Gene loss and lineage specific restriction-modification systems associated with niche differentiation in the *Campylobacter jejuni* Sequence Type 403 clonal complex.

Laura Morley[1], Alan McNally[1], Konrad Paszkiewicz[2], Jukka Corander[3], Guillaume Méric[4], Samuel K. Sheppard[4,5,6], Jochen Blom[7], Georgina Manning[1*]

[1]Pathogen Research Group, School of Science and Technology, Nottingham Trent University, Nottingham, UK, NG11 8NS;

[2]Department of Biosciences, University of Exeter, Exeter, UK

[3]Department of Mathematics and Statistics, University of Helsinki, P.O. Box 68, FI-00014 University of Helsinki, Finland;

[4]College of Medicine, Institute of Life Science, Swansea University, Swansea, SA2 8PP;

[5]MRC CLIMB Consortium, Swansea University, Institute of Life Science, Swansea, SA2 8PP, UK

[6]Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

[7]Bioinformatics and Systems Biology, Justus Liebig University, Giessen, Germany

[*]Corresponding author: Dr Georgina Manning, Pathogen Research Group, Nottingham Trent University, Clifton Lane, Nottingham, NG11 8NS, georgina.manning@ntu.ac.uk, 0044 1158483373.

Running title: Gene gain and gene loss in niche restriction of a *C. jejunii* lineage

Abstract

Campylobacter jejuni is a highly diverse species of bacteria commonly associated with infectious intestinal disease of humans and zoonotic carriage in poultry, cattle, pigs, and other animals. The species contains a large number of distinct clonal complexes that vary from host generalist lineages commonly found in poultry livestock and human disease cases, to host-adapted specialised lineages primarily associated with livestock or poultry. Here we present novel data on the ST-403 clonal complex of *C. jejuni*, a lineage that has not been reported in avian hosts. Our data show this lineage exhibits a distinctive pattern of intra-lineage recombination that is accompanied by the presence of lineage specific restriction-modification systems. Furthermore we show that the ST-403 complex has undergone gene decay at a number of loci. Our data provides a putative link between the lack of association with avian hosts of *C. jejuni* ST403 and with both gene gain and gene loss through non-sense mutations in coding sequences of genes resulting in pseudogene formation.

Introduction

Campylobacter is a common component of the gut microbiota of many avian and mammalian species where it is often considered a commensal organism as it is typically carried without obvious disease symptoms. While diarrhoeal infections are rarely recorded in animals (1, 2), they are extremely common in humans where the majority of infections are caused by *Campylobacter jejuni* (3). Human *C. jejuni* infection can originate in multiple reservoirs but it is known that a large proportion of human *C. jejuni* cases are attributed to chicken (4), typically through handling of raw meat, cross contamination or direct consumption of undercooked meat. However this does not account for all cases of campylobacteriosis and it is clear that isolates from other sources and species can infect humans.

The ubiquity of *Campylobacter* poses interesting questions about its ecology and infection biology (5). *C. jejuni* and *C. coli* have been isolated from numerous avian and mammalian species including food production animals such as poultry, pigs and cattle (6, 7), as well as in companion animals including cats and dogs (2). Wild birds, faecally contaminated ground and surface waters, and drinking water are also reservoirs for *C. coli* and *C. jejuni* (8, 9). While both species are widely distributed, disease-causing *C. coli* is most commonly associated with food production mammals, especially pigs, but improving knowledge of the population structure and evolution of these organisms is challenging some of the traditional ideas. It is now clear that both *C. jejuni* and *C. coli* are frequently isolated from multiple species (10, 11) and understanding lineage distribution across multiple hosts is an important current objective in *Campylobacter* research.

Multi-locus sequence typing (MLST) has been performed on a large number of *C. jejuni* isolates from clinical samples, veterinary sources, abattoir surveys and environmental sources (4, 9, 10, 12). These studies have revealed the existence of host restricted and host generalist lineages (5) with considerable overlap of some lineages (Sequence type, ST complexes) that

are found in both animal and clinical samples (4, 12). This host associated genetic structuring has formed the basis of quantitative attribution studies that estimate the relative contribution of different reservoir hosts to human disease (12, 13). Generalist lineages have also been used to investigate cryptic niche structure (5) and host-specific signals of genetic import in *Campylobacter* (14). However, less work has focussed specifically on host-restricted lineages, such as the ST-403 complex (10). A previous MLST study of UK abattoir isolates found of a lack of poultry isolates within this ST complex, with 89% from pigs (16/18, the remaining two being associated with cattle).

Lateral gene transfer plays a significant role in bacterial evolution, with the gain of DNA from another lineage potentially conferring novel function and driving bacterial evolution (15, 16). *Campylobacter* is usually considered to be a highly recombinogenic organism (17, 18) with homologous recombination introducing as much as eight times more DNA polymorphism than mutation alone. Over time, recombination between lineages has the potential to blur the boundaries between clonal complexes or even between *C. jejuni* and *C. coli* (19).

The aim of this study is to investigate *C. jejuni* ST-403 complex isolates at the genome level. We report the lack of any ST-403 complex strains isolated from avian host species, the primary reservoir of *C. jejuni*. This lineage exhibits a specific core-genome recombination pattern with little apparent exchange of DNA outside of the ST403 complex lineage. This is possibly the result of lineage-specific restriction-modification systems. In addition, a number of loci present in a large number of non ST-403 complex *C. jejuni* isolates were shown to have undergone lineage specific decay and pseudogenisation, a mechanism previously not reported in hypothesised niche restriction events in *Campylobacter*. Together our data provide information on evolutionary events that have contributed to the formation of a lineage of *C. jejuni* that is seemingly not colonising avian hosts.

Material and Methods

Bacterial strains and growth conditions

The thirteen *C. jejuni* ST403-complex strains used in this study are listed in Table 1. *Campylobacter* strains were stored at -80°C in Mueller-Hinton broth containing 20% (v/v) glycerol until required. *Campylobacter* strains were cultured from -80°C freezer stocks onto mCCDA (Oxoid) and incubated for 48 hours microaerobically in a gas jar with the addition of a CampyGen sachet (Oxoid, UK) at 37°C prior to use.

DNA Extraction and genome sequencing

Genomic DNA was prepared from overnight agar cultures by harvesting the entire plate growth, re-suspension in sterile PBS, and then classical phenol:chloroform extraction using phase lock tubes (5Prime). Sequencing was performed on the Illumina Hiseq 2500 platform using 100bp paired-end sequencing. De novo assemblies were performed using Velvet (20) and improved using the PAGIT suite of programmes (21). Genomes were annotated using PROKKA (22).

Phylogenetic inference

For population structure analyses, the 13 ST-403 clonal complex genomes were augmented with a dataset of 126 genomes of *C. jejuni* and 60 *C. coli* genomes previously published and characterised (14). Core genome alignments were produced using MAFFT (23) on 595 genes that were present to 80% nucleotide level identity in every individual genome (5) and concatenated to produce a core genome (24). Trees were reconstructed using an approximation of the maximum likelihood algorithm implemented in FastTree2 (25).

Comparative genomics

The thirteen ST-403 complex genomes were aligned using Mugsy, with a phylogenetic tree constructed from the extracted SNPs using FastTree. Comparative genomics of the ST-403

112 complex was performed using EDGAR (26). Iterative BLAST searches were conducted in
113 EDGAR to produce a pan-genome for the 13 genomes listed in Table 1 and a further 21
114 reference genomes of *C. jejuni* and *C. coli* (Table S1). The resulting pan-genome was
115 subsequently filtered to identify coding sequences unique to ST-403 complex (present in
116 100% of ST-493 genomes and 0% non-ST-403 genomes with an 80% nucleotide identity
117 cutoff), and coding sequences absent or divergent from ST-403 complex (present in 0% ST-
118 403 genomes and > 20% non-ST-403 genomes with an 8% nucleotide identity cutoff). The
119 putative function of these regions was determined by BLASTx against the entire NCBI non-
120 redundant database. Loci identified as ST-403 complex unique and ST-403 complex absent
121 were validated by searching for their presence within the entire BIGSdb *Campylobacter*
122 database by BLASTn, using the default parameters in BIGSdb.

123 **Recombination analysis**

124 To estimate the amount of recombination in the core genome of ST-403 complex strains in
125 relation to the remaining *C. jejuni* population we used the BratNextGen software (27) on the
126 core genome alignment of all 139 *C. jejuni* genomes used in our phylogenetic inference. A
127 total of 20 iterations of HMM parameter estimation were performed and significant (p-value
128 not exceeding 5%) recombinations were obtained with 100 parallel permutation runs
129 executed on a cluster computer. The negligible changes in HMM parameter values observed
130 already after approximately 30% of the iterations indicated sufficient convergence in the
131 estimation procedure.

132

133 **Results**

134

135 ***C. jejuni* ST-403 clonal complex is a distinct lineage within the species with no**
136 **catalogued avian isolation**

We examined the host source of ST-403 complex isolates in the *Campylobacter* MLST database (http://pubmlst.org/campylobacter/). A total of 278 ST-403 complex isolates, representing 1.22% of the entire database (accessed 19/08/2014), were composed of 173 from human clinical cases, 82 from pigs and 23 from cattle, with no isolates from avian sources recorded in the database. Core genome alignments were produced using thirteen ST-403 complex isolates and 186 previously published genomes of *C. jejuni* and *C. coli* (14). The resulting maximum likelihood phylogeny showed that the ST-403 clonal complex is a *C. jejuni* lineage that sits clearly within the *C. jejuni* species (Figure 1), despite the absence of any catalogued isolations from avian hosts. A separate alignment of the ST-403 complex genomes identified 2,831 SNPs across the clade, with pig, human and cattle isolates intermixed (Fig 2).

Identification of lineage specific restriction-modification systems in the ST-403 complex

We sought to determine the presence of clade-specific genes that may underpin the observed absence of isolation from avian hosts. EDGAR was used to create a pan-genome of the thirteen ST-403 complex strains, and twenty-one reference *C. jejuni* and *C. coli* genomes (Table S1). The pan-genome was mined to determine loci unique to the ST-403 complex strains, and any identified loci were then searched for across the entire *Campylobacter* BIGSdb genomic database to confirm their restriction to the ST-403 complex. From this analysis a total of ten ST-403 complex unique loci were identified (Table 2). Of the ten ST-403 complex unique CDS, seven putatively encoded hypothetical proteins and one encoded a putative Recombination F protein. The remaining CDS encoded two putative type II restriction-modification systems, R.HinP1I restriction endonuclease and Modification Methylase HhaI, and R.Pab1 restriction endonuclease. BLASTx comparisons showed the former R-M system to be orthologous to a system found in *Helicobacter cinaedi*, and the latter R-M system to be orthologous to a system found in *Helicobacter pylori*.

Given the presence of these unique R-M systems across the entire ST-403 complex lineage we sought to determine if there was an accompanying effect on the levels of detectable core genome recombination within ST-403 complex strains. BRATNextGen was used to detect recombination events across the *C. jejuni* core genome alignment constructed for phylogenetic testing (Fig 3). The resulting recombination profile shows a distinctive pattern of recombination events in the ST-403 complex that is composed primarily of intra-lineage events. Phylogenetic trees were reconstructed on the core genome alignment with all recombination removed (Supp Fig 1), and on the regions identified as recombination events in the ST-403 complex (Supp Fig 2). Both phylogenies show tight clustering of the ST-403 complex strains with 0.964 bootstrap support for the clustering of the ST403 recombining regions. Combined these data suggest that the recombination occurring in the ST-403 complex is predominantly lineage-specific.

Evidence of gene decay in the *C. jejuni* ST-403 complex

Further analysis of the pan-genome identified a total of fourteen loci that were absent or divergent in every ST-403 complex genome and present in at least 20% of the non-ST-403 complex genomes included in the analysis (Table 3). To allow a more detailed comparison of the nature of absence or divergent loci, the sequence for each was extracted from a relevant reference genome sequence and used to perform pairwise BLAST comparisons against each of the ST-403 complex genomes. This confirmed that six of the loci were completely absent from all of the CC403 genomes. More importantly the remaining eight loci all showed patterns of pseudogenisation and gene decay across the ST-403 complex with five of those loci containing identical pseudogenisation events in every genome (Table 3). Loci C8J_0199 – 0200 had been merged into a single ORF by mutation removing the stop codon delineating the two ORFs in the reference genomes leading to single polypeptide, followed by a second mutation just downstream introducing a stop codon, whilst the other four loci contained

multiple stop codon mutations which were common across all the ST-403 complex genomes, and a single locus containing a deletion common across the lineage. The remaining three loci contained multiple insertions, deletions and SNPs which varied across the ST-403 complex but which could result in the loss of gene function of that CDS across the lineage. As the 5 loci showing conserved patterns of pseudogenisation represent ST-403 unique alleles of those CDS, we searched for matching alleles in the entire BIGSdb *Campylobacter* database using BLAST. Among the isolates contained in the BIGS database, no alleles that matched these 5 loci with >70% nucleotide identity over >50% of the sequence length contained identical mutations to those in the ST-403 isolates, further suggesting that these evolutionary events are associated with the ST-403 complex.

Discussion

In this study we investigated the *C. jejuni* ST-403 complex, a lineage of *C. jejuni* that has never reportedly been isolated from an avian host. In our initial MLST study (10) identifying this lineage, 16 of the 18 ST-403 complex isolates were from pigs with the other two from bovine sources, leading to the hypothesis that this was a pig-adapted clone. Subsequent mining of the MLST database has revealed the presence of isolates from other sources including cattle, and a large number of human isolates within the ST-403 complex, the majority of which were isolated from the Dutch Antilles (28). This indicates that isolates from this complex have the capacity to cause human disease. However the most important observation is that no ST-403 complex isolates from poultry have been recorded in PubMLST, suggesting that the ST403 complex may be less well adapted to avian hosts, or represents a lineage of *C. jejuni* that has not evolved the ability to colonise avian hosts as well as the many other *C. jejuni* lineages.

211 Recent studies of the population structure and ecology of *Campylobacter* have indicated the
212 presence of generalist lineages such as the ST-21 and ST-45 complexes, which contain
213 isolates from multiple sources, as well as specialist lineages, such as the ST-61 and ST-42
214 complexes that have been reported to be associated with cattle (9, 10, 29), or the ST-354, ST-
215 443, ST-353 and ST-257 complexes that are associated with poultry (9). It is also known that
216 within the generalist lineages there are sublineages with evidence of host association (5, 14)
217 indicating that in some cases adaptation to a particular host might still be occurring. It is
218 possible therefore that ST-403 complex represents another specialist lineage of *C. jejuni* that
219 has evolved to become less suited to colonisation of the avian host. This would seem more
220 plausible than the possibility that ST-403 has not evolved the ability to colonise avian hosts,
221 given its central position in the *C. jejuni* species phylogeny, as this would require multiple
222 lineages of *C. jejuni* sharing MRCA with ST-403 independently evolving to become efficient
223 avian colonisers whilst ST-403 did not.
224 We investigated clade-specific loci and identified three restriction-modification (R-M) loci
225 that were unique to the ST-403 complex. Strain specific R-M systems have been reported
226 previously in *Campylobacter jejuni* strains 81116 (30), ATCC43431 (31) and 81176 (32) and
227 are thought to contribute to the apparent recombination and transformation restriction that has
228 hindered genetic manipulation of this organism for some time. Mutagenesis of the Type IIG
229 R-M enzyme Cj1051c in NCTC11168 increased this strain's ability to take up plasmid DNA,
230 including that from *E. coli* (33). Single nucleotide polymorphisms in known R-M systems in
231 the Japanese ST-4526 clone are thought to be responsible for the reduced uptake of plasmid
232 DNA when compared to NCTC11168 (34) as well as contributing to the ability of this clone
233 to thrive in Japan. In other organisms such as *Neisseria meningitidis* R-M systems have also
234 been reported to play key roles in the formation of structured phylogenetic clades and
235 patterns of recombination (35). The distinct recombination pattern of the ST-403 complex

236 isolates showed within lineage recombination, as evidenced by the tight phylogenetic
237 clustering of the recombinant regions. Recombination is thought to play an important role in
238 niche adaptation and acquisition of a host signature (36). However this recombination
239 appears to be restricted as it was recently reported that two major generalist lineages (ST-21
240 and ST-45 complexes) have limited recombination with each other, but readily recombine
241 with other specialist, host-adapted lineages (5).
242 Besides the presence of ST-403 complex specific R-M systems there are a number of coding
243 sequences that are absent from isolates within this complex or have degraded when compared
244 with homologues in other *C. jejuni* strains. It is not possible here to determine if the genes
245 were present in the common ST-403 complex ancestor and were deleted through time, or
246 were never present in the ancestral lineage. However the high prevalence of these absent
247 genes across *C. jejuni* and some *C. coli* clades suggests that they have most likely been lost in
248 the CC403 lineage through time, a hypothesis supported by the presence of a central deletion
249 in locus C8J_0806 which is identical across the ST-403 complex. What is clear is that the ST-
250 403 complex shows signs of lineage-specific mutations in distinct loci. There are several
251 possible explanations for these findings but one possible evolutionary scenario would be that
252 the selection pressure at these loci changed with a move away from an avian host reservoir
253 and that mutations resulting in loss of function have increased in the ST-403 complex
254 because they do not influence fitness. Three loci appear to be undergoing a similar process
255 with multiple independent deletions and mutations across the ST-403 complex, possibly
256 suggesting that the process is ongoing. Interestingly, none of these loci have clearly
257 identifiable functions which one may associate with avian colonisation or indeed niche
258 adaptation, such as those described for cattle-associated *C. jejuni* lineages (14), but
259 predominantly encode hypothetical proteins.

Functional investigation may improve understanding of the possible role of the ST403 complex pseudogenised genes in adaptation away from avian hosts. Rather than a simple case of no longer being able to colonise birds, it may be that loss of these loci results in reduced competition with other lineages, low colonisation numbers, or reduced ability to survive transmission outside the host. Furthermore, evidence of acquisition of specific R-M systems and the pseudogenisation and loss of several loci suggest that both the loss and gain of specific loci may be associated with adaptation to a restricted host set in *C. jejuni*. The combination of both gene gain and adaptive gene loss are known to have played a role in niche-adaptation in other enteric bacterial pathogens (37, 38). Further work will be necessary to quantify the influence of host, pathogen and environmental factors on colonization of different host species by *C. jejuni* and the increasing availability of bacterial genomes and understanding of gene function will provide a basis for future investigation.

Acknowledgements

LM was funded by Nottingham Trent University PhD studentship. AM is supported by Royal Society (IE121459). SKS is supported by a Wellcome Trust Career Development Fellowship, with additional funding from the BBSRC (BB/I02464X/1) and MRC-CLIMB (MR/L015080/1). JC is supported by ERC grant 239784 and AoF grant 251170. DNA Sequencing was carried out at the Exeter Sequencing Service which is supported by the following grants: Wellcome Trust Institutional Strategic Support Fund (WT097835MF), Wellcome Trust Multi User Equipment Award (WT101650MA) and BBSRC LOLA award (BB/K003240/1).

References

1. **Janssen R, Krogfelt KA, Cawthraw SA, van Pelt W, Wagenaar JA, Owen RJ**. 2008. Host-pathogen interactions in *Campylobacter* infections: the host perspective. Clin. Microbiol. Rev. **21**:505–18.

2. **Young KT, Davis LM, Dirita VJ**. 2007. *Campylobacter jejuni*: molecular biology and pathogenesis. Nat. Rev. **5**:665–679.

3. **Gillespie I, O'Brien S, Frost J, Adak G, Horby P, Swan A, Painter M, Neal K CSSSC**. 2002. A case-case comparison of *Campylobacter coli* and *Campylobacter jejuni* infection: A tool for generating hypotheses. Emerg. Infect. Dis. **8**:937–42.

4. **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.

5. **Sheppard SK, Cheng L, Méric 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, Hänninen ML, Parkhill J, Hanage WP, Corander J**. 2014. Cryptic ecology among host generalist *Campylobacter jejuni* in domestic animals. Mol. Ecol. **23**:2442–51.

6. **Strachan NJ, MacRae M, Thomson A, Rotariu O, Ogden ID, Forbes KJ**. 2012. Source attribution, prevalence and enumeration of *Campylobacter* spp. from retail liver. Int. J. Food Microbiol. **153**:234–236.

7. **Strachan NJ, Gormley FJ, Rotariu O, Ogden ID, Miller G, Dunn GM, Sheppard SK, Dallas JF, Reid TM, Howie H, Maiden MC, Forbes KJ**. 2009. Attribution of *Campylobacter* infections in northeast Scotland to specific sources by use of multilocus sequence typing. J. Infect. Dis. **199**:1205–1208.

8. **Szewzyk U, Szewzyk R, Manz W, Schleiffer KH**. 2000. Microbiological safety of drinking water. Annu. Rev. Microbiol. **54**:81–127.

9. **Sheppard SK, Colles FM, McCarthy ND, Strachan NJ, Ogden ID, Forbes KJ, Dallas JF, Maiden MC**. 2011. Niche segregation and genetic structure of *Campylobacter jejuni* populations from wild and agricultural host species. Mol. Ecol. **20**:3484–3490.

10. **Manning G, Dowson CG, Bagnall MC, Ahmed IH, West M, Newell DG**. 2003. Multilocus sequence typing for comparison of veterinary and human isolates of *Campylobacter jejuni*. Appl. Environ. Microbiol. **69**:6370–6379.

11. **Milnes AS, Stewart I, Clifton-Hadley FA, Davies RH, Newell DG, Sayers AR, Cheasty T, Cassar C, Ridley A, Cook AJC, Evans SJ, Teale CJ, Smith RP, McNally A, Toszeghy M, Futter R, Kay A, Paiba GA**. 2008. Intestinal carriage of verocytotoxigenic *Escherichia coli* O157, *Salmonella*, thermophilic *Campylobacter* and *Yersinia enterocolitica*, in cattle, sheep and pigs at slaughter in Great Britain during 2003. Epidemiol. Infect. **136**:739–751.

12. **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.

13. **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.

14. **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 **16**:11923–11927.

15. **Tang J, Hanage WP, Fraser C, Corander J**. 2009. Identifying currents in the gene pool for bacterial populations using an integrative approach. PLOS Comput. Biol. **5**:e1000455.

16. **Feil EJ, Maynard Smith J, Enright MC, Spratt BG**. 2000. Estimating the recombinational parameters in *Sterptococcus pneumoniae* from multilocus sequence typing data. Genetics **154**:1439–1450.

17. **Fearnhead P, Smith NG, Barrigas M, Fox A, French N**. 2005. Analysis of recombination in *Campylobacter jejuni* from MLST population data. J. Mol. Evol. **61**:333–340.

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

19. **Sheppard SK, McCarthy ND, Falush D, Maiden MC**. 2008. Convergence of *Campylobacter* species: implications for bacterial evolution. Science **320**:237–239.

20. **Zerbino D, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18**:821–829.

21. **Swain MT, Tsai IJ, Assefa SA, Newbold C, Berriman M, Otto TD**. 2012. A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. Nat Protoc **7**:1260–1284.

22. **Seemann T**. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics **30**:2068–9.

23. **Katoh K, Standley D**. 2013. MAFFT multiple seuqnce alingment software version 7: Improvements in performance and usability. Mol. Biol. Evol. **30**:772–80.

24. **Jolley KS, Maiden MC**. 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics **10**:595.

25. **Price M, Dehal P AA**. 2010. FastTree2 - approximately maximum-likelihood trees for large alignments. PLoS One **5**:e9490.

26. **Blom J, Albaum S, Doppmeier D, Puhler A, Vorholter F, Zakrzewski M, Goesmann A**. 2009. EDGAR: A software for the comparative analysis of promkaryotic genomes. BMC Bioinformatics **20**:154.

27. **Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, Corander J**. 2012. Detection of recombination events in bacterial genomes from large population samples. Nucleic Acids Res. **40**:e6.

28. **Duim B, Godschalk PC, van den Braak N, Dingle KE, Dijkstra JR, Leyde E, van der Plas J, Colles FM, Endtz HP, Wagenaar JA, Maiden MC, van Belkum A**. 2003. Molecular evidence for dissemination of unique *Campylobacter jejuni* clones in Curacao, Netherlands Antilles. J. Clin. Microbiol. **41**:5593–5597.

29. **Dingle KE, Colles FM, Wareing DR, Ure R, Fox AJ, Bolton FE, Bootsma HJ, Willems RJ, Urwin R, Maiden MC**. 2001. Multilocus sequence typing system for *Campylobacter jejuni*. J. Clin. Microbiol. **39**:14–23.

30. **Ahmed IH, Manning G, Wassenaar TM, Cawthraw S, Newell DG**. 2002. Identification of genetic differences between two *Campylobacter jejuni* strains with different colonization potentials. Microbiology **148**:1203–1212.

31. **Poly F, Threadgill D, Stintzi A**. 2004. Identification of *Campylobacter jejuni* ATCC 43431-specific genes by whole microbial genome comparisons. J. Bacteriol. **186**:4781–4795.

32. **Poly F, Threadgill D, Stintzi A**. 2005. Genomic diversity in *Campylobacter jejuni*: identification of C. jejuni 81-176-specific genes. J. Clin. Microbiol. **43**:2330–2338.

33. **Holt J, Grant A, Coward C, Maskell D QJ**. 2012. Identification of Cj1051c as a major determinant for the restriction barrier of *Campylobacter jejuni* strain NCTC11168. Appl. Environ. Microbiol. **78**:7841–8.

34. **Asakura H, Bruggemann H, Sheppard S, Ekawa T, Meyer T, Yamamoto S IS**. 2012. Molecular evidence for the thriving of *Campylobacter jejuni* ST-4256 in Japan. PLoS One **11**:e48394.

35. **Budroni S, Siena E, Dunning-Hotopp J, Seib K, Serruto D, Nofroni C, Comanducci M, Riley D, Daughery S, Angiuoli S, Covacci A, Pizza M, Rappuoli R, Moxon ER, Tettelin H MD**. 2011. *Neisseria meningiditis* is structured in clades associated with restriction modification systems that modulate homologous recombination. Proc Natl Acad Sci U S A **108**:4494–9.

36. **McCarthy ND, Colles FM, Dingle KE, Bagnall MC, Manning G, Maiden MC, Falush D**. 2007. Host-associated genetic import in *Campylobacter jejuni*. Emerg. Infect. Dis. **13**:267–272.

37. **Reuter S, Connor TR, Barquist L, Walker D, Feltwell T, Harris SR, Fookes M, Hall ME, Petty NK, Fuchs TM, Corander J, Dufour M, Ringwood T, Savin C, Bouchier C, Martin L, Miettinen M, Shubin M, Riehm JM, Laukkanen-Ninios R, Sihvonen LM, Siitonen A, Skurnik M, Falcão JP, Fukushima H, Scholz HC, Prentice MB, Wren BW, Parkhill J, Carniel E, Achtman M, McNally A, Thomson NR**. 2014. Parallel independent evolution of pathogenicity within the genus *Yersinia*. Proc Natl Acad Sci U S A **111**:6768–6773.

38. **Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, Churcher C, Mungall KL, Bentley SD, Holden MT, Sebaihia M, Baker S, Basham D, Brooks K, Chillingworth T, Connerton P, Cronin A, Davis P, Davies RM, Dowd L, White N, Farrar J, Feltwell T, Hamlin N, Haque A, Hien TT, Holroyd S, Jagels K, Krogh A, Larsen TS, Leather S, Moule S, O'Gaora P, Parry C, Quail M, Rutherford K, Simmonds M, Skelton J, Stevens K, Whitehead S, Barrell BG**. 2001. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. Nature **413**:848–852.

39. **Zeng X, Xu F, Mo Y LJ**. 2013. Identification and characterisation of a periplasmic trilactone esterase, Cee, revealed unique features of ferric enterobactin acquistion in *Campylobacter*. Mol. Microbiol. **87**:594–608.

Table 1: A list of the strains used for comparative genomic analysis in this study. Raw data for the bacterial strains sequenced as part of this study have been deposited in the ENA under accession number ERP006801.

Strain	Species	ST-Complex	Source	Sequenced
857	*C. jejuni*	403	Pig	This study
549.1	*C. jejuni*	403	Pig	This study
623	*C. jejuni*	403	Pig	This study
304	*C. jejuni*	403	Pig	This study
484	*C. jejuni*	403	Pig	This study
444	*C .jejuni*	403	Pig	This study
88	*C. jejuni*	403	Cow	(14)
1779	*C. jejuin*	403	Dog	(14)
2208	*C. jejuni*	403	Human	(14)
2226	*C. jejuni*	403	Human	(14)
2362	*C. jejuni*	403	Environmental	(14)
2455	*C. jejuni*	403	Human	(14)
ATCC33560	*C. jejuni*	403	Cow	(39)

Table 2: Loci unique to the *C. jejuni* ST-403 complex.

CDS[+]	**Putative function**[++]	**Orthologues**[+++]
cje135_06701	Hypothetical protein	None outside of ST-403 complex
cje135_06696	Hypothetical protein	None outside of ST-403 complex
CJ857_00839*	Hypothetical protein	None outside of ST-403 complex
cje135_03870	R.HinP1 Restriction Endonuclease	*H. cinaedi* CCug18818
cje135_03865	Modification methylase Hhal	
CJ857_01361*	Hypothetical protein	Cc 317-04/90-3
CJ857_01649*	Hypothetical protein	Weak similarity with Cjj LMG23223
cje135_02353	Hypothetical protein	Cc LMG23336/*H. bilis* ATCC43879/*H.cinaedi* PAGU611
cje135_02348	R.Pab1 restriction endonuclease	*H.pylori* 51
cje135_02293	Recombination protein F	*H.pullorum* MIT98-5489

[+]CDS are annotated according to the genome annotation of the ST-403 reference strain ATCC33560, except those marked *, which are relative to our strain 857 due to ambiguous annotation in ATCC3560.

[++]Putative function is that ascribed to the CDS by Pfam and BlastP searches

[+++]Orthologues are as determined by BlastP against the entire Blast nrDatabase

Table 3: Characterisation of Loci absent from ST-403 complex *C. jejuni*

CDS[+]	**Putative function**	**ST-403 complex mutation**
C8J_0199-200*	Protease/IgA1 protease domain family/Serine protease	Genes merged by SNP and pseudogenised by stop codon
C8J-_0806*	Hypothetical protein (Seryl-tRNA synthetase domain; provisional endonuclease subunit domain)	Pseudogenised-central deletion in CDS
C8J-_0815*	Hypothetical protein (cytochrome C oxidase cbb3 subunit)	Pseudogenised by stop codons
C8J-_0628*	Hypothetical protein (potassium transporting ATPase subunit)	Pseudogenised by stop codons
C8J-_0466*	Putative outer membrane protein (assembly complex/hemolysin activation/secretion protein)	Pseudogenised by stop codons
CJE0296	Conserved domain protein (MCP-domain signal transduction protein)	Multiple insertions and deletions varying across lineage
CJE0392	Hypothetical protein	Absent
CJE0660	Hypothetical protein	Multiple deletions across lineage.
CJE0659	Putative membrane protein/ putative dicarboxylate carrier protein MatC/ putative integral membrane protein	Absent
C8J_0033	Hypothetical protein (gamma-glutamyltranspeptidase)	Absent
C8J_0392	Hypothetical protein	Entire or central deletion across lineage
Cj1158c	Hypothetical protein (putative small hydrophobic protein/small integral membrane protein)	Absent
C8J_1559	Hypothetical protein	Multiple deletions and SNPs across lineage
C8J_0239	Probable methyl accepting chemotaxis protein signalling domain	Absent

[+]CDS as annotated in appropriate reference genome

*Mutations that are identical across every ST-403 complex isolate

436 **Figure legends**

437 Figure 1: Maximum likelihood core genome phylogeny of 139 *C. jejuni* and 60 *C. coli*
438 isolates. Isolates from the previously identified distinct *C. coli* clades (5), as well as pig *C.*
439 *coli*, and the ST-403 complex strains are identified in the legend.

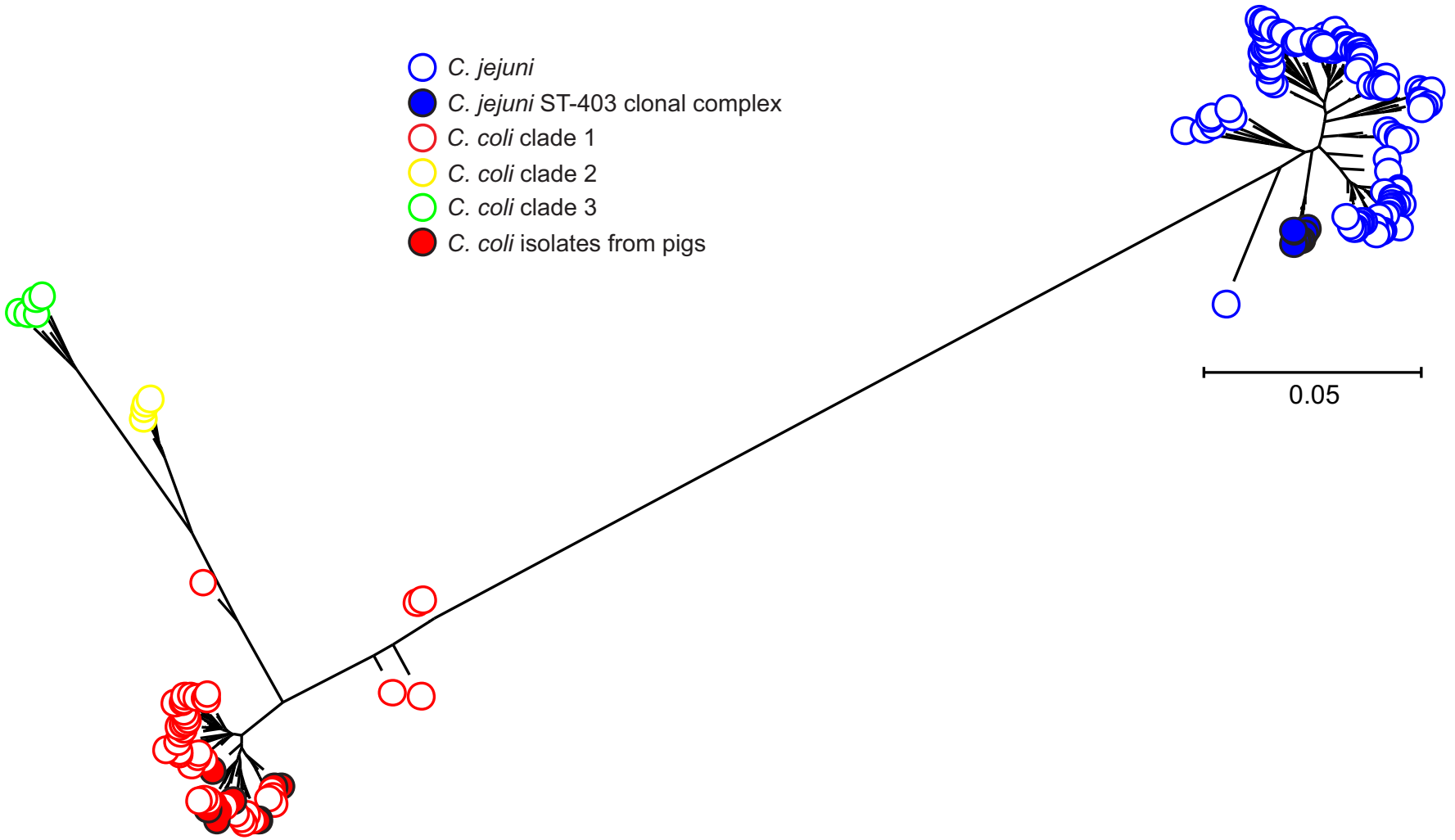
440

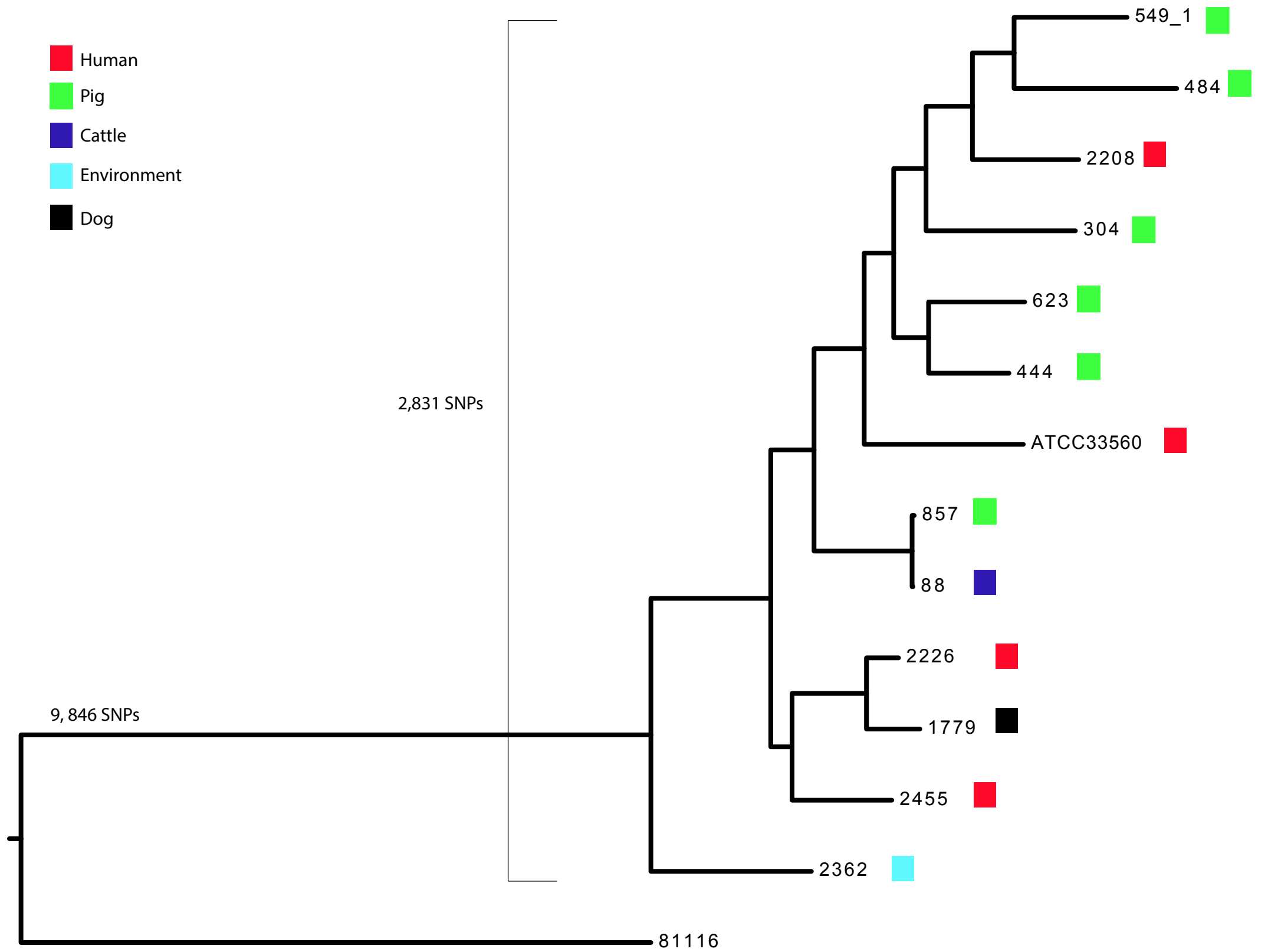
441 Figure 2: SNP based phylogeny of the ST-403 complex. Isolates are colour coded according
442 to their environmental source. The SNP distance between the 81116 rooted outlier and the
443 ST-403 complex is indicated, as is the SNP distance range observed across the ST-403
444 complex.

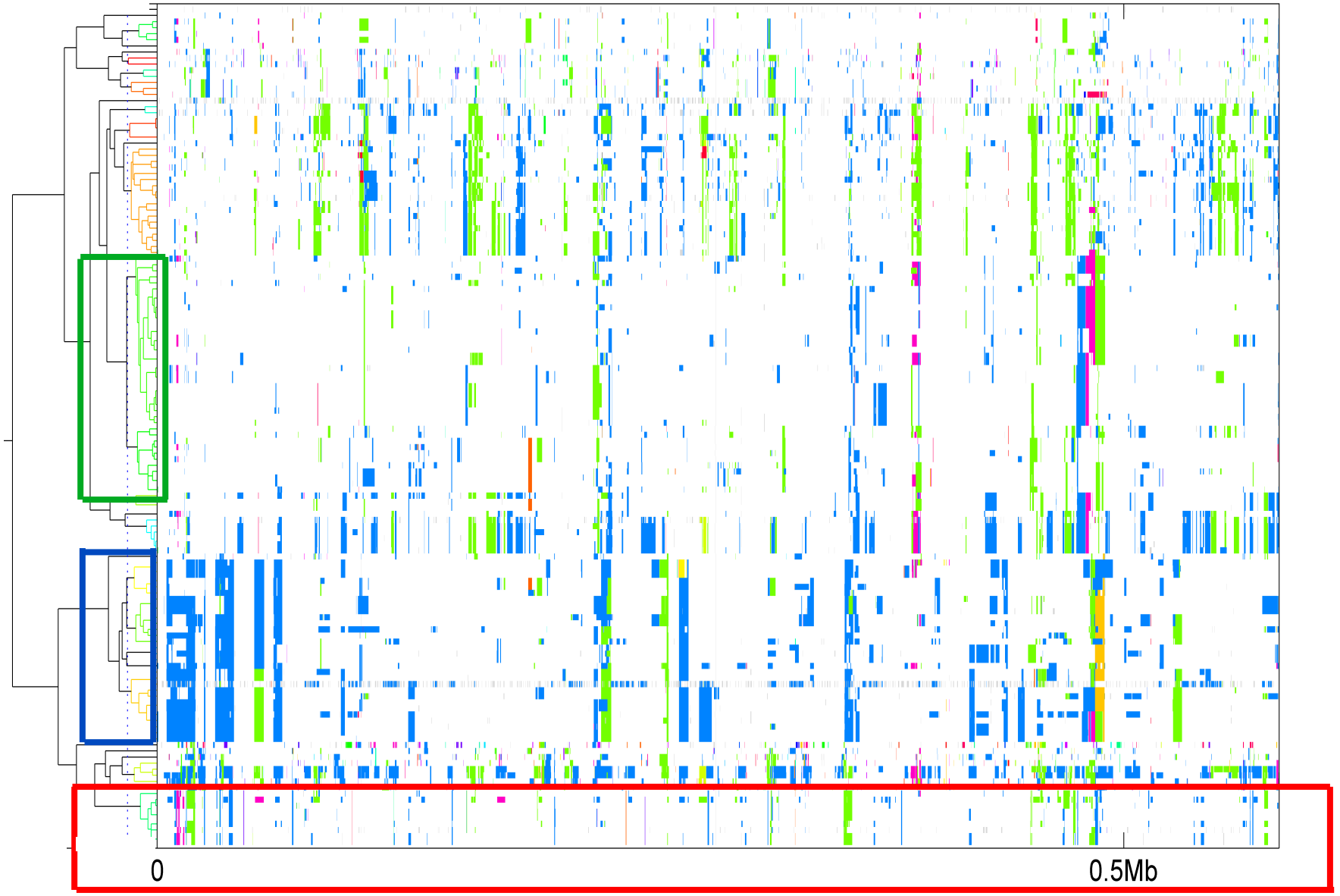
445

446 Figure 3: Visualized output of BRATNextGen analysis of the core genome alignment of 139
447 C. jejuni. On the left a clustering tree of the 139 isolates is shown based on proportion of
448 shared ancestry through recombination. Coloring of the branches indicates cluster
449 membership and significant recombinations are indicated by colored rectangles on the right.
450 Shared color in the same column implies that the recombination segments in the respective
451 isolates correspond to a shared origin. A contiguous single-colored rectangle along the
452 genome represents a single inferred recombination event. The colors indicate the cluster in
453 which the corresponding recombined genome segment has the highest frequency. For
454 convenience the clusters corresponding to the darker blue and green hues are indicated by the
455 blue and green boxes respectively. The ST-403 complex genomes are indicated by the red
456 box.

457


C. jejuni
C. jejuni ST-403 clonal complex
C. coli clade 1
C. coli clade 2
C. coli clade 3
C. coli isolates from pigs
0.05


Human
Pig
Cattle
Environment
Dog
549_1
484
2208
304
623
444
ATCC33560
857
88
2226
1779
2455
2362
81116
2,831 SNPs
9, 846 SNPs


0
0.5Mb