology (27, 31), we identified genetic elements, in the form of 30-bp words, that were significantly over-represented in clinical isolates compared to farm chicken isolates from ST-21 and ST-45 clonal complexes (Figure 2). A total of 2749 and 633 words were identified in ST-21 and ST-45 clonal complex isolates respectively, with $p<5x10^{-4}$ when compared to the null distribution based on the population structure (Figure S1). All words within ST-45

complex isolates and 68% of those in ST-21 complex isolates were mapped to coding sequences from the annotated *C. jejuni* NCTC11168 reference genome (Figure 2). The mapped words were then classified into corresponding genetic variations (SNP, indel or entire gene gain/loss) by examining how they were located in gene-by-gene alignments. In total there were 419 words in the ST-21 complex and 607 words in the ST-45 complex that corresponded to disease-associated SNPs, all of which represented variation in homologous DNA sequence rather than indels or entire locus presence/absence. Not all of the words could be classified into corresponding genetic variations (SNP, indel or locus presence/absence), which can be explained if the words are in a contig that is incomplete or not able to be mapped to a gene-by-gene alignment, or if an associated word corresponds to a combination of SNPs (29). In this study, around 10% of the words could be explained by combination of two SNPs in a single word, and were removed from the analysis.

For the ST-21 clonal complex, from the 2749 disease-associated words, we discovered 23 SNPs in 10 genes (Figure 3A, Table 1, Table S3, Table S5), showing a 34-46% frequency increase (41% on average) from farmed chicken to clinical isolates potentially through several repeated evolutionary events. Comparatively, for the ST-45 clonal complex, from the 633 disease-associated words, we discovered 47 SNPs in 9 genes (Figure 3B, Table 1, Table S3, Table S5), showing a 32-34% frequency increase (34% on average) from farmed chicken to clinical isolates. The frequency increase of the disease-associated alleles found in the ST-21 clonal complex was replicated in a cross-validation dataset of 24 ST-21 human disease strains (Figure S3): all of the 23 SNPs showed 19-64% (40% on average) frequency increase compared to the farm strains.

There was no overlap of associated genes between isolates from the two host generalist clonal complexes, suggesting that the adaptive signatures may differ in these two lineages. Genes associated with human disease in the ST-21 clonal complex isolates included *kpsC* and *kpsD*, which are members of the capsular polysaccharides biosynthetic pathway and contribute to adhesion and biofilm formation (34). Variation in the *nuoK* gene, associated with increased survival in ST-21 complex isolates, encodes a subunit of the Nuo flavodoxin:quinone oxidoreductase, which is involved in energy conserving electron transfer from 2-oxoacids to oxygen or other electron acceptors (Weerakoon and Olson, 2008). *nuo* gene expression changes in response to hyperosmotic stress and in stringent response mutants (35, 36). Interestingly, the stringent response has been suggested to be important for surviving oxygen

stresses (36). Human disease in ST-45 clonal complex isolates was associated with genes, including *cj1373* and *cj1375*, which putatively encode efflux proteins involved in detoxification and antimicrobial resistance (Table 1) (37, 38). Genes associated with clinical isolates in the ST-45 clonal complex were from similar or neighbouring regions (Table 1), while the disease-associated words clustered together when mapped on the reference *C. jejuni* NCTC11168 genome sequence, suggesting co-acquisition or selection of these elements. None of these genes was located with known phase variable elements, a known mechanism of rapid variation and adaptation for *C. jejuni* (39).

Clustering and linkage disequilibrium of associated elements

The extent of clustering of the disease-associated words was examined by comparing an observed distribution of distance between successive disease-associated SNPs with expected distribution calculated from randomly selected SNPs in the genome (Figure 4A, 4B). Most of the disease-associated SNPs were clustered with each other, both in ST-21 and ST-45 clonal complexes, suggesting strong linkage between associated SNPs. We then plotted linkage disequilibrium (LD) coefficient r^2 values between the disease-associated SNPs and the average LD decay in SNPs in core genes for both ST-21 and ST-45 complexes (Figure 4C, 4D).

For the ST-21 clonal complex, there was a complete linkage between SNPs in *cj0694* and *cj1048c* or *cj1049c* (located >300 kb apart on the genome), which substantially deviated from the average LD decay. In addition, there were two other highly-linked combinations across >100kb (SNPs in *lpxD* and *cj0694* or *cj1049c*), that clearly deviated from the average LD decay. These results suggest long-distance interaction of loci that could be functionally interdependent. Cj0694 is a SurA-like membrane bound periplasmic facing chaperone (PpiD homologue) that is implicated in periplasmic or outer membrane protein folding (40), while Cj1048 and Cj1049 are implicated in lysine biosynthesis and excretion respectively (Table S1). Such a high linkage between SNPs separated by >100kb has recently been reported in a genomic study of *Staphylococcus aureus* (41) but is even more surprising here due to the higher rate of recombination in *C. jejuni* which should reduce linkage between distant sites. Distributions of r^2were significantly different between core SNPs with ≥ 25% minor allele frequency (i.e., minimum among the disease-associated SNPs) and the disease associated SNPs in the ST-21 clonal complex (Figure S2) ($p< 0.0005$, Kolmogorov-Smirnov test). The proportion of pairs of disease-associated SNPs with $r^2>0.8$ was significantly higher than that

of core SNPs for those with ≥ 25% minor allele frequency (25% vs 14%, $p<0.005$, Fisher's exact test). For the ST-45 clonal complex, all pairs of disease-associated SNPs had a value of $r^2>0.7$. This reflects a long haplotype block, over-represented in clinical isolates, spanning about 19kb that is formed by the SNPs (Figure 4D). Around 16% (7 out of 45, 6 of which are clinical isolates) of the strains have all of these linked disease-associated SNPs (Figure 3B).

While much of the linkage between SNPs occurs within genes, the linkage of SNPs in large haploblocks is potentially influenced by the sequencing technology used in this study. Specifically, as genome files comprise multiple contiguous sequence assemblies (contigs), it is conceivable that linkage analysis could be influenced by the beginning and end point of a contig, and be missed. To account for this we quantified the frequency of genomes carrying pairs of SNPs in high LD on the same contigs. Examination of inter-locus pairs of SNPs, that showed $r^2>0.9$, and were located in different genes (Table 1), revealed that the frequency was approximately 9% and 91% on average (SD = 19% and 7%) in ST-21 and ST-45 clonal complexes, respectively. This indicated that even in ST-45 clonal complex isolates, in which the longest haplotype block was found, approximately 9% of the strains had the inter-locus pairs of SNPs in high LD that were present in different contigs (Figure S6).

Frequency of disease-associated words in other clonal complexes

The frequency change of disease-associated words from farm to human disease was examined in nine other clonal complexes, namely: ST-48, ST-257, ST-353, ST-354, ST-443, ST-52, ST-574, ST-607 and ST-828 complexes. No single disease-associated SNP increased in frequency from farm to human disease in all of the clonal complexes (Figure 3C). There was little consistency between SNPs that showed more than a 10% increase in ST-21 and ST-45 complexes but some increased in other complexes. Among the disease-associated SNPs found in ST-21, those in *hypD* showed ≥ 50% frequency increase in ST-353 and ST-607 isolates, although there is a small sample size for the latter (4 farm and 6 disease strains). In the ST-353 clonal complex, one of the SNPs in *cj1161c* (putative cation-transporting ATPase) showed a 45% frequency increase. Among the SNPs in the large disease-associated haplotype block found in ST-45, only the SNPs in *cj1364* show a 30% frequency increase in ST-607. The chicken-associated ST-353 and ST-257 clonal complexes show a similar overall pattern of SNP frequency change. Among the most consistent signals of frequency increase of disease-associated SNPs across clonal complexes are those in *cj1414c* (*kpsC*) which is required for capsular polysaccharides (CPS) on the cell surface. This SNP was

nonsynonymous, and showed frequency increase in four other clonal complexes outside ST-21 (ST-45, ST-48, ST-353, and ST-354). The pooled odds ratio among the four other clonal complexes is 7.1 (P = 0.006, 95% CI: 1.4 - 35.9, Mantel-Haenszel test), indicating significant positive association across the clonal complexes, although it is not the case in others.

Examination of possible confounding sampling factors

Despite the statistical stringency of our approach for identifying associated elements, there are possible confounding factors, which we examined. First, it has been estimated that 60% to 80% of clinical infections result from the consumption of contaminated chicken. The remainder originates in other reservoirs including other livestock, wild birds or environmental sources (8, 18). We examined the exact match of associated elements (100% sequence identity on the total 30-bp length of the element) in 354 genomes from chicken, cattle and wild birds, and observed that source was irrelevant to the association signal (Kruskal-Wallis test, p=0.10; Figure S5A). Second, sampling of clinical and poultry isolates was performed over time with 96% of our samples were obtained within 7 years (2005-2011). In the clinical isolates genomes from our dataset (209 genomes with isolation date information), the presence of associated words in individual isolates was independent of the year of isolation (Kruskal-Wallis test, p=0.64; Figure S5B).

Formate dehydrogenase (FdhA) activity is dependent on Cj1377

The *cj1377c* and *selA* genes were identified in our association mapping study as having clinical-associated SNPs that are enriched from farm to humans. The *cj1377c* gene is of unknown function but shares a divergent promoter region with the *selAB* selenocysteine synthase operon. The only predicted selenocysteine-containing protein in *C. jejuni* reference strain NCTC11168 is formate dehydrogenase (FDH), specifically the alpha subunit FdhA (Cj1512). Therefore the selenocysteine synthesis pathway can be hypothesised to be specific for FdhA activity, and in the absence of selenocysteine, FdhA would either not be translated, or produce a non-functional truncated product, as the selenocysteine codon becomes a stop codon. We hypothesised here that the ferredoxin-like protein encoded by *cj1377c* might have an electron transfer function affecting selenocysteine synthesis or FDH activity directly. To investigate the role of this protein in formate metabolism and to contextualise its association with survival throughout poultry processing, we engineered a deletion mutant of *cj1377c* in the reference strain NCTC11168, and assessed it for FdhA activity using the formate-dependent reduction of the artificial electron acceptor methyl viologen, a reaction specifically

catalysed by the FdhA subunit (Figure 5A). The Δ*cj1377c* mutant showed no growth defect in complex media compared to isogenic wildtype *C. jejuni* NCTC11168 (data not shown), but FdhA activity was completely abolished compared to wild-type. This result implicates Cj1377 as being involved in the biogenesis of FdhA itself, probably via a reductive function related to selenium incorporation.

Having tested the utility of the knock-out approach in a reference strain, as an indicator of the link to *cj1377c* absence, we measured methyl viologen linked FdhA activity in intact cells for 16 ST-21 and ST-45 complex isolates from farmed chicken and clinical cases with homologous sequence variation at this gene (Figure 5A). Allelic variations produced measurable differences in FdhA activity. In the ST-45 complex, the average FdhA activity for clinical isolates was lower than for farmed chicken isolates. This observation is consistent with the association of variation in the *cj1377c* genes with survival through poultry processing in the ST-45 complex, which could indicate that allelic variation associated with variation in FdhA activity is an important survival mechanism. Previous studies have shown that an FdhA mutant showed reduced chicken colonisation abilities (42, 43), consistent with a physiological trade-off between chicken colonisation and stress resistance in *C. jejuni,* similar to that hypothesised for *E. coli* (44).

Evidence of a role for Cj1368-70 in a nucleotide salvage pathway

Sequence variation in the *cj1368* gene was found to be disease-associated in our analysis (Table 1). *cj1368* forms a 3-gene operon with *cj1369* and *cj1370*, all being transcribed from a single promoter. As operon structuring often indicates a related physiological role, the function of these 3 genes was investigated together. Bioinformatic predictions suggest that: (i) *cj1368* encodes a radical S-adenosyl methionine (SAM) family protein which uses an iron-sulphur cluster to generate a 5-deoxadenosyl radical that could be used in various metabolic pathways; (ii) *cj1369* encodes a sulphate/Xn/Ur-type membrane transporter which may transport sulphate/xanthine/uracil compounds; (iii) *cj1370* encodes a type I phosphoribosyl transferase (PRTase) implicated in nucleotide salvage. We therefore investigated whether this 3-gene operon could represent a nucleotide salvage pathway in *C. jejuni*. Single mutants of *cj1368*, *cj1369* and *cj1370*, and triple mutant Δ*cj1368-70*, were constructed in *C. jejuni* strain NCTC11168 and assayed for their putative role in exogenous nucleotide salvage using phenotypic resistance to toxic nucleotide analogues. Initial growth curves showed that Δ*cj1368* and the triple mutant Δ*cj1368-70* have a small but significant growth defect in MH

broth at standard conditions (unpaired *t*-test; p=0.002 and p=0.004, respectively) after 8 hours of growth. Δ*cj1369* showed a mild but non-significant defect (p=0.07) and Δ*cj1370* showed no defect (Figure 5B). Of the toxic nucleotide analogues tested, wildtype *C. jejuni* 11168 and isogenic mutant Δ*cj1368-70* were susceptible to 6-Mercaptopurine (MP) and 6-Thioguanine (TG), but not 8-Azaxanthine monohydrate, 2,6-Diaminopurine or 5-Fluorouracil (Figure S4A). Further growth experiments with wildtype and triple mutant Δ*cj1368-70* showed that, despite Δ*cj1368-70* displaying a significant growth defect in controls, the mutant grew significantly better (i.e. is more resistant than) wildtype in the presence of either 0.1mM MP or TG (Figure 5C). A full set of disk diffusion assays were then performed with MP and TG, which showed that each single mutant, and the triple mutant, was completely resistant to MP and TG at 100 mM nucleotide on the disc (Figure S4B).

In vitro growth under varied oxygen tensions suggests NuoK is required for aerobic survival

Response to oxygen was amongst the functional categories of genes associated with clinical isolates, and potentially survival through the poultry processing chain (Table 1). In order to investigate whether specific associated genes could play a role in responding to oxygen, we generated the defined deletion mutants Δ*nuoK* (*CJM1_1505*, *cj1569c*) and Δ*fumC* (*CJM1_1325*, *cj1364c*) in the *C. jejuni* M1 background. (43) reported that addition of formate was necessary for isolation of certain *nuo* mutants in *C. jejuni*. However, we were able to isolate the *nuoK* mutant on BHI media alone. We compared the growth of the *nuoK* and *fumC* mutants with their isogenic wild-type *C. jejuni* M1 strain in batch cultures at variable atmospheric oxygen tensions; oxygen-limited (5% v/v oxygen in the gas phase with minimal headspace and without shaking), microaerobic (10% v/v O_2 with shaking) and aerobic (20.9% v/v O_2 with shaking) (Figure 6A). Interestingly, while neither mutant displayed a growth defect under oxygen-limited conditions compared to wild-type, both displayed a mild defect in microaerobic growth, and in particular the *nuoK* mutant had highly attenuated survival at aerobic oxygen concentrations. This result highlights the possibility that natural variation at the *nuoK* and to a lesser extent the *fumC* loci could play a role in variable responses to oxygen in natural populations of *C. jejuni*.

The role of NuoK in oxygen-linked respiration of 2-oxoacids

In ST-21 and 37 ST-45 clonal complex isolates that were used in the GWAS and phenotypical testing, there were 7 distinct *nuoK* alleles, predicted to encode 4 different

protein variants. Interestingly, some alleles were more frequently found in clinical or farm isolates (Figure 6B). Four of these alleles (1, 3, 4, and 9) were specific to the ST-21 complex and three (2, 6, and 7) were found exclusively in the ST-45 complex (Table S4). NuoK is a proton-translocating subunit of the inner membrane respiratory complex I, referred to as NDH-1 or the Nuo complex (Figure 6C). In most bacteria the function of complex I is to link NADH oxidation to the reduction of quinone in electron transport chain (ETC) for energy conservation; however the *C. jejuni* genome lacks the genes encoding *nuoE* and *nuoF*, responsible for NADH dehydrogenase activity (43, 45). Instead, 2 unique subunits, Cj1574 and Cj1575, are present which mediate electron flow into the ETC from reduced flavodoxin, not NADH, via complex I (Figure 6C), as evidenced by previous studies with various *nuo* mutants (43). In *C. jejuni*, flavodoxin is reduced by 2-oxoglutarate:acceptor oxidoreductase (Oor) and possibly by pyruvate:acceptor oxidoreductase (Por) enzymes (43, 46). Thus the function of the Nuo complex in *C. jejuni* is to link the respiration of 2-oxo acids to the ETC via reduction of flavodoxin (Figure 6C).

A defined *nuoK* deletion mutant in reference strain M1 was assayed for its ability to respire 2-oxoacids by measuring the rate of 2-oxoacid dependent oxygen consumption in an oxygen electrode (Figure 6D). The *nuoK* mutant showed significantly decreased, but not abolished, respiration with 2-oxoglutarate, suggesting the NuoK subunit is not absolutely essential to the function of complex I. In contrast, pyruvate respiration was only slightly decreased in the *nuoK* mutant. A *fumC* citric acid cycle (CAC) mutant showed no significant reduction in 2-oxoacid respiration compared to the isogenic *C. jejuni* M1 wildtype.

Discussion

Evidence from large MLST datasets in this study and others (8, 14, 47) show that some *Campylobacter* genotypes increase in frequency as they pass from the reservoir chicken host to human infection. Here we used a GWAS approach to investigate genetic variation that was differentially associated with isolates from poultry processing and clinical infection. This was related to the bacteria's capacity to survive outside of the host through the poultry processing chain. We analysed isolates from the ST-21 and ST-45 clonal complexes separately, both of which are common throughout the poultry processing poultry processing chain and in clinical disease. To minimize the potential confounding effects of the strong population structure in *C.*

jejuni, we used a method (27, 31) which adjusts for the effect of relatedness between individual strains in the clonal genealogy compared to the null distribution of expected associations within each clonal complex. This allowed the identification of genetic elements that are significantly over-represented among clinical *C. jejuni.* These elements, which increased in frequency through processing, were mapped to known virulence and candidate survival genes.

Among the 70 disease-associated SNPs, around 75% were synonymous. While sequence variants linked to changes in protein sequences are simpler to interpret in relation to functional variation, there are several explanations for the abundance of synonymous SNPs among clinical isolates. First, the patterns of variation across bacterial genomes in features such as gene order, distribution of coding sequences on leading and lagging strands, GC skew, and codon usage are consistent with selection operating on sequence features other than maintenance of the protein sequences encoded (48). These interactions are likely to be important in complex phenotypes such as survival that will involve multiple genes, and the occurrence of pervasive selection pressures across much of the genome has been previously described in the genus *Campylobacter* (49). Second, disease-associated SNPs can be in strong linkage disequilibrium with synonymous sequence variation. In this case, it is expected that all linked SNPs will be associated with disease irrespective of which confers the functional advantage. The presence of large clinical associated haploblocks is clear in ST-45 complex isolates (Figure 4D). Third, it is possible that some non-synonymous SNPs are recorded as synonymous due to frame-shifts or misinterpretation of start codons.

Human disease-associated sequence variation can provide indirect information about the complex environmental stresses imposed on *C. jejuni* through the many steps of the poultry processing chain, and how conditions select for particular *C. jejuni* lineages that infect humans after the consumption of contaminated meat. Among the genes with genetic signatures of human disease association, and potentially survival adaptation, were those associated with formate metabolism, which occurs on epithelial surfaces within the animal host (43, 50). One gene associated with survival through processing was *cj1377c*, originally annotated as a "putative ferredoxin", but which was found to be involved in formate metabolism in this study. Formate oxidation provides electrons for *C. jejuni* respiration and is abundant in the gut environment of hosts where it is an excreted product of the resident microbiota (43, 50, 51). Formate is oxidised by the FDH complex and its electrons donated to

the menaquinone pool. A *C. jejuni* NCTC11168 FDH null mutant showed reduced colonisation in chicken infection models, particularly when combined with the absence of hydrogenase (43). We demonstrated that a *cj1377c* mutant totally lacked FdhA activity. Given its genomic context of sharing a palindromic promoter with the selenocysteine synthesis enzyme *selA*, we conclude that the ferredoxin Cj1377 has a redox function relating to selenocysteine incorporation into FdhA. FDH activity was not significantly different between a small subset of farm and clinical isolates. However, we observed a trend consistent with clinical isolates having reduced FDH activity.

In this study we also discovered that the disease-associated genes *cj1368-70* had possible functions in nucleotide salvage. The function of these genes in environmental survival or host colonisation may be to increase bacterial adaptability by allowing the efficient utilisation of nucleotides to supplement *de novo* synthesis for replication. On the assumption that mutants in this pathway would be unable to take up nucleotides from the environment, they should be resistant to toxic analogues of such nucleotides. Each single mutant and a triple mutant were shown to be resistant to 6-Mercaptopurine and 6-Thioguanine, supporting a role in nucleotide salvage. In addition, growth curves under standard conditions showed Δ*cj1368* (and the triple mutant), but not Δ*cj1369* or Δ*cj1370*, had a significant growth defect. We postulate that while Cj1369 and Cj1370 have specific functions as a permease and transferase for nucleotide uptake, the radical SAM enzyme Cj1368, although clearly involved in this pathway, may also participate in other metabolic pathways. This may explain why this mutant has an additional growth defect. Along with *cj1377c* and its role in formate metabolism, the association of *cj1368* with disease in our GWAS analysis could indicate a broader importance of metabolic plasticity for the survival through poultry processing and/or the subsequent infection of humans.

We also identified two proteins important in oxidative energy conservation showing signals of association to human disease; NuoK and FumC. The *nuoK* gene encodes a membrane-bound subunit of the 14 subunit oxidoreductase complex I, which in *C. jejuni*, unlike classical complex I NADH dehydrogenases, transfers electrons from reduced flavodoxin, formed from 2-oxoacid oxidation by Oor (and possibly Por) enzymes, to the respiratory chain (43, 46) (Figure 6C). Por and Oor, which convert pyruvate to acetyl-CoA and 2-oxoglutarate to succinyl-CoA, respectively, are oxygen sensitive Fe-cluster enzymes, usually found in obligate anaerobes, which replace the oxygen stable pyruvate and 2-oxoglutarate

dehydrogenases of aerobes. This has been proposed to partially explain the microaerophilic nature of *C. jejuni* (45, 46, 52). Although *nuoK* gene presence did not vary significantly in prevalence between farm and clinical isolates, different alleles of the gene were differentially distributed with sample source (Figure 6B). The growth of the *nuoK* mutant showed attenuated survival with increasing oxygen, and the *nuoK* mutant had significantly lower rates of 2-oxoacid respiration, confirming NuoK is an important component of the flavodoxin oxidising complex I, but perhaps not absolutely essential (Figure 6A and 6D). FumC (fumarase) is responsible for the hydration of fumarate to malate in the citric acid cycle (CAC) and thus a mutant in this enzyme will have an incomplete CAC, which should affect growth (Figure 6C). It was surprising that a *fumC* deletion mutant only displayed a mild growth defect, but this highlights the flexible metabolism of *C. jejuni*, which is able to use alternative anaplerotic pathways to replenish CAC intermediates, especially C4-acids. Thus in rich media, where numerous metabolites and intermediates are available, a *fumC* mutant may not be expected to be excessively growth attenuated. Variation at the *nuoK* and *fumC* loci throughout the poultry processing chain could indicate the importance of an adaptable utilisation of available respiratory and metabolic intermediates.

Sequence variation in other genes was also significantly associated with clinical isolates but their phenotypical relevance to survival is not necessarily clear. For example, modulation of growth at various temperatures is also likely to be an important trait for survival through poultry processing. *lpxD*, associated with survival in ST-21 complex isolates, is involved in temperature-regulated membrane remodelling directed by the lipid A-modifying N-acyltransferase enzyme. Different alleles of *lpxD* add chains of varying lengths of heat stable N-linked fatty acyl chains during lipid A biosynthesis, which could play a role in survival in a wider temperature range (53). Additional indications of the stresses associated with the poultry production can be inferred from the association of *glmS* with survival in ST-21 complex isolates. GlmS, encoded by *cj1366c*, is a cell wall biosynthesis ribozyme essential for cell viability and is produced in response to changes in pH (54, 55) and has a role in biofilm formation (56), potentially eliciting a bacterial response to acid stress (57). Capsular polysaccharide (CPS) genes, *kpsC/cj1414c* and *kpsD/cj1444c* are required for *Campylobacter* to form a capsule that plays an important role in its interaction with the host and the wider environment. Biosynthesis of the CPS is controlled by a large cluster of genes (*cj1413c - cj1448c; 58, 59-61)* and is involved in serum resistance and invasion of epithelial cells (62-64).

The inconsistency of disease-associated elements among *C. jejuni* and *C. coli* lineages, as well as between ST-21 and ST-45 complexes, suggests that genomic changes that promote functional variation among strains are not consistent across the species. Elements associated with clinical isolates will not only represent those that confer a fitness advantage to the various pressures encountered in the poultry processing chain, but also virulence genes that are directly associated with human infection. The numerous genomic variations promoted by this complex landscape of varying environmental pressures are difficult to characterize. However, the absence of a consistent signal of disease-association across lineages implies that survival/infection strategies may differ between strains, despite convergence towards phenotypes related to survival through processing.

Phenotypic differences between ST-21 and ST-45 complex isolates include differential metabolic abilities (65) and cell invasiveness (66). Furthermore, ST-45 complex isolates are commonly sampled from a variety of sources including agricultural animal and wild bird faeces and riparian sources (67, 68). The observed divergences in disease-associated genetic variation between these clonal complexes could reflect different interactions with the selective conditions throughout the poultry processing chain, which comprises a series of sudden selective bottlenecks. Understanding the functional traits associated with the survival of *Campylobacter* through processing has important implications for developing targeted interventions to control the contamination of retail meat. This work identifies candidate genes involved in zoonotic transmission of a pathogen to humans from an agricultural reservoir, and demonstrates that GWAS studies in bacteria can be applied to unravel the genetic basis of complex phenotypes.

Materials and Methods

Isolates

The initial *Campylobacter* isolate dataset comprised 5556 archived samples (http://pubmlst.org/campylobacter/) from large published MLST studies (8, 14, 69) representing three sampling points: farm/caeca; carcass/retail poultry and clinical. A total of 1719 farm/caeca isolates were cultured and typed from 17 UK broiler chicken flocks in June and November 2008 including chicken faeces and caecal swabs - from 25-31 day old birds

and at evisceration in the abattoir (70, 71). Carcass and retail poultry samples comprised 1372 samples collected after carcass chilling (72) and from retail poultry meat (69). Clinical isolates were from a previous sampling of human campylobacteriosis cases in the UK, as well as unpublished genomes, representing reported cases of human disease from the John Radcliffe Hospital, Oxford in 2008 (73) and a comprehensive survey of clinical isolates from all 28 diagnostic laboratories in the 15 health board regions in Scotland (14).

Genome sequencing and assembly

A total of 600 *Campylobacter* isolates were chosen for whole genome sequencing to represent various stages of the poultry processing chain and human infection cases (Table S1). All samples were cultured on mCCDA plates and sequenced as described previously (12, 27). Briefly, bacterial isolates were subcultured and grown overnight in a microaerophilic workstation (5% CO_2, 5% O_2, 3% H_2 and 87% N_2) at 42°C on Columbia Blood Agar (CBA) plates with 5% defibrinated horse blood (Oxoid, Basingstoke, UK). Colonies were picked onto fresh CBA plates and genomic DNA extraction was carried out using the QIAamp® DNA Mini Kit (Qiagen GmbH, Hilden, Germany) according to the manufacturer's instructions. DNA was eluted in 100-200 µl of the supplied buffer and stored at -20°C. Oxfordshire clinical isolates were cultured and DNA prepared as previously described (73).

Genome sequencing was performed using an Illumina HiSeq at the Wellcome Trust Sanger Institute, using the standard Illumina Indexing protocol involving fragmentation of 2 µg genomic DNA by acoustic shearing to enrich for 200 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) and a qPCR was used for final quantification of DNA sequencing libraries. Contiguous sequences (contigs) were assembled *de novo* using Velvet (74). Assembled genome files were archived in the Dryad repository (doi:10.5061/dryad.8t80s). Raw reads are available on the European Nucleotide Archive (ENA) and the Short Read Archive (SRA) (Table S1 for accession numbers).

Contiguous assemblies of whole genome sequences were individually archived on the web-based database platform BIGSdb (75). Briefly, individual genes from the *C. jejuni* strain NCTC11168 reference genome were locally aligned to all *Campylobacter* genomes using

default BLAST parameters implemented in BIGSdb. A gene was considered present when the local alignment had at least 70% sequence identity on at least 50% of the sequence length. This allowed gene discovery, sequence export and gene-by-gene alignments using MUSCLE (76), as previously described (77, 78).

Background population structure and clonal genealogy

The genome-wide association mapping approach infers statistically significant associations of genetic elements over-represented in one of two compared phenotype groups. To account for the clonal ancestry signal, the strength of each association is compared to its expectation under a simple model of evolution along the branches of the clonal genealogy which represents the background population structure. Clonal genealogies for ST-21 and ST-45 clonal complexes were inferred separately by ClonalFrame (79) which differentiates mutation and recombination events on each branch of the tree based on the density of polymorphisms. The program was run with 10,000 burn-in iterations followed by 10,000 sampling iterations for gene-by-gene alignments of core genes in ST-21 and ST-45 clonal complexes, separately.

Genome-wide association mapping

We adopt a similar approach to previously published genome-wide association studies (27, 31). Briefly, for each genome, the presence or absence of unique 30bp 'words' on the forward or reverse strand of any contiguous DNA sequence (or "contig"), was examined. This word-based method has the advantage that it detects both homologous and non-homologous sequence variation without requiring sequence alignments, accounting for frequent gain and loss of genetic material in bacterial genomes. An association score was calculated for each word as $a+d-(b+c)$, where a and b are the number of clinical isolates in which the word is present or absent, respectively; and c and d are the number of farm isolates in which the word is present or absent, respectively. To test significance of association of each word after controlling for the effect of population structure and clonal inheritance of genetic variants, the method computed p values by comparing the observed association score with a null distribution of the score (Figure S1) as detailed above. The null distribution was created by a Monte Carlo simulation with 10^6 replicates in which words were simulated to evolve through a process of gain and loss along the branches of a ClonalFrame phylogeny. The process of gain and loss was modelled so that the presence or absence of a word changed by any genetic mechanism on a branch with length d according to continuous-time Markov chain with a

probability of $1 - \frac{(1+\exp(-2dr))}{2}$; where r is rate (27), and an inverse of total branch length was used as the rate parameter. The null model assumes that presence/absence of a word is randomly changed in the phylogeny irrespective of the phenotype. It is expected that false positives are also included in the results by multiple testing (31). To account for multiple testing, only words with a p value below 5×10^{-4}, were considered as targets for further examination and experimental testing, and were mapped on the *C. jejuni* strain NCTC11168 reference genome as previously described (27, 31).

Statistical validation

In the analysis of the ST-21 clonal complex, the original dataset contained 117 UK human disease isolates. Although they were sampled only in UK (mostly in Oxford) and contained closely related strains, this large sample size allowed us to prepare two datasets for discovery and cross-validation of the genome-wide association mapping. For discovery we selected 20 clinical isolates consisting of 14 strains randomly selected from Oxford and all of the other 6 strains from the rest of the UK. For validation, we selected 24 human disease strains sampled from various lineages that were different from the 20 isolates (Figure S3). This dataset was prepared to be as independent as possible of the discovery dataset, and was examined to test whether the results of the discovery dataset were replicated and validated in terms of frequency increase of disease-associated genetic variations from farm to human disease. Similar cross-validation was not possible for the ST-45 isolates due to the limited sample size.

Disease-associated SNP clustering and comparison with the average linkage disequilibrium decay

The distance between successive disease-associated SNPs was compared with an expected distribution in the genomes of ST-21 and ST-45 clonal complexes, separately. The expected distribution was calculated based on randomly selecting SNPs with missing frequency <50% from all genes in the genomes. The same number of SNPs as the observed disease-associated SNPs was sampled 100 times, and distances between successive SNPs were calculated. The observed and expected distributions were illustrated together in the base 10 logarithmic scale by ggplot2 (80).

The linkage disequilibrium coefficient r^2, which measures correlation of alleles at two loci (81), was calculated between the associated SNPs in ST-21 and ST-45 clonal complexes,

separately. r^2 was also calculated between SNPs in core non-associated genes. Only bi-allelic SNPs without missing data were used for these calculations of r^2. Average r^2 values were then plotted against inter-SNP distances rounded to the nearest ten.

Consistency of association in other *Campylobacter* clonal complexes

For disease-associated genetic variations found above, we examined changes in their frequency from farmed chicken to human disease in isolates from nine clonal complexes including: ST-48 (4 farm and 28 clinical isolates), ST-257 (14 farm and 35 clinical), ST-353 (10 farm and 28 clinical), ST-354 (9 farm and 16 clinical), ST-443 (5 farm and 15 clinical), ST-52 (4 farm and 10 clinical), ST-574 (7 farm and 12 clinical), ST-607 (4 farm and 6 clinical) and ST-828 (42 farm and 52 clinical); in addition to the ST-21 and ST-45 complexes. We used gene-by-gene alignments of farm and disease strains in all 11 clonal complexes. We visualized them as a heatmap by using a function in the GMD package for R (82). To examine the consistency of the disease-associated genetic variation across different clonal complexes, we used the Mantel-Haenszel method (83) to calculate the pooled odds ratio and test its significance.

Generation of defined mutants of associated genes in *C. jejuni* reference strains

Nineteen genes containing genetic elements significantly associated with survival through the poultry processing (p<0.0005) were considered as candidates for further functional characterization using defined mutants (Table 1). Almost half (8) of the genes containing associated words were co-located on the chromosome in a 20 kbp region with poorly defined predicted functions. These, along with *nuoK* - which has a known role in oxygen response, were chosen for generation of defined mutants. Mutagenesis was performed in *C. jejuni* strain M1 (84) to generate Δ*nuoK* (*CJM1_1505*, *cj1569c*) and Δ*fumC* (*CJM1_1325*, *cj1364c*) deletion strains. Defined gene deletion mutants were obtained after allelic replacement of the target gene with a chloramphenicol (*cat*) resistance cassette, as described earlier (85). Briefly, the *cat* cassette was amplified by PCR from pCC027 (86) and the 5' and 3' flanking regions of the target gene were amplified by PCR from *C. jejuni* M1 genomic DNA. The PCR primers used to amplify the target gene flanking regions contain extensions complementary to the *cat* cassette. The *cat* cassette was integrated between the gene flanking regions in an overlap PCR without primers, and further amplified in a second round PCR in the presence of primers that amplify the whole fragment. The overlap PCR product was subsequently used for electroporation (87) of *C. jejuni* M1 to obtain first generation defined gene deletion

mutants. Genomic DNA of first generation gene deletion mutants was subsequently used for natural transformation (87) of M1 wild-type and the gene deletion was selected, yielding the gene deletion mutants used in functional assays. In addition, the M1 wild-type was processed in parallel through the natural transformation procedure without any added mutagenic DNA to obtain a 'coupled' wild-type strain. This was done to reduce the genetic variation between the wild-type strain and defined mutant strains.

For formate metabolism and nucleotide salvage assays, additional deletion mutants in *cj1377c* and *cj1368, cj1369, cj1370* and *cj1368-70*, were generated in *C. jejuni* NCTC11168 as follows. The gene was inactivated *in vitro* by deletion of most of the coding region and insertion of a kanamycin resistance cassette using the Gibson assembly method (88). Briefly, ~400 bp upstream and downstream gene flanks F1 and F2 were amplified using primers F1R1 and F2R2, respectively, with adapters homologous to either the kanamycin cassette amplified from pJMK30, or the ends of HincII linearised pGEM3ZF. An isothermal assembly reaction specifically anneals all 4 fragments together to yield the mutant plasmid. Wild-type *C. jejuni* NCTC11168 was transformed by electroporation and mutants, arising by double homologous recombination, selected for by kanamycin resistance. Correct insertion of the kanamycin cassette was confirmed by PCR. Primers and vectors used for all constructs are listed in Table S2.

Variable oxygen tension growth assays

Growth of the defined M1 mutants at variable oxygen concentrations was conducted as follows. Strains were grown from glycerol stocks on Columbia base agar plates, containing 5% v/v defibrinated horse blood and 10 μg ml^{-1} vancomycin and amphotericin B, overnight under standard microaerobic conditions (37°C, 10% v/v 0_2, 5% v/v CO_2, 85% v/v N_2). A total of 30 ml Muller-Hinton (MH) broth cultures were inoculated from plates and grown overnight under microaerophilic conditions with agitation. From these cultures, new 50 ml MH cultures were inoculated at an OD 600 nm of approximately 0.05 and transferred to orbital shakers at 160 rpm in either a microaerophilic (gas atmosphere as above) or fully aerobic 37°C incubator. For oxygen-limited growth, the 50 ml cultures were contained in ~50 ml flasks with minimal head-space in a 5% v/v O_2 5% v/v CO_2, 90% v/v N_2 atmosphere with slow orbital shaking (50 rpm) to severely reduce oxygen transfer. Samples were taken every 2 hours and the OD 600 nm measured to monitor growth.

Toxic nucleotide analogue growth curves and disk diffusion assays

For growth curves, overnight cultures of *C. jejuni* were adjusted to an OD_{600nm} of 0.1 in MH broth and growth monitored by sampling every 2 hours. For disk diffusion assays, overnight cultures of *C. jejuni* were used to seed MH agar to an OD of approximately 0.1 that was quickly poured and allowed to set. Sterile 5 mm filter paper disks were placed on the agar surface and 5 µl of concentrated nucleotide was added. The inhibition diameter was measured after 3 days incubation in standard microaerophilic conditions at 37°C. The toxic nucleotide analogues AZ (8-Azaxanthine monohydrate, Sigma), DP (2,6-Diaminopurine, Alfa Aesar), FU (5-Fluorouracil, Sigma), MP (6-Mercaptopurine, Sigma) and TG (6-Thioguanine, Sigma) were solubilised in DMSO and used at a final concentration of 0.1 mM for growth curves and 100 mM for disk diffusions. DMSO controls were used.

Determination of FdhA enzyme activity

FdhA-dependent formate oxidation was directly assayed by a methyl-viologen linked spectrophotometric assay. Overnight cultures were grown to an OD_{600nm} of at least 0.75 and concentrated 25-50 fold by centrifugation, washed and resuspended in 20 mM sodium phosphate buffer pH 7.5. The intact cell preparations were held on ice and total protein concentration was determined by Lowry assay in triplicate. An anaerobic cuvette containing 780 µl of 20 mM sodium phosphate buffer pH 7.5, 100 µl of 10 mM methyl-viologen and 100 µl of whole-cell suspension was sparged with argon for 6 minutes and placed into a Shimadzu recording spectrophotometer set at 37°C. The sample was zeroed at 585 nm and absorbance monitored for 10 s to ensure no background rate. The reaction rate was then measured for 180 s after addition of 20 µl 1 M sodium formate (argon sparged). FdhA activity was calculated as nmol of methyl-viologen reduced per min per mg of total protein. The experiment was performed as two biological replicates for each strain with three technical replicates.

Substrate-dependent oxygen respiration rates

Respiration of formate and the 2-oxoacids pyruvate and 2-oxoglutarate was measured in terms of dissolved oxygen consumption in a Clark-type oxygen electrode (Rank Brothers Ltd., Cambridge, UK). The electrode was first calibrated with air-saturated 20 mM sodium phosphate buffer pH 7.5, with 100% saturation assumed to be 220 µM O_2. The zero oxygen baseline was determined by the addition of sodium dithionite. The chamber was maintained at 37°C with constant stirring, and kept sealed with an airtight plug. Concentrated whole cell

suspensions were prepared, as above, and 50 µl of cells added to 2 ml of 20 mM sodium phosphate buffer pH 7.5 by injection through a central pore in the airtight plug. Once the background rate stabilised, 20 µl of substrate, either 1 M sodium formate, pyruvate or 2-oxoglutarate, was added and the rate of oxygen consumption measured over at least 1 minute. The total protein concentration of the whole cell preparations was determined by Lowry assay and the specific rate of substrate-linked oxygen consumption calculated as nmol of oxygen consumed per min per mg of total protein. All assays were performed in triplicate.

Funding information

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 grants 088786/C/09/Z and 098051. KY was supported by a JSPS Research Fellowship for Young Scientists. GM was supported by a NISCHR Health Research Fellowship (HF-14-13). AT was supported by a BBSRC DTP PhD studentship. XD was supported by BBSRC grant BB/L023458/1 and NIHR grant HPRU-2012-10080. SdV was supported by BBSRC grant RG66581.

Acknowledgments

Authors declare no competing financial interests. Computational calculations were performed at the Human Genome Center of the Institute of Medical Science (University of Tokyo, Japan) and at HPC Wales (UK).

Availability of Data and Materials

Assembled genome files are archived in the Dryad repository (doi:10.5061/dryad.8t80s). Raw reads are available on the European Nucleotide Archive (ENA) and the Short Read Archive (SRA) (see Table S1 for accession numbers).

Author contributions

KY and GM contributed equally. KY, GM, XD and SKS conceived the study. KY, GM, SM, AJT, SPWdV, BP, LM, AT, SK, CDB, AG, DM, DJK and SKS designed experiments, generated and analysed results. BP, HW, AV, AR, AJC, FMC, NMcC, DH, JEB, KAJ, MCJM, SDB and JP, contributed bacterial samples and sequenced whole genomes. KY, GM, AJT, DJK and SKS wrote the manuscript. All authors helped in the interpretation of results and commented on the manuscript before submission.

Figure and table legends

Figure 1. Survival of *Campylobacter* lineages through the poultry processing chain. Each line represents prevalence of *C. jejuni* for different STs from (A) the two main host-generalist lineages ST-21 and ST-45 clonal complex (blue and red lines, respectively), and (B) other major clonal complexes. In panel B, increasing or decreasing prevalence throughout the poultry processing chain was indicated with pink and green lines, respectively. The source isolation information of 7,343 isolates from the pubMLST database (as of June 2013) was examined, with a total of 1,497 farm/caecum isolates, 1,256 abattoir/retail meat isolates and 5,941 clinical isolates. Lineages were shown when constituting at least 5% prevalence in at least one of the three process stages, which amounted to 5,428 isolates in total. Of these, a total of 1,464 isolates were from ST-21 complex, 842 from ST-45, 949 from ST-257, 355 from ST-48, 284 from ST-354, 308 from ST-574, 313 from ST-443, 235 from ST-573, 204 from ST-661, 140 from ST-61, 125 from ST-464, 105 from ST-607, 55 from ST-658 and 49 from ST-1034 complex.

Figure 2. Location of poultry processing chain survival-associated elements in *C. jejuni* ST-21 and ST-45 clonal complex isolate genomes. Circular genomic map from the *C. jejuni* reference strain NCTC11168, with black lines indicating annotated coding regions. Numbers indicate positions along the genome in Mbp. The map is overlaid with genetic elements ('words') resulting from the genome-wide association study with a statistical increase in prevalence in clinical isolates compared to chicken faeces/caecum isolates, for ST-21 clonal complex isolates (red) and ST-45 clonal complex isolates (blue).

Figure 3. Distribution of sequence variants associated with the survival of ST-21 and ST-45 complex isolates through the poultry processing chain. ClonalFrame genealogy and distribution of disease-associated alleles in the isolates used for the association mapping analysis of (A) ST-21 clonal complex and (B) ST-45 clonal complex. Isolates from chicken faeces/caecum are indicated with grey circles and clinical isolates with red circles. Gene names with corresponding associated SNPs, are based on the *C. jejuni* strain NCTC11168 nomenclature. (C) Changes in frequency of the associated alleles shown in panels A and B in *C. jejuni* ST-21, ST-48, ST-443, ST-257, ST-574, ST-52, ST-353, ST-607, ST-354 and ST-45 clonal complexes, and *C. coli* ST-828 clonal complex. The phylogeny above the plot is

based on representative isolates from each clonal complex. The red colours in the heatmap indicate a frequency increase from farm to clinical, while the blue colours indicate frequency decrease, as shown in the colour legend at the bottom of the plot. The grey colours indicate that frequency of the disease-associated nucleotide is 0% or 100% in both farm and human disease isolates.

Figure 4. Genomic distance and linkage disequilibrium of SNPs associated with survival from farm to human disease in *C. jejuni* ST-21 and ST-45 complexes. Observed distribution of distance between successive disease-associated SNPs of (A) ST-21 and (B) ST-45 complexes compared with expected distribution in the genome. Linkage disequilibrium (LD) was calculated between the disease-associated SNPs of (C) ST-21 and (D) ST-45 complexes compared with the average LD decay in the core SNPs. The y-axis is the linkage disequilibrium coefficient (r^2).

Figure 5. Phenotypic investigation of genotypes associated with survival through the poultry processing chain. (A) Comparison of formate dehydrogenase activity between *C. jejuni* strain NCTC11168, an isogenic Δ*cj1377c* knock-out mutant, and a selection of farm (n=9) and clinical isolates (n=7) used in the genetic association, bars indicate average distributions for each condition. (B) Growth of defined nucleotide salvage *C. jejuni* strain NCTC11168 mutants under standard microaerobic conditions, represented by the OD_{600nm} 8 hours after inoculation (n=3). (C) Growth of *C. jejuni* strain NCTC11168 wildtype and isogenic triple mutant Δ*cj1368-70* in the presence either 0.1mM 6-Mercaptopurine (MP), 0.1mM 6-Thioguanine (TP), an equivalent volume of DMSO or no-addition control. Values represent the average OD_{600nm} 6 hours after inoculation (n=3). Statistical significance was analyses using unpaired *t*-tests with * $p<0.05$, ** $p<0.01$

Figure 6. Effects of *nuoK* and *fumC* deletion on aerobic growth and 2-oxoacid respiration. (A) Growth of *C. jejuni* strain M1 wildtype and isogenic *nuoK* and *fumC* mutants under various oxygen atmospheres. Values represent the average OD_{600nm} after 12 hours (for oxygen limited and microaerobic) or 6 hours (for aerobic) (n = 2). (B) Distribution of *nuoK* allelic types; the number of clinical (red bars) and farm (blue bars) isolates harbouring particular alleles is shown. The allelic type numbers are arbitrary and indicate different nucleotide sequences at the *nuoK* locus. (C) Physiological roles of NuoK and FumC. NuoK (dark blue) is a proton-translocating subunit of the 14-subunit Nuo complex, which

oxidises reduced flavodoxin derived from 2-oxoacids by Oor (solid lines) and possibly Por (dashed lines) enzymes. Cj1574 and Cj1575 are two unique subunits that replace the NADH dehydrogenase components in conventional Nuo complexes (orange). FumC forms part of the CAC and converts fumarate to malate. (D) Oxygen-linked respiration rates of 2-oxoacids by *C. jejuni* strain M1 wildtype and isogenic *nuoK* and *fumC* mutants as measured by oxygen electrode. The control substrate formate was used to show that these mutants had similar formate respiration rates as the wild-type. Values represent the average of 3 independent experiments. Statistical significance was analyses using unpaired *t*-tests with * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Figure S1. Null and empirical distributions of the association score in ST-45 clonal complex.

Figure S2. Distribution of squared correlation coefficient (linkage disequilibrium r^2) values between SNPs across *C. jejuni* genomic regions longer than 100 kb. (A) Core and (B) disease-associated SNPs with 25% minor allele frequency are used for calculation.

Figure S3. Clinical strains from ST-21 clonal complex analysed in this study. Clinical strains used for discovery and replication in GWAS are coloured in red and pink, respectively. The tree was constructed from the core genes in ST-21 clonal complex by using FastTree.

Figure S4. Sensitivity of *C. jejuni* NCTC11168 and nucleotide salvage mutants to toxic nucleotide analogues. (A) *C. jejuni* NCTC11168 wildtype was assessed for its sensitivity to the toxic nucleotide analogues AZ (8-Azaxanthine monohydrate), DP (2,6-Diaminopurine), FU (5-Fluorouracil), MP (6-Mercaptopurine) and TG (6-Thioguanine) with a DMSO control. Of those tested, sensitivity was only seen towards MP and TG. * $p<0.05$, *** $p<0.001$ (B) Disc diffusion assays were used to evaluate the resistance of nucleotide salvage mutants *cj1368-70* to 6-Mercaptopurine (MP) and 6-Thioguanine (TP). Each single mutant, and a triple mutant, were completely resistant compared to wildtype.

Figure S5. Prevalence of associated words in *Campylobacter* genomes from various sources. (A) Prevalence on 354 genomes from cattle (n=43) chicken (n=300) and wild birds

(n=11). Genomes have been previously published or are part of this study. (B) Prevalence on the 209 clinical genomes sampled before 2010, in 2010, and in 2011.

Figure S6. Frequency of strains carrying pairs of SNPs in high LD on the same contigs. For each inter-locus pair of SNPs with r^2>0.9 among genes in Table 1, frequency of strains that have the pair on the same contigs was calculated and shown in Y-axis. The 47 and 45 isolates of ST-21 and ST-45 clonal complexes used for GWAS discovery were examined, respectively.

Table 1. Genes containing associated elements and their predicted functions and functional categories.

Table S1. Description of sequenced isolates used in this study.

Table S2. Primers and vectors used for mutant construction in this study.

Table S3. Summary of the classification of the survival-associated words.

Table S4. **Allelic types and predicted protein variants of the nuoK gene in 44 ST-21 clonal complex and 37 ST-45 clonal complex isolates used in the GWAS and phenotypical testing**.

Table S5. List and description of disease-associated SNPs identified in this study.

References

1. **Taylor LH, Latham SM, Woolhouse ME.** 2001. Risk factors for human disease emergence. Philos Trans R Soc Lond B Biol Sci **356:**983-989.
2. **Karesh WB, Dobson A, Lloyd-Smith JO, Lubroth J, Dixon MA, Bennett M, Aldrich S, Harrington T, Formenty P, Loh EH, Machalaba CC, Thomas MJ, Heymann DL.** 2012. Ecology of zoonoses: natural and unnatural histories. Lancet **380:**1936-1945.
3. **Mead PS, Slutsker L, Griffin PM, Tauxe RV.** 1999. Food-related illness and death in the united states reply to dr. hedberg. Emerg Infect Dis **5:**841-842.
4. **Garcia S, Heredia NL.** 2013. Campylobacter. *In* Labbé RG, García S (ed.), Guide to Foodborne Pathogens. John Wiley & Sons, Oxford.

5. **Buzby JC, Roberts T.** 1997. Economic costs and trade impacts of microbial foodborne illness. World Health Stat Q **50:**57-66.
6. **Friedman CR, Hoekstra RM, Samuel M, Marcus R, Bender J, Shiferaw B, Reddy S, Ahuja SD, Helfrick DL, Hardnett F, Carter M, Anderson B, Tauxe RV.** 2004. Risk factors for sporadic Campylobacter infection in the United States: A case-control study in FoodNet sites. Clin Infect Dis **38 Suppl 3:**S285-296.
7. **Neimann J, Engberg J, Molbak K, Wegener HC.** 2003. A case-control study of risk factors for sporadic campylobacter infections in Denmark. Epidemiol Infect **130:**353-366.
8. **Sheppard SK, Dallas JF, Strachan NJ, MacRae M, McCarthy ND, Wilson DJ, Gormley FJ, Falush D, Ogden ID, Maiden MC, Forbes KJ.** 2009. Campylobacter genotyping to determine the source of human infection. Clin Infect Dis **48:**1072-1078.
9. **Allen VM, Bull SA, Corry JE, Domingue G, Jorgensen F, Frost JA, Whyte R, Gonzalez A, Elviss N, Humphrey TJ.** 2007. Campylobacter spp. contamination of chicken carcasses during processing in relation to flock colonisation. Int J Food Microbiol **113:**54-61.
10. **Herman L, Heyndrickx M, Grijspeerdt K, Vandekerchove D, Rollier I, De Zutter L.** 2003. Routes for Campylobacter contamination of poultry meat: epidemiological study from hatchery to slaughterhouse. Epidemiol Infect **131:**1169-1180.
11. **Klein G, Beckmann L, Vollmer HM, Bartelt E.** 2007. Predominant strains of thermophilic Campylobacter spp. in a German poultry slaughterhouse. Int J Food Microbiol **117:**324-328.
12. **Sheppard SK, Cheng L, Meric G, de Haan CP, Llarena AK, Marttinen P, Vidal A, Ridley A, Clifton-Hadley F, Connor TR, Strachan NJ, Forbes K, Colles FM, Jolley KA, Bentley SD, Maiden MC, Hanninen ML, Parkhill J, Hanage WP, Corander J.** 2014. Cryptic ecology among host generalist Campylobacter jejuni in domestic animals. Mol Ecol **23:**2442-2451.
13. **Sheppard SK, Colles F, Richardson J, Cody AJ, Elson R, Lawson A, Brick G, Meldrum R, Little CL, Owen RJ, Maiden MC, McCarthy ND.** 2010. Host association of Campylobacter genotypes transcends geographic variation. Appl Environ Microbiol **76:**5269-5277.
14. **Sheppard SK, Dallas JF, MacRae M, McCarthy ND, Sproston EL, Gormley FJ, Strachan NJ, Ogden ID, Maiden MC, Forbes KJ.** 2009. Campylobacter genotypes from food animals, environmental sources and clinical disease in Scotland 2005/6. Int J Food Microbiol **134:**96-103.
15. **Duim B, Wassenaar TM, Rigter A, Wagenaar J.** 1999. High-resolution genotyping of Campylobacter strains isolated from poultry and humans with amplified fragment length polymorphism fingerprinting. Appl Environ Microbiol **65:**2369-2375.
16. **Fitzgerald C, Stanley K, Andrew S, Jones K.** 2001. Use of pulsed-field gel electrophoresis and flagellin gene typing in identifying clonal groups of Campylobacter jejuni and Campylobacter coli in farm and clinical environments. Appl Environ Microbiol **67:**1429-1436.
17. **Hanninen ML, Perko-Makela P, Pitkala A, Rautelin H.** 2000. A three-year study of Campylobacter jejuni genotypes in humans with domestically acquired infections and in chicken samples from the Helsinki area. J Clin Microbiol **38:**1998-2000.
18. **Wilson DJ, Gabriel E, Leatherbarrow AJ, Cheesbrough J, Gee S, Bolton E, Fox A, Fearnhead P, Hart CA, Diggle PJ.** 2008. Tracing the source of campylobacteriosis. Plos Genet **4:**e1000203.

19. **Hastings R, Colles FM, McCarthy ND, Maiden MC, Sheppard SK.** 2011. Campylobacter genotypes from poultry transportation crates indicate a source of contamination and transmission. J Appl Microbiol **110:**266-276.
20. **Johnsen G, Kruse H, Hofshagen M.** 2006. Genotyping of Campylobacter jejuni from broiler carcasses and slaughterhouse environment by amplified fragment length polymorphism. Poult Sci **85:**2278-2284.
21. **Newell DG, Shreeve JE, Toszeghy M, Domingue G, Bull S, Humphrey T, Mead G.** 2001. Changes in the carriage of Campylobacter strains by poultry carcasses during processing in abattoirs. Appl Environ Microbiol **67:**2636-2640.
22. **Colles FM, McCarthy ND, Sheppard SK, Layton R, Maiden MC.** 2010. Comparison of Campylobacter populations isolated from a free-range broiler flock before and after slaughter. Int J Food Microbiol **137:**259-264.
23. **Butzler JP.** 2004. Campylobacter, from obscurity to celebrity. Clin Microbiol Infect **10:**868-876.
24. **Solomon EB, Hoover DG.** 2004. Inactivation of Campylobacter jejuni by high hydrostatic pressure. Lett Appl Microbiol **38:**505-509.
25. **Murphy C, Carroll C, Jordan KN.** 2006. Environmental survival mechanisms of the foodborne pathogen Campylobacter jejuni. J Appl Microbiol **100:**623-632.
26. **Park SF.** 2002. The physiology of Campylobacter species and its relevance to their role as foodborne pathogens. Int J Food Microbiol **74:**177-188.
27. **Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, Bentley SD, Maiden MC, Parkhill J, Falush D.** 2013. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. Proc Natl Acad Sci U S A **110:**11923-11927.
28. **Chewapreecha C, Marttinen P, Croucher NJ, Salter SJ, Harris SR, Mather AE, Hanage WP, Goldblatt D, Nosten FH, Turner C, Turner P, Bentley SD, Parkhill J.** 2014. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. Plos Genet **10:**e1004547.
29. **Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, Sloan TJ, Williams P, Endres JL, Bayles KW, Fey PD, Yajjala VK, Widhelm T, Hawkins E, Lewis K, Parfett S, Scowen L, Peacock SJ, Holden M, Wilson D, Read TD, van den Elsen J, Priest NK, Feil EJ, Hurst LD, Josefsson E, Massey RC.** 2014. Predicting the virulence of MRSA from its genome sequence. Genome Res **24:**839-849.
30. **Alam MT, Petit RA, 3rd, Crispell EK, Thornton TA, Conneely KN, Jiang Y, Satola SW, Read TD.** 2014. Dissecting vancomycin-intermediate resistance in staphylococcus aureus using genome-wide association. Genome Biol Evol **6:**1174-1185.
31. **Pascoe B, Meric G, Murray S, Yahara K, Mageiros L, Bowen R, Jones NH, Jeeves RE, Lappin-Scott HM, Asakura H, Sheppard SK.** 2015. Enhanced biofilm formation and multi-host transmission evolve from divergent genetic backgrounds in Campylobacter jejuni. Environ Microbiol.
32. **McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN.** 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet **9:**356-369.
33. **Falush D, Bowden R.** 2006. Genome-wide association mapping in bacteria? Trends Microbiol **14:**353-355.
34. **Karlyshev AV, Linton D, Gregson NA, Lastovica AJ, Wren BW.** 2000. Genetic and biochemical evidence of a Campylobacter jejuni capsular polysaccharide that accounts for Penner serotype specificity. Mol Microbiol **35:**529-541.

35. **Gaynor EC, Wells DH, MacKichan JK, Falkow S.** 2005. The Campylobacter jejuni stringent response controls specific stress survival and virulence-associated phenotypes. Mol Microbiol **56:**8-27.
36. **Cameron A, Frirdich E, Huynh S, Parker CT, Gaynor EC.** 2012. Hyperosmotic stress response of Campylobacter jejuni. J Bacteriol **194:**6116-6130.
37. **Li XZ, Nikaido H.** 2009. Efflux-mediated drug resistance in bacteria: an update. Drugs **69:**1555-1623.
38. **Jeon B, Wang Y, Hao H, Barton YW, Zhang Q.** 2011. Contribution of CmeG to antibiotic and oxidative stress resistance in Campylobacter jejuni. J Antimicrob Chemother **66:**79-85.
39. **Bayliss CD, Bidmos FA, Anjum A, Manchev VT, Richards RL, Grossier JP, Wooldridge KG, Ketley JM, Barrow PA, Jones MA, Tretyakov MV.** 2012. Phase variable genes of Campylobacter jejuni exhibit high mutation rates and specific mutational patterns but mutability is not the major determinant of population structure during host colonization. Nucleic Acids Res **40:**5876-5889.
40. **Kale A, Phansopa C, Suwannachart C, Craven CJ, Rafferty JB, Kelly DJ.** 2011. The virulence factor PEB4 (Cj0596) and the periplasmic protein Cj1289 are two structurally related SurA-like chaperones in the human pathogen Campylobacter jejuni. J Biol Chem **286:**21254-21265.
41. **Everitt RG, Didelot X, Batty EM, Miller RR, Knox K, Young BC, Bowden R, Auton A, Votintseva A, Larner-Svensson H, Charlesworth J, Golubchik T, Ip CL, Godwin H, Fung R, Peto TE, Walker AS, Crook DW, Wilson DJ.** 2014. Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus. Nat Commun **5:**3956.
42. **Islam Z, van Belkum A, Wagenaar JA, Cody AJ, de Boer AG, Sarker SK, Jacobs BC, Talukder KA, Endtz HP.** 2014. Comparative population structure analysis of Campylobacter jejuni from human and poultry origin in Bangladesh. Eur J Clin Microbiol Infect Dis.
43. **Weerakoon DR, Borden NJ, Goodson CM, Grimes J, Olson JW.** 2009. The role of respiratory donor enzymes in Campylobacter jejuni host colonization and physiology. Microb Pathog **47:**8-15.
44. **Ferenci T.** 2005. Maintaining a healthy SPANC balance through regulatory and mutational adaptation. Mol Microbiol **57:**1-8.
45. **Kelly DJ.** 2008. Complexity and versatility in the physiology and metabolism of Campylobacter jejuni. ASM Press, Washington, DC, USA.
46. **Kendall JJ, Barrero-Tobon AM, Hendrixson DR, Kelly DJ.** 2014. Hemerythrins in the microaerophilic bacterium Campylobacter jejuni help protect key iron-sulphur cluster enzymes from oxidative damage. Environ Microbiol **16:**1105-1121.
47. **Colles FM, McCarthy ND, Bliss CM, Layton R, Maiden MC.** 2015. The long-term dynamics of Campylobacter colonizing a free-range broiler breeder flock: an observational study. Environ Microbiol **17:**938-946.
48. **Bentley SD, Parkhill J.** 2004. Comparative genomic structure of prokaryotes. Annu Rev Genet **38:**771-792.
49. **Lefebure T, Stanhope MJ.** 2009. Pervasive, genome-wide positive selection leading to functional divergence in the bacterial genus Campylobacter. Genome Res **19:**1224-1232.
50. **Myers JD, Kelly DJ.** 2005. A sulphite respiration system in the chemoheterotrophic human pathogen Campylobacter jejuni. Microbiology **151:**233-242.
51. **Bereswill S, Fischer A, Plickert R, Haag LM, Otto B, Kuhl AA, Dasti JI, Zautner AE, Munoz M, Loddenkemper C, Gross U, Gobel UB, Heimesaat MM.** 2011.

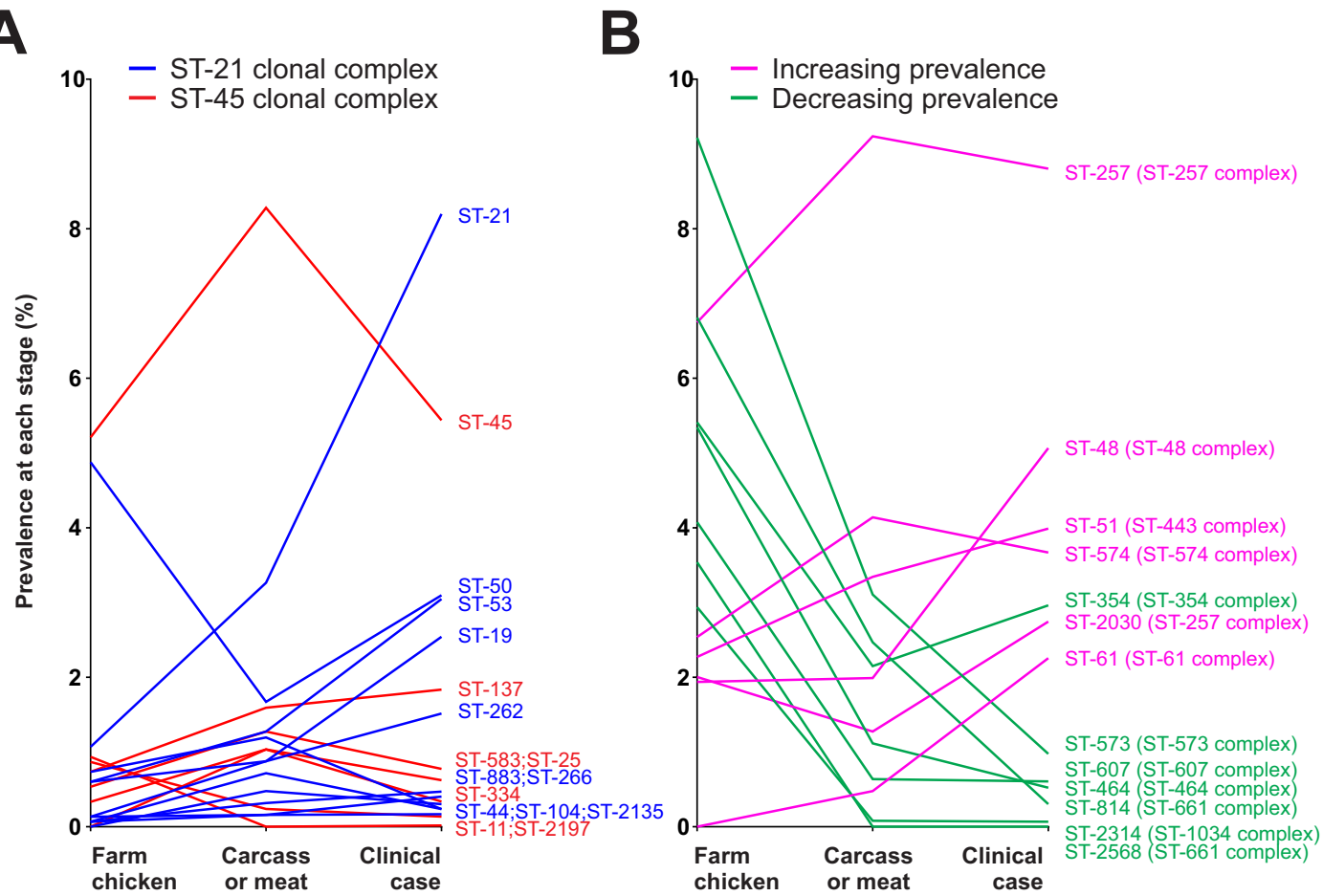
Novel murine infection models provide deep insights into the "menage a trois" of Campylobacter jejuni, microbiota and host innate immunity. PLoS One **6:**e20953.

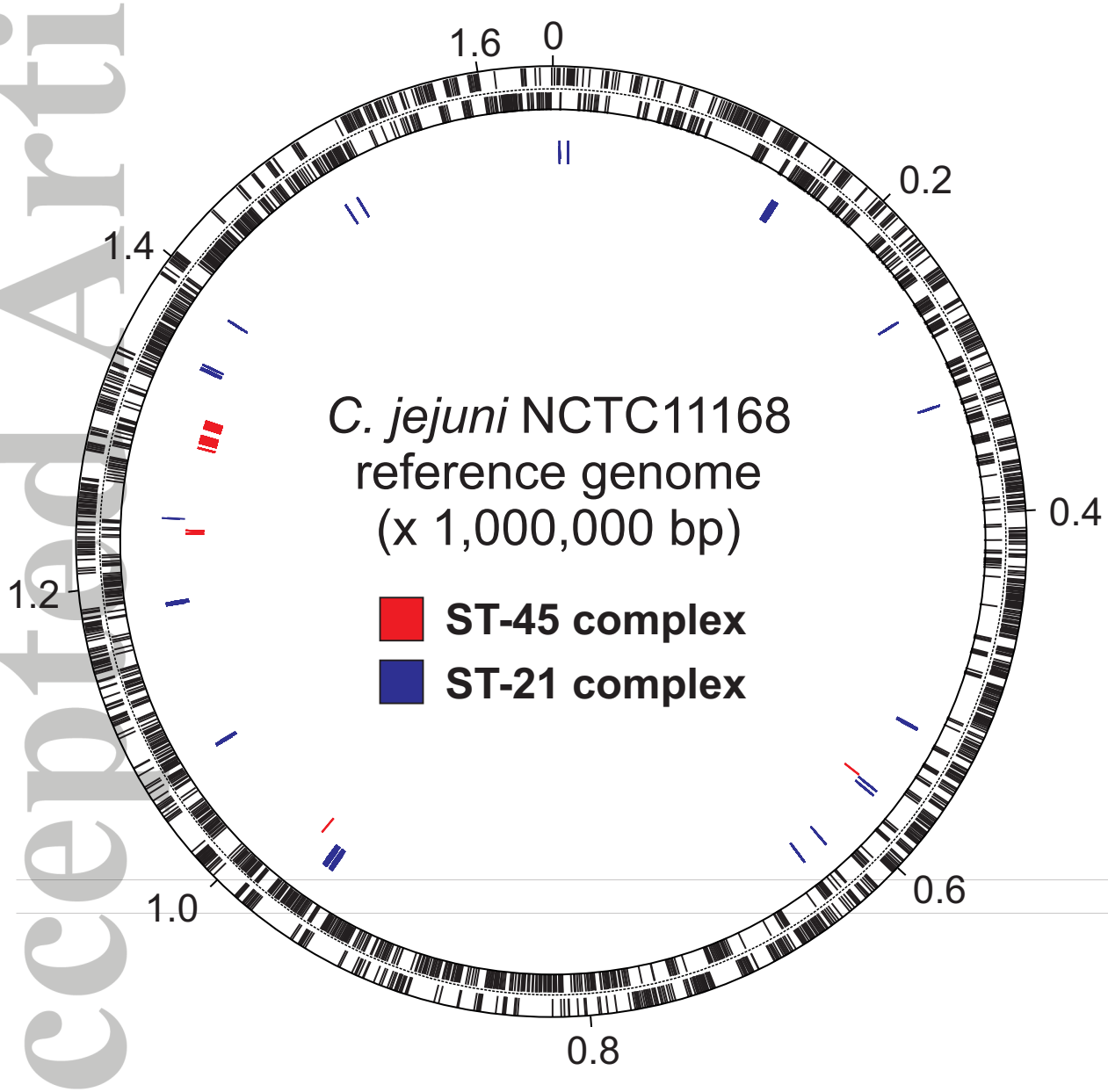
52. **Pan N, Imlay JA.** 2001. How does oxygen inhibit central metabolism in the obligate anaerobe Bacteroides thetaiotaomicron. Mol Microbiol **39:**1562-1571.
53. **Li Y, Powell DA, Shaffer SA, Rasko DA, Pelletier MR, Leszyk JD, Scott AJ, Masoudi A, Goodlett DR, Wang X, Raetz CR, Ernst RK.** 2012. LPS remodeling is an evolved survival strategy for bacteria. Proc Natl Acad Sci U S A **109:**8716-8721.
54. **Cochrane JC, Lipchock SV, Smith KD, Strobel SA.** 2009. Structural and chemical basis for glucosamine 6-phosphate binding and activation of the glmS ribozyme. Biochemistry **48:**3239-3246.
55. **Klein DJ, Ferre-D'Amare AR.** 2006. Structural basis of glmS ribozyme activation by glucosamine-6-phosphate. Science **313:**1752-1756.
56. **Yeom J, Lee Y, Park W.** 2012. Effects of non-ionic solute stresses on biofilm formation and lipopolysaccharide production in Escherichia coli O157:H7. Res Microbiol **163:**258-267.
57. **Budin-Verneuil A, Pichereau V, Auffray Y, Ehrlich DS, Maguin E.** 2005. Proteomic characterization of the acid tolerance response in Lactococcus lactis MG1363. Proteomics **5:**4794-4807.
58. **Karlyshev AV, McCrossan MV, Wren BW.** 2001. Demonstration of polysaccharide capsule in Campylobacter jejuni using electron microscopy. Infect Immun **69:**5921-5924.
59. **Young NM, Brisson JR, Kelly J, Watson DC, Tessier L, Lanthier PH, Jarrell HC, Cadotte N, St Michael F, Aberg E, Szymanski CM.** 2002. Structure of the N-linked glycan present on multiple glycoproteins in the Gram-negative bacterium, Campylobacter jejuni. J Biol Chem **277:**42530-42539.
60. **Karlyshev AV, Champion OL, Churcher C, Brisson JR, Jarrell HC, Gilbert M, Brochu D, St Michael F, Li J, Wakarchuk WW, Goodhead I, Sanders M, Stevens K, White B, Parkhill J, Wren BW, Szymanski CM.** 2005. Analysis of Campylobacter jejuni capsular loci reveals multiple mechanisms for the generation of structural diversity and the ability to form complex heptoses. Mol Microbiol **55:**90-103.
61. **Karlyshev AV, Quail MA, Parkhill J, Wren BW.** 2013. Unusual features in organisation of capsular polysaccharide-related genes of C. jejuni strain X. Gene **522:**37-45.
62. **Bacon DJ, Szymanski CM, Burr DH, Silver RP, Alm RA, Guerry P.** 2001. A phase-variable capsule is involved in virulence of Campylobacter jejuni 81-176. Mol Microbiol **40:**769-777.
63. **Guerry P, Poly F, Riddle M, Maue AC, Chen YH, Monteiro MA.** 2012. Campylobacter polysaccharide capsules: virulence and vaccines. Frontiers in cellular and infection microbiology **2:**7.
64. **van Alphen LB, Wenzel CQ, Richards MR, Fodor C, Ashmus RA, Stahl M, Karlyshev AV, Wren BW, Stintzi A, Miller WG, Lowary TL, Szymanski CM.** 2014. Biological roles of the O-methyl phosphoramidate capsule modification in Campylobacter jejuni. PLoS One **9:**e87051.
65. **de Haan CP, Llarena AK, Revez J, Hanninen ML.** 2012. Association of Campylobacter jejuni metabolic traits with multilocus sequence types. Appl Environ Microbiol **78:**5550-5554.
66. **Habib I, Louwen R, Uyttendaele M, Houf K, Vandenberg O, Nieuwenhuis EE, Miller WG, van Belkum A, De Zutter L.** 2009. Correlation between genotypic diversity, lipooligosaccharide gene locus class variation, and caco-2 cell invasion

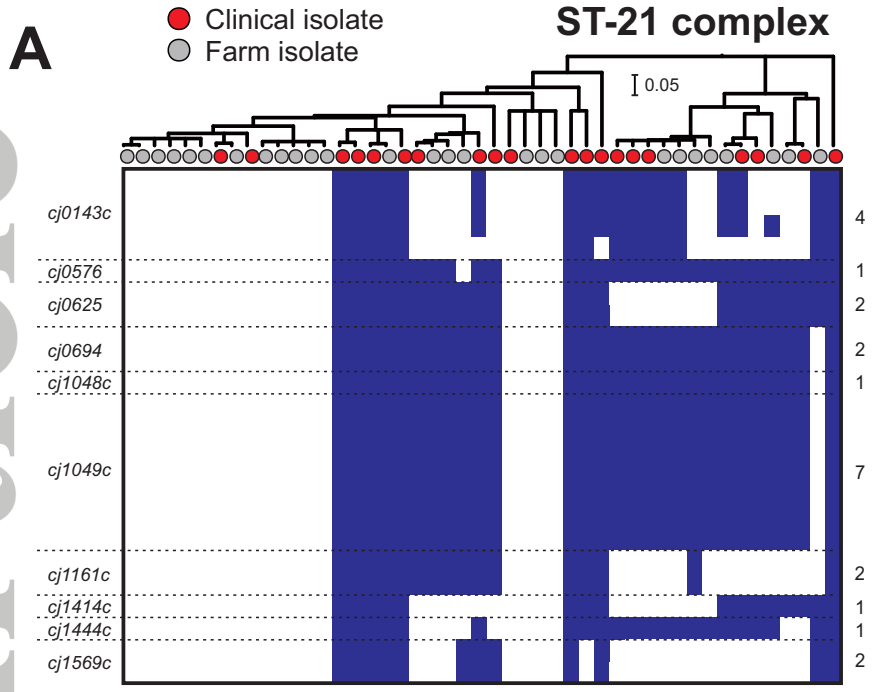
potential of Campylobacter jejuni isolates from chicken meat and humans: contribution to virulotyping. Appl Environ Microbiol **75:**4277-4288.

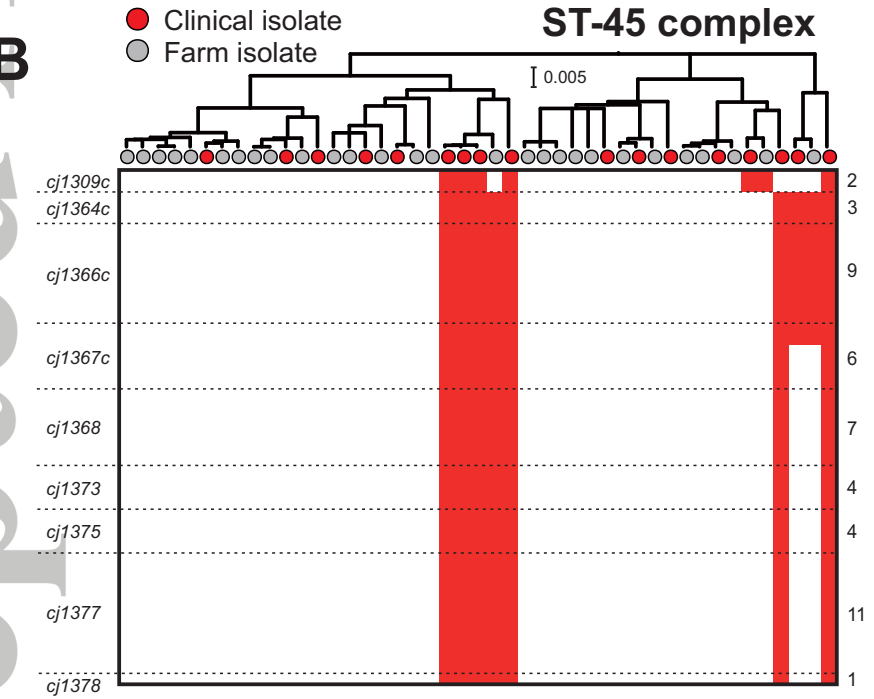
67. **Griekspoor P, Olsen B, Waldenstrom J.** 2009. Campylobacter jejuni in penguins, Antarctica. Emerg Infect Dis **15:**847-848.
68. **Sopwith W, Birtles A, Matthews M, Fox A, Gee S, Painter M, Regan M, Syed Q, Bolton E.** 2008. Identification of potential environmentally adapted Campylobacter jejuni strain, United Kingdom. Emerg Infect Dis **14:**1769-1773.
69. **Wimalarathna HM, Richardson JF, Lawson AJ, Elson R, Meldrum R, Little CL, Maiden MC, McCarthy ND, Sheppard SK.** 2013. Widespread acquisition of antimicrobial resistance among Campylobacter isolates from UK retail poultry and evidence for clonal expansion of resistant lineages. BMC Microbiol **13:**160.
70. **Lawes JR, Vidal A, Clifton-Hadley FA, Sayers R, Rodgers J, Snow L, Evans SJ, Powell LF.** 2012. Investigation of prevalence and risk factors for Campylobacter in broiler flocks at slaughter: results from a UK survey. Epidemiol Infect **140:**1725-1737.
71. **Vidal AB, Rodgers J, Arnold M, Clifton-Hadley F.** 2013. Comparison of different sampling strategies and laboratory methods for the detection of C. jejuni and C. coli from broiler flocks at primary production. Zoonoses and public health **60:**412-425.
72. **Powell LF, Lawes JR, Clifton-Hadley FA, Rodgers J, Harris K, Evans SJ, Vidal A.** 2012. The prevalence of Campylobacter spp. in broiler flocks and on broiler carcases, and the risks associated with highly contaminated carcases. Epidemiol Infect **140:**2233-2246.
73. **Cody AJ, McCarthy ND, Jansen van Rensburg M, Isinkaye T, Bentley SD, Parkhill J, Dingle KE, Bowler IC, Jolley KA, Maiden MC.** 2013. Real-time genomic epidemiological evaluation of human Campylobacter isolates by use of whole-genome multilocus sequence typing. J Clin Microbiol **51:**2526-2534.
74. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res **18:**821-829.
75. **Jolley KA, Maiden MC.** 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics **11:**595.
76. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32:**1792-1797.
77. **Meric G, Kemsley EK, Falush D, Saggers EJ, Lucchini S.** 2013. Phylogenetic distribution of traits associated with plant colonization in Escherichia coli. Environ Microbiol **15:**487-501.
78. **Sheppard SK, Jolley KA, Maiden MCJ.** 2012. A Gene-By-Gene Approach to Bacterial Population Genomics: Whole Genome MLST of Campylobacter. Genes **3:**261-277.
79. **Didelot X, Falush D.** 2007. Inference of bacterial microevolution using multilocus sequence data. Genetics **175:**1251-1266.
80. **Wickham H.** 2010. ggplot2: Elegant Graphics for Data Analysis. Journal of Statistical Software **35**.
81. **Hill WG, Robertson A.** 1968. Linkage desequilibrium in finite populations. Theor Appl Genet **38:**226-231.
82. **Zhao X, Sandelin A.** 2012. GMD: measuring the distance between histograms with applications on high-throughput sequencing reads. Bioinformatics **28:**1164-1165.
83. **Mantel N, Haenszel W.** 1959. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst **22:**719-748.

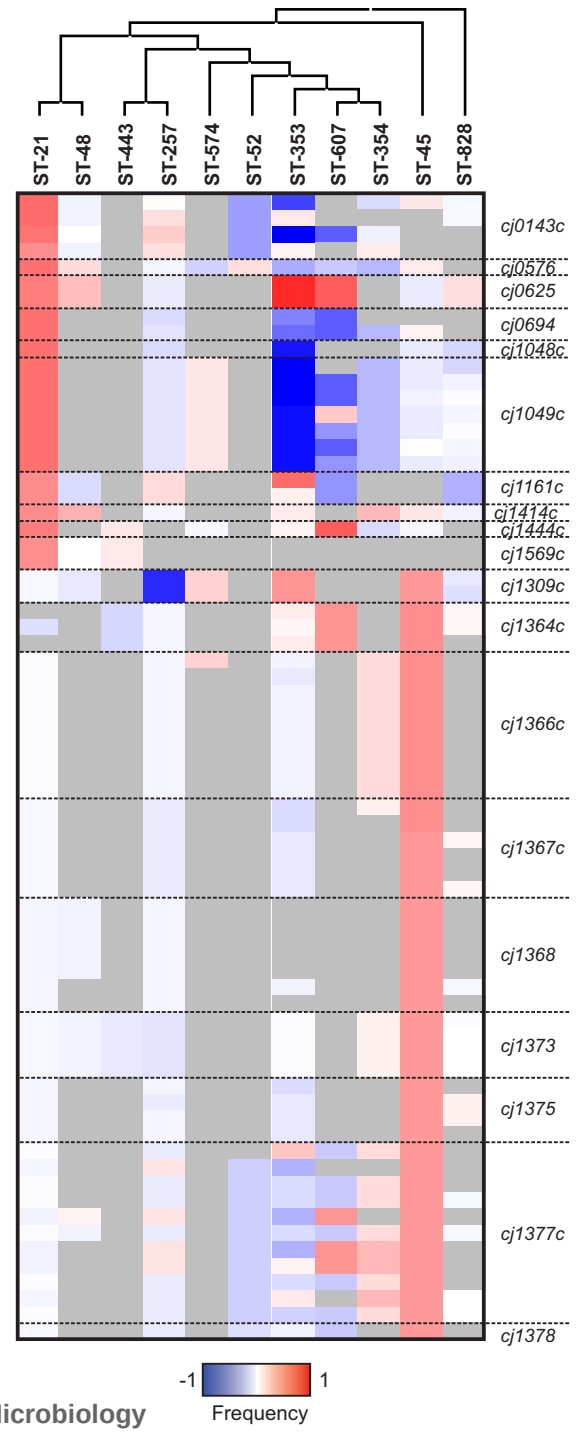
84. **Friis C, Wassenaar TM, Javed MA, Snipen L, Lagesen K, Hallin PF, Newell DG, Toszeghy M, Ridley A, Manning G, Ussery DW.** 2010. Genomic characterization of Campylobacter jejuni strain M1. PLoS One **5:**e12253.
85. **de Vries SP, Gupta S, Baig A, L'Heureux J, Pont E, Wolanska DP, Maskell DJ, Grant AJ.** 2015. Motility defects in Campylobacter jejuni defined gene deletion mutants caused by second-site mutations. Microbiology **161:**2316-2327.
86. **Coward C, van Diemen PM, Conlan AJ, Gog JR, Stevens MP, Jones MA, Maskell DJ.** 2008. Competing isogenic Campylobacter strains exhibit variable population structures in vivo. Appl Environ Microbiol **74:**3857-3867.
87. **Holt JP, Grant AJ, Coward C, Maskell DJ, Quinlan JJ.** 2012. Identification of Cj1051c as a major determinant for the restriction barrier of Campylobacter jejuni strain NCTC11168. Appl Environ Microbiol **78:**7841-7848.
88. **Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA, 3rd, Smith HO.** 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods **6:**343-345.

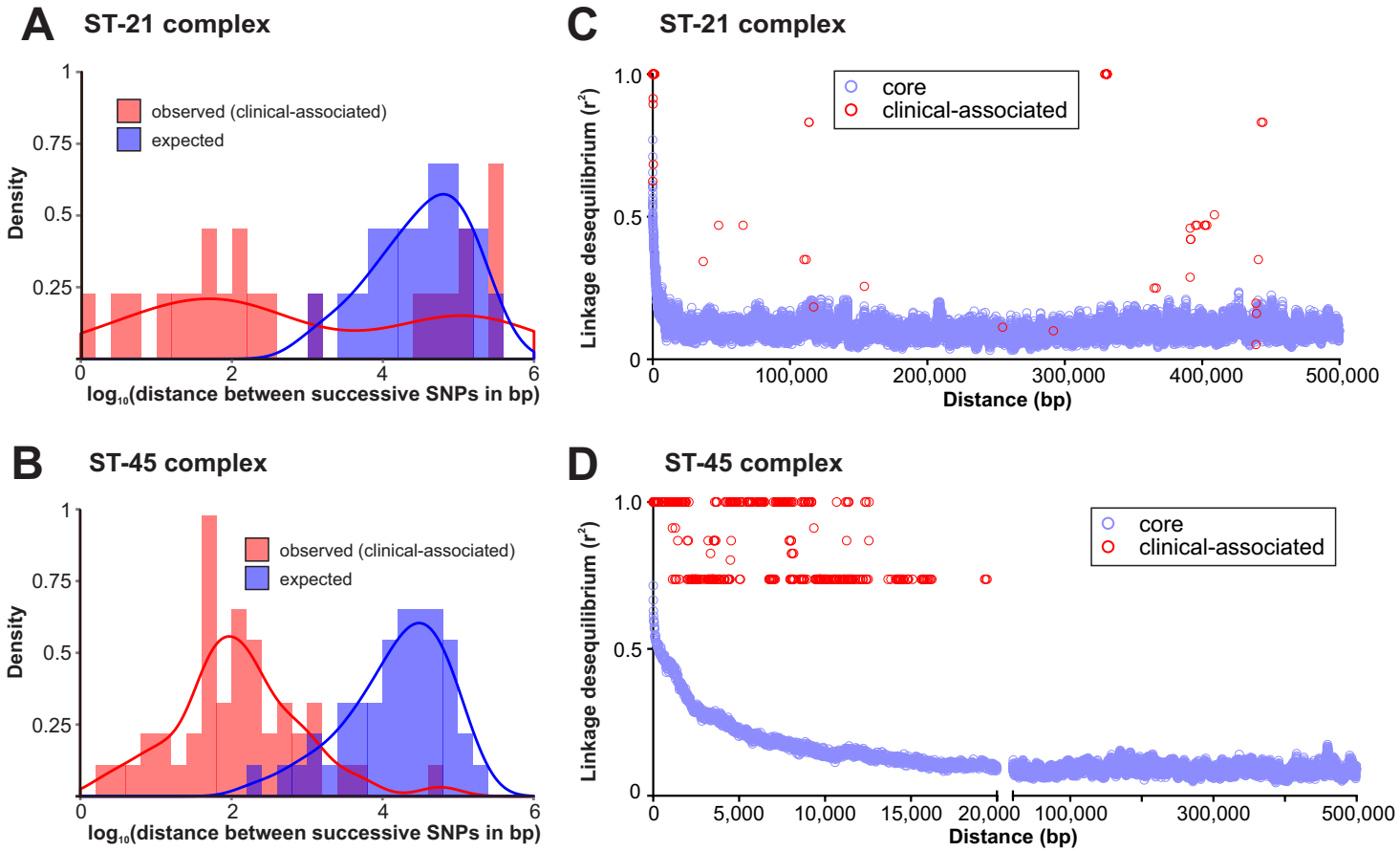


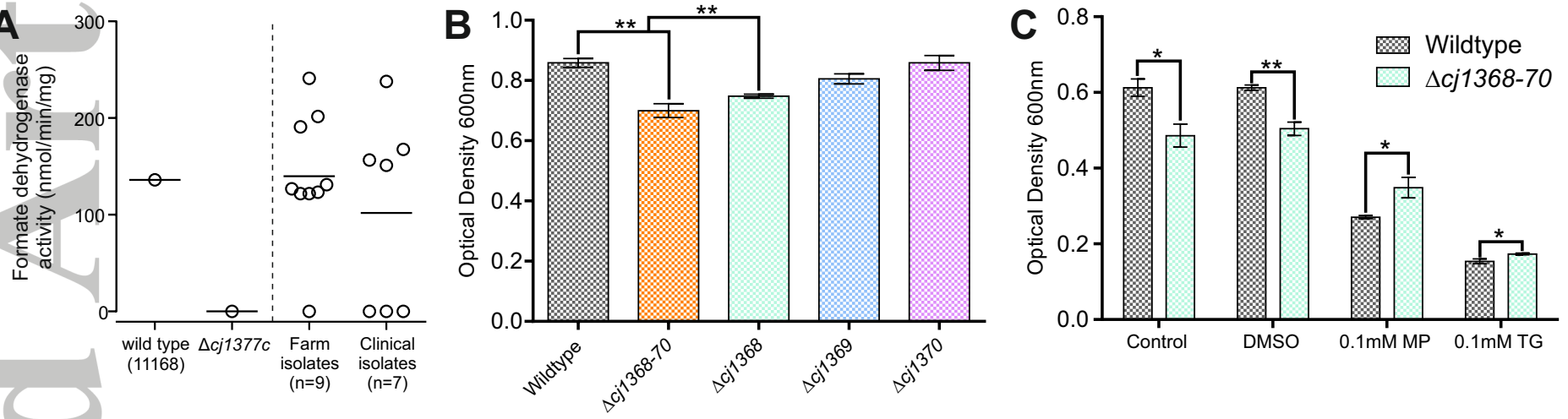


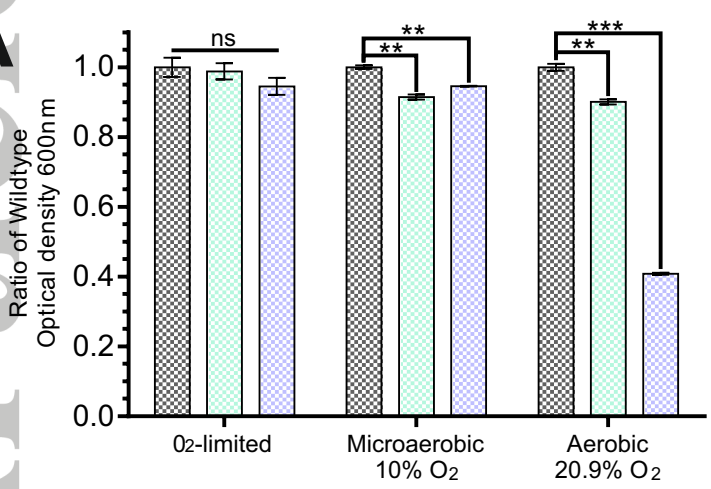


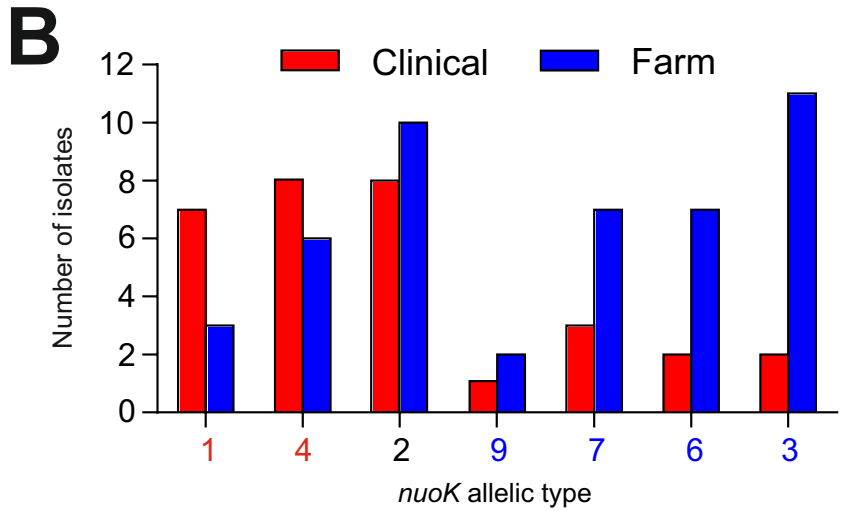


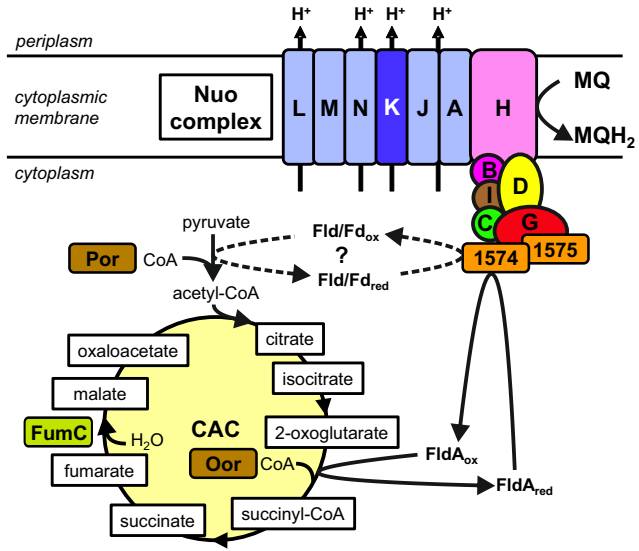












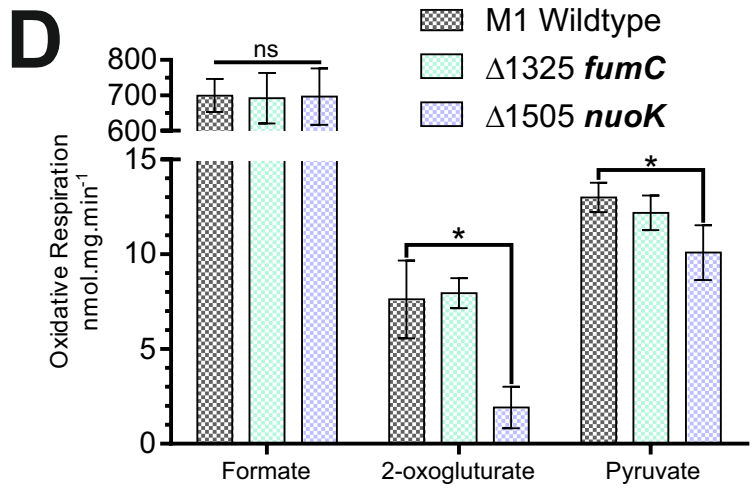


Table 1. Genes containing associated elements and their predicted functions and functional categories.

Gene name	Alias	Predicted function[a]	Transcriptional unit number[b]	Genomic position[c]	Associated in	UniProt identifier	Gene ontology (GO)	Gene ontology (cellular component)	Gene ontology IDs
cj0143c	*znuA*	Putative periplasmic solute binding protein for ABC transport system	55	145,616	ST-21	Q0PBZ4	metal ion binding; metal ion transport	-	GO:0046872; GO:0030001
cj0576	*lpxD*	UDP-3-O-acylglucosamine N-acyltransferase (EC 2.3.1.-)	226	537,023	ST-21	Q9PHU0	lipid A biosynthetic process; transferase activity, transferring acyl groups other than amino-acyl groups	-	GO:0009245; GO:0016747
cj0625	*hypD*	Hydrogenase isoenzymes formation protein	238	585,102	ST-21	Q0PAP2	metal ion binding	-	GO:0046872
cj0694	*ppiD*	Putative periplasmic protein	264	651,043	ST-21	Q0PAI5	isomerase activity	-	GO:0016853
cj1048c	*dapE*	Succinyl-diaminopimelate desuccinylase (SDAP desuccinylase) (EC 3.5.1.18) (N-succinyl-LL-2,6-diaminoheptanedioate amidohydrolase)	395	980,554	ST-21	Q0P9K4	cobalt ion binding; diaminopimelate biosynthetic process; lysine biosynthetic process via diaminopimelate; metallopeptidase activity; succinyl-diaminopimelate desuccinylase activity; zinc ion binding	-	GO:0050897; GO:0019877; GO:0009089; GO:0008237; GO:0009014; GO:0008270
cj1049c	-	Putative LysE family transporter protein	395	981,655	ST-21	Q0P9K3	amino acid transport; integral component of membrane; plasma membrane	integral component of membrane; plasma membrane	GO:0006865; GO:0016021; GO:0005886
cj1051c	*cjeI*	Restriction modification enzyme	395	982,991	ST-21	Q0P9K1	DNA binding; DNA methylation; N-methyltransferase activity	-	GO:0003677; GO:0006306; GO:0008170
cj1161c	-	Putative cation-transporting ATPase	428	1,091,795	ST-21	Q0P995	ATP binding; cation-transporting ATPase activity; integral component of membrane; metal ion binding; metal ion transport	integral component of membrane	GO:0005524; GO:0019829; GO:0016021; GO:0046872; GO:0030001
cj1309c	-	Uncharacterized protein	492	1,238,581	ST-45	Q0P8U8	-	-	0
cj1364c	*fumC*	Fumarate hydratase class II (Fumarase C) (EC 4.2.1.2)	509	1,296,244	ST-45	O69294	fumarate hydratase activity; fumarate metabolic process; tricarboxylic acid cycle; tricarboxylic acid cycle enzyme complex	tricarboxylic acid cycle enzyme complex	GO:0004333; GO:0006106; GO:0006099; GO:0045239
cj1366c	*glmS*	Glutamine--fructose-6-phosphate aminotransferase [isomerizing] (EC 2.6.1.16) (D-fructose-6-phosphate amidotransferase) (GFAT) (Glucosamine-6-phosphate synthase) (Hexosephosphate aminotransferase) (L-glutamine--D-fructose-6-phosphate amidotransferase)	509	1,300,819	ST-45	Q9PMT4	carbohydrate binding; carbohydrate biosynthetic process; cytoplasm; glutamine-fructose-6-phosphate transaminase (isomerizing) activity; glutamine metabolic process	cytoplasm	GO:0030246; GO:0016051; GO:0005737; GO:0006541; GO:0004360
cj1367c	-	Putative nucleotidyltransferase	509	1,302,620	ST-45	Q0P8P2	transferase activity	-	GO:0016740
cj1368	-	Putative radical SAM domain protein	510	1,305,112	ST-45	Q0P8P1	4 iron, 4 sulfur cluster binding; menaquinone biosynthetic process; transferase activity, transferring alkyl or aryl (other than methyl) groups	-	GO:0051539; GO:0009234; GO:0016765
cj1373	-	Putative ntegral membrane protein	510	1,309,284	ST-45	Q0P8N6	integral component of membrane	integral component of membrane	GO:0016021
cj1375	-	Putative multidrug efflux transporter	512	1,312,555	ST-45	Q0P8N4	integral component of membrane; plasma membrane; transmembrane transport; transporter activity	integral component of membrane; plasma membrane	GO:0016021; GO:0005886; GO:0055085; GO:0005215
cj1377c	-	Putative ferredoxin	513	1,314,649	ST-45	Q0P8N2	iron-sulfur cluster binding	-	GO:0051536
cj1378	*selA*	L-seryl-tRNA(Sec) selenium transferase (EC 2.9.1.1) (Selenocysteine synthase) (Sec synthase) (Selenocysteinyl-tRNA(Sec) synthase)	514	1,316,388	ST-45	Q9PMS2	cytoplasm; L-seryl-tRNASec selenium transferase activity; pyridoxal phosphate binding; selenocysteine incorporation; selenocysteinyl-tRNA(Sec) biosynthetic process	cytoplasm	GO:0004125; GO:0005737; GO:0030170; GO:0001514; GO:0097056
cj1414c	*kpsC*	Capsule polysaccharide modification protein	529	1,346,283	ST-21	Q0P8K0	polysaccharide biosynthetic process; polysaccharide transport	-	GO:0000271; GO:0015774
cj1444c	*kpsD*	Capsule polysaccharide export system periplasmic protein	532	1,383,486	ST-21	Q0P8H0	membrane; polysaccharide transmembrane transporter activity	membrane	GO:0016020; GO:0015159
cj1569c	*nuoK*	NADH-quinone oxidoreductase subunit K (EC 1.6.99.5) (NADH dehydrogenase I subunit K) (NDH-1 subunit K)	578	1,501,081	ST-21	Q0P859	ATP synthesis coupled electron transport; integral component of membrane; NADH dehydrogenase (quinone) activity; plasma membrane; quinone binding; transport	integral component of membrane; plasma membrane	GO:0042773; GO:0050136; GO:0016021; GO:0005886; GO:0048038; GO:0006810

a. As defined on the UniProt database
b. In the *C. jejuni* NCTC11168 reference genome according to ProOpDB. Two identical numbers reflect co-transcription of the corresponding genes in the same transcriptional unit.
c. Starting position on the *C. jejuni* NCTC11168 genome sequence