K-mer-based High-throughput Analysis of the Adaptive Potential of *Campylobacter*

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

vorgelegt von

Lennard Epping

Berlin
Oktober 2021

Erstgutachter:	Prof. Dr. Rosario M. Piro
Zweitgutachter:	Prof. Dr. Dr. h.c. mult. Lothar H. Wieler

Tag der Disputation: 15.02.2022

Abstract

Several species of the genus *Campylobacter* (*C.*) are zoonotic pathogens, especially *C. jejuni* and *C. coli*, the leading causes of foodborne diseases worldwide. Although both species colonize many hosts including poultry, livestock and wild animals, persistence mechanisms enabling the bacteria to adapt towards new ecological niches are not yet fully understood. In this work, novel *k-mer*-based methods enabling high-throughput analysis of whole-genome sequencing (WGS) data of *C. jejuni* and *C. coli* have been developed, extended and applied to investigate the adaptive potential of the distinct species towards different ecological niches and changing environments.

A *k-mer*-based microbial genome-wide association study (GWAS) was set up to identify host-specific genomic signatures of *C. jejuni* isolated from chicken, cattle, pig and clinical human samples. GWAS revealed a strong association of both, the core and the accessory genome of *C. jejuni*, with distinct host animal species. Moreover, multiple adaptive trajectories defining the evolution of *C. jejuni* lifestyle preferences in different ecosystems were identified.

In a second approach, WGS data of *Campylobacter* isolates that showed ambiguous probing results using different polymerase chain reaction (PCR)-based species classification methods during routine-diagnostics were investigated. The *Campylobacter* genomes were analyzed with respect to their genomic make-up. For this purpose, a *k-mer*-based method was developed in order to identify recombination events between *C. jejuni* and *C. coli*. The identified genes encode proteins that were commonly associated with important pathways involved in chromosome maintenance and DNA repair, membrane transport and stress defense.

Overall, the results presented in this work promote molecular surveillance and rapid diagnostics of *Campylobacter*. In addition, host-specific allelic variants identified among different phylogenetic backgrounds might serve as important maker genes in future source attribution models for fast and precise retrograde outbreak investigation along the food chain.

Acknowledgements

First and foremost, I would like to thank my supervisor Dr. Torsten Semmler for providing the opportunity to work on this exciting project in his research group. Thank you for your continuous support, the freedom you guaranteed throughout all my projects and for always taking the time to discuss my ideas and questions. Further, I would like to thank Prof. Dr. Rosario M. Piro for excellent supervision and guidance implementing a genome-wide association study and his independent views. Many thanks belong to Prof. Dr. Dr. h.c. mult. Lothar H. Wieler for supervising my work and for sharing his knowledge about zoonotic diseases and disease surveillance.

I would like to thank Dr. Birgit Walther for her constructive reviews of my research work during my thesis and many other projects. Additionally, I would like to thank Sara Hetzel for her tireless efforts for scientifically proofreading of my thesis and support during this time. Furthermore, I would like to thank Silver Wolf for his best efforts proofreading my thesis.

I wish to thank all my amazing colleagues who participated in my projects, in particular Sumeet Tiwari, Silver Wolf, Felix Hartkopf, Vanessa Johanns, Charlotte Huber and Anne Kauter for the plenty and profound discussions and their excellent collaborative efforts. Further, many thanks to Johanna von Wachsmann and Mustafa Helal who were great students and did an excellent job.

Last but not least, I would like to thank my family and friends, in particular my parents and brother, for their unconditional support. I am thankful to know that they stand by me whatever happens.

Contents

List of Figures

List of Tables

Abbreviations

AMR antimicrobial resistance

ANI average nucleotide identity

BAPS Bayesian analysis of population structure

BC before Christ

BfR German Federal Institute for Risk Assessment

BLAST basic local alignment search tool

CC clonal complex

CDCV common disease common variant

CDS coding sequence

CNV copy number variant

COG clusters of orthologous genes

DDH DNA-DNA hybridization

EFSA European Food Safety Authority

GTR general time-reversible

GWAS genome-wide association study

HGT horizontal gene transfer

HMM Hidden Markov Model

I/O input/output

indels insertions or deletions

IS insertion elements

LD linkage disequilibrium

LMM linear-mixed model

MALDI-TOF matrix-assisted laser desorption/ionization-time-of-flight

MLST multilocus sequence typing

mPCR multiplex polymerase chain reaction

MSA multiple sequence alignment

MST minimum spanning tree

NCBI National Center for Biotechnology Information

NGS next-generation sequencing

NP nondeterministic polynomial time

ONT Oxford Nanopore Technologies

PCA principal component analysis

PCR polymerase chain reaction

qPCR quantitative polymerase chain reaction

RAxML randomized axelerated maximum likelihood

RM system restriction modification system

RNA ribonucleic acid

rRNA ribosomal ribonucleic acid

RSS residual sum of squares

SNP single nucleotide polymorphism

spp. species pluralis

SRA sequence read archive

ST sequence type

t-SNE t-Distributed Stochastic Neighbor Embedding

TE transposable elements

tRNA transfer ribosomal ribonucleic acid

USDA United States Department of Agriculture

WGS whole-genome sequencing

1 Introduction

Campylobacter is the most common cause of food-borne infections worldwide. The bacterium is a commensal of the gut microbiota of many wild and livestock animals including poultry, cattle, pigs and wild birds [1–4]. Through the consumption of contaminated (chicken) meat, water, raw-milk or other food products, *Campylobacter* can be transmitted to humans, where it commonly causes acute symptoms of gastroenteritis [5]. Patients suffering from Campylobacteriosis are mostly infected by either *Campylobacter jejuni* (90%) or *Campylobacter coli* (5%-10%) [6–8].

In high-income countries, *Campylobacter* has become the leading cause of bacterial food-born infections during the last two decades [9]. Although most of the clinical symptoms caused by *Campylobacter* are self-limiting, serious consequences in terms of post-infectious complications, including life-threatening events such as bacteremia or Guillain-Barrè syndrome, are possible [10, 11]. The global incidence and prevalence of food-borne *Campylobacteriosis* have increased in both developed and developing countries over the last 10 years [5, 12, 13]. Apart from the individual burden for the patients' health, the economic burden is considerable according to the European Food Safety Authority (EFSA) and the United States Department of Agriculture (USDA) [14, 15]. In 2014 the costs associated with *Campylobacter* infections have been estimated to be around 2.4 billion € in the European Union alone [16, 17].

Various studies regarding the overall population structure and adaptive abilities of *Campylobacter* towards certain niches have been conducted [18–21]. However, despite recent achievements in understanding niche and host adaptation processes of distinct *C. jejuni* lineages, the overall knowledge on the subject is still scarce, especially compared to the insights available for *Escherichia coli* or *Salmonella* species [22].

In order to prevent and combat infectious diseases caused by *Campylobacter*, it is necessary to investigate the adaptive potential of this bacteria in livestock as well as in strains of (human) clinical origin. In the recent past, molecular typing of *Campylobacter* frequently included allelic profiling of seven housekeeping genes [23]. Previous research showed that certain combinations of allelic variants of those genes seemed to be associated with a particular host species [24]. However, the adaptive potential towards specific

ecological niches of *Campylobacter* cannot be explained based on seven housekeeping genes. Thus, enhancing the knowledge on adaptation processes by use of genome-wide association studies to identify host-specific genomic signatures would support the development of diagnostic markers, source attribution models and enhances the capabilities of rapid outbreak detection.

Furthermore, analysis based on the seven housekeeping genes also revealed an exchange of genetic material between *C. jejuni* and *C. coli*, leading to the hypothesis of frequent recombination between both species as a strong marker for ongoing evolution [25]. The high gene-flow from *C. jejuni* to *C. coli* not only indicates important adaptation mechanisms between both species, but also has particular implications for the diagnostic and zoonoses surveillance programs.

This thesis is focused on generating in-depth knowledge about the population structure of *C. jejuni* and its host-adaptation mechanisms including the recombination potential between *C. coli* and *C. jejuni*. For this purpose, hundreds of genomes were compared using a population-based method from human genetics that has been recently adapted for bacterial genomics. In order to evaluate these aspects with high resolution, whole genomes of all samples were analyzed. To overcome the computational complexity of genomic data and to investigate the entire genome sequence of *Campylobacter*, efficient computational methods have been utilized and tailored to serve the aims of this study. In order to generate an overview of all recombinant regions and associated genes, a novel *in silico* approach has been developed and applied within the scope of this work.

Following the introduction, the biological background, centered around the general population structure of *Campylobacter*, is described in detail (chapter 2). The third chapter focuses on the computational aspects of population-based GWAS and its harnessing for microbial research. It also includes an introduction to the general concepts of sequence-based methods in bioinformatics research. Chapter 4 provides an overview of current *in silico* methods and a detailed description of the workflows designed and implemented to analyze the genomic data in this work. An in-depth analysis of the population structure of *C. jejuni* and results from the GWAS focused on factors likely associated with niche and host specificity of *C. jejuni* is presented in chapter 5. Chapter 6 contains the results and discussion of inter-species recombination of *C. jejuni* and *C. coli* based on a newly

developed computational workflow. Finally, in chapter 7 the results of this work are summarized and conclusions are drawn.

2 Biological background

Campylobacter is a Gram-negative bacterium and typically appears as curved, S-shaped rods with bipolar flagella (Figure 2.1). In the year 2000, the complete genome of *C. jejuni* was sequenced, revealing a 1.7 Mbp long circular sequence that included 1,654 genes with an average GC content (guanine-cytosine content within a DNA sequence) of 30.6% [26]. Of these, 1,343 genes were considered as the core genome, which is defined as the entirety of genes shared by all genomes under consideration. These genes are commonly considered to be housekeeping genes, holding important roles in the general life cycle and metabolic processes of an organism. The remaining genes are classified as accessory genes. Consequently, the accessory genome includes all genes that are present in equal or less than 99% of the individuals of a species, also all unique genes associated with few or even singleton genomes [27, 28]. Those genes may help the microbes to adapt towards novel ecological niches [27]. Accessory genes are often located on mobile genetic elements such as plasmids which can be horizontally transferred between strains or even different bacterial species. Plasmids are circular extra-chromosomal fragments of various lengths, and beyond others - tend to carry antimicrobial resistance (AMR) genes.

The evolution of bacterial species and particular strains is influenced by changes in the environment or the necessity to adapt to a novel ecological niche. Three distinct concepts are commonly involved in these evolutionary processes [29]:

- Genetic drift: Random fluctuations in the frequencies of alleles from generation to generation due to stochastic events.
- Bottleneck effect: Sharp reduction in the size of a population due to environmental changes.
- Selection pressure: Development of variants that provide an individual with an increased chance of survival facing changes in the environment.

In the following section, the evolutionary history of *Campylobacter*, as well as molecular typing concepts, are described to explain the current population structure of *C. jejuni* and *C. coli*. Furthermore, the general concepts of horizontal gene transfer (HGT) and

recombination are explained with respect to their impact on the niche adaption potential of bacteria.

Figure 2.1: Colored image of *C. jejuni* made from an electron microscope at the Robert Koch Institute. *C. jejuni* shows the typical S-shape rods with their bipolar flagella. The picture was provided by Doreen Weigelt - Robert Koch Institute, Department of Advanced Light and Electron Microscopy.

The following subsections are based on material published in Chapter 4 in the book "Fighting *Campylobacter* Infections" (Current Topics in Microbiology and Immunology):

Epping, L., Antão, E. M., & Semmler, T. (2021). Population Biology and Comparative Genomics of *Campylobacter* Species. In: Backert S. (ed) Fighting Campylobacter Infections. Current Topics in Microbiology and Immunology, vol. 431, Springer, Cham, Switzerland. Pages 59-78.

2.1 A brief history of *Campylobacter* evolution

The term molecular clock was introduced in 1962 by Linus Pauling and Emile Zuckerkandl in order to measure the evolution rates of different organisms [30]. The concept is based on counting the number of mutations within the nucleotide sequences of different species in relation to a defined time unit, e.g. number of genetic changes per generation or number of genetic changes per year.

In order to estimate the molecular clock for bacteria, Ochman & Wilson developed an approach based on the 16S ribosomal ribonucleic acid (rRNA) encoding gene, since this sequence of ribosomal RNA is highly conserved in many bacterial species[31]. To "calibrate" the molecular clock, the ancestral diversification of *E. coli* and *Salmonella typhimurium* was used as the foundation mutation rate of this approach. As a result, 1% divergence in the 16S rRNA nucleotides per 50 million years for ancestral diversification was calculated. Applying this particular model directly to the genus *Campylobacter*, the diversification of *C. coli* from *C. jejuni* started 10 million years ago. Moreover, regular clade formation of *C. coli* began around 2.5 million years ago [32]. However, the molecular clock approach might not sufficiently describe evolutionary processes for rather rapidly evolving genera like *Campylobacter*, which shows twice as many genetic recombination events than *de novo* mutations. In addition, a multi-host lifestyle promotes adaptive evolution which is known to increase the accumulation of mutations [33].

Taking this into account, Wilson et al. developed a new approach using a more rapid molecular clock to estimate divergence within the genus *Campylobacter*, dating the onset of individual species diversification for *C. coli* and *C. jejuni* roughly to 6,580 years ago [34]. In the context of the development of human economics, this estimation fits into

the time of the Neolithic Revolution (first agricultural revolution) when people began domesticating animals. The Neolithic revolution started in the Middle East around 12,000 - 10,000 before Christ (BC) and spread through central Europe 5,000 - 3,000 BC. The subsequent and rapid changes in the lifestyle of mankind also provided multiple possibilities for bacteria to colonize novel niches while establishing novel transmission routes for both, commensal and pathogenic bacteria [35]. The divergence of *C. coli* clades was estimated to have happened 1,700 - 1,000 years ago, while clonal complexes of *C. jejuni* just started to evolve 400 years ago (Figure 2.2).

While the separation of the genus *Campylobacter* was most likely enabled by the agricultural revolutions thousands of years ago, methods of the modern food industry, globalization and environmental changes represent novel evolutionary niches and selection pressures for bacteria in general [36–38]. Recent research revealed a new trend: Gene flow analysis between *C. coli* and *C. jejuni* showed a large number of introgression from *C. jejuni* into the two most important *C. coli* lineages. Additionally, these results indicate that separation of the species did not necessarily lead to a recombination barrier and might lead to a convergence of *C. coli* towards *C. jejuni* (Figure 2.2).

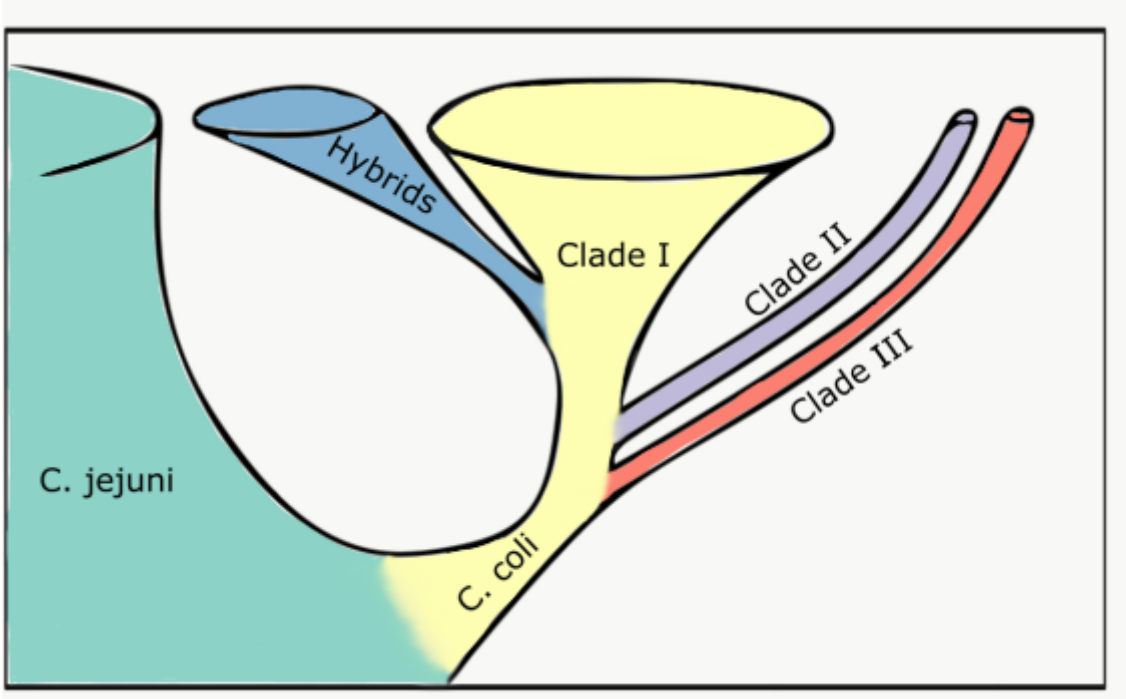


Figure 2.2: Evolutionary scheme of *C. jejuni* and *C. coli* adapted from Sheppard et al. [25]. *C. coli* and *C. jejuni* diverged into two species, which was followed by additional splits within *C. coli* most likely caused by substantially different ecological niches for distinct sub-populations. Nowadays, the species *C. coli* is differentiated into three main clades (I-III) [25, 36]. Recent research identified recombination between strains from *C. coli* belonging to clade I and *C. jejuni*, which has lead to the formation of *C. coli* hybrid strains with large incorporated genetic elements of *C. jejuni* [25, 39]. This figure was originally created for publication in the article "Population Biology and Comparative Genomics of *Campylobacter* Species" [40].

2.2 Mechanisms of host adaptation

It has been estimated that more than 60% of all currently known human pathogens originated from animal hosts [41]. Therefore, it is pivotal to understand the mechanisms involved in host and niche adaptation of zoonotic bacteria. The decreasing cost associated with the employment of next-generation sequencing (NGS) technologies led to a considerable increase in population-scale studies based on whole-genome data. Many of these studies have investigated host-specific factors in different zoonotic and pathogenic

bacteria [42–44]. Several bacterial species such as *Staphylococcus aureus*, *Salmonella enterica* and *C. jejuni* include multiple host-associated coexisting lineages within their population [29]. The diversity within a bacterial population plays an important role in colonization of novel ecological niches, which is driven by two main mechanisms: mutation and recombination. However, not every host-associated genetic variation increases the adaptive potential, some might randomly occur as a result of a bottleneck effect or genetic drift. Furthermore, various bacteria remove genes from their genomes during adaptation towards a particular host or niche. This process is known as reductive evolution [45]. Commonly, reductive evolution affects genes that are not essential for the survival in the new host [45]. As a result, it has been widely observed that genomes of host-restricted bacterial lineages become more clonal while decreasing their genome sizes [46].

2.2.1 Mutation rate

Mutations are changes of the genomes caused by DNA replication errors or may be induced by chemical, biological or physical factors such as antibiotic agents, deaminating agents and UV-light [47]. These natural changes occur in both coding sequence (CDS) and intergenic regions: Non-silent mutations in the CDS alter the amino acid sequence of the encoded protein, possibly affecting its structure, purpose or functionality. Thus, mutations enhance the opportunities for the development of novel evolutionary trajectories by, for example, influencing biosynthesis pathways or AMRs. Single nucleotide polymorphisms (SNPs), short insertions or deletions (indels) in CDS can also cause loss of function, which for instance, might increase the tolerance towards certain environmental factors [48]. For example, loss of function mutations in the *slt* (peptidoglycan-recycling enzyme) gene of *E. coli* are known to enhance the ethanol tolerance of the bacteria by changing the cell wall structure [49]. Furthermore, mutations in non-coding regions such as non-coding ribonucleic acids (RNAs) genes or transcription-factor binding sites might influence the gene expression. Westerman et al., for example, proved that changes in the function of non-coding RNAs influenced the expression of invasion-associated effectors and virulence genes in *Salmonella enterica* during interactions with a new host [50].

Bacterial mutation rates, commonly measured in mutations per site per year, can vary

substantially. Table 2.1 gives an overview of some common example species with mutation rates ranging from $1.44 * 10^{-7}$ (*E. coli*) to $1.85 * 10^{-5}$ (*H. pylori*) mutation per site per year. By comparisons, bacterial populations associated with high mutation rates might adapt more easily to novel ecological niches or hosts. However, mutation rates are commonly rather low, since these changes are often deleterious [29].

2.2.2 Recombination

Besides mutations of existing CDS or non-coding regions, adaptation of bacteria is often highly influenced by the acquisition of novel genes from other bacteria or environmental sources through HGT and recombination. Recombination promotes a more rapid adaptation to changing environments than random mutations, due to the speed and effectiveness of the process [51, 52]. The transfer of genetic information between different bacteria is regarded as the driving force for evolution and adaptation [53]. To measure recombination within different bacterial species, Michiel Vos and Xavier Didelot defined recombination rates as "the ratio of nucleotide changes as the result of recombination relative to point mutation (r/m)" [54]. They demonstrated that r/m varies widely between bacterial species (Table 2.1). Species like *E. coli* and *S. aureus* have a low relative recombination frequency leading to a highly clonal population structure [55]. Bacteria like *H. pylori* are known for extremely high frequency of genetic exchanges and have a non-clonal, star-like population structure [56]. *C. jejuni* is known to have a moderate to high r/m which results in a weakly clonal population structure [57].

Table 2.1: Overview of mutation and recombination rates of different bacterial species. Mutation rates are measured in mutations per site per year and recombination rates as "the ratio of nucleotide changes as a result of recombination relative to point mutations(r/m)" [54].

Species	**Mutation Rate**	**r/m**	**Reference**
S. aureus	$2.05 * 10^{-6}$	0.0 - 0.6	[58]
E. coli	$1.44 * 10^{-7}$	0.03 - 2.0	[59]
C. jejuni	$3.23 * 10^{-5}$	1.7 - 2.8	[34, 54]
H. pylori	$0.91 * 10^{-5}$ - $1.85 * 10^{-5}$	13.6	[54]

The transfer among and within bacterial populations is achieved by three fundamental mechanisms summarized as follows [60] Figure 2.3:

- Conjugation: In the process of conjugation, DNA is transmitted directly from a donor bacterium to a recipient bacterium through a protein structure named pilus.
- Transformation: Describes the uptake of DNA fragments from the environment. This process is often driven by the exchange of plasmidial sequences and has high clinical relevance e.g. for the transmission of antibiotic resistance between bacteria which can evolve into multidrug-resistant pathogens.
- Transduction: The process of DNA being transferred bacteria through a vector e.g. a virus or phage.

These mechanisms can either lead to the exchange of complete loci of gene cassettes or to the formation of mosaic alleles that share sequence content of evolutionary different genetic backgrounds within the same gene.

In general, strains adapt towards new hosts with the acquisition of beneficial genes or positive mutations. During this process other genetic variations in the neighborhood of those beneficial genes or mutations will also be distributed among these lineages [61]. This effect is called genetic hitchhiking.

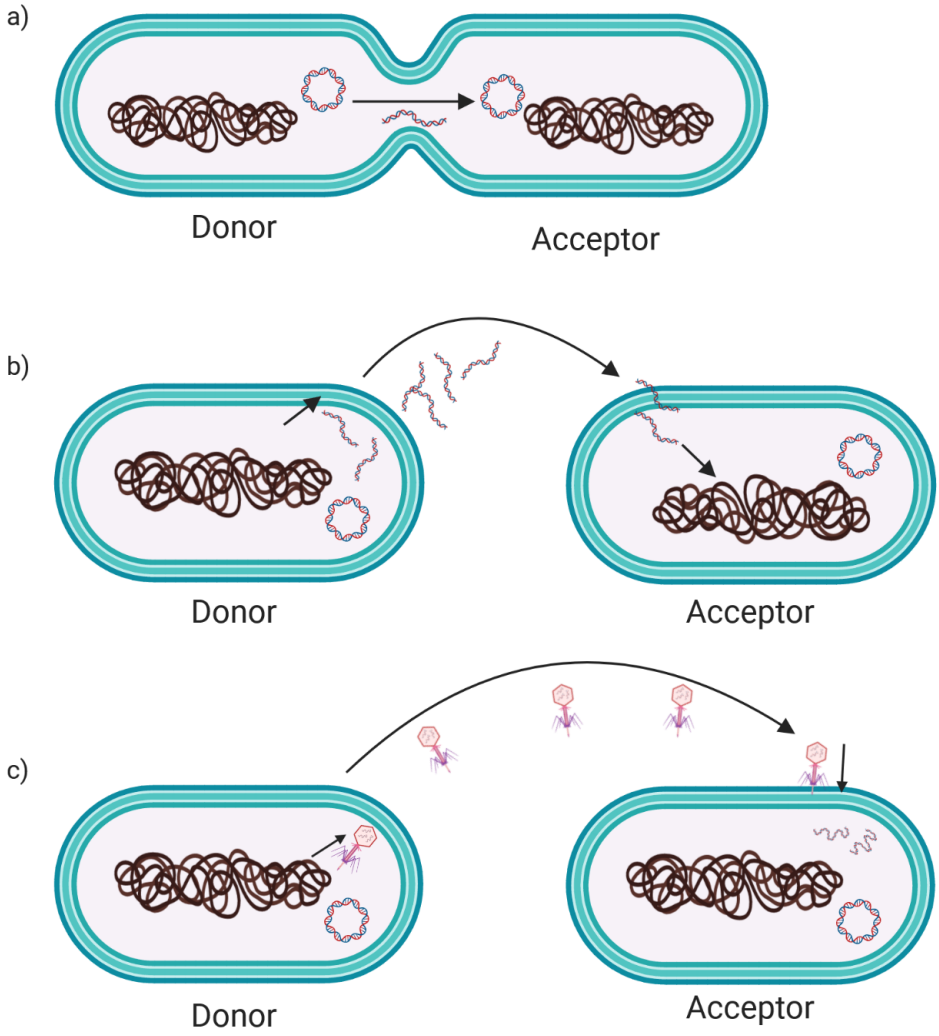


Figure 2.3: Schematic representation of the three different mechanisms of genetic transfer between bacteria. The sub-figures illustrate a) process of conjugation; b) transformation; c) transduction through a phage. This figure was adapted from [29] and created with BioRender (April 2021).

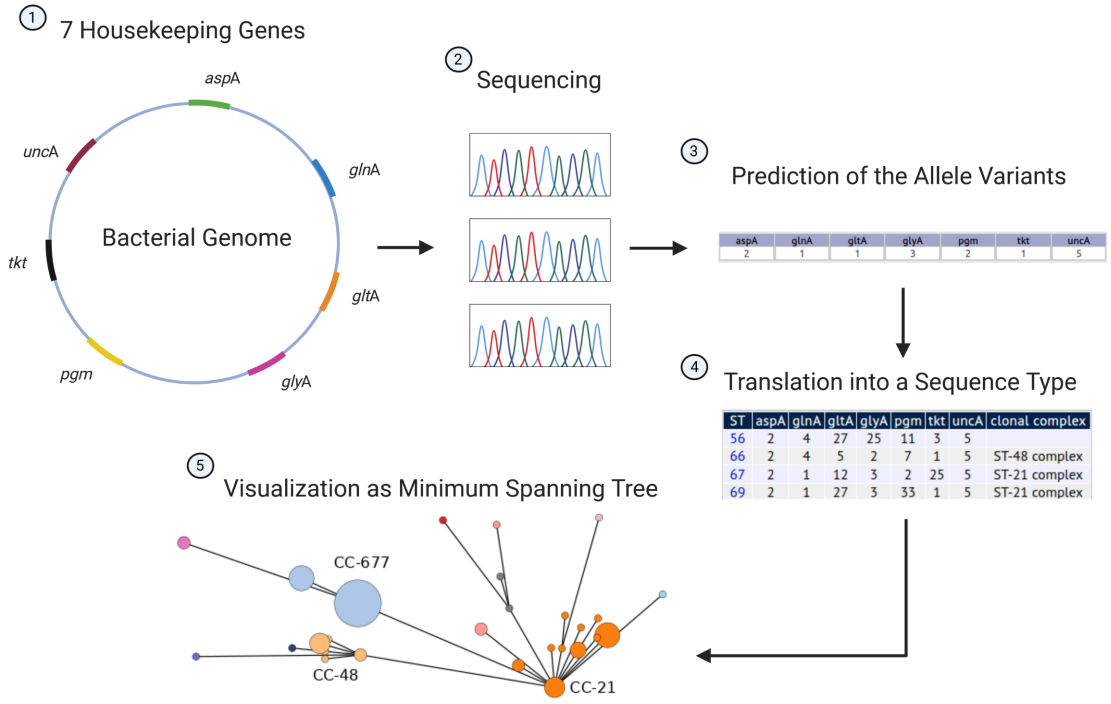
2.3 Concept of multilocus sequence typing (MLST)

Since *Campylobacter* species pluralis (spp.) became more and more relevant for public health, high-throughput molecular typing gained an important role in surveillance programs and outbreak control [62]. Most importantly, multilocus sequence typing (MLST) and NGS provided a generic approach for this matter and additionally had and still have a massive impact on understanding the population structure of *Campylobacter*. MLST is a molecular genetic classification scheme based the profile of the allelic variants of seven housekeeping genes which have been used to classify bacteria into related or more distant lineages [63]. For *C. jejuni* and *C. coli* the same MLST scheme is utilized, which characterizes allelic variants of the same orthologous loci in both species, enabling the possibility to directly compare the genetic relationship of both species with each other [57, 64].

MLST is known as a robust and generic method used to characterize bacterial isolates on a molecular level. The approach was originally introduced in 1998 by Maiden et al. to characterize and discriminate the species *Neisseria meningitidis* [63]. Since then, more and more schemes for several other bacterial species have been developed. They all use the same basic concept (Figure 2.4) of selecting seven representative slowly evolving housekeeping genes from different locations of the genome. Each variant of each of these loci is then assigned a unique consecutive number. The combination of the numbers of typed variants is represented as a vector or allelic profile that is translated into a species-specific sequence type (ST) (Figure 2.4). Strains with similar STs can be grouped as a clonal complex (CC), which commonly represents a lineage potentially derived from one common ancestor [65]. For *C. jejuni* and *C. coli*, ST profiles within at least four identical genes are grouped in the same CC [57].

For *C. jejuni* and *C. coli*, a common MLST scheme was developed by Dingle et al. that allows to directly compare the phylogenetic relationships of both species [23, 57]. The scheme is based on the housekeeping genes *asp*A (aspartase A), *gln*A (glutamine synthetase), *glt*A (citrate synthase), *gly*A (serine hydroxymethyltransferase), *pgm* (phosphoglucomutase), *tkt* (transketolase), and *unc*A (ATP synthase α subunit). Since *Campylobacter* is a highly recombining species, it was important to include genes that are not located within the same recombination site. In order to ensure this, the min-

imum distance between two loci was set to 70 kb [57]. The current MLST scheme for *C. jejuni*, *C. coli* and other *Campylobacter* species is available on the PubMLST website (`https://pubmlst.org/campylobacter/`).



ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA	clonal complex
56	2	4	27	25	11	3	5	
66	2	4	5	2	7	1	5	ST-48 complex
67	2	1	12	3	2	25	5	ST-21 complex
69	2	1	27	3	33	1	5	ST-21 complex

Figure 2.4: Schematic overview of MLST in bacteria. Seven particular housekeeping genes (1) from the bacterial genome are sequenced either by Sanger sequencing or extracted from whole-genome sequencing (WGS) data (2). The gene sequence is compared against an allele database in order to predict the exact allele number (3). As a result, a vector containing seven numbers can be used to assign a strain to a specific ST (4) and to directly compare the distances of the different strains by visualizing them within a minimum spanning tree (MST) (5). Each node represents a specific ST and closely connected nodes can be summarized as a clonal complex (CC), indicated by the identically colored nodes. The size of a node correlates with the number of strains encoding the same ST within a study set or population. This figure was created with BioRender (April 2021).

MLST data enables researchers to investigate phylogenetic relationships between different genomes and to correlate this analysis with relevant observations. As a result,

pathogenic characteristics associated with specific clusters of genomes, representing pathogenic strains of clinical relevance, have been identified together with their geographic distributions [66, 67].

For *C. jejuni*, it has been shown that specific MLSTs are significantly associated with the presence in certain hosts and that several MLST strains of *C. jejuni* and *C. coli* are frequently identified in human clinical isolates [68, 69] . However, the seven genes included in the MLST scheme of *Campylobacter* with their lengths of around 500 bp only cover 0.002% of the average genome size of the species [32]. Consequently, the overall epidemiological resolution is insufficient for in-depth comparative investigations of distinct bacterial attributes and presences. Utilizing MLST data solely, for instance, is insufficient for outbreak investigations, which rely on genomic approaches providing higher resolution [62, 70, 71].

2.4 The concept of pan-genomics

Computational pan-genomic research was initially introduced in 2005 by Tettelin et al. [28]. In the following years, this novel and holistic approach generated analysis concepts with a set of tools to gain in-depth knowledge about bacterial evolution in novel ecological niches, strain transmission events and adaptive processes induced by selective pressure [27]. In general, the pan-genome is defined as the entirety of DNA sequences of each organism in a set of genomes under consideration. Further differentiation of the pan-genome is provided by the terms core- and accessory genome. The core genome encompasses sequences that are present in all of the genomes of a specific population or a study set [27]. Most of these genes are essential "housekeeping" genes, since they are pivotal to maintain the general cell life cycle and metabolic processes. The accessory genome (also: flexible or dispensable genome) covers all DNA sequences that are present in less than 99% of the genomes under consideration, including unique genes. Those sequences may help the organism to adapt towards new ecological niches and can be acquired through HGT (see Section 2.2.2) [72]. Based on the set theory, the pan-genome is the *union* of the gene sets of all strains of a species, whereas the core genome is the *intersection* of these gene sets (Figure 2.5). The computational determination of the pan-genome is considered as a nondeterministic polynomial time (NP)-hard problem. Thus, computational calculation of the pan-genome increases exponentially with an

increasing number of strains within the study set [73].

With the decrease in costs, WGS has become more frequently used for routine surveillance of different bacterial pathogens. By applying the concept of pan-genomics based on WGS data, different core genome or whole-genome MLST (cgMLST/wgMLST) have been developed. Instead of only utilizing seven housekeeping genes, the cgMLST scheme for *C. jejuni* and *C. coli* incorporates 1,343 different gene loci across the whole genome [74]. These methods all provide a high genomic resolution, which is preferable for outbreak investigations. In addition, these approaches harbor a huge potential to study evolutionary mechanisms.

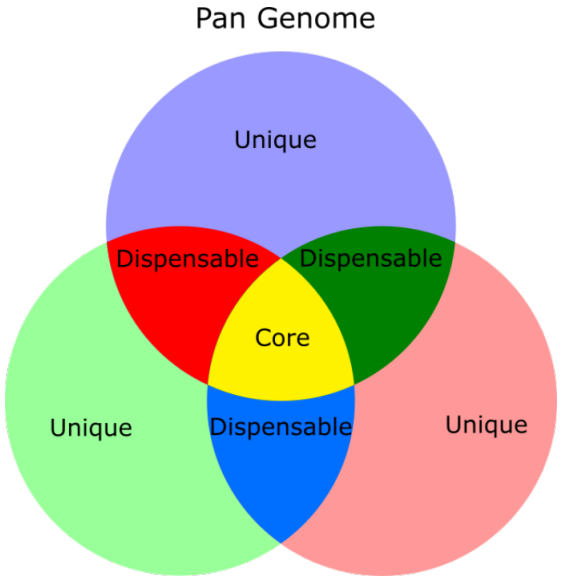


Figure 2.5: Visualization for the concept of the pan-genome based on three distinct (bacterial) genomes. Each circle represents the complete gene set of one genome.

2.5 Population genomics of *Campylobacter*

With the advent of high-throughput NGS experiments, epidemiological studies have improved with more detailed and complex analyses. In addition, novel methods for comparative genomics generated more in-depth knowledge about the population structure of microbial organisms than ever before. In the following section, the population structure of *C. jejuni* and *C. coli* are described.

2.5.1 General population structure

C. jejuni is a natural commensal of the gut microbiota in a wide range of mammalian hosts and birds. However, *C. jejuni* also seems to be associated with environmental reservoirs such as water [75]. This multi-host lifestyle is reflected by the broad diversity of the genomes, which has been detected even by the low-resolution method of MLST. Based on phylogenetic analyses (Figure 2.6) resulting from a concatenated alignment of the alleles of the genes used for cgMLST, *C. jejuni* forms a weak clonal complex structure [57, 76].

CC-45 and CC-21 harbor the most relevant clinical strains and those frequently associated with outbreak events and represent 24% and 9% of the genomes available at PubMLST (`https://pubmlst.org/`), highlighting their general clinical importance. Isolates belonging to these CCs are known to be "host-generalists", which have the ability to rapidly switch between cattle, chicken or humans [68, 69], making them a dangerous threat when acting as a contaminant of food products. Pascoe et al. showed that the geographical signatures in *Campylobacter* are relatively weak and "identical" host-associated lineages have emerged all over the world [77]. However, the frequency of specific STs can change between distinct countries: isolates belonging to the ST-22 (CC-22), for instance, have been predominately identified in Finland [78], while isolates associated with ST-4526 (CC-21) occur in Japan [79]. *C. jejuni* isolates belonging to ST-190 (CC-21) and ST-474 (CC-48) were observed to emerge rapidly in New Zealand [80, 81].

Besides the major lineages of clinical relevance (CC-21 and CC-45), further lineages are known to be frequently associated with isolates of specific host species (Table 2.2): for instance, isolates CC-42 and CC-61 are commonly associated with cattle and sheep [82], whereas several other STs and CCs are associated with isolates from chicken, namely CC-257, CC-353 or CC-443 [43, 83]. Those strains can also cause transient infections in humans through the consumption of contaminated food products. Isolates representing other lineages including CC-177 and CC-682 are often associated with samples from wild birds - a known cause for subsequent water-born *Campylobacter* infections [81, 84].

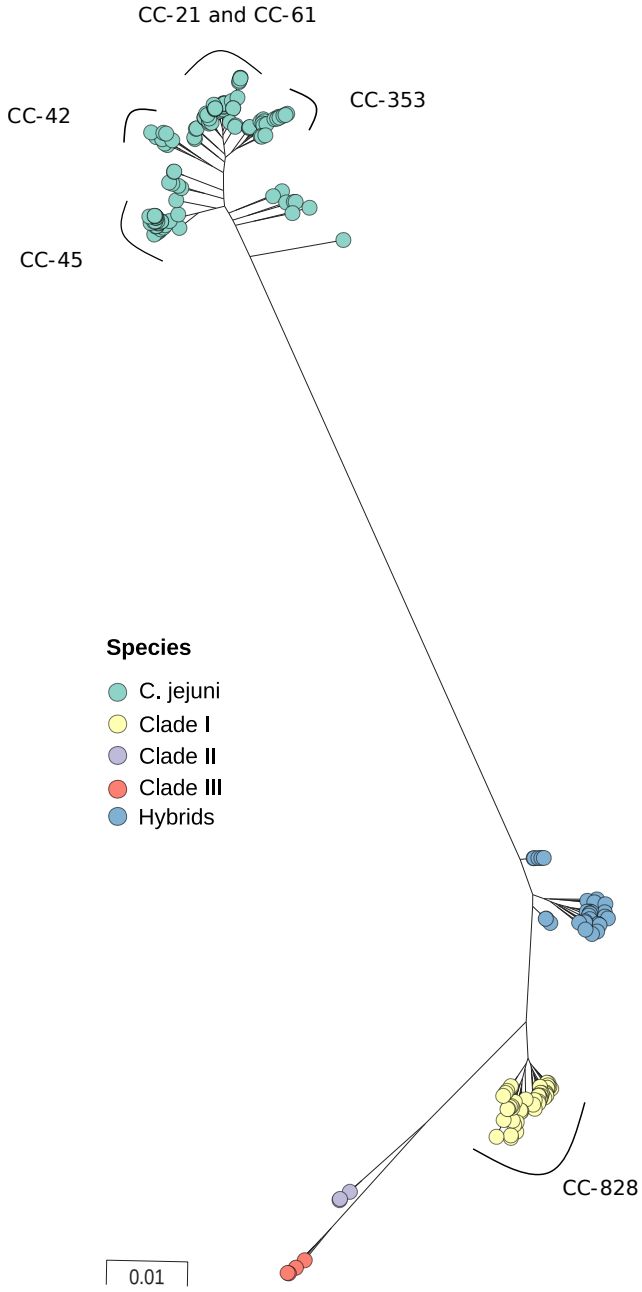

CC-21 and CC-61
CC-42
CC-353
CC-45
Species
C. jejuni
Clade I
Clade II
Clade III
Hybrids
CC-828
0.01

Figure 2.6: Illustration of the core genome phylogeny based on 874 genes with 123,223 variable SNP sites of *C. coli*, *C. jejuni* and *C. coli/C. jejuni* hybrid strains. *C. jejuni* (turquoise) shows a diverse lineage-specific population structure with CC-21 and CC-45 (host-generalists), CC-42 and CC-61 (predominantly isolated from cattle) and CC-353 (chicken associated). *C. coli* consists of three clades with Clade I (yellow; from clinical and farm-related sources), Clade II (purple) and Clade III (red, both from waterfowl and water samples). Clade I mainly consists of genomes belonging to CC-828. "Hybrid genomes" with high DNA introgression from *C. jejuni* are colored in blue. Data was taken from [36],[18], [39] and the phylogenetic tree was created with FastTree v2.1 [85]. This figure was originally created for publication in the article "Population Biology and Comparative Genomics of *Campylobacter* Species" [40].

The local diversity of *C. jejuni* within a single barn or herd needs to be addressed as well: Reports showed, for example, that more than 10 distinct CCs might occur in the same chicken flock [86, 87], sometimes even comprising distinct accessory gene sets.

Even among isolates that share the same ST, the genetic diversity can vary widely: isolates of ST-230, ST-267 and ST-677 that have been identified during a clinical outbreak in Finland, showed allelic differences in up to 40 out of 1,200 genes within each ST based on WGS data. Additionally, isolates of ST-45 carry up to 400 different allelic variants within their core genome [88]. In general, *C. jejuni* has a strong host-genotype relationship based on MLST as well as WGS data, which will be analyzed in section 5.

In contrast to *C. jejuni*, *C. coli* forms three distinct clades (I-III) (Figure 2.2 and Figure 2.6), each of them colonizing a distinct ecological niche. *C. coli* clade I is associated with agricultural origin, whereas *C. coli* clades II and III are mostly found in water-associated environmental sources [25, 36, 89]. At present, around 81% of the *C. coli* genomes listed in PubMLST belong to clonal complex CC-828 and are part of clade I. The high proportion of clade I/CC-828 reflects the clinical relevance and prevalence of this lineage worldwide [90–93]. The second-most predominant clonal complex (CC-1150), which comprises around 5% of the genomes submitted to PubMLST, is also integrated into clade I. Of note, genomes of clade I are less diverse compared

Table 2.2: Summary of lifestyle preference for exemplary *C. jejuni* clonal complexes.

Clonal Complex	**Lifestyle Preference**	**Reference**
21	host-generalist	[68, 69]
45	host-generalist	[68, 69]
48	host-generalist	[43]
42	cattle/sheep	[82]
61	cattle	[82]
257	chicken	[43, 83]
353	chicken	[43, 83]
443	chicken	[43, 83]
177	wild bird	[81, 84]
682	wild bird	[81, 84]

to genomes of *C. coli* clade II, III or the general population of *C. jejuni* [23, 94, 95]. The sparse amount of variation within the housekeeping genes as well as the limited lineage diversification most likely indicates a recent bottleneck event and thus an early phase of lineage separation within the *C. coli* population [95]. The formation of the three distinct clades might be a result of the colonization of distinct ecological niches by *C. coli*. The ecological separation probably forms a recombination barrier between the clades [95]. In addition, recombination between *C. coli* clade I and *C. jejuni* was shown by several studies [25, 32, 39, 95] - events that resulted in hybrid strains. Contributions and discovery of hybrid strains have also been made in the scope of this work and are described and discussed throughout section 6.

3 Computational background

3.1 Concept of *k-mer*-based methods

Different concepts based on *k-mer* frequencies are currently used within computational genomics, often as a key method in several bioinformatics applications: recent genome assembly tools, for instance, rely on *k-mers* to identify overlaps by overlap-layout-consensus methods (i.e. Celera assembler [96] and Canu [97]) or building a de Bruijn graph (i.e. Velvet Assembler [98] and SPAdes [99]). Further, *k-mer* tools allow a reference-free analysis of genomic sequences e.g. in metagenome studies with tools like kraken [100], kraken2 [101] or kmerFinder [102–104] and alignment-free estimations of similarities or distances between thousands of input sequences using mash [105].

K-mers are defined as all possible substrings of a fixed length k of a string s. Each *k-mer* overlaps with the previous and subsequent *k-mer* of a string in *k-1* characters. Therefore, the total number of *k-mers* (n) from a string s of length L can be calculated as:

$$n = L - k + 1 \tag{1}$$

In computational genetics, the complexity of *k-mer* counting is reduced since an input string s is commonly a combination of an alphabet based on one or more DNA nucleotides (A, C, G, T). In this context, 4^k is the maximum number of different *k-mers* of length k. In general, counting *k-mers* in a string is a simple and straightforward problem by processing each sequence character by character.

While using *k-mer*-based methods for WGS analysis it is important to choose the optimal *k-mer* size. A shorter *k-mer* size helps to keep the memory requirements low, but is also very likely to be present in a lot of sequences. A longer *k-mer* size on the other hand will increase the memory usage but is likely to provide *k-mers* that are unique for a sequence which will avoid problems in repeat-rich regions. Longer *k-mers* decrease the number of correct *k-mers* present in the data as they are prone to be excluded due to “errors” or SNPs in the sequence, since *k-mer*-based approaches rely on exact sequence comparisons [106]. A long *k-mer* that includes an "error" might introduce an

unconnected node within a *de novo* assembly approach or cannot be identified by a metagenome analysis. In other words, small *k-mers* are more sensitive whereas longer *k-mers* are more specific [101, 106] .

K-mer-based assembly tools such as Velvet or SPAdes try to address the problem to choose an appropriate *k-mer* size by iteratively increasing the length of *k-mers* and combining the assembly results based on different sizes in a consensus method [99, 107]. Studies utilizing metagenomic tools such as kraken2 often estimated an fixed *k-mer* size experimentally by either optimization with regards to the accuracy of the prediction or the computational resources required by the software [101].

K-mer-based approaches can be seen as alignment-free methods for sequence comparison that utilize *k-mer* abundance distributions and offer several advantages contrary to alignment-based methods [108]. Alignment-based methods have major computational drawbacks with increasing sequence length or dataset size, while *k-mer*-based methods scale well with large datasets. *K-mer*-based methods have a lower complexity compared to alignment-based methods, such as local and global alignments as implemented in basic local alignment search tool (BLAST) [109, 110] or Clustal W [109]. Alignment-based methods commonly have a computational complexity of O(mn) to align two sequences of length n and m and quickly require large amounts of computational resources for larger sequences. *K-mer* counting and comparison based on frequencies generally scale linearly with increasing sequence length [111]. Furthermore, *k-mer*-based approaches are made on exact comparisons of subsequences and do not rely on heuristics or underlying substitution or evolutionary models [112], which makes them easy to use for downstream and statistical analyses. On the downside, *k-mers* are less suitable to compare highly different sequences as they are not able to deal with stochastic sequence variations, homology and cannot determine recombination events directly [108].

3.1.1 *K-mer* counting concepts

In order to compare *k-mer* frequencies or abundances of two or more genome sequences, *k-mers* within each sequence have to be counted. With millions or billions of sequencing reads and increasing *k-mer* sizes, *k-mer* counting becomes computational challenging and requires considerable working capacity as well as large amounts of memory

resources. Therefore, previous research focused on improving *k-mer* counting methods through the implementation of disk-based approaches as opposed to in-memory/internal memory approaches, since disk space is commonly more salutary than memory expansion. In general *k-mer* counting is done based on two concepts, **Hashing** or **Sorting**:

Hashing: The naive *k-mer* counting approach uses a data structure in form of a dictionary/hashtable that is stored in memory and in which unique *k-mers* are represented as keys and their counts as values [113]. The position of a *k-mer* within the hashtable is defined by a hash function. If a *k-mer* already exists the count gets incremented by 1. In case the *k-mer* occurs the first time, it is initialized at the position with a count of 1. Under the circumstances that a *k-mer* does not exist in the hashtable, but the position is already occupied (collision), a probing strategy (quadratic probing, double hashing or linear probing), also called open addressing, is used in order to effectively calculate a new position. Marçais et al. [114] developed memory-efficient *k-mer* counting approaches based on a lock-free hash table to overcome the limitations of memory for this kind of *k-mer* counting. Instead of doubling the size of the hashtable once it is full, the hashtable is written to the disk and merged with the intermediate *k-mer* counts. To further reduce computational time and memory usage, Bloom filters can be used to remove *k-mers* that only occur once per dataset (singletons) [115].

Sorting: In a sorting approach, every single *k-mer* in the dataset is determined and stored in a list. Afterwards the list is sorted, which causes identical *kmers* to be shifted to adjacent positions. The difference between the index of the position of a first and a last identical *k-mer* are thus its total number of occurrences [113]. In order to further reduce memory space and redundancy of the *k-mer* counting, several tools such as MSPKmerCounter, KMC2 or DSK utilize the concept of "minimizers" [116]. Minimizer describes the lexicographically smallest substring (*m-mer*), where $m < k$. As neighboring pairs of *k-mers* share $k - 1$ bases it is most likely that adjacent *k-mers* share the same minimizer sequence. Thus, adjacent *k-mers* with the same minimizer can be represented by one "super *k-mer*" of more than k characters, leading to a significant reduction of memory or disk space in order to store *k-mers*[117].

3.2 Genome-wide association studies

Genome-wide association studies (GWAS) have been widely applied since the first GWAS was published in 2005 [118]. GWAS are widely used on data from human genetics and have revolutionized the process of genetic disease detection. Currently, more than 50,000 associations of particular loci with common human diseases have been identified, providing in-depth insights into biological mechanisms and architecture of genetic predispositions [119]. Type 2 diabetes mellitus (T2DM), inflammatory bowel disease or even schizophrenia have, besides many other illnesses, been successfully linked to genetic markers through GWAS [120, 121]. Identification of genetic markers support the translation of biological knowledge directly to medical applications. The intuitive usage of variants discovered by GWAS is either to use markers for early detection and prevention of genetic diseases or to use genetic variants to classify subtypes of diseases, e.g. in cancer or diabetes mellitus [122]. Additionally, GWAS results can also lead to the discovery of novel drug targets, which might reduce the cost and improve the success rate [123]. Target genes and disease-causing pathways identified by a GWAS can be investigated in downstream experiments to enhance development of novel therapeutical strategies [123].

The GWAS concept is based on the assumption of the common disease common variant (CDCV) hypothesis, where a risk of a certain disease depends on many high-frequency but low-effect SNPs [124, 125]. GWAS represents a typical top-down approach, providing several major advantages over bottom-up approaches, for instance, gene knock-out experiments, where the knock-out effect on the consecutive phenotype appearance of a particular loss of function gene variant is observed. GWAS do not rely on a predefined working hypothesis, since all genomic regions can be tested simultaneously. Genetic variants occur naturally in the study population and any measurable phenotype can be tested. This methodology enables researchers to test hundreds of thousands to millions of genetic variants to identify risk factors that are linked with a certain phenotype.

Apart from SNPs, other types of genetic variation like copy number variants (CNVs) can also be linked with disease susceptibility [126].

The basic workflow of GWAS consists of 5 steps shown in Figure 3.1:

1. Collection of case and control samples: In order to perform a reliable GWAS an appropriate minimum sample size is required. Human genetic research with GWAS commonly include at least 10^3 up to 10^6 samples [127], whereas microbial GWAS may consider only 100 isolates per group, as suggested by Lees et al. [128].

2. Extraction of sequence variants: All genome sequence variations in a sample set under consideration are identified. In the case of human genetic research the variants are predominately encoded by SNPs, whereas results for bacteria often include SNPs in the core genome as well as variations of the accessory genome. Recently, *k-mers* have been used to set up GWAS to analyze WGS data from bacterial genomes [128].

3. Use of regression analysis to investigate phenotype-genotype correlation: A regression analysis will be performed according to the phenotypic data and population structure. A continuous phenotype commonly requires a linear regression, whereas binary phenotypes are analyzed with a logistic regression (see details in Section 3.2.2).

4. p-value calculation of the β coefficient: In order to identify reliable associations, the impact of the beta coefficient on the slope of the regression curve is calculated. The stronger the associations, the lower its p-value [129].

5. Determination of the genomic position of significant variants: The significant genetic variants identified in the previous step are mapped to the genome structure to allow further downstream analysis, including biological interpretation of the results [130].

Bacteria or other microbial organisms have interesting heritable phenotypes that can be investigated by GWAS in the same manner [131]. Due to the low genetic variance within the human genome in comparison to the genetic variance of microbial organisms, GWAS is commonly focused on the detection and effects of SNPs. Adaptation mechanisms in microbial organisms, especially bacteria depend on a variety of individuals SNPs but also on complete genes or gene operons that have been derived through HGT. Therefore, microbial GWAS needs to be adapted in several ways.

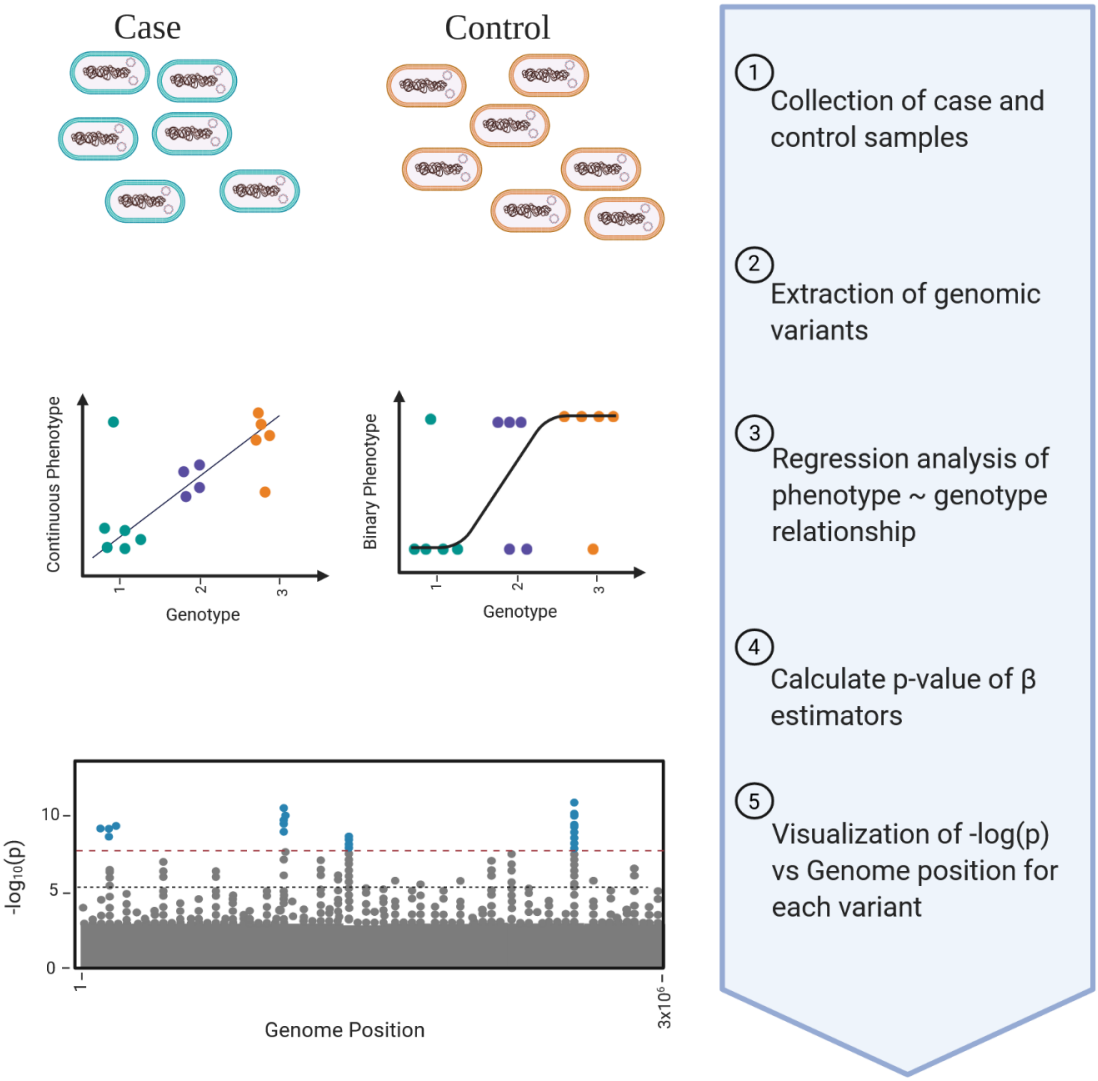


Figure 3.1: Conceptual GWAS workflow. The basic workflow of GWAS consists of 5 steps: (1) Collection of case and control samples; (2) Extraction of sequence variants; (3) Regression analysis to investigate phenotype-genotype correlation; (4) p-value calculation of the β-coefficient; (5) Determination of the genomic position of significant variant. This figure was created with BioRender (April 2021).

3.2.1 The era of microbial *GWAS*

Initial GWAS performed during research in human and microbial genetics commonly involved large microarrays to capture single SNPs among the genomes under consider-

ation. Most of the currently known bacteria vary widely in terms of diversity, genome size, gene content and mutation rates. Due to the constantly decreasing WGS costs, the amount and availability of WGS data of different bacterial populations accompanied by detailed phenotypic information such as antibiotic resistance, virulence or host specificity, has rapidly increased [132]. Establishing microbial GWAS based on whole-genome data provides an adequate possibility to study the contributions of bacterial population variation with these traits. The first microbial GWAS was performed by Falush and Bowden in 2006 [133], and since then microbial GWAS has become a new subject area of genomic research in microbiology. To date (2021), publications on bacterial GWAS results are still rare, especially when compared with the wide employment of GWAS in the field of human genetics [134]. Successful application of GWAS in bacterial research identified genetic determinants in *S. aureus* that are associated with the skeletal disease pyomyositis [135], the severity of pneumococcal meningitis caused by *Streptococcus pneumoniae* [136] or the detection of genetic signatures for extra-intestinal virulence in *E. coli* [137]. Despite the linkage of risk factors with certain diseases carried by different bacterial genera and species, microbial GWAS have been commonly used in order to identify antibiotic resistance associated genes and mechanisms, for instance in *Mycobacterium tuberculosis* [138] and *Plasmodium falciparum* [139].

To apply the general concept of GWAS frequently used in human genetics to microbiological research, four major issues need to be addressed thoroughly:

1. High diversity within bacterial populations. As explained above, GWAS in human genetic research are focused on allelic markers such as SNPs that can be identified on multiple genetic backgrounds [140]. Bacterial genomes similarly develop point mutations, small insertions or deletions as part of their natural evolution, but sequence diversity within a bacterial species can vary widely [134]. This circumstances makes it a nontrivial task to select a suitable reference sequence for SNP calling for a whole bacterial population [141]. In addition, bacteria tend to have a flexible genome, which includes, besides the core genome content, also genetic information which is not present in all genomes of the population under consideration: The entirety of these optional sequences are referred to as accessory genome (detailed information are provided in Section 2.4).

GWAS performed with bacterial populations were originally either based on SNP analysis of the core genome or correlation of the presence or absence of certain genes of the accessory genome content with a phenotype. In 2013 a first *k-mer* approach was published by Sheppard et al. providing the possibility to apply GWAS on the core and accessory genome content of the *C. jejuni* population [18]. Their aim was to estimate the statistical association of presence or absence of *k-mers* of *C. jejuni* genomes adapted to cattle or chicken hosts. For this purpose, a fixed *k-mer* length of 30 bp was used. To overcome obstacles in terms of population diversity (see next paragraph), Sheppard et al. utilized a Monte-Carlo simulation to predict to gain and loss of certain *k-mers* along phylogentic branches [18]. However, due to their iterative character, Monte-Carlo simulations need a considerable computation time, which limits their usability with respect to large scale genome comparisons [128].

2. Strong population structures. Diversity within the bacterial population often leads to a strong population stratification [142]. Overall adjacent phylogenetic subgroups (including: lineages or clades) within a bacterial population have a closer genomic relationship than more distantly clustering members of the same population [142]. As a result, specific allele variants can occur frequently within the same genetic background [140]. This population structure effect might lead to a lineage-specific association of a sequence or an allelic variant rather than to a causal phenotype-genotype link [143, 144]. For instance, assume an ancestral bacterial genome carried a mutation that was causally linked to a specific phenotype. Many or even all of the genomes representing the descendants from the original strains harbor the same mutation, which is still associated with the phenotype of the bacteria. However, since the original mutation affecting the ancestor, many other genomic events occurred within the specific clade/lineage of the bacterial population. These genetic changes could appear equally as being associated with the phenotype when applying the GWAS [143].

3. Variation of linkage disequilibrium (LD). LD describes the non-random association of alleles at different loci in a population under consideration [145]. In general, the level of LD is higher between physically closely related alleles in a genome. However, the LD can be reduced by recombination events, especially in distant alleles. Recombination in eukaryotic genomes occurs only once in the germ cell, involving two homologous

chromosomes during crossing-over, the exchange step of meiosis [146]. Therefore, both, the recombination rate and levels of LD between two arbitrary loci stay rather constant considering the genomes of the human population. In comparison to humans, recombination among bacterial cells is a repetitive process with stretches of DNA fragments of different sizes. The extensive exchange of genetic material can either occur regularly over time or due to specific adaptation processes (Section 2.2). As a result the level of LD may vary between the same loci within different lineage of the same bacterial population. Of note, recombination does not occur in all bacterial species to the same extent. The rates of DNA exchange can range from none to high frequencies depending on the species or even the lineage, resulting in strictly clonal to panmictic (all individuals are potential ancestors) population structures (see Section 2.2.2). Representative species showing these differences are for example *Staphylococcus aureus* (strictly clonal) or *H. pylori* (panmictic). Bacterial species with high recombination rates, are often associated with a low LD, whereas species without any recombination are in complete LD.

4. Natural selection relative to genetic drift. Novel bacterial phenotypes are commonly shaped by natural selection, for instance positive directional selection driving AMR [147]. Environmental events resulting in a sharp reduction of the population size are commonly known as "bottleneck events" [148]. While bacteria experience genetic drift (particularly in frequently bottlenecked populations), many interesting traits (e.g. resistance, virulence, host-association) have evolved recently and under strong positive selection. In addition, these bacterial traits might be associated with mutations having a large effect sizes on the novel phenotypes [142]). Consequently, a relatively low number of bacterial genomes should be sufficient to identify causal mutations in such cases [142, 149].

3.2.2 Correction for lineage effects

Regression analyses are a set of statistical modeling methods used to asses the relationships between a dependent variable (observed effect or experimental outcome) and one or more independent variables/features [150]. Generally, regression analyses are used for two main purposes, either forecasting specific events, in which the dependent

variable is calculated based on a given model, or to infer causal and associated relationships between independent and dependent variables [150]. Within a GWAS approach, regression is used to detect associations between phenotypes and genotypes that may have a causal effect [131].

In the following the dependent/observed variables are described as the vector $Y = (y_1, y_2, ..., y_n)$ with number of observations (n) and the independent variables are given as vector $X = (x_1, x_2, ..., x_n)$. In case of GWAS the independent variables, described by X, can be the frequencies, counts or presence/absence of a SNP, *k-mer* or gene. β_0 is the expected mean value of Y when all $x_i = 0$ for $i = 1..n$, also known as the intercept. The slope of the regression curve is given by β_1 and the error terms are given by vector $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_n)$.

The regression method of choice depends on the type of the observed/depended variables. The easiest way is to construct a simple linear regression model which can be used for continuous phenotypes. A linear regression can be described as follows (Figure 3.2):

$$Y = f(x) = \beta_0 + \beta_1 \cdot X + \epsilon \tag{2}$$

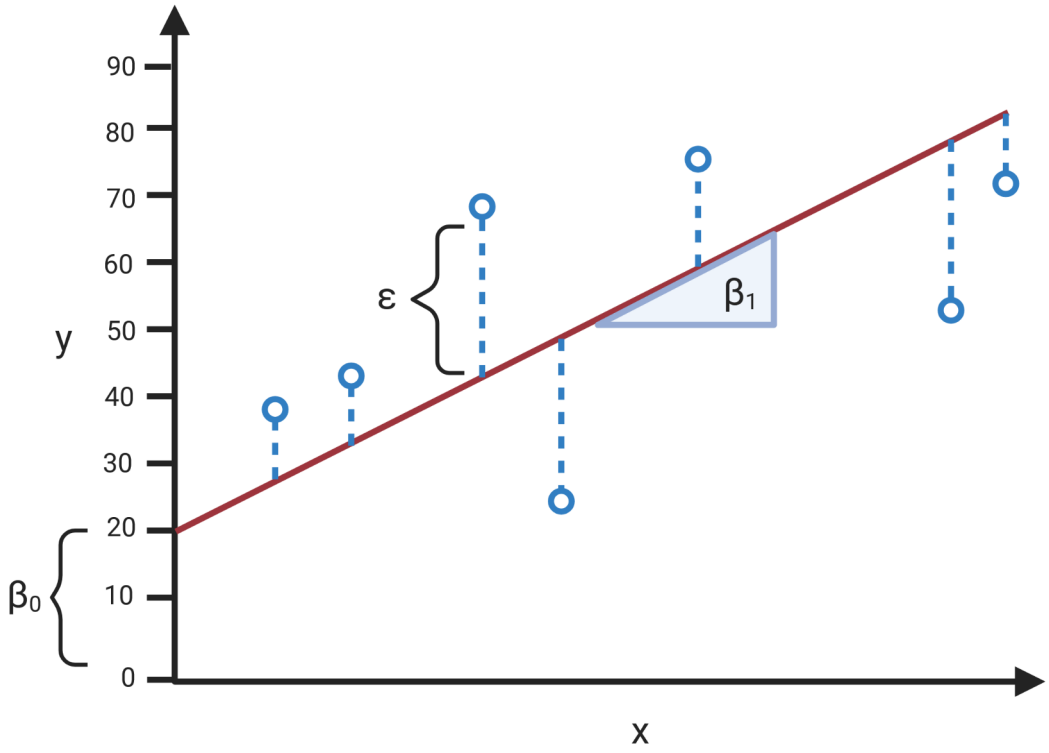


Figure 3.2: Schematic visualization of a linear regression model for continuous y values. The blue dots represent the data points that were used to calculate the regression line (red). β_0 describes the intercept, β_1 the slope of the regression line and ϵ the error terms.

The regression coefficients for the linear regression model is estimated by minimizing the error term $\epsilon = Y - \beta_0 - X \cdot \beta_1$. The minimization of the residual sum of squares (RSS) is defined as:

$$\hat{\beta} = \arg\min \mathrm{RSS}(\beta_0, \beta_1) \tag{3}$$

where RSS is defined as:

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{4}$$

In order to calculate a GWAS with respect to binary phenotypes such as AMR, pathogenicity or association with a certain host, a logistic regression instead of a linear regression needs to be applied. A logistic regression calculates the odds of association of a given

phenotype and can be described as:

$$log\left(\frac{P(X)}{1-P(X)}\right) = \beta_0 + \beta_1 \cdot X + \epsilon \tag{5}$$

$$\frac{P(X)}{1-P(X)} = e^{\beta_0+\beta_1 \cdot X+\epsilon} \tag{6}$$

where P(X) is defined as the sigmoid function, that transforms a real number to a value between [0,1]:

$$P(X) = \frac{e^{\beta_0+\beta_1 \cdot X}}{1+e^{\beta_0+\beta_1 \cdot X}} \tag{7}$$

The regression coefficients in the logistic regression model is estimated by minimizing the posterior probability, which can be done by maximizing the likelihood L estimator:

$$\hat{\beta} = \arg\max \mathrm{L}(\beta_0, \beta_1) \tag{8}$$

where the L can be written as the product of probabilities for both binary outcomes:

$$L(\beta_0, \beta_1) = \prod_{i:y_i=1} P(x_i) + \prod_{i:y_i=0} 1 - P(x_i) \tag{9}$$

In microbiological research, a typical approach would link specific genetic markers (*k-mers*) with phenotypical traits, i.e. infection or incubation time, growth rates or gravity of illness [134]. In order to correct for lineage effects/population stratification an accurate phylogenetic tree is required, which is combined with a principal component analysis (PCA) to adjust for fixed effects within the regression model [143]. A PCA reduces the data to lower dimensionality(principal components) via geometrical projections with the goal to summarize all features [151].

If leading principal components are included as fixed-effect covariates in the association model of a GWAS, the lineage effect can be corrected. A similar effect can be achieved by excluding closely related individuals, however this often comes with a significant loss in statistical power [152, 153]

Another way to correct for the lineage effects and confounding factors in the linear regression model, is to incorporate a similarity matrix, called kinship matrix. A kinship matrix can be calculated with alignment-free methods such as mash [105], which saves time and resources for computational steps such as whole-genome alignments or construction of an exact phylogenetic tree. The linear mixed regression model is constructed as follows:

$$Y = f(x) = \beta_0 + \beta_1 \cdot X + u \cdot K + \epsilon \tag{10}$$

where u is an unknown vector of random effect coefficients, with mean $E(u) = 0$ and K is the kinship matrix, incorporating phylogenetic distances or similarity scores to measure genetic covariance between individuals.

4 Methods

Chapter four is structured into two parts. At first, the composition of the datasets, the metadata associated with the original isolates and information on the sequencing results of the datasets are described. The second part provides detailed information on the *in silico* methods and workflows used to process and analyze the data.

4.1 Datasets used in this thesis

4.1.1 Dataset 1: PAC-Campy strain collection

Dataset 1 was collected within the German research consortium "PAC-Campy - Preventing and Combating Campylobacter Infections: On Track towards a One Health Approach" (http://www.pac-campy.de/), funded by the Federal Ministry of Education and Research. One goal of the PAC-Campy research consortium was to identify host-specific determinants of *C. jejuni* to improve outbreak investigations and source attribution models. For this purpose, a uniform stratified random collection comprising 324 *C. jejuni* isolates obtained from samples of four different sources, including human [n=96], chicken [n=102], cattle [n=98] and pig [n=28] originating from the 16 federal states of Germany between 2010 and 2017 (appendix Table A.1) was subjected to WGS. Since *C. jejuni* is not particularly known as a common commensal residing in the porcine gut, availability of isolates representing this origin was limited. The set of genomes was complemented by whole genome data of further 166 isolates from a Canadian study which included *C. jejuni* from cattle [n=39], chicken [n=12], human clinical cases [n=40], environmental [n=54] and other animal [n=21] origins [19]. The original purpose of the Canadian study was to identify diagnostic markers which can be used for rapid screening approaches detecting *C. jejuni* subtypes that pose an increased risk to human health [19].

DNA for WGS was prepared using the PureLink Genomic DNA Mini Kit (Thermo Fisher Scientific Inc., Waltham, Massachusetts, USA) or the DNeasy Blood & Tissue Kit (QIAGEN, Hilden, Germany). WGS sequencing libraries were generated with the Nextera XT (Illumina Inc., San Diego, CA) library kit following the manufacturer's instructions. Sequencing was performed on a MiSeq sequencer (MiSeq Reagent Kit v.3; Illumina Inc., San Diego, CA) resulting in 300 bp paired-end reads and an average

coverage of 80x, and on a HiSeq 1500 using a PE Rapid Cluster Kit v2 and Rapid SBS Kit v2 (500 cycles; Illumina Inc., San Diego, CA) resulting in 250 bp paired-end reads and an average coverage of 80x [154].

The DNA of strain BfR-CA-14430 was prepared for MinION sequencing using the QIAamp DNA Mini Kit (QIAGEN) and was further concentrated by precipitation with 0.3 M sodium acetate pH 5 and 0.7 volume isopropanol at room temperature for 30 min. After centrifugation and washing of the precipitate with 70% ice-cold ethanol, the DNA was dissolved in Tris buffer pH 7.5. The quality of the DNA was evaluated by spectral analysis (NanoDrop Spectrophotometer, Thermo Fisher Scientific, USA) and the concentration was fluorimetrically adjusted to 110 ng/µl by Qubit 3.0 Fluorometer (dsDNA BR Assay Kit; Invitrogen, USA). DNA samples of BfR-Ca-11439 were additionally controlled for lack of sheering products < 20 kb on a 0.8% agarose gel [155].

4.1.2 Dataset 2: *Campylobacter* obtained during food-chain monitoring

During routine monitoring and zoonosis surveillance along the food chain, more than 4000 *Campylobacter* isolates were sampled by the German Federal State Laboratories between January 2016 and December 2018. *C. jejuni* and *C. coli* field strains were isolated from different food matrices and animal samples by the Federal State Laboratories according to ISO 10272 [156].

For species verification, real-time polymerase chain reaction (PCR) [157] was employed, with the genes *map*A (encoding a fitness factor relevant for chicken colonization) specific for *C. jejuni* and *ceu*E (encoding a factor enhancing iron acquisition abilities) for *C. coli* [157–159]. However, several samples showed ambiguous PCR results. In this study, 37 of these isolates originating from various animal and food matrices (chicken meat [n=9], duck meat [n=1], eggs [n=9], turkey cecum [n=12], turkey meat [n=5], turkey skin [n=1]), have been investigated. Further isolates from eggs [n=8] were investigated as well, as the prevalence of ambiguous quantitative polymerase chain reaction (qPCR) results from eggs were surprisingly high [39].

To evaluate phylogenetic relationships of the 45 genomes described above with the *Campylobacter* population, 21 additional genomes with unambiguous qPCR results from the German Federal Institute for Risk Assessment (BfR) and 247 closed *Campylobacter*

genomes available from National Center for Biotechnology Information (NCBI) were utilized (appendix Table A.2).

Isolates were paired-end sequenced on the Illumina MiSeq (2×301 cycles) or the NextSeq (2×151 cycles) platform using the MiSeq v3 (600 cycles) reagent kit or the NextSeq 500/550 Mid Output kit v2.5 (300 cycles), respectively. The sequences were published within the BioProject No. PRJNA595957, BioSample No. SAMN13577876 - SAMN13577920, sequence read archive (SRA) accession No. SRR10698060 - SRR10698104 at NCBI (appendix Table A.2).

Wet-lab workflows for qPCR and matrix-assisted laser desorption/ionization-time-of-flight (MALDI-TOF) were performed by Julia Golz and Kerstin Stingl at the BfR. In order to enhance understanding of the interdisciplinary approach set-up within the PAC-Campy research network, a brief overview of the methods used is provided in the appendix A.3.

4.2 *In silico* methods

4.2.1 Whole-genome sequence analysis

WGS data analysis performed in this work is based on the *de novo* assembly of each genome of dataset 1 and 2. In order to prepare the Illumina raw read data of both datasets, several pre-processing steps had to be performed. Pre-processing of the sequencing data included quality control by use of fastQC [160] v0.11.7 and standardized quality and adapter trimming of all reads by flexbar v3.0.3 [161]. Additionally, all paired-end reads were scanned for reads which most likely arose from contamination by Kraken2 v2.0.8-beta [101].

The pre-processed reads were *de novo* assembled with SPAdes v3.11.1 [99]. SPAdes utilizes a Bayesian approach, namely BayesHammer [162], for read error correction of Illumina paired-end reads. The assembly process itself is based on a de Bruijn graph approach [163] that applies the idea of iteratively increasing *k-mer* sizes in order to construct a graph.

In addition to short-read sequencing, long-read sequencing with a MinION from Oxford Nanopore Technologies (ONT) was performed for the reference strain BfR-CA-14430.

Combining data from both technologies is a sufficient and frequently used approach to reconstruct closed bacterial genomes and plasmids with high accuracy [164].

In order to perform a hybrid assembly based on both sequencing technologies, long-read data were quality filtered by removing all reads with an average base quality Phred score < 8. As described for the Illumina paired-end reads, contamination control was done by Kraken2 as well for long-read data. Construction of the hybrid assembly was performed by Unicycler v0.4.7 [164], that utilizes SPAdes for a short-read de novo assembly in the first step. The resulting contigs of the draft genome are merged by use of miniasm [165] based on ONT long-read data and circularized in an additional step. Finally, the resulting assembly is post-processed in order to polish sequencing errors from ONT data by Pilon [166]. A comparison of results obtained from Illumina short-read sequencing in combination with PacBio and ONT long-read sequencing with various assembly tools and further genetic characteristics such as virulence and antibiotic resistance genes of the strain BfR-CA-14430 has been performed during this work. Further information on the comparative approach is provided by our study published the BMC journal Gut Pathogens:

Epping L, Golz JC, Knüver MT, Huber C, Thürmer A, Wieler LH, Stingl K, Semmler T. Comparison of different technologies for the decipherment of the whole genome sequence of *Campylobacter jejuni* BfR-CA-14430. Gut Pathogens. 2019 Dec 1;11(1):59.

4.2.2 *In silico* MLST

In silico determination of MLST and CCs for *C. jejuni* and *C. coli* was performed employing a BLAST-based pipeline (`https://github.com/tseemann/mlst`) considering seven housekeeping genes (*asp*A, *gln*A, *glt*A, *gly*A, *pgm*, *tkt* and *unc*A) [57, 64]. MLST profiles, allelic variants and CC information were downloaded from `https://pubmlst.org/campylobacter`. The allelic profiles from dataset 1 and dataset 2 were used to calculate a MST which was consecutively visualized by use of Grapetree [167].

4.2.3 Gene prediction

De novo assembled genomes have been further inspected by *in silico* prediction of CDSs and additional non-coding genetic elements like transfer ribosomal ribonucleic acids

(tRNAs), non-coding RNAs, and signal leader peptides using the pipeline for prokaryotic genome annotation (Prokka) [168]. Prokka combines several different approaches and state-of-the-art tools such as Hidden Markov Models (HMMs), BLAST and Prodigal (PROkaryotic DYnamic programming Gene-finding ALgorithm) [169]. As a result, an annotated version of each genome was generated ready to be used in further downstream analysis shown in the next section (Section 4.2.4).

Additionally, Prokka provides the utility to create a customized BLAST-based database for non-generic tasks. As *Campylobacter* is a non-model organism, this functionality of Prokka provides the possibility to generate species and genus-specific annotation databases. For this purpose, closed and well annotated reference genomes from NCBI were used as input.

4.2.4 Pan-genome analysis

The pan-genome (Section 2.4), including the core and accessory genome of dataset 1 as well as dataset 2 were assessed by use of the salable and rapid pan-genome method Roary v3.13.0 [170]. Since this pipeline requires annotated genomes as input data, annotations were assigned by use of Prokka (Section 4.2.3). The working process of Roary can be summarized as follows: Initially, a pre-clustering of all predicted protein sequences from the annotated draft genomes is performed by CD-hit [171]. The implementation of this step has significantly improved the runtime of the tool [170]. Then a more accurate all-against-all BLAST comparison is accomplished to identify both, orthologous core- and accessory genes. Both steps strongly depend on the threshold chosen for sequence identity to define orthologous genes [170]. Since dataset 1 consists of *C. jejuni* isolates only, the minimum percent of sequence identity was set to 95% for orthologous genes. The isolates sampled along the food-chain comprised by dataset 2 contain *C. jejuni* as well as *C. coli* genomes. Accordingly, the threshold for orthologous gene detection was set to 80% sequence identity.

Identified core genes were further processed to generate a core genome alignment. The runtime to calculate a multiple sequence alignment (MSA) increases exponentially with $O(k^n)$, where k is the total number and n the length of the sequences. Each core gene is separately aligned by the heuristic approach MAFFT [172] and the gene-wise alignments

are concatenated to the complete core genome alignment afterwards.

Besides the core genome alignment, Roary generates a gene-presence-absence table that can be translated into a binary matrix M representing the accessory genome profiles for all genomes. The size of the binary matrix M is calculated with $n \cdot m$, where n is the number of genomes and m the number of genes. T-Distributed Stochastic Neighbor Embedding (t-SNE) was used to transform the high-dimensional structure of the matrix M into a 2-dimensional space. In general, t-SNE is a non-parametric and non-linear approach that calculates the euclidean distances between data points in a high-dimensional space. Distance matrices are then transformed to a probability score utilizing a t-distribution for all the input values. Data points that actually appear closely together in the high-dimensional space, have an increased impact on the final probability score [173]. The same concept is used to calculate the probability score for the data values in the low-dimensional space. By a gradient descent function, t-SNE minimizes the difference between the probability scores of both spaces while reordering the values in the two-dimensional space, in iterative steps.

To investigate similarities of accessory genome profiles of dataset 1, Rtsne v0.15 was used together with the Barnes-Hut algorithm in R v.3.4.1. The transformed data was later visualized in microreact in form of a network [174].

4.2.5 Prediction of phylogenetic relationships

Correct prediction of genetic relationships and phylogenetic relatedness between genomes is a key step for population and outbreak studies in molecular microbiology [175]. Two different approaches to define and classify species boundaries are commonly used and applied in scope of this work for in-depth strain and lineage interference analysis of *Campylobacter*.

Since the 1960's, the wet-lab molecular biological DNA-DNA hybridization (DDH) technique has been used for taxonomic classification purposes among bacteria. DDH makes use of the hybridization reaction of a DNA single-strand in order to identify the similarity between a probe and a reference DNA [176]. DNA strands with a high degree of similarity form more hydrogen bonds of complementary base pairs than samples with lower sequence similarities. The DDH approach used 70% similarity as a threshold

for DNA belonging to the same bacterial species [177]. Nowadays, average nucleotide identity (ANI) analysis based on WGS data is the most commonly used method for *in silico* species verification based on WGS data [178]. Species boundaries are defined by an ANI value of 95% sequence identity [176]. In this work, FastANI v1.3 [179], a tool based on Mashmap [180], was utilized. Since ANI is calculated pairwise between all genomes, FastANI uses a heuristic to efficiently calculate those values. Results were visualized by the R package "pretty heatmap" v1.0.12 [181]. However, an overview on overall genome similarities does not provide further information about evolutionary relationships between different lineages or genomes. Therefore, a high-resolution phylogenetic tree based on the core genome alignments was constructed by use of randomized axelerated maximum likelihood (RAxML) v8.2.10 [182, 183].

As building a phylogenetic tree is an NP-hard problem in terms of its computational complexity [184, 185], RAxML provides a maximum likelihood approach to estimate an optimal solution. The RAxML algorithm can be summarized in two major steps:

In the first step, an initial parsimony tree is calculated by dnapars (`http://evolution.genetics.washington.edu`). Genomes are grouped together so that the number of evolutionary changes are minimized. Parsimony trees commonly have a reliable likelihood value [186], which makes them a useful starting point for further optimization in the second step.

In order to utilize the likelihood approach, an underlying DNA evolution model for nucleotide substitutions is required. Here, the general time-reversible (GTR) model from Miura et al. was employed together with an optimized gamma model of rate heterogeneity (Figure 4.1) [187, 188]. The phylogenetic tree was optimized by maximizing the likelihood value: To achieve such an optimal result, subtrees or branches in a defined neighborhood of a particular branch are rearranged. The robust and stable likelihood tree is generated by an integrated hill climbing algorithm in combination with 100 bootstraps to determine the optimal phylogenetic structure.

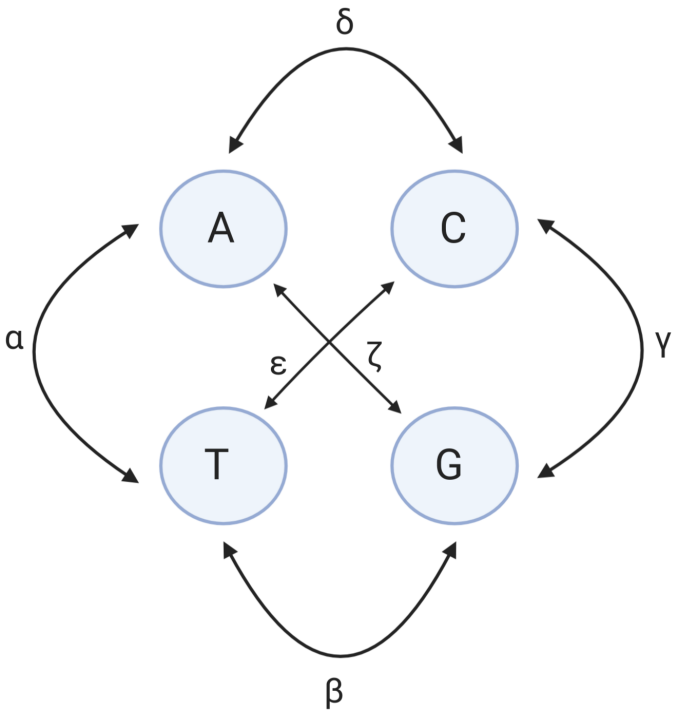


Figure 4.1: Visualization of the GTR model used for phylogenetic tree construction. The nodes represent the nucleotides Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). Substitution rates between bases can differ, which is indicated by Greek letters. All outgoing substitution rates of a node have to sum up to 1.

Since recombination events are a known problem for phylogenetic tree building algorithms, leading to an overestimation of branch lengths, the phylogenetic trees were corrected for recombination events by ClonalFrameML v1.12 [189].

Bacterial lineages within the population structure of dataset 1 were further investigated by a Bayesian analysis of population structure (BAPS) [190]. BAPS works under the assumption that the overall gene flow within a population is limited through boundaries between subspecies. The genetic population structure is identified by describing molecular variations for each subpopulation with a joint probability distribution over the observed sequence sites or loci using Bayesian models. Here, BAPS was applied with hierarchical clustering (hierBAPS), implemented in the R package RhierBAPS v1.0.1 [191].

4.2.6 *C. jejuni* lifestyle classification

In order to facilitate statistical comparisons, a set of closely-related *C. jejuni* lineages were defined as host-specific if more than 50% of their genomes building the respective BAPS cluster were associated with isolates from a specific animal origin (e.g. cattle or chicken) while each of the other (animal) origin should be represented by less than 10% of the genomes under consideration. Potential host-generalist lineages were assumed when more than 25% of the genomes represented in a corresponding BAPS cluster were from human clinical cases while at least two distinct animal origins are represented by more than 10% of the genomes each [154].

4.2.7 *K-mer* counting algorithms

Within the scope of this work, *k-mers* were applied in two different manners and therefore different *k-mer* counting frameworks were used. The overall concept of *k-mer* counting is described in Section 3.1. How general *k-mer*-based concepts were implemented in bioinformatic tools for usage in this work is explained in the following section.

For the first dataset, *k-mers* were used for a GWAS in order to discover genomic signatures that might be associated with a certain trait. For this purpose, *k-mers* of different lengths (9 to 100 bp) were used to identify SNPs among CDS belonging to the core genome as well as genes of the accessory genome. To efficiently count *k-mers* of different sizes and combine them in a data structure that is suitable for a GWAS, a tool for frequency-based string mining, fsm-lite (`https://github.com/nvalimak/fsm-lite`) implemented with the c++ lsdsl-lite-2.0.3 library, was used. This tool utilizes the sdsl library and was tailored for GWAS tool Seer/pyseer [128, 192] used for this work.

For the second dataset, *k-mers* were utilized to discover inter-species recombination between *C. jejuni* and *C. coli*. For this task, the KMC3 [193] and its related toolbox were used. KMC3 is a fast and memory efficient software to count *k-mers*. It takes FASTA or FASTQ formatted files as input to construct a *k-mer* count database and provides statistics that can be manipulated downstream. In addition, KMC3 tools provide a variety of operations to filter and manipulate the *k-mer* database including different filter options, intersections and unions of two databases or complex set operations.

The *k-mer* count algorithm from KMC3 uses two steps to process the input sequence data: First, it splits up the *k-mers* into bins, based on hash values, saving the intermediate results on disk, which was done by using a modified concept of minimizer [116]. In a second stage, these bins are sorted using a parallelized version of "radix sort" for large datasets [194]. The fact that the bins are stored on disk rather than kept in memory is the key reason why KMC3 is so memory-efficient. However, the memory efficiency is bought at the expense of an increased input/output (I/O) overhead [194].

4.2.8 Genome-wide association study

To further investigate dataset 1, a k-mer-based GWAS with pyseer v.1.1.2 [192] was used to study the host specificity of *C. jejuni* on whole-genome data (Figure 4.2). Pyseer provided the possibility to discover genomic alterations within the core as well as in the accessory genome. *K-mers* of variable length (9 to 100 bp) from 490 genomes were counted by fsm-lite v1.0 (https://github.com/nvalimak/fsm-lite). By applying pyseer, *k-mer* counts of genomes representing different *C. jejuni* lifestyles (generalist, cattle-specialist, chicken-specialist or pig-specialist) were compared. Each group was compared against the combined *k-mer* counts from all other genomes with respect to their phylogenetic structure by using a linear-mixed model (LMM) as mathematical background to calculate the associations. For each comparison, significant *k-mers* were filtered by an individually calculated threshold (based on the Bonferroni correction) for the lineage corrected p-value obtained from pyseer and split into two groups based on their direction of effect. Significantly associated *k-mers* were mapped by bwa v0.7.17 [195] against a representative reference genome of each lifestyle group used in the study to identify putative lifestyle-specific factors, genes and consecutive gene loci.

Implementation of a consensus GWAS approach

Due to the uneven natural appearances of *C. jejuni* in different hosts (Section 4.1), the comparison of different sample sets would lead to the problem of highly imbalanced groups and accordingly to a high false positive rate of the microbial GWAS. In this work the number of genomes within each lifestyle associated groups (case groups) was always smaller than the control group. The naïve approach to balance the case and control group would be to uniformly down-sample the control group. However, since

the phylogenetic structure of a bacterial population has huge impact on the results of a microbial GWAS (Section 3.2.1), the naïve approach was not sufficient here. In order to address the problem of the highly imbalanced groups with respect to the population structure of bacterial species a bootstrapping approach based on a proportional stratified random sampling has been conceptualized and implemented in this work:

With this approach, the number of genomes in the control group was down-sampled towards the number of genomes in the lifestyle preference group. To ensure that the underlying population structure of the bacterial species is still given, a uniform sampling was done separately within each BAPS cluster that was derived from the calculated phylogeny. BAPS clusters that contain a high number of genomes were also represented with more genomes in the sub-sampled control group based on the proportion of particular a cluster within the original dataset. As a result, the number of genomes compared within a single GWAS bootstrapping run was balanced and due to stratification each lineage of the population was also covered by at least one genome per GWAS run. Furthermore, the proportions of the different phylogentic lineages within the population were constant and did not vary from run to run.

To generate reliable and robust results, the bootstrapping approach was repeated 100 times for each lifestyle preferences group. Genes identified by *k-mer* mapping were stored within each bootstrapping iteration in order to create a consensus microbial GWAS in the end. Genes identified in at least 90% of these tests were selected as candidates.

Of note, putative genes important for host-generalist lineages were not affected by the approach, since both, the host-generalist and the host-specialist groups contained an equal number of isolates. For this purpose, only genes associated with an average -log(p-value) $\geq$ 80 were selected. The determined set of genes was further analyzed considering functional annotations and metabolic pathways using EggNog v.4.5.1 [196]. EggNog provides a database with pre-computed orthologues groups and phylogenies in order to transfer functional information for an input set of genes.

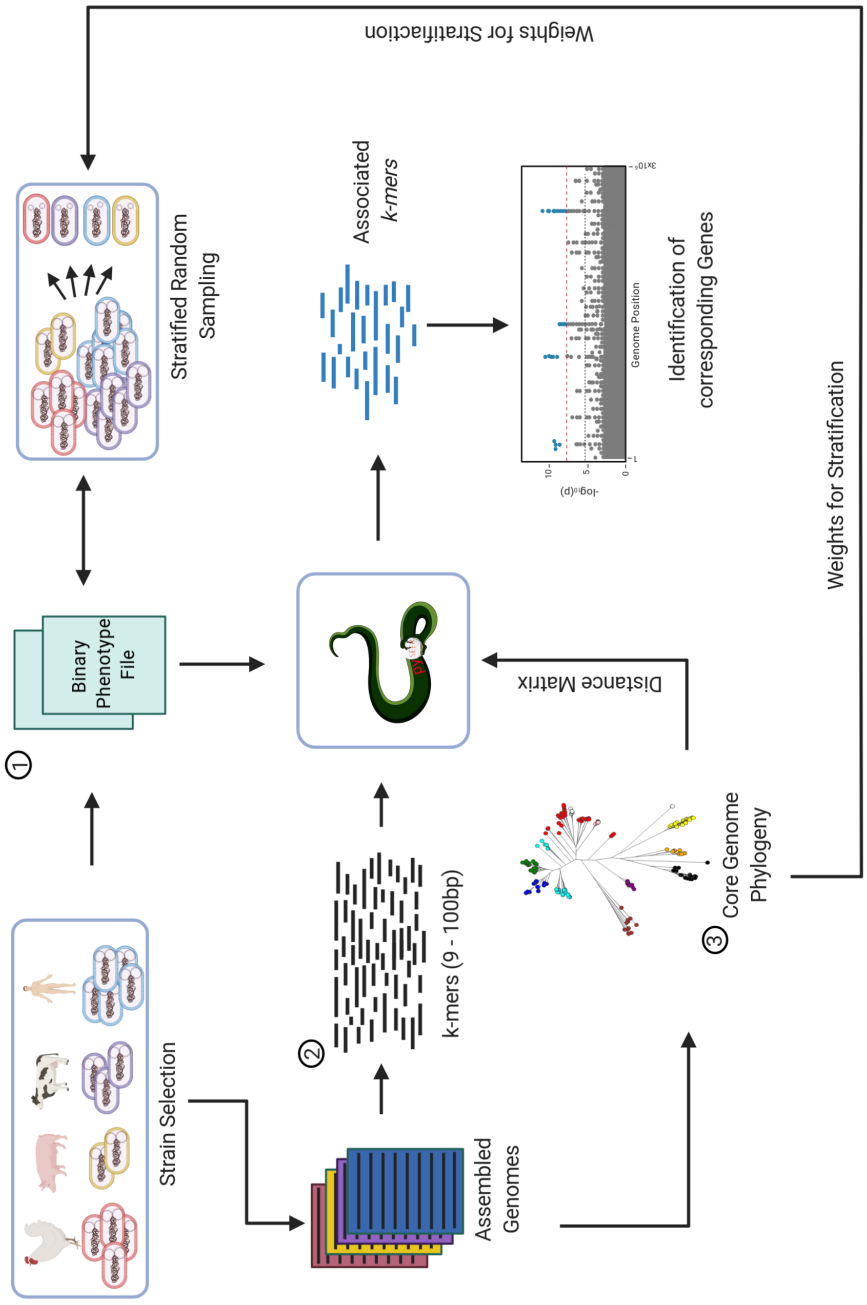


Figure 4.2: Schematic overview of the GWAS workflow for this work. Pyseer requires three input files: 1) a phenotypic file encoding the groups that will be compared, 2) *k-mers* counted based on assembled draft genome and 3) a distance matrix based on a phylogeny in order to correct for lineage effects. Stratified random sampling is applied to subsample inputs used in the phenotypic file. Output of the workflow are significant associated *k-mers*. This figure was created with BioRender (April 2021).

4.2.9 Identification of intra- and inter-species recombination events

Since intra-species recombination holds a major role for evolution as well as niche and host adaptation in general, recombination events were further analyzed within dataset 1. BratNextGen [197] was used to reconstruct putative recombination events for the 490 *C. jejuni* genomes for the PAC-Campy dataset. For this purpose, the core genome alignment was utilized. Parameter estimation was performed by an HMM-based approach with 20 iterations. Significant recombination events (p-value ≤ 0.05) were obtained by permutation testing incorporating 100 parallel iterations.

Implementation of inter-species recombination detection

As described in Section 2.5.1, recombination has not only been observed separately within individual *Campylobacter* species, but also between different *Campylobacter* species. It has been proposed that gene flow from *C. jejuni* towards *C. coli* is an ongoing trend within the *Campylobacter* genus [25]. In this work, a *k-mer*-based workflow was developed in order to identify genomic elements that are transferred between *C. jejuni* and *C. coli*

Database construction

To detect DNA sequence recombination sites among *C. jejuni* and *C. coli* genomes in dataset 2, two distinct *k-mer*-based databases for *C. jejuni* and *C. coli* were precomputed. Besides 21 genomes with unambiguous qPCR results (Section 4.1.2) from the BfR, 247 closed *Campylobacter* genomes available from NCBI were used to set up the initial approach.

In order to identify genetic elements from *C. jejuni* that recombined into the *C. coli* population, the *k-mer* databases were generated using KMC v3.1.1 [193]. The databases contain all *k-mers* present in at least 95% of the *C. jejuni* genomes and all *k-mers* present in at least 5% of the *C. coli* genomes. Thresholds of 95% and 5% have been chosen since gene transfer in terms of recombination has been reported to occur from *C. jejuni* to *C. coli*, not vice-versa [25, 36]. *C. jejuni* was considered as the donor species, whereas *C. coli* was the acceptor species during the following workflow. The *k-mer* size was set to 16 and 31, but can be adjusted using a user-defined parameter for further usage.

Implementation of the pipeline to identify genes of *C. jejuni* origin in *C. coli* (dataset 2)

The pipeline, "Relative *k-mer* Project" (RKP), was primarily implemented in Python and Bash. The workflow combines and connects the output of several tools to explore *k-mers* that are involved in recombination events (Figure 4.3). A *k-mer* counting is performed on assembled draft genomes given as input data (dataset 2): First, all *k-mers* of these genomes are calculated based on a fixed *k-mer* size, which was already used to build the databases. This step is performed by KMC3 v.3.1.1. The *k-mers* identified in the draft genome assemblies of dataset 2 are intersected with the *k-mer* database representing the donor species (*C. jejuni*). In a "cleaning step", the *k-mers* present in both *Campylobacter* species (e.g. the core genome shared by both species) is subtracted from the dateset using the database of the acceptor species (*C. coli*) by use of KMC3 tools. In order to identify genes and adjacent loci in *C. coli* that putatively originated from *C. jejuni*, the resulting *k-mers* are mapped against a well characterized *C. jejuni* reference genome (NCTC 11168) by Bowtie2 v.2.3.5 [198]. The mapping results are further processed with samtools v.1.10 [199] and bedtools v.2.29.2 [200] in order to extract putative genes associated with recombination events with at least 20% *k-mer* coverage. The gene-wise *k-mer* coverage was automatically visualized by the pheatmap package v.1.0.12 in R v.3.6. Additionally, mapping information was analyzed to identify regions within the genome that are transmitted as complete loci. Regions that are consecutively covered by *k-mers* are identified by a sliding window approach using a maximum gap size of 100 bp.

The whole pipeline is freely and publicly available under a GPLv3 license at `https://gitlab.com/microbial_genomics/relative-kmer-project` and can easily be installed by use of conda with all required dependencies.

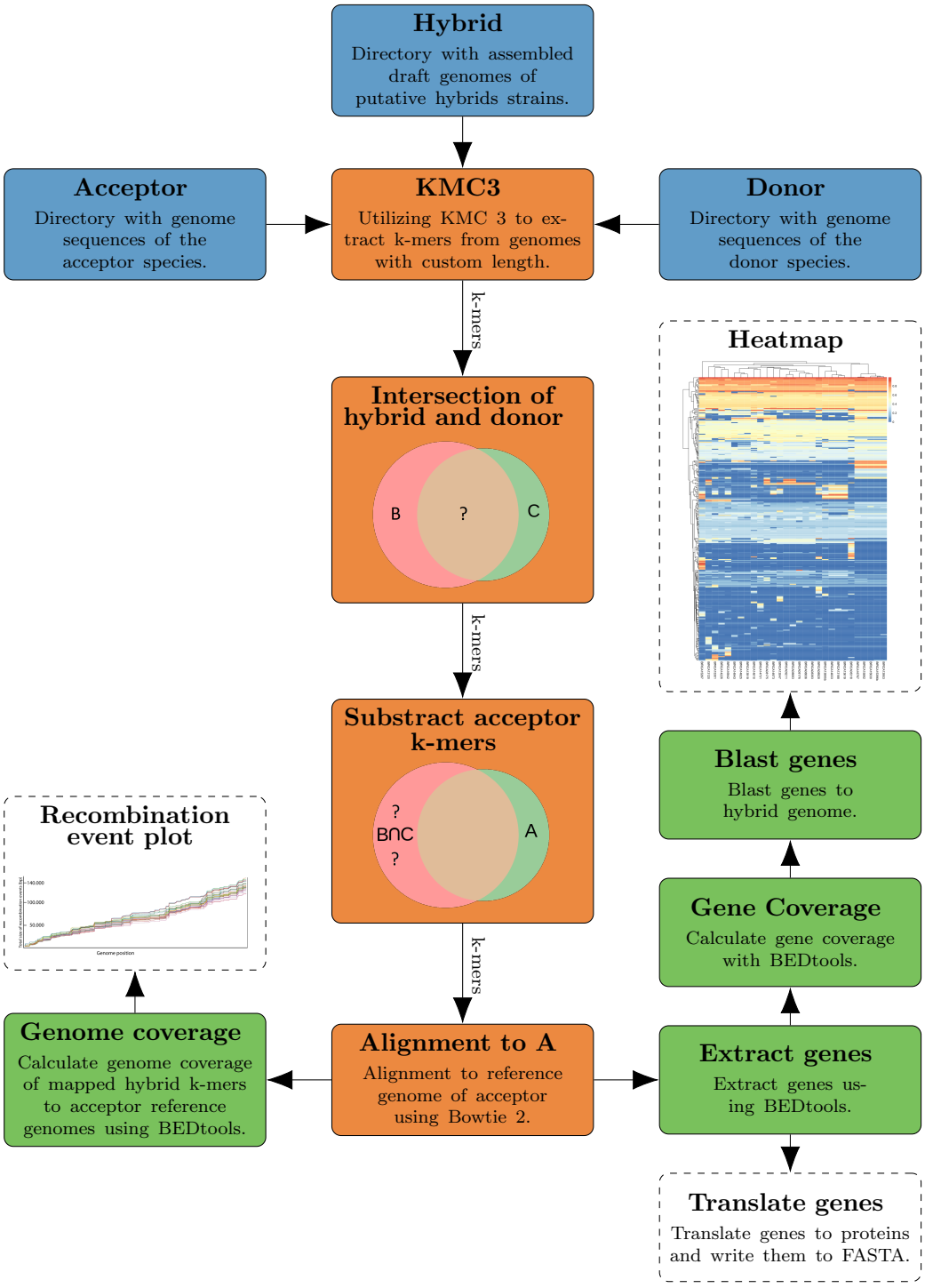


Figure 4.3: Flowchart representing a *k-mer*-based approach to detect recombination sites between two bacterial species: Blue colored fields describe the input data, orange fields show mathematical operations performed in order to identify relevant *k-mers*, green boxes symbolize post-processing steps of the *k-mers* and black dotted lines display the tool's output.

5 Identification of host-associated sequence determinants

The following chapter is based on material of a published article:

Epping, L., Walther, B., Piro, R. M., Knüver, M. T., Huber, C., Thürmer, A., Flieger, A., Fruth, A., Janecko, N., Wieler, L. H., Stingl, K. & Semmler, T. (2021). Genome-wide insights into population structure and host specificity of *Campylobacter jejuni*. Scientific reports, 11(1), 1-15.

5.1 Background

Previous *Campylobacter* research was focused on factors likely influencing the niche adaptation abilities of *C. jejuni* to certain growth conditions, especially host species, with an emphasis on poultry and cattle [18, 201, 202]. Nonetheless, novel bioinformatic methods and tools, including GWAS, proved their potential to identify genetic factors promoting host adaptation and/or pathogenicity on a genomic scale in *C. jejuni* only recently [18–20, 202, 203]. Accessory genes encoding factors facilitating the bacterial vitamin B5- biosynthesis pathway were identified as being associated with the host species cattle and its typical diet [18], while proteins enhancing iron acquisition abilities of the bacteria during infection seem to be often harbored by isolates obtained from human clinical samples [19]. Many of these studies implemented a gene-by-gene approach for population-scale analysis or focused mainly on strains related to clonal complex CC-45 [18, 19, 203], a phylogenetic background known for its frequent association with cases of human diseases worldwide [202, 204–206]. Besides baseline typing in order to define STs and CCs, most of the GWAS have been predominantly focused on the variable set of genes commonly addressed as accessory genes. Changes among (essential) core genes (i.e. basic cellular and regulatory functions) within the *C. jejuni* population might reflect adaptation towards a particular bacterial lifestyle as well. Core genome alterations probably play an important role to overcome specific host-associated intestinal stress conditions [207, 208], while other alterations might enable certain lineages to cope with colonization inhibitors or even diets associated with gastrointestinal tracts of a much broader range of host species [22]. A recent GWAS indicated that the worldwide intensified cattle farming for meat production was accompanied by a timeline of genomic events increasing host adaptation of certain *C. jejuni* lineages to cattle [21].

The aim of this study was to generate in-depth insights into the current population structure, host specificity and outbreak potential of *C. jejuni* in Germany using a stratified random sampling approach combined with GWAS considering all nucleotide substrings of length k (*k-mers*). For this purpose, genomes obtained from *C. jejuni* of human, chicken, cattle, and pig origin in Germany were complemented with further *C. jejuni* genomes publicly available from a recent Canadian study including similar sample origins [19] in order to limit spatial or temporal effects on the study outcomes.

5.2 Relationships on a pan-genomic level

In this work, 490 genomes of *C. jejuni* isolates from different animal, human and environmental origins from Germany and Canada have been analyzed. In order to classify those isolates with respect to their lifestyle classification (Section 4.2.6), the pan-genome was analyzed during an initial step. The average genome size of the set was 1,690,635 bp including 1,747 ± 106 genes on average. In total 1,111 were associated with the core genome, whereas 7,250 additional genes represent the accessory gene content.

Core- and accessory genome: phylogenetic structure and organization of the *C. jejuni* population

A multiple sequence alignment of the 1,111 core genes was used to calculate a phylogenetic representation of the 490 genomes (Figure 5.1). Overall, 15 distinct branches (1-15) were identified using BAPS clustering. Major BAPS clusters incorporating more than 15 genomes (11/15) were further evaluated with respect to their respective CCs, original sample source and lifestyle classification (Table 5.1).

Apart from lifestyle classification, a comparison of geographic origins between isolates from Germany and the 166 *C. jejuni* isolates of the previous Canadian study was performed in order to enhance the general representativeness of the data used for this work. German and Canadian isolates were represented among nearly all BAPS clusters, indicating limited geographical effect within the sample selection process (Figure 5.1). Notably, BAPS cluster 15 represents environmental isolates from Canada, as this origin has not been considered during the sampling approach in Germany.

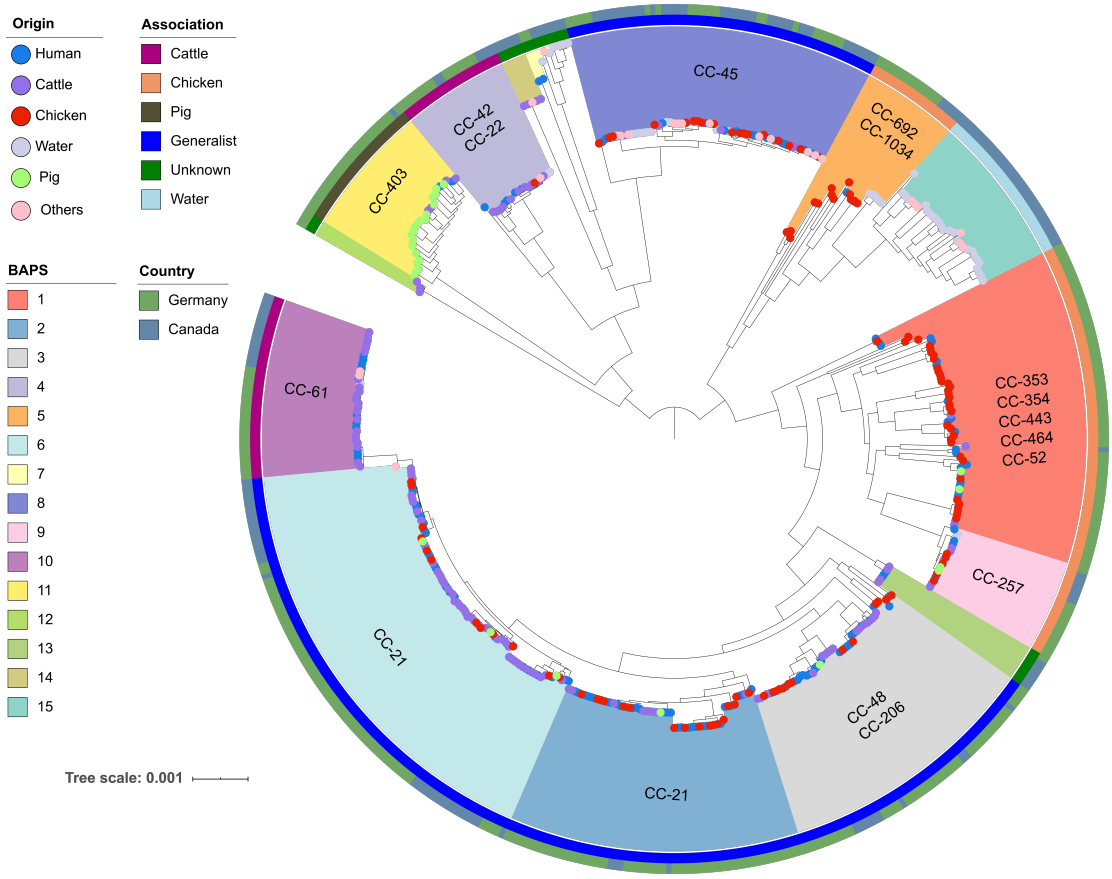


Figure 5.1: Phylogenetic structure of the *C. jejuni* population across all genomes. Leavs are colored by the origin of each sample. Colors in the inner ring represent the BAPS clusters used in this study, whereas the colors of the second ring stand for a certain lifestyle preference (e.g. host association). The outer ring color indicates the country of origin. This image was created and published in [154] during this work.

Distribution of sample origins for the major BAPS clusters are classified regarding their lifestyle preference with criteria defined in Section 4.2.6. Absolute and relative abundance of sample origins across each BAPS cluster is visualized in Figure 5.2.

BAPS cluster 1 includes 62 genomes with isolates from CC-353, CC-354, CC-443, CC-464 and CC-52 and the 19 isolates from BAPS cluster 5, predominantly represented by isolates of CC-1034 and CC-692, were classified as chicken-specific *C. jejuni* genomes, based on this criteria. Notably, BAPS cluster 5 and BAPS cluster 15, with Canadian samples from water-born environments of *C. jejuni*, show a close phylogenetic relation-

Table 5.1: BAPS clusters, original sample source and lifestyle preferences of 490 *C. jejuni* genomes [154].

			Sample origin						**Lifestyle classification**		
			chicken	**human**	**cattle**	**pig**	**environment**	**other**			
BAPS	**n**	**CC**	**n**	**n**	**n**	**n**	**n**	**n**	**this study**	**previous studies**	**reference**
1	**62**	CC-353	36	19	5	2	0	0	chicken	chicken	[43]
		CC-354							chicken	chicken	[43]
		CC-443							chicken	chicken	[43]
		CC-464							chicken	chicken	[43, 209]
		CC-52							chicken	chicken	[209]
		ST-2274							**chicken**	chicken	-
2	**57**	CC-21	17	30	9	1	0	0	generalist	chicken	[43, 210]
3	**52**	CC-48	12	23	16	1	0	0	generalist	generalist	[43]
		CC-206							generalist	generalist	[43]
		CC-21							generalist	generalist	[43, 210]
4	**21**	CC-42	1	5	11	0	3	1	cattle	cattle	[43]
		CC-22							**cattle**	-	-
5	**19**	CC-1034	13	1	0	0	5	0	chicken	wild birds, turkey	[81]
		CC-692							chicken	wild birds, turkey	[211]
6	**86**	CC-21	12	23	47	3	0	1	generalist	generalist	[43]
7	**9**	CC-677	0	2	0	0	6	1	n.a.	-	-
		CC-682							n.a.	-	-
8	**60**	CC-45	18	14	6	0	12	10	generalist	generalist	[210]
		CC-283							generalist	generalist	[43]
9	**18**	CC-257	5	4	5	2	2	0	chicken	chicken	[83]
10	**35**	CC-61	0	10	22	0	1	2	cattle	cattle	[83]
11	**26**	CC-403	0	2	4	20	0	0	pig	pig	[212]
12	**3**	ST-3098	0	0	3	0	0	0	n.a.	-	-
		ST-10282							n.a.	-	-
13	**7**	ST-922	0	3	4	0	0	0	n.a.	-	-
14	**5**	CC-508	0	0	4	0	1	0	n.a.		-
15	**30**	ST-693	0	0	0	0	24	6	n.a.	-	-
		ST-996							n.a.	-	-
		ST-1030							n.a.	-	-
		ST-1206							n.a.	-	-

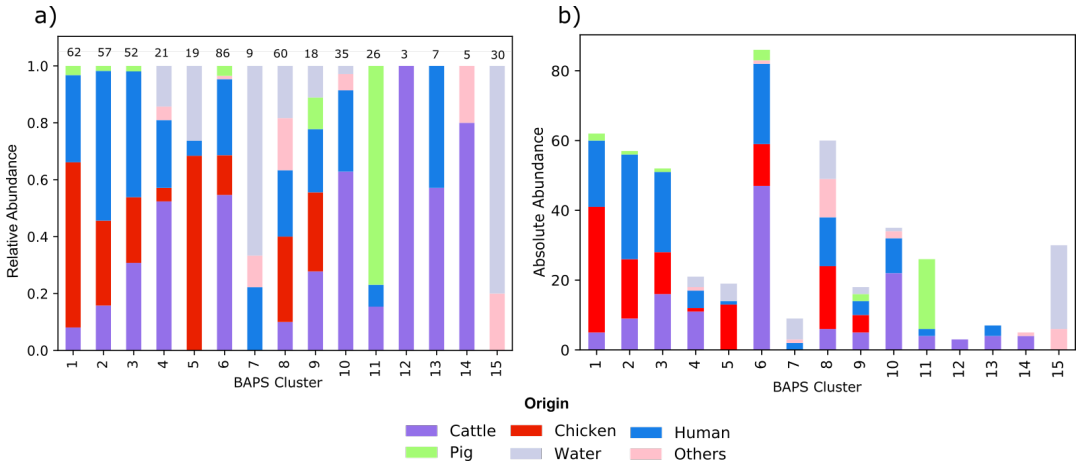


Figure 5.2: Stacked bar plots visualizing the relative a) and absolute b) abundance of sample origins distributed on each BAPS cluster. Plot b) shows the total amount of isolates per BAPS cluster on top of each bar. Proportions are coloured by their origin. This image was created and published in [154] during this work.

ship (Figure 5.1). *C. jejuni* genomes assigned to BAPS cluster 4 with 21 genomes, mainly represented by CC-42 and C-22, as well as BAPS cluster 10 with 35 genomes mainly identified as CC-61, are considered as cattle-specific strains with respect to lifestyle preference criteria (Table 5.1). The distribution of isolate origins of BAPS cluster 11 genomes (e.g. CC-403) lead to its classification as specific for pigs (Table 5.1).

Besides identification of host-specialist lineages, several BAPS clusters incorporate strains with a host-generalist lifestyle. Those lineages are represented by BAPS cluster 2 with 57 *C. jejuni* genomes (mainly CC-21), BAPS cluster 3 with 52 genomes (mainly CC-48 and CC-206), BAPS cluster 6 (CC-31) and BAPS cluster 8 including 60 genomes (mostly from CC-45 and CC-283). Although BAPS clusters 2, 3, 6 and 8 contain over 66% of *C. jejuni* genomes isolated human clinical samples in dataset 1, a human-specific lineage was not identified.

BAPS cluster 9, which consists of 18 genomes (mainly CC-257), includes samples of cattle (28%), chicken (28%), human (22%), pig (11%) and other (11%) origins and therefore failed the inclusion criteria for either host-specific or even –generalist lineages in this study. In addition, the genomes (n=30) of BAPS cluster 15 mostly represent

environmental isolate origins (80% water and sewage) which are associated with multiple STs (Table 5.1). In the following, the genomes of BAPS cluster 15 and those of BAPS clusters with genomes from less than 15 isolates have not been analyzed with respect to host specificity and were only used as a control group in our study.

As shown in Table 5.1, the predicted lifestyle preferences for each BAPS cluster have been evaluated based on previously published studies and data: CC-353, CC-354, CC-443, CC-464 and CC-52 have been frequently reported as chicken-specific lineages, whereas CC-42 and CC-61 generally occur as cattle-specific strains. CC-403 has been identified as pig-associated in one former study. Additionally to existing knowledge, the whole genome approach used in this work revealed host-specific lifestyle of the lineages from CC-22 for cattle and ST-2274 in chicken hosts. Furthermore, strains assigned to CC-21, CC-45 and CC-48 have been reported as host-generalist lineages and thus are concordant with the lifestyle prediction of BAPS clusters 2, 3, 6 and 8.

The relationships of sampling origins, host association and BAPS classifications derived from whole-genome sequence analysis were also visualized on the low-resolution MLST scheme by building a MST (Figure 5.3). The strain origins mirror the sampling distributions analyzed based on the phylogenetic clustering of Figure 5.1 (Figure 5.3 a). The BAPS clusters are concordantly assigned, since connected nodes in the MST are most likely assigned to the same BAPS cluster (Figure 5.3 b). The lifestyle preference assigned based on sample origins within the BAPS clusters perfectly fit with the CC annotation shown in Figure 5.1 as well as summarized results from previous studies in Table 5.1 (Figure 5.3 c).

Further evaluation of the phylogenetic structure (Figure 5.1) suggests that strains from BAPS clusters assigned as host-specific for cattle (BAPS 4 including CC-42 and CC-22; BAPS 10, CC-61) or pigs (BAPS 11, CC-403) are more clonal as strains assigned to chicken-specific BAPS clusters such as BAPS 5, including CC-1034 and CC-692. BAPS clusters 8 (CC-45 and CC-283), BPAS clusters 2 (CC-21), BAPS cluster 3 (CC-48 and CC-206) and BAPS cluster 6 (CC-21) show a more diverse population structure than lineages associated with cattle and pig.

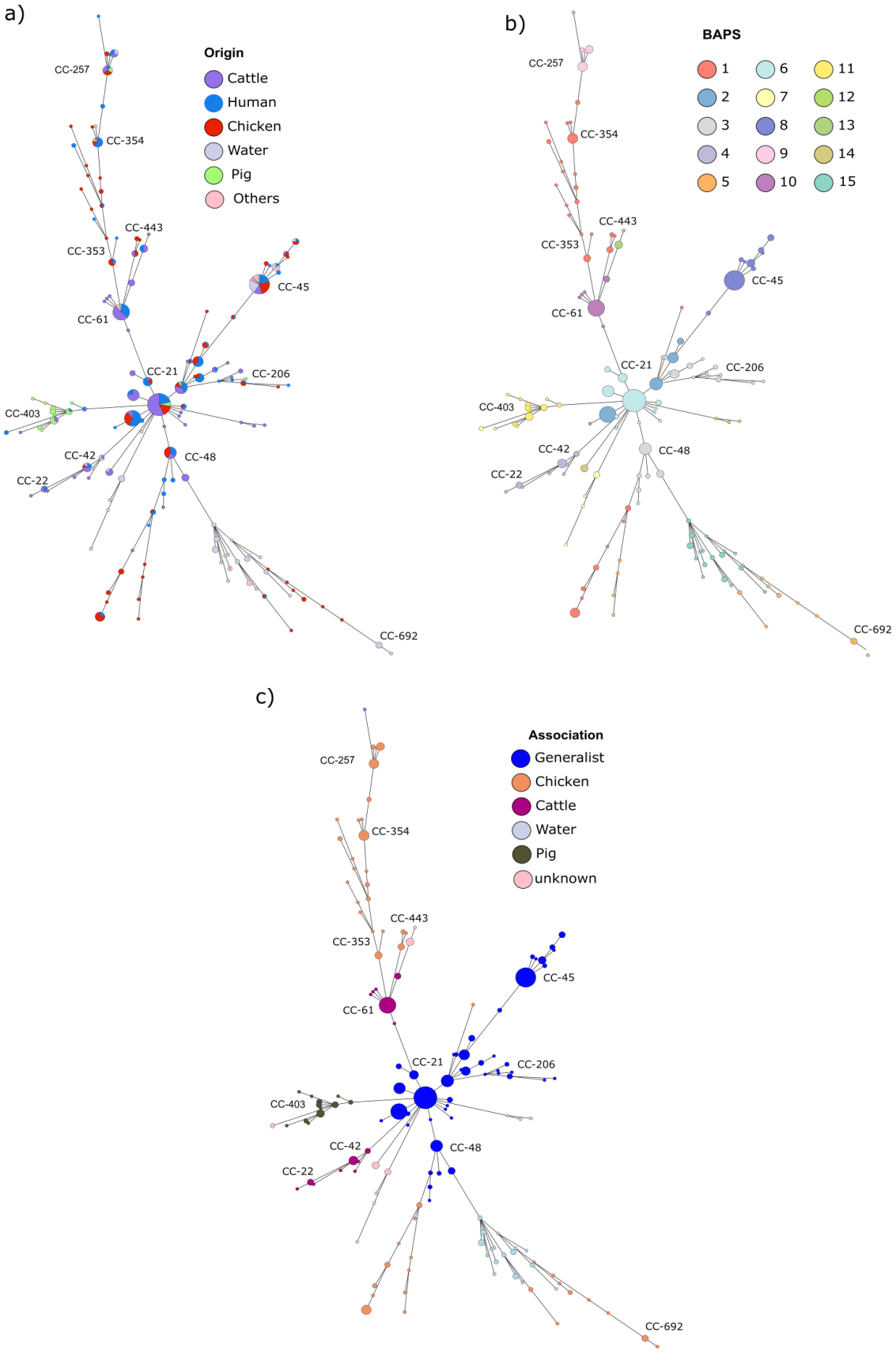

a)
Origin
Cattle
Human
Chicken
Water
Pig
Others
CC-257
CC-354
CC-443
CC-353
CC-45
CC-61
CC-21
CC-206
CC-403
CC-42
CC-48
CC-22
CC-692
b)
BAPS
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
CC-257
CC-354
CC-443
CC-353
CC-45
CC-61
CC-21
CC-206
CC-403
CC-42
CC-48
CC-22
CC-692
c)
Association
Generalist
Chicken
Cattle
Water
Pig
unknown
CC-257
CC-354
CC-443
CC-353
CC-45
CC-61
CC-21
CC-206
CC-403
CC-42
CC-48
CC-22
CC-692

Figure 5.3: Minimum spanning trees illustrate the relationship between different MLSTs based on the 7 housekeeping genes coloured by a) sampling source, b) source association by CC and c) BAPS cluster. This image was created and published in [154] during this work.

Lineages with the same lifestyle preference are not necessarily phylogenetically related to each other. For instance, cattle-related BAPS cluster 4 is more closely related to host-generalist BAPS cluster 6 than to the other cattle-related lineage BAPS cluster 10. The same observation holds true for chicken-specific lineages, as BAPS cluster 1 shows closer relationships to strains from host-generalist BAPS cluster 6, whereas chicken-specific BAPS cluster 5 seems to have an independent branch with respect to host-generalist lineages and are more closely related to water-borne isolates from Canada (Figure 5.1). These observations clearly reject the hypothesis of a common evolutionary background for host-specific lineages considering the host species represented here. The core genome phylogeny of the *C. jejuni* strains representing host-generalist lineages shows that BAPS cluster 8 seems to have evolved from a completely independent genomic background. Other host-generalist lineages, for instance, those of BAPS clusters 2, 3 and 6, seem to be phylogenetically linked to each other, at least to some degree (Figure 5.1).

Lifestyle assignments and phylogenetic relationships in this work were so far described based on the core genome phylogeny. In order to link these results based on the distribution of accessory genes as a variable set of the *C. jejuni* population, t-SNE was utilized (Figure 5.4). The t-SNE plots mirror the results derived from the core genome phylogeny and show that each BAPS cluster has a unique set of accessory genes. Each BAPS cluster is represented by its own group and also host-specialist lineages such as cattle-associated BAPS clusters 4 and 10 or chicken-assigned BAPS clusters 1 and 5 do not seem to share an extensive amount of genes of accessory genome content (Figure 5.4 b).

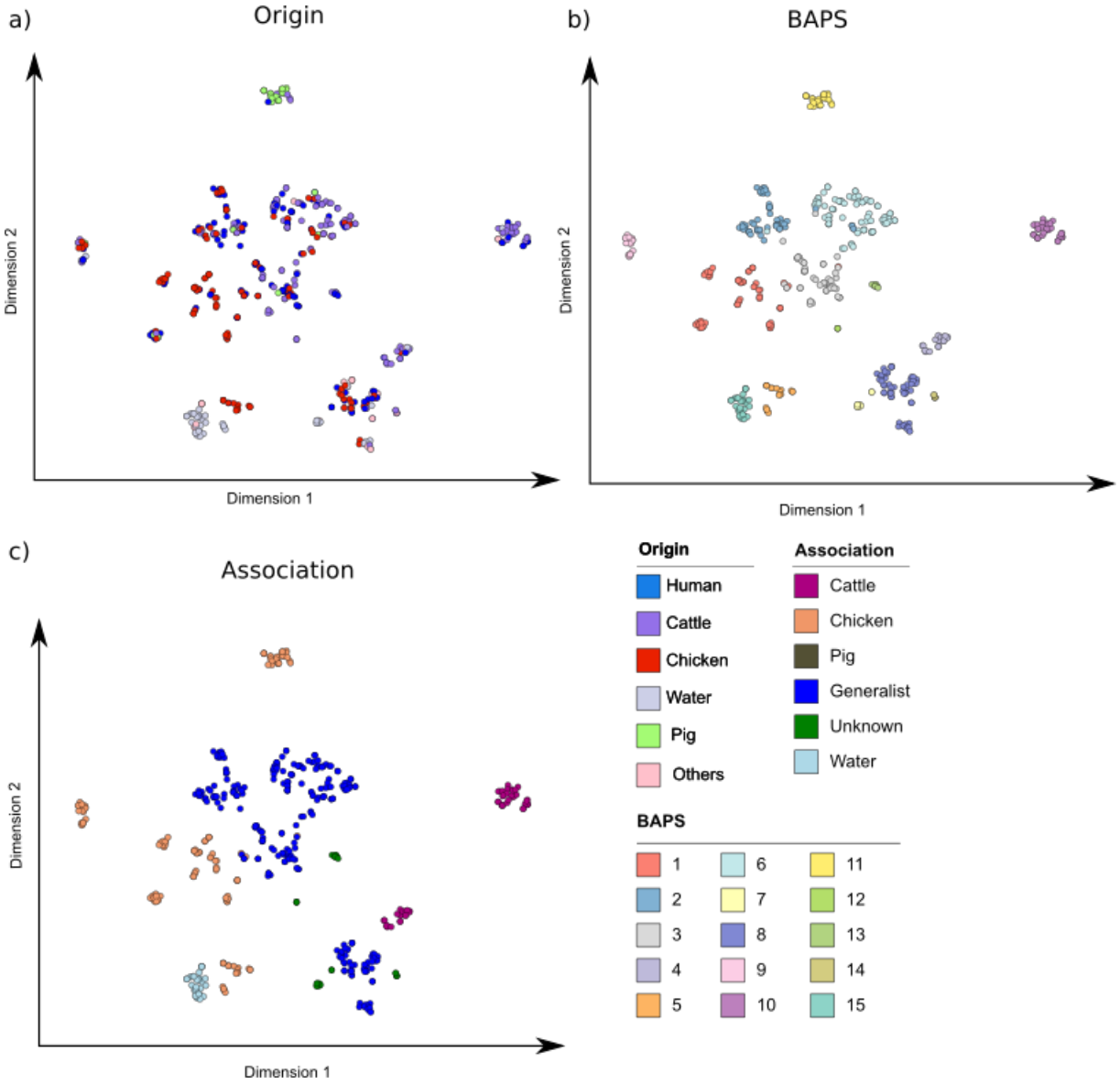


Figure 5.4: t-SNE plots of the accessory genome profile in 2-dimensional space. Colors represent a) the sampling origin, b) the BAPS clusters and c) the host association. This image was created and published in [154] during this work.

5.3 Effect of stratified random sampling

In order to reduce the false positive rate of the GWAS caused by the problem of highly imbalanced groups, a consensus approach has been applied to identify significantly associated *k-mers*. Those *k-mers* were mapped to annotated reference genomes, representing

lifestyle preferences, after adjusting for multiple testing. Regions of the genome covered by significantly associated *k-mers* putatively promote a specific lifestyle preference. A visualization of the effect on the set of genes that have been identified is provided in Figure 5.5. The left column of the figure shows the results of a single GWAS run without stratified random sampling for chicken, cattle and pig-associated genomes. The right side of the figure illustrates the same plot colored by genes that have been identified by the consensus GWAS in at least 90 out of 100 sampling iterations. For host-generalists the procedure was not applied since groups have been equally balanced. The plots show a clear reduction of genes that have been commonly associated with small amounts of *k-mers* with unlikely allele frequencies and low p-values. The remaining genes for cattle, chicken and pig are the top hits within the range of the expected allele frequencies for each host, based on the number of samples used in each group. Of note, genes classified by our analysis included genes belonging to the core as well as to the accessory genome.

The expected allele frequency of significant *k-mers* is associated with the proportion of genomes assigned to a particular lifestyle (e.g. chicken, cattle, host-generalist) in relation to the total number of samples.

Overall, genes identified by *k-mers* in more clonal lineages (pig and cattle associated *C. jejuni* genomes) showed a denser point distribution around the expected allele frequency than the results obtained for the genomes representing chicken- or host-generalist lineages that are more diverse in their population structure (Figure 5.1, Figure 5.5). An overview of the number of genes identified with a single GWAS run and the consensus GWAS approach is presented in Table 5.2.

Table 5.2: Overview of number of genes identified with the default GWAS and the consensus GWAS approach. Additionally the expected allele frequency for each lifestyle group is listed.

Lifestyle Preference	**Number of Genes GWAS**	**Number of Genes Consensus GWAS**	**Expected Allele Frequency**
Pig	939	127	0.06
Cattle	549	207	0.11
Chicken	483	42	0.20
Generalist	617	120 (top hits*)	0.52

With the consensus GWAS approach, 127 genes with significantly associated *k-mers* from genomes assigned to the pig-specific *C. jejuni* lineages (BAPS cluster 11) have been identified. Additionally, 207 genes of lineages associated with the two cattle-specific BAPS (4 and 10) clusters were identified by considering the *k-mer* abundance. For the chicken-specialist lineages (BAPS clusters 1, 5 and 9), the *k-mer* abundance distribution analysis revealed significant differences for 42 genes. GWAS analysis of strains associated with a host generalist lifestyle showed 120 genes that seem to incorporate *k-mers* associated with this particular lifestyle. All genes identified for each lifestyle are listed in the appendix (Table A.4-Table A.7).

Genes identified for each lifestyle preference can be classified into three categories:

- core genes, which are likely to incorporate important allelic variants
- accessory genes present in different host-specific lineages
- accessory genes with an almost unique presence in a certain genomic background

In order to assess the putative host-specific importance of the allelic variants, genes under consideration have been checked for synonymous mutations by comparing their amino acid sequences.

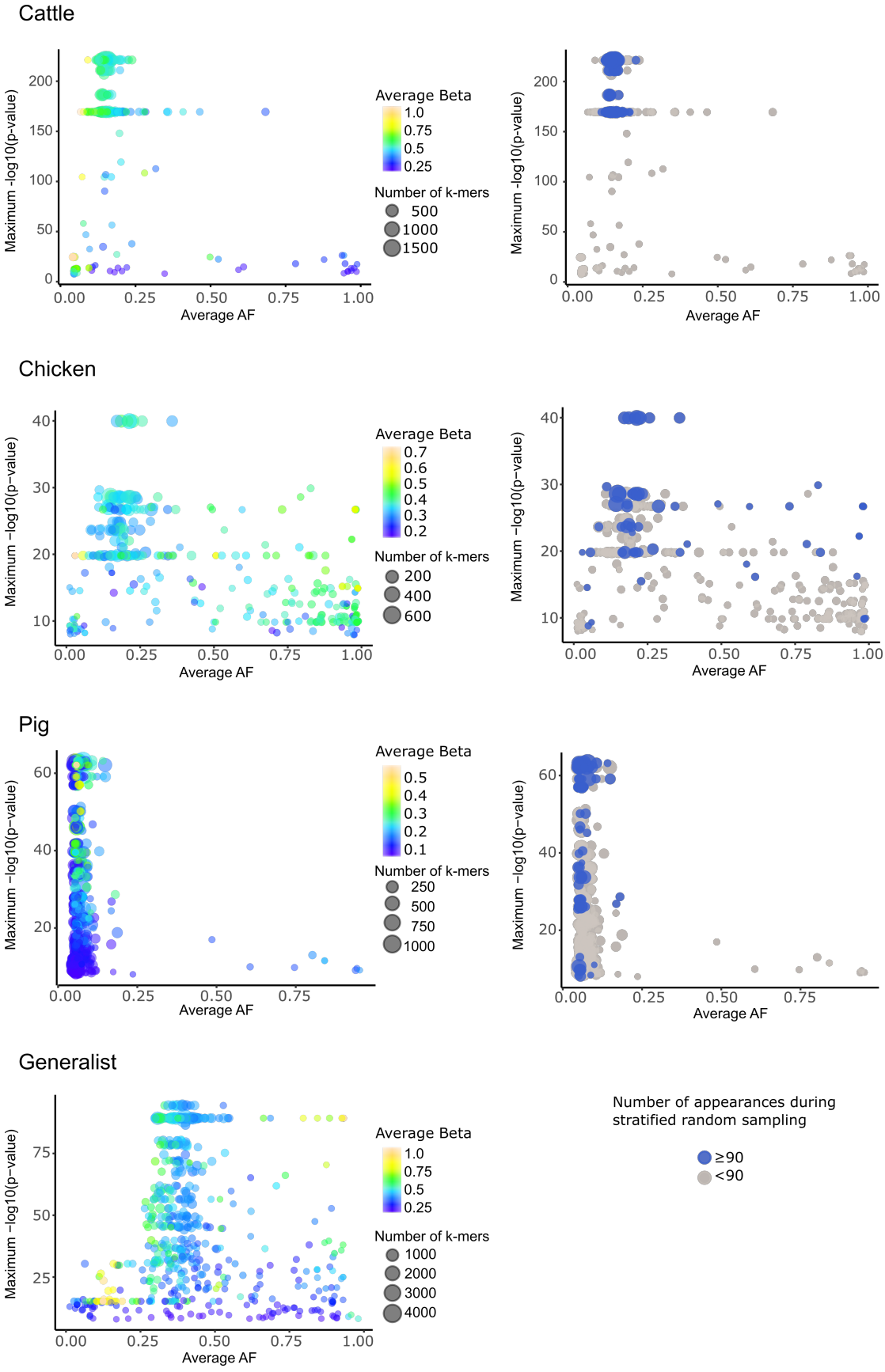

Cattle
Maximum −log10(p−value)
Average AF
Average Beta
1.0
0.75
0.5
0.25
Number of k-mers
500
1000
1500
Chicken
Average Beta
0.7
0.6
0.5
0.4
0.3
0.2
Number of k-mers
200
400
600
Pig
Average Beta
0.5
0.4
0.3
0.2
0.1
Number of k-mers
250
500
750
1000
Generalist
Average Beta
1.0
0.75
0.5
0.25
Number of k-mers
1000
2000
3000
4000
Number of appearances during stratified random sampling
≥90
<90

Figure 5.5: Dotplots representing genes derived by mapping significantly associated *k-mers*. The x-axis shows the average allele frequency of *k-mers* mapped to the particular genes and y-axis shows the maximum -log(p-value) of those *k-mers*. In the left column results obtained without adjusting for highly imbalanced groups are plotted. Dots are colored according to the slope (average beta) and the dot size indicates the number of *k-mers* mapped to a gene. The right column shows the results that were obtained by significantly associated *k-mers* with the stratified random sampling approach in at least 90% of the runs (blue). Genes found in less than 90% (grey) of the consensus GWAS iterations are colored in grey. This image was created and published in [154] during this work.

5.4 Host-specific signatures

Accessory genes and allelic variants of the core genome associated with *C. jejuni* lineages assigned as pig-specific

Genomes associated with a pig lifestyle of CC-403 (BAPS cluster 11), showed 21,781 significantly associated *k-mers* that mapped to 49 accessory genes as well as covering 78 allelic variants of the core genome content (Table A.6). Among the accessory genes 14 were exclusively associated with this particular genomic background (Table 5.3).

Three of these accessory genes (A6J90_06670, A6J90_06675, A6J90_02350) are part of transcription units encoding for type II restriction modification systems (RM systems). Two additional genes encode for the restriction subunit (R) of the host-specific determinant (*hsd*R; A6J90_08990) of a type I RM system. The remaining 8/14 genes were annotated as hypothetical or putative proteins without any specific functional information of homologous genes available in NCBI GenBank (17.06.2020).

Considering the *k-mer* results for genes belonging to the core genome, nucleotide changes leading to actual effects with respect to host adaptation capabilities of certain lineages are difficult to pinpoint. For this reason, selected loci of interest were checked for changes inducing amino acid sequence variation. Alterations for the predicted amino acid sequences associated with the capability of *C. jejuni* to synthesize vitamins and en-

zyme co-factors such as thiamine-phosphate pyrophosphorylase (*ten*I) and pyridoxine'-phosphate synthetase (*dxs*) (Figure 5.6). In addition, the predicted amino acid sequence for a putative membrane protein-encoding open reading frame (Cj1484) was found to be altered in pig hosts (Figure 5.6).

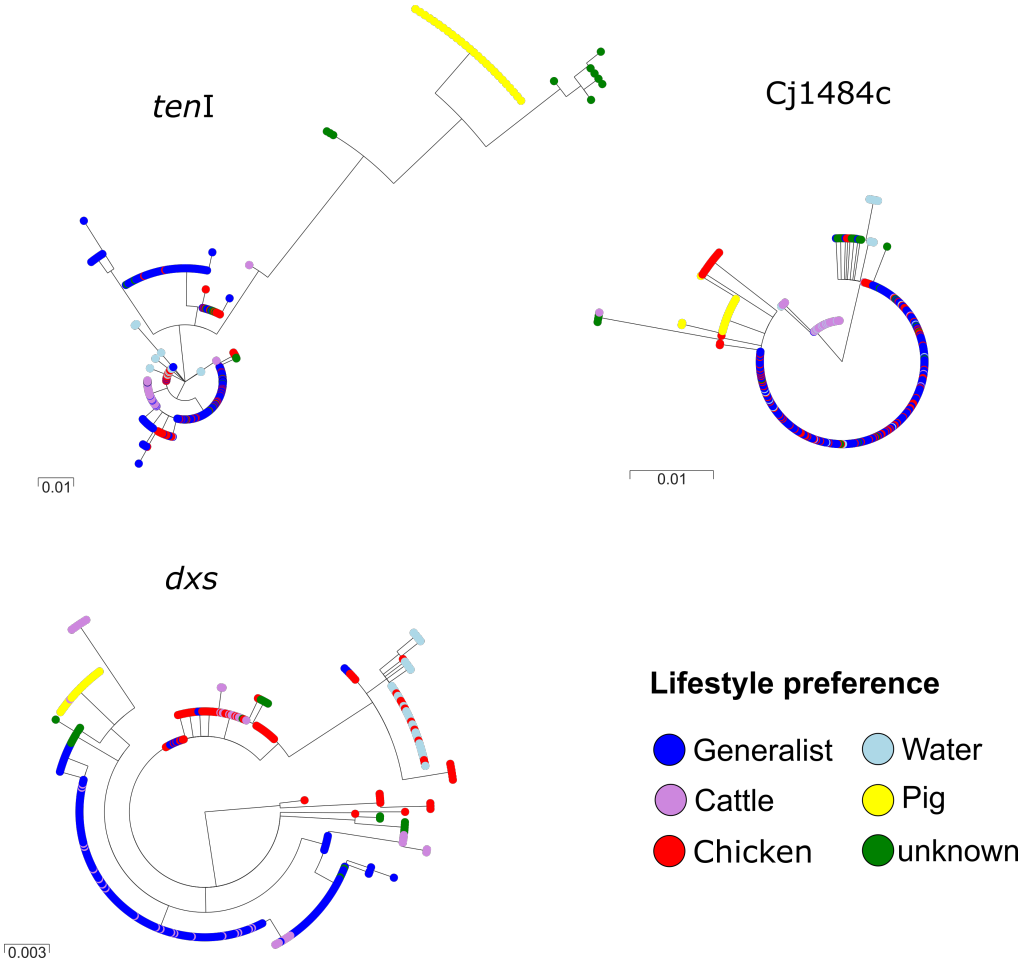


Figure 5.6: Phylogenetic tree of predicted amino acid sequences variants encoded by *ten*I, Cj484c and *dxs* (selected from Table 5.3) that show lifestyle-associated variants in different phylogenetic lineages originating from different genetic and geographic backgrounds (Figure 5.1). Lifestyle preferences are colored in purple (cattle), red (chicken), blue (host-generalists) and yellow (pig).

Table 5.3: Selected accessory genes and allelic variants of the core genome content associated with the host pig. The table was created and published in [154] during this work.

Locus Tag[a]	Gene	Predicted function	BfR-CA-14430[b]	COG[c]	COG Description	# pig[d]	# cattle[d]	# chicken[d]	# host generalists[d]	# other[d]	Accessory Variant[e]
A6J90_00190	-	putative protein	-	-	-	25	0	0	0	0	A
A6J90_00195	-	hypothetical protein	-	S	function unknown	26	0	0	0	*0*	A
A6J90_00200	-	hypothetical protein	-	-	-	26	1	0	0	*0*	A
A6J90_00270	-	putative protein	-	-	-	26	0	0	0	0	A
A6J90_00275	*dpn*A	DNA methylase	-	L	replication, recombination and repair	26	0	0	0	0	A
A6J90_01490	-	putative protein	-	-	-	26	0	0	0	0	A
A6J90_01500/ A6J90_01505	-	hypothetical protein	-	V	defense mechanisms	25	0	0	0	0	A
A6J90_02340	-	undecaprenyl-diphosphooligosaccharide-protein glycotransferase	-	-	-	25	0	0	0	0	A
A6J90_02350	-	R Pab1 restriction endonuclease	-	L	replication, recombination and repair	25	0	0	0	0	A
A6J90_06670	-	type II restriction endonuclease	-	L	replication, recombination and repair	26	0	0	0	1	A
A6J90_06675	*hha*IM	cytosine-specificmethyltransferase NlaX	-	H	coenzyme transport and metabolism	26	0	0	0	1	A

Continued on next page

Table 5.3 Continued from previous page

Locus Tag[a]	Gene	Predicted function	BfR-CA-14430[b]	COG[c]	COG Description	# pig[d]	# cattle[d]	# chicken[d]	# host generalists[d]	# other[d]	Accessory Variant[e]
A6J90_08990	*hsd*R	type I restriction enzyme EcoR124II R protein	-	V	defense mechanisms	26	0	1	0	0	A
A6J90_01640	-	hypothetical protein	-	-	-	26	0	0	0	0	A
A6J90_02350	*sua*5	hypothetical protein	-	J	translation, ribosomal structure and biogenesis	26	0	0	0	0	A
Cj0321	*dxs*	1-deoxy-D-xylulose-5-phosphate synthase	298,748	H	coenzyme transport	26	56	90	255	63	V
Cj1043c	*ten*I	thiamine-phosphate pyrophosphorylase	991,366	H	coenzyme transport and metabolism	26	56	90	255	63	V
Cj1484c	-	putative membraneprotein	1,428,185	-	-	26	56	90	255	63	V

End of table

[a] Locus tag for accessory genes based on *C. jejuni* reference genome CP022076.1 (NCBI accession). Locus tags for allelic variants of the core genome refer to *C. jejuni* strain NCTC11168 (NCBI accession: AL111168.1)

[b] Position of core genes in the reference strain BfR-CA-14430

[c] Clusters of orthologous groups (`http://clovr.org/docs/clusters-of-orthologous-groups-cogs/`)

[d] Number of genomes assigned to a particular lifestyle carrying the gene or allelic variant (pig, cattle, chicken, host generalists, others)

[e] Accessory (A) indicates that a gene belongs to the accessory genome content of *C. jejuni*. Variant (V) indicates a specific allelic variant of the core genome content.

Accessory genes and allelic variants of the core genome associated with *C. jejuni* lineages assigned as cattle-specific

GWAS analysis for cattle-assigned *C. jejuni* genomes, revealed 66,491 significantly associated *k-mers* mapping to 71 accessory genes and 136 core gene variants (appendix Table A.4). Mirroring the observation based on the accessory genome content (Figure 5.4), particular accessory genes that are representative for both major cattle-associated lineages of BAPS clusters 4 and 10 were not identified.

16 of the accessory genes belong to the same region of 9.9 kb size in *C. jejuni* genomes from BAPS cluster 10 (CC-61). Homology search in NCBI revealed a known locus with reference IDs from NCTC13261_01705 up to NCTC13261_01720, among others encoding for two predicted integrases, a putative protease, a HicA-HicB toxin/antitoxin system inhibiting the transfer of mRNA in case of nutrient limitation, a protein known to be involved in extracytoplasmatic stress response (*Yaf*Q) and a *Rep*A for plasmid DNA repair (Table 5.4).

Several genes identified within the core genome showed identical variants on nucleotide level (Table 5.4). Among these a 9.7 kb region consisting of 9 adjacent genes that encode for a ribosomal complex were identified. Genomic variants in the genes encoding for the DNA polymerase III subunit alpha (*dna*E) and the signal recognition particle protein (*ffh*) show non-synonymous substitutions, illustrated in a gene-wise phylogeny based on the amino acid sequence in Figure 5.7. Further genes of this cassette such as the transcriptional regulator C (*ars*C), phospho-2-dehydro-3-deoxyheptonate aldolase (*aro*F), 5-hydroxyisourate hydrolase (*ura*H), 50S ribosomal protein L19 (*rpl*S), tRNA (guanine-N(1)-)-methyltransferase (*trm*D), ribosome maturation factor (*rim*M) and 30S ribosomal protein S16 (*rps*P), showed identical allelic variants in both cattle assigned BAPS cluster 4 and 10 as well as in the host-generalist BAPS cluster 8. Allelic variants of *ura*H, *ars*C, *rpl*S and *rps*P lead to SNPs with synonymous changes and thus do not affect the amino acid sequence. Nonetheless, these SNPs might be used as markers to probably indicate evolutionary processes. The high conservation of the amino acid sequence among several lineages illustrates their biological importance as housekeeping genes for *C. jejuni*.

Adaptation towards cattle hosts seems to have happened on several independent loci within the *C. jejuni* genome, as additional non-synonymous, cattle-specific allelic variants were identified on several genome positions (Table 5.4). This includes genes encoding for a putative methyltransferase domain protein (Cj0495), a putative protein-disulfide oxidoreductase (*dsb*I) and a putative HAD-superfamily hydrolase (Cj1233). A comparison of amino acid sequence for gene Cj0495 is shown in Figure 5.7, again revealing direct influence on the resulting protein.

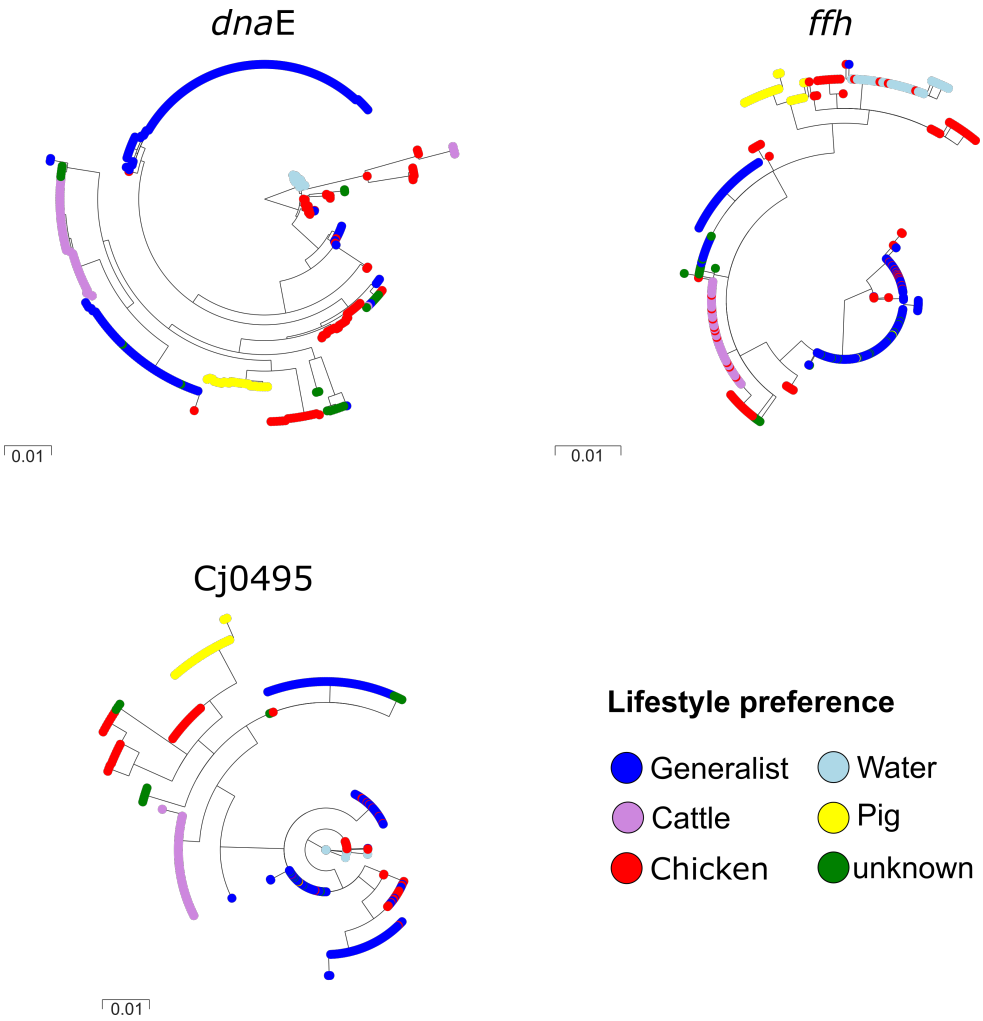


Figure 5.7: Phylogenetic tree of predicted amino acid sequences variants encoded by *dna*E, *ffh*, Cj0495 (selected from Table 5.4) that show lifestyle-associated variants in different phylogenetic lineages originating from different genetic and geographic backgrounds (Figure 5.1). Lifestyle preferences are colored in purple (cattle), red (chicken), blue (host-generalists) and yellow (pig).

Table 5.4: Selected accessory genes and allelic variants of the core genome content associated with the host cattle. The table was created and published in [154] during this work.

Locus Tag[a]	Gene	Predicted function	BfR-CA-14430[b]	COG[c]	COG Description	# pig[d]	# cattle[d]	# chicken[d]	# host generalists[d]	# other[d]	Accessory Variant[e]
Cj0718	*dnaE*	DNA polymerase III, alpha chain	679,065	L	replication, recombination and repair	26	56	90	255	*63*	V
Cj0717	*arsC*	putative ArsC family protein	678,288	P	inorganic ion transport and metabolism	26	56	90	255	*63*	V
Cj0716	*aroF*	putative phospho-2-dehydro-3-deoxyhep-tonate aldolase	678.951	E	amino acid transport and metabolism	26	56	90	255	*63*	V
Cj0715	*uraH*	transthyretin-like periplasmic protein	676,514	S	function unknown	26	56	90	255	63	V
Cj0714	*rplS*	50S ribosomal protein L19	676,024	J	translation, ribosomal structure and biogenesis	26	56	90	255	63	V
Cj0713	*trmD*	tRNA (guanine-N1)-methyltransferase	675,309	J	translation, ribosomal structure and biogenesis	26	56	90	255	63	V
Cj0712	*rim*M	putative 16S rRNA processing protein	674,773	J	translation, ribosomal structure and biogenesis	26	56	90	255	63	V

Continued on next page

Table 5.4 Continued from previous page

Locus Tag[a]	Gene	Predicted function	BfR-CA-14430[b]	COG[c]	COG Description	# pig[d]	# cattle[d]	# chicken[d]	# host generalists[d]	# other[d]	Accessory Variant[e]
Cj0710	*rps*P	30S ribosomal protein S16	674308	J	translation, ribosomal structure and biogenesis	26	56	90	255	63	V
Cj0709	*ffh*	signal recognition particle protein	672,906	U	intracellular trafficking, secretion, and vesicular transport	26	56	90	255	*63*	V
Cj0495	-	tRNA methyltransferase	465,764	J	translation, ribosomal structure and biogenesis	26	56	90	255	63	V
Cj0017c	*dsb*I	disulfid-deoxidoreductase	825,673	C	energy production and conversion	26	56	90	255	63	V
Cj1233	-	HAD-superfamily hydrolase	1175101	S	function unknown	26	56	90	255	63	V
_01705	-	putative periplasmic protein	-	-	-	35	38	0	193	43	A
_01706	-	RelE/ParE family plasmid stabilization system	-	S	function unknown	35	20	0	0	4	A
_01707	–	hypothetical protein	-	-	–	35	0	0	0	0	A
_01708	-	hypothetical protein	-	-	-	35	0	0	0	0	A
_01709	-	acyl carrier protein	-	K	transcription	34	0	0	0	0	A
_01710	-	DnaB-like protein helicase-like protein	-		-	35	0	0	0	0	A

Continued on next page

Table 5.4 Continued from previous page

Locus Tag[a]	Gene	Predicted function	BfR-CA-14430[b]	COG[c]	COG Description	# pig[d]	# cattle[d]	# chicken[d]	# host generalists[d]	# other[d]	Accessory Variant[e]
_01712	-	hypothetical protein	-	-	-	34	7	0	1	4	A
_01713	-	hypothetical protein	-	-	-	35	0	0	1	0	A
_01714	-	helix-turn-heelix domain-containing	-	-	-	35	19	0	1	4	A
_01716	-	putative protein	-	-	-	35	0	0	0	0	A
_01717	*hic*B	antitoxin HicB	-	S	function unknown	34	14	0	1	4	A
_01718		hypothetical protein	-	N	cell motility	35	20	0	0	4	A
_01719	*hic*A	probable mRNA interferase toxin HicA	-		-	35	20	0	0	4	A
_01720	-	integrase	-	L	replication, recombination and repair	35	20	0	1	4	A

End of table

[a] Locus tags for accessory genes based on *C. jejuni* reference strain NCTC13265 genome LR134498.1 (NCBI accession). Locus tags for allelic variants of the core genome refer to *C. jejuni* strain NCTC11168 (NCBI accession: AL111168.1)

[b] Position of core genes in the reference strain BfR-CA-14430

[c] Clusters of orthologous groups (`http://clovr.org/docs/clusters-of-orthologous-groups-cogs/`)

[d] Number of genomes assigned to a particular lifestyle carrying the gene or allelic variant (pig, cattle, chicken, host generalists, others)

[e] Accessory (A) indicates that a gene belongs to the accessory genome content of *C. jejuni*. Variant (V) indicates a specific allelic variant of the core genome content.

Accessory genes and allelic variants of the core genome associated with *C. jejuni* lineages assigned as chicken-specific

The broad phylogenetic diversification within and across chicken-associated *C. jejuni* lineages from BAPS clusters 1, 5 and 9 (e.g. CC-353 and CC-1034) resulted in less specific signatures associated with this particular host, when compared with other host-specific lineages for cattle, pig and even host-generalist (Figure 5.1),. In total, 5,712 *k-mers* of *C. jejuni* genome sequences were identified by GWAS to be significantly associated with chicken hosts and were mapped to 17 accessory genes and 25 core gene variants accordingly (Table A.5).

None of the accessory genes were present in all chicken strains, however a *Tra*G-like protein (NCTC13265_01618) of the type IV secretion system [213] was detected among the accessory genomes in 59/90 chicken-associated genomes (Table 5.5). *Tra*G-like proteins are known to play a crucial role in the conjugative transfer of plasmids [214]. Additionally, two genes for putative proteins (NCTC13265_01627, NCTC13265_01633) of unknown function are carried by 66 and 68 of the 90 chicken associated strains, respectively (Table 5.5).

Similar to cattle- and pig-associated *C. jejuni* lineages, several genes within the core genome content of chicken-assigned strains show identical non-synonymous allelic variants on independent positions within the genome. The gene *rpo*B that encodes for the RNA polymerase subunit B showed a specific amino acid sequence in most of the genomes from BAPS clusters 1, 5 and 9 (Figure 5.8). Beyond these, BAPS clusters 1 and 5 carried the same allelic variant of the flagella basal body rod protein (*flg*B) that was also identified within host-generalist strains from BAPS cluster 2 (CC-21) and is represented in chicken associated BAPS cluster 9 by a closely related variant (Figure 5.8). Additionally, identical allelic variants of genes in metabolic pathways such as *pyc*B as part of the pyruvate carboxylase are carried by most genomes of BAPS clusters 1 and 9 (Table 5.5).

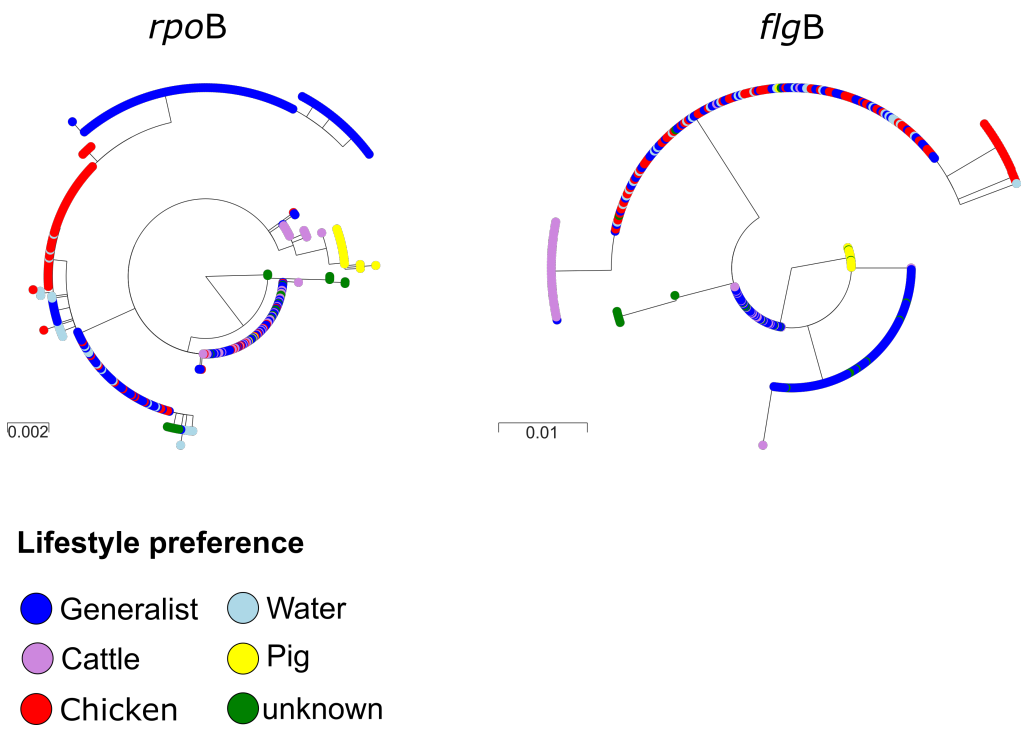


Figure 5.8: Phylogenetic tree of predicted amino acid sequences variants encoded by *rpo*B and *flg*B (selected from Table 5.5) showing lifestyle-associated variants in different phylogenetic lineages originating from different genetic and geographic backgrounds (Figure 5.1). Lifestyle preferences are colored in purple (cattle), red (chicken), blue (host-generalists) and yellow (pig).

Table 5.5: Selected accessory genes and allelic variants of the core genome content associated with the host chicken. The table was created and published in [154] during this work.

Locus Tag[a]	Gene	Predicted function	BfR-CA-14430[b]	COG[c]	COG Description	# pig[d]	# cattle[d]	# chicken[d]	# host generalists[d]	# other[d]	Accessory/ Variante[]
Cj0933c	*pyc*B	putative pyruvatecarboxylase B subunit	882.094	C	Energy productionand conversion	26	56	90	255	63	V
Cj0478	*rpo*B	DNA-directed RNApolymerase beta chain	444.215	K	Transcription	26	56	90	255	63	V
Cj0528c	*flg*B	flagellar basal-body rod protein	495.238	N	Cell Motility	26	56	90	255	63	V
_01618	*tra*G	conjugal transferprotein TraG	-	U	Intracellular trafficking and secretion	1	1	59	1	13	A
_01627	-	putative Protein	-	-	-	3	0	66	1	7	A
_01633	-	putative protein	-	-	-	3	0	68	0	0	A

[a] Locus tags for accessory genes based on *C. jejuni* reference strain NCTC13265 genome LR134498.1 (NCBI accession). Locus tags for allelic variants of the core genome refer to *C. jejuni* strain NCTC11168 (NCBI accession: AL111168.1)

[b] Position of core genes in the reference strain BfR-CA-14430

[c] Clusters of orthologous groups (`http://clovr.org/docs/clusters-of-orthologous-groups-cogs/`)

[d] Number of genomes assigned to a particular lifestyle carrying the gene or allelic variant (pig, cattle, chicken, host generalists, others)

[e] Accessory (A) indicates that a gene belongs to the accessory genome content of *C. jejuni*. Variant (V) indicates a specific allelic variant of the core genome content.

Independent adaptation of host-generalist lineages

The GWAS for the host-generalist *C. jejuni* lineages, results in 37,339 significantly associated *k-mers* that were mapped to 33 accessory genes and revealed allelic variants of 87 core genes (Table A.7). Similar to cattle-assigned BAPS clusters, particular accessory genes exclusively assigned to all host-generalist BAPS clusters have not been identified. In addition, a multitude of different allelic variants assigned to the core genome were identified for BAPS cluster 8 when compared with the genomes of the more closely related lineages of clusters 2, 3 and 6 (appendix Table A.7). Again, this observation mirrors the independent phylogenetic background of those lineages (Figure 5.1). Among the identified core genes, genes with identical and closely related variants between BAPS clusters 2, 3, 6 and 8 have also been detected (Table 5.6). These include genes such as cell division gene *fts*X, 50 S ribosomal genes *rpl*S and *rps*P and a gene coding for a HP0268 domain-containing protein (Cj0459c) (Table 5.6). An evaluation of variant assignments based on the amino acid sequence of those genes, revealed non-synonymous variants in the sequence of *fts*X (Figure 5.9) in all host-generalist lineages. 50S ribosomal genes *rpl*S and *rps*P have already been identified in cattle associated lineages and show a high level of conservation on protein level in the whole population and Cj0459c is known as a nicking endonuclease and purine-specific ribonuclease conserved domain in *H. pylori* [215].

BAPS clusters 2, 3 and 6 that show a similar phylogenetic background, emphasize identical allelic variants on nucleotide and amino acid sequences in several genes (Table 5.6). Those genes are broadly distributed across the genome of *C. jejuni* and seem to be independent of each other. The genes are associated with multiple metabolic pathways such as 1-deoxy-D-xylulose-5-phosphate synthase (*dxs*), cysteine synthase B (*cys*M) and phosphoenolpyruvate carboxykinase (pckA). Also the genes *dna*E and *ffh* (Figure 5.7), already described as part of a transpinal pathway in cattle-associated strains, show an identical pattern for these host-generalist strains. However, also other transitional-associated genes such as *rpo*D for the RNA polymerase sigma factor and substrate transport functions like *ybi*T for the putative ABC transporter ATP-binding protein have been identified (Figure 5.9).

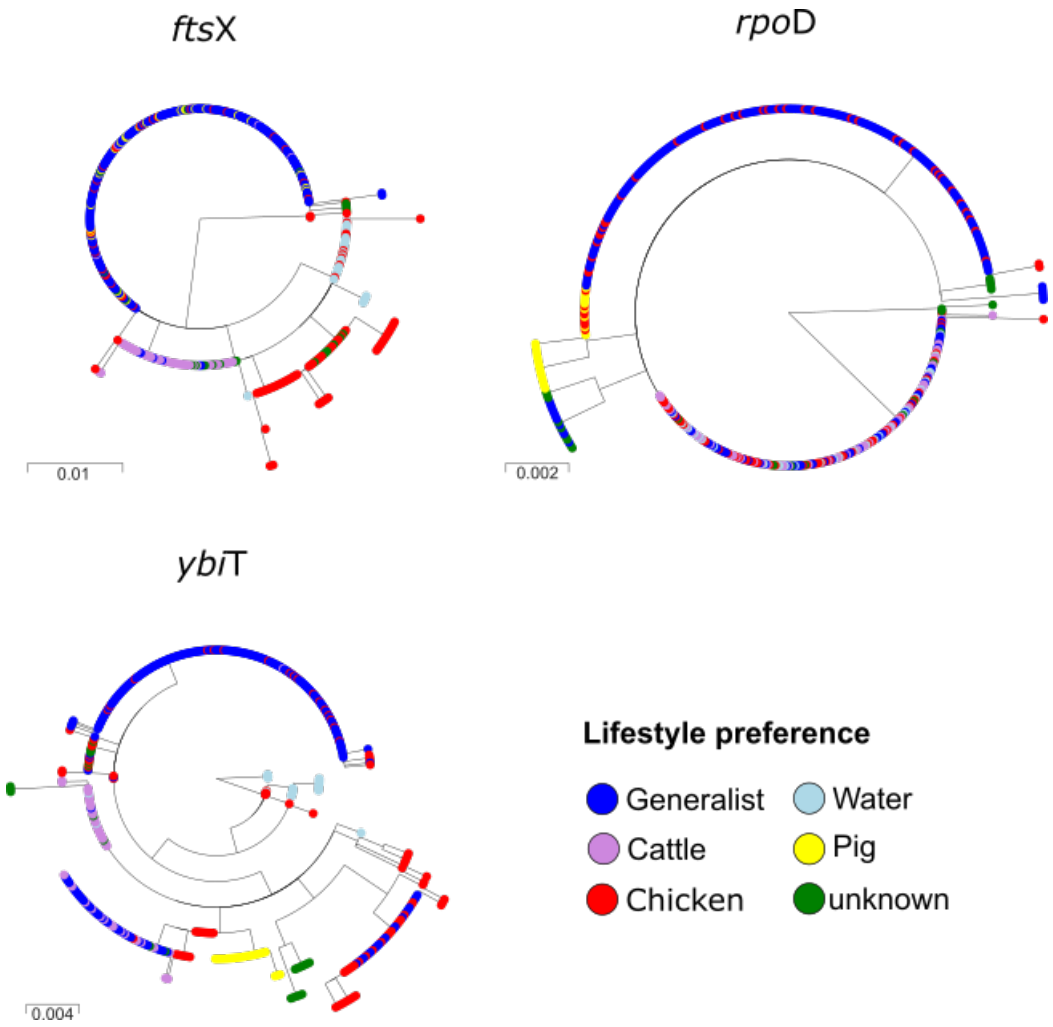


Figure 5.9: Phylogenetic tree of predicted amino acid sequences variants encoded by *fts*X, *rpo*D, *ybi*T (selected from Table 5.6) that show lifestyle-associated variants in different phylogenetic lineages originating from different genetic and geographic backgrounds. Lifestyle preferences are colored in purple (cattle), red (chicken), blue (host-generalists) and yellow (pig).

Table 5.6: Selected accessory genes and allelic variants of the core genome content associated with host-generalism. The table was created and published in [154] during this work.

Locus Tag[a]	Gene	Predicted function	BfR-CA-14430[b]	COG[c]	COG Description	# pig[d]	# cattle[d]	# chicken[d]	# host generalists[d]	# other[d]	Accessory/ Variant[e]
Cj1276c	*fts*X	cell division protein FtsX	1.223.530	D	Cell cycle control,cell division,chromosome partitioning	26	56	90	255	63	C
Cj0459c	-	conserved hypothetical protein (32.5% identical to HP0268)	428.984	-	-	26	56	90	255	63	C
Cj0321	*dxs*	1-deoxy-D-xylulose-5-phosphate synthase	296.904	H	Coenzyme transportand metabolism	26	56	90	255	63	C
Cj0912c	*cys*M	belongs to the cysteinesynthase cystathionine beta-synthase family	862.739	E	Amino acid transportand metabolism	26	56	90	255	63	C
Cj1001	*rpo*D	RNA polymerasesigma factor RpoD	945,528	K	Transcription	26	56	90	255	63	C
Cj0426	*ybi*T	abc transporter atp-binding protein	393,511	S	Function unknown	26	56	90	255	63	C
Cj0932c	*pck*A	phosphoenolpyruvate carboxykinase (ATP)	880.507	H	Coenzyme transport and metabolism	26	56	90	255	63	C

[a] Locus tags for allelic variants of the core genome refer to *C. jejuni* strain NCTC11168 (NCBI accession: AL111168.1)

[b] Position of core genes in the reference strain BfR-CA-14430

[c] Clusters of orthologous groups (`http://clovr.org/docs/clusters-of-orthologous-groups-cogs/`)

[d] Number of genomes assigned to a particular lifestyle carrying the gene or allelic variant (pig, cattle, chicken, host generalists, others)

[e] Accessory (A) indicates that a gene belongs to the accessory genome content of *C. jejuni*. Variant (V) indicates a specific allelic variant of the core genome content.

5.5 Recombination barriers within *Campylobacter jejuni*

Since recombination is one of the fundamental mechanisms behind adaptation of bacteria populations towards a specific ecological niche, a recombination analysis was performed by BratNextGen. As a result, recombination patterns and events within distinct phylogenetic *C. jejuni* lineages were observed and show the putative influence of recombination on the adaptation of the core genome to a particular lifestyle preference.

Significant recombination events identified here were visualized in Figure 5.10. Recombination profiles of BAPS clusters assigned as pig- and cattle-specific as well as host-generalist groups strictly mirroring their lineages showing mainly intra-lineages recombination events. In general, pig-associated BAPS cluster 11 (CC-403) and cattle-specific BAPS cluster 4 (CC-41; CC-22) show very limited amounts of recombinant sites with other strains of the population. This possibly indicates the presence of lineage-specific recombination barriers. Contrarily, cattle-assigned BAPS cluster 10 (CC-61) showed several recombination events with the phylogenetic closely related host generalist BAPS clusters 2, 3 and 6. Between both cattle-associated BAPS clusters only one significantly assigned recombinant event was detected and led to the assumption of an independent evolution. The analysis revealed that chicken-associated lineages (BAPS clusters 1, 5 and 9) were prone to trade-off genetic material with each other and further host-generalist lineages (e.g. CC's belong to BAPS clusters 2, 3 and 6; Figure 5.10). In addition, the host-generalist strains belonging to BAPS clusters 2, 3 and 6 seem to be highly recombinant across the lineage boundaries within the core genome and exchange genetic material with chicken-associated strains as well as cattle-associated BAPS cluster 10 to some extent. However, genetic exchange between BAPS cluster 8 and other host-generalists lineages seems to be rare, due to potential recombination boundaries.

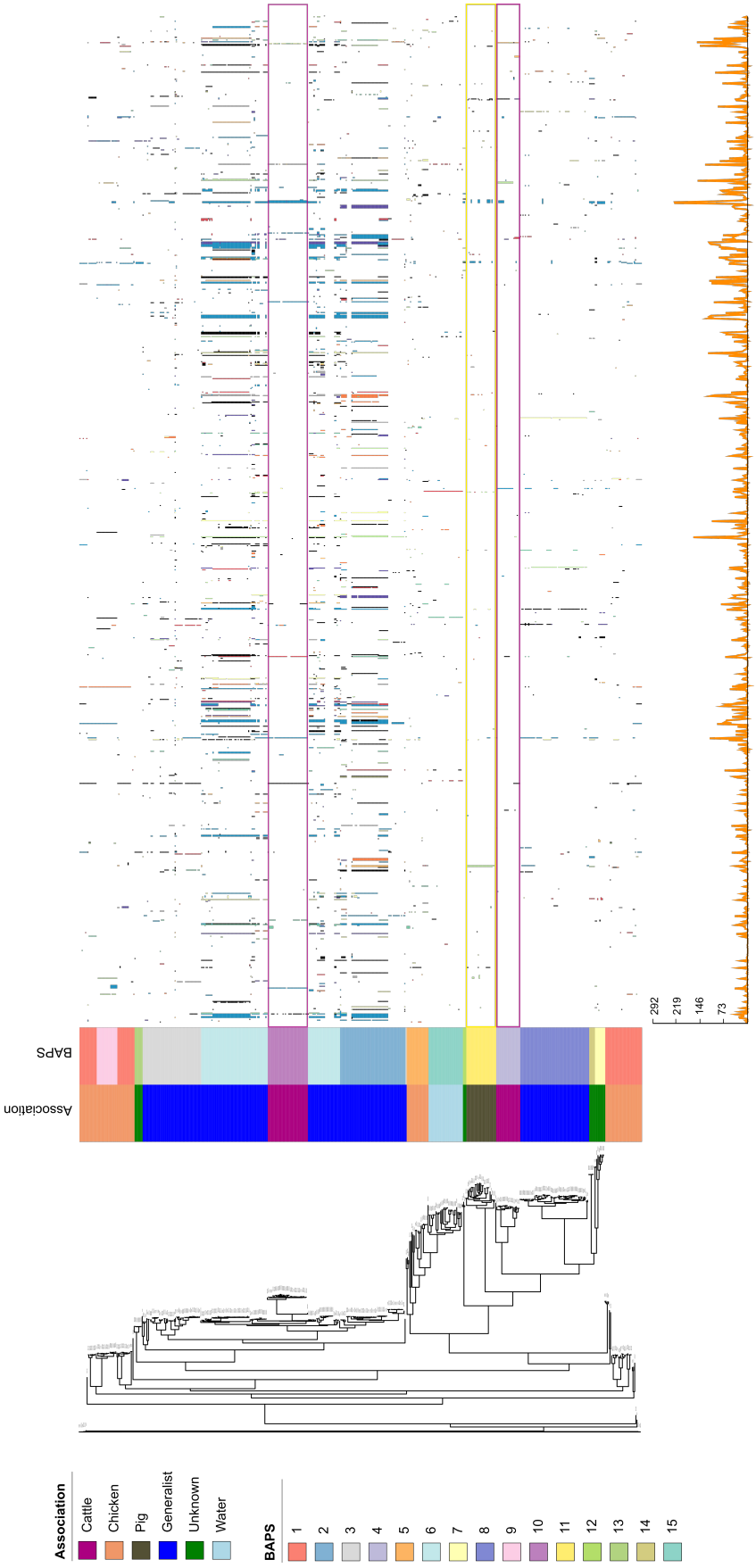


Figure 5.10: Recombination analysis of the core genome of 490 *C. jejuni* genomes. Left side shows the core genome phylogeny in combination with metadata information of host association and BAPS cluster classification. Recombination events are shown by vertical colored lines and blocks on the right side. Blocks with the same color at the same position indicate the same recombination event. Otherwise, the color was chosen arbitrarily. Recombination profiles of cattle (purple) and pig (yellow) are covered by a colored frame. The line graph on the bottom shows the number of recombination events per line. This image was created and published in [154] during this work.

5.6 Discussion

Despite recent achievements in understanding the mechanisms fostering niche and host adaptation processes of distinct *C. jejuni* lineages [18–21], the overall knowledge on the subject is still scarce, especially when compared to the insights on this topic available for *E. coli* or *Salmonella* species [22].

This work provides high-resolution insights into the population structure of *C. jejuni* utilizing various aspects of WGS data. Lifestyle preferences of *C. jejuni* lineages have been defined based on Bayesian clustering of the population structure, that confirmed previous knowledge of MLST data and identified novel host-specific strains. Furthermore, stratified random sampling was used to construct a consensus GWAS based on pyseer [192] for bacteria, in order to gain in-depth knowledge about accessory genes and allelic variants of core genes that might reflect adaptation of *C. jejuni* towards a certain host (pig, cattle, chicken) or a host-generalist lifestyle. Especially, in the core genome content, a broad set of genes with host-specific variants resulting in differences in the predicted amino acid sequences (e.g. *dna*E, *ffh*, *pyc*B, *rpo*B or *fts*X; see Figure 5.6-Figure 5.9 and Table 5.3-Table 5.6) have been found, even in phylogenetically independent lineages that are associated with the same host. This provides evidence for host-adaptive genetic signatures of those genes [29].

Lifestyle preference classification

Over the last decades, assumptions of host association of *C. jejuni* lineage have been made based on MLST data. In this work, the existing associations have been confirmed and extended by whole genome data to generate high-resolution insights into the population structure of *C. jejuni*. Genetic variation is known to be essential for evolutionary change [216]. The genetic diversity between strains assigned to different lifestyles differs vastly. Chicken and host-generalist lineages of *C. jejuni* show a broad variety of genetic backgrounds and diversities [57], whereas strains associated with cattle or pig hosts seem to be more clonal and less diverse. Previous studies assumed that the tight clonal structure of the cattle-associated lineages CC-42 and CC-61 resulted from a more recent onset of the colonization of cattle by *C. jejuni* and therefore might reflect a bottleneck event in its evolution [21, 29]. Considering the limited diversity of pig-specific strains

(Figure 5.1) in BAPS cluster 11 (mainly CC-403) and the fact that *C. jejuni* is not primarily isolated in pigs yields to a similar assumption for *C. jejuni* in pig host.

Clustering of phylogenetic lineages based on WGS data was used resulting in 15 distinct BAPS clusters within the *C. jejuni* population. Each BAPS cluster was analyzed with respect to the sampling origin of each isolate and confirmed the overall outline of the population structure defined by MLST data (Figure 5.1 and Figure 5.2) and previous studies made with WGS data [43, 217]. The predicted lifestyle preference of isolates within particular BAPS clusters (e.g host-specialist or host-generalist) suggested here, was verified by known lifestyle preferences of CCs with BAPS clusters of previous studies (Table 5.1) [43, 81, 83]. Cattle-associated clusters mainly include isolates assigned to CC-42 and CC-61, chicken-associated BAPS clusters include CC-453 and CC-692 and the pig-specific isolates mainly cover CC-403 [43, 81, 83]. Besides known CC-host associations, the WGS approach identifies novel CCs and STs associated with certain hosts, e.g. CC-22 was reported to be adapted towards cattle and strains typed as ST-2274 and were mainly isolated from chicken (Table 5.1).

Microbial GWAS

Throughout the last years GWAS has been adapted and improved in order to study genetic variation in microbial organisms. Since 2016, bacterial GWAS has become an adequate method to study genotypic alterations associated with certain phenotypic traits based on large-scale WGS data [131, 140], including studies on *C. jejuni* [18, 19, 202, 203]. Most studies performed with *Campylobacter* focused on clinically relevant strains CC-21 and CC-45 and utilized a gene-by-gene comparison for the GWAS approach in order to identify significantly associated accessory genes with respect to the phenotypic traits. A further study from Sheppard et al. used a *k-mer* based GWAS for the first time with a fixed *k-mer* length analyzing cattle and chicken isolates from CC-45 [18]. In this work a GWAS based on *k-mers* of different lengths has been used to analyze *C. jejuni* strains from chicken, cattle, pig and further isolate sources such as human clinical cases or environmental samples from Germany and Canada. This results in a representative sample set for the whole population of *C. jejuni* (Figure 5.1).

The GWAS acts like a “sieve” that only allows a subset of mutations to persist and

become the observable differences between groups of genomes [216] not only showing differences in the accessory genome, but also associating imported allelic variants within the core genome of *C. jejuni*. Several non-synonymous allelic variants of genes were identified in independent phylogenetic lineages associated with a particular lifestyle or ecological niche. Consequently, gene variants have most likely not been transmitted between those lineages, but have been acquired independently, a process that is also known as homoplasy, likely to reflect the adaptation to a particular ecological niche and/or host [216, 218, 219].

Allelic variants of core genes that have been discovered for pig, cattle and chicken specific *C. jejuni* strains were evaluated by showing the same alteration within the encoded amino acid sequences even in far distantly related BAPS clusters/lineages, e.g. BAPS 4 and 10 in cattle or BAPS 1, 5 and 9 in chicken. Thus, evidence is provided for the role of these allelic variants in (niche) adaptation processes [29]. In host-generalist *C. jejuni* strains, allelic variants of core genes, such as *fts*X identified in CC-21 and CC-45, could be used as diagnostic markers, since these variants occurred independently of both, phylogenetic background and geographic origin (Figure 5.1, Figure 5.9) and analysis of the recombination patterns (Figure 5.10). In order to prove the independent acquisition of the variants, a recombination analysis was performed. This lacked notable recombination events accountable for most of the changed core genome sites identified. Variants have been acquired constantly through independent evolution, leading to the assumption of homoplasy once again. However, more research on the subject including isolates covering a broader time span is clearly needed to gain insight into the bacterial evolution of *C. jejuni* [21].

Cattle-associated *C. jejuni* - From Central Europe to Canada

Both cattle-associated lineages (CC-42/CC-22 and CC-61) represented by BAPS clusters 4 and 10 incorporated an independent set of accessory genes which have been observed in Figure 5.4 and were confirmed with the *k-mer* based GWAS. This could be a result of evolutionary independent colonization events for the cattle host by *Campylobacter* [32]. This hypothesis is also supported by a limited amount of recombination events within the core genome (Figure 5.10). A most relevant previous study on recombination events between distinct *C. jejuni* lineages showed that - beyond others - the lack

of opportunities due to physical distance (ecological reasons) or other mechanical/functional incapability of DNA exchange might play a role here [43]. This was shown in experimental settings on host-generalists that seem not to be highly recombinant in nature but frequently exchange DNA when in physical contact. In conclusion, colonizing the same host does not necessarily result in interaction or direct contact of different lineages as they might prefer different (sub-)niches of the cattle gut, which would limit the opportunity of DNA exchange between lineages [43]. Additionally, structure and composition of the gut microbiome might play a role here and little is known about putative lineage-specific differences among *C. jejuni*, e.g. with respect to strategies for attachment to host cell tissue [220, 221].

In BAPS cluster 10, a plasmid-associated gene cassette that has been potentially acquired through HGT and shown to play an important role in adoption under stress conditions [222–224], has been identified. This region includes various genes such as a phage-associated integrase (*xer*C), a putative protease (*prot*A), a gene known to be involved in extracytoplasmatic stress response (*yaf*Q), a plasmid DNA repair gene (*rep*A) and genes associated with a HicA-HicB toxin/antitoxin system inhibiting the transfer of mRNA in case of limited nutrient availability.

From a historical point of view, it seems that these lineages have likely been spread through import and export of dairy cattle. This hypothesis is supported by the historical import and export data from North America. For example, Holstein Friesian cows were exported from Friesland and North Holland to Canada (e.g. Ontario) at least since the 1880's where it has become the most common dairy breed in Canada (>90%) nowadays, according to the Canadian Dairy Information Centre (`https://www.dairyinfo.gc.ca`). Since then, numerous more animals have been imported by North America to increase breeding success.

In the core genome, genes with identical allelic variants have been discovered in both cattle clusters (BAPS clusters 4 and 10) by the consensus GWAS (Table 5.4). Several of those changes in corresponding amino acid sequences have been confirmed (Table 5.4, Figure 5.7). A putative cattle-specific allelic variant of the DNA polymerase III subunit alpha encoded by *dna*E might contribute to the adaptation towards cattle hosts, as variants of these genes have been shown to increase the overall mutation rate in *E. coli*

[225, 226]. An increased mutation rate is one of the key factors for niche adaptation in evolution [29] and supports the hypothesis of a more recent transmission of *C. jejuni* to this host [21]. Furthermore, the gene *ffh* (Figure 5.7) also shows cattle-associated variants in the amino acid sequence. It encodes a signal recognition particle protein as part of the signal recognition particle (SRP) pathway GTPases that mediates the co-translational targeting of membrane and secretory proteins to the bacterial membrane [227]. This indicates an adaptation of transport processes, too. In *E. coli*, the SRP system plays an important role in membrane protein biosynthesis, and previous research indicated that Ffh is involved in the regulation of membrane protein translation [228]. Notably, a further GTPase (FlhF) possessing an active domain most similar to Ffh, was discovered to be involved in flagellar gene regulation and biosynthesis in *C. jejuni* [229]. Mutation of these genes might have also evolved independently, leading once more to the assumption of homoplasy. *Ffh* has been already been described as a homoplasic gene on nucleotide level in cattle-associated genomes in a recent study by Mourkas et al. [21].

Pig as novel niche for *C. jejuni*

Most of the pig-associated *C. jejuni* genomes represented by BAPS cluster 11 (CC-403 and ST-1942) carry a unique set of accessory genes that encode for an RM system of type I and type II (Table 5.3). These might be involved in lineage-specific recombination boundaries while shielding the bacteria from introgression [79, 230, 231]. This hypothesis is supported by the recombination analysis (Figure 5.10) that shows an internal-lineage recombination pattern, which has been previously noted for CC-403 and its related STs [212].

A former study showed amino acid variants in the *ten*I gene that likely affect the thiamine metabolism in *C. jejuni* lineages associated with cattle [21]. In this work, further non-synonymous SNPs were identified within the same gene in pig-associated BAPS cluster 11. In addition, the gene *dxs*, encoding a putative thiamine-dependent synthase (1-deoxyxylulose-5-phosphate synthase) affecting the same pathway, also shows pig-specific alterations. In general, this underlines the importance of this metabolic pathway for host adaptation of *C. jejuni*. Amino acid changes detected for final aromatase (TenI), needed in thiamine biosynthesis, seemed rather extensive (Figure 5.6),

likely indicating functional alterations or even loss-of-function of the enzyme, which would be interesting to characterize in the future. Since industrial diets for pigs are generally supplemented with thiamine [232], reduction or even shut-off of the pathway might be beneficial for pig-specialized *C. jejuni* lineages.

Chicken association

Chicken-associated *C. jejuni* lineages are represented by three BAPS clusters (1, 5 and 9). In most of these strains, accessory genes for a putative conjugative transfer protein (*Tra*G-like), which is commonly linked to a type IV secretion system essential for DNA transfer in bacterial conjugation [233, 234], were identified by mapping the significantly associated *k-mers*. Within the core genome content of chicken-specific strains, alterations of the amino acid sequence were observed in several genes (Figure 5.8, Table 5.5, appendix Table A.5). Those genes include the *pyc*B gene, encoding the second subunit of the anaplerotic and glucogenic pyruvate carboxylase in *C. jejuni* [235], indicating specific adaptation of a basal metabolic pathway. Furthermore, the housekeeping gene *rpo*B also shows chicken-specific variations on nucleotide level leading to change of the resulting protein. Besides the frequent use of the nucleotide sequence of *rpoB* to investigate the genetic relationships of the *Campylobacter* genus [236], it has been shown that mutation of *rpo*B enhances growth at 42.2 ℃ compared to the wildtype in *E. coli* [237]. Since the body temperature of poultry is commonly between 39 and 42 ℃ [238] and therefore vastly differs from other hosts such as cattle or pig, temperature–induced adaptive changes likely play an important role here.

Host-generalist lineages

The host-generalist *C. jejuni* strains are divided into several main lineages represented by BAPS clusters 2, 3, 6 (mainly CC-21 and CC-48) and BAPS cluster 8 (CC-45). Lineages from CC-21/CC-48 show a phylogenetically independent background in comparison to CC-45. This difference is clearly shown in the accessory genome profiles as all lineages carry a different set of genes, which confirms previous results from Yahara et al., who have tracked these lineages from the chicken flock down the meat production chain to clinical strains from humans [202]. The accessory gene profiles visualized in Figure 5.4 provide evidence that each BAPS cluster carries its own set of genes indepen-

dently of geographical location or sample origin, i.e. animal host, human clinical sample or environment. However, host-generalist BAPS clusters seem to have a larger pool of accessory genes possibly reflecting a repertoire of genes that might be used in order to survive in different hosts or within the environment [48, 239]. Those genes might be exchanged through HGT between the strains, as the recombination analysis (Figure 5.10) provides evidence that host-generalist lineages, especially BAPS clusters 2, 3 and 6, are highly prone to DNA exchange within their core genome. Natural transformation and recombination between host-generalist lineages enhance adaptive possibilities needed to survive in the environment, in different animal species and humans. Since genomes of *C. jejuni* isolated from clinical human samples were often identified within the BAPS clusters 2, 3, 6 and 8, they probably harbor allelic variants that enhance their potential to cause Campylobacteriosis. Due to the overall adaptive potential and ongoing evolvement of host-generalist lineages and their frequent contact with the human gut, it seems possible that some *C. jejuni* strains will become human commensals in the future.

Mapping of significantly associated *k-mers* of host-generalist strains reveal several nucleotide variants of transcriptional regulators and ribosomal genes e.g. for *ars*C, *rpl*S and *rps*P. These genes show a highly conserved amino acid sequences within the whole *C. jejuni* species. Nonetheless, a variant of the gene *fts*X (Figure 5.9) encoding a cell division protein, shows a host-generalist specific amino acid sequence. This might reflect a difference in stress response, as Riedel et al. showed that *fts*X transcription appears in *C. lari* after exposure to heat stress [240]. As already assumed for cattle associates strains, identical allelic variants of genes in host-generalists might be a result of convergent evolution (homoplasy). Since BAPS 8 seems to have an independent phylogenetic background when compared to 2, 3 and the recombination analysis suggests a limited number of recombination events between both major groups. Host-generalist BAPS cluster 8 (CC-45) shared several allelic synonymous as well as non-synonymous variants belonging to the core genome (e.g. *rpl*S, *trm*D, *rim*M and *rps*P (Table 5.6), with the cattle-specific BAPS clusters 4 and 10 suggesting a phylogenetic relationship between host-generalists and cattle-specialists [21].

6 Reference-free identification of inter species recombination

The following chapter is based on material of a published article:

Golz JC*, Epping L*, Knüver MT, Borowiak M, Hartkopf F, Deneke C, Malorny B, Semmler T & Stingl K. Whole genome sequencing reveals extended natural transformation in *Campylobacter* impacting diagnostics and the pathogens adaptive potential. Scientific reports. 2020 Feb 28;10(1):1-2. (*These authors contributed equally)

6.1 Background

Around one-third of *Campylobacter* infections can be directly linked to handling, preparation and consumption of broiler meat [15]. In 2017, there have been 250,161 confirmed Campylobacteriosis cases within the European Union, which confirms the outstanding public health importance of this particular gastro-enteric disease. As a consequence, routine surveillance within the food chain has been implemented to monitor contamination by *Campylobacter* in food products [241]. Between 2016 and 2018, more than 4,000 *Campylobacter* isolates were sampled and screened by the German Federal State Laboratories. qPCR was employed in order to discriminate between *C. jejuni* and *C. coli* with *map*A and *ceu*E as species markers, respectively [39]. During that verification process, ambiguous results have been detected, presumptively indicating interspecies transfer of genetic material can occur between *C. jejuni* and *C. coli* [25]. For this chapter, WGS data were analyzed in order to identify putative recombinant regions in 37 *C. coli* isolates. To achieve high-resolution across the whole genome, a novel *k-mer*-based workflow has been developed and applied as proof of concept (Section 4.2.9).

6.2 *C. coli* yielding ambiguous results using a species-specific qPCR

While performing qPCR for species deterioration of *C. jejuni* and *C. coli*, 37 out of 4,335 *Campylobacter* isolates showed ambiguous PCR results. A qPCR result was classified as ambiguous when amplification of both target genes employed to discriminate *C. jejuni* and *C. coli* yielded comparable cycle threshold values or the amplification process lacked a detectable result. In total, 31 genomes (31/37) showed an amplification product for both species-specific target regions, *map*A (*C. jejuni*) and *ceu*E (*C. coli*). For 4

Table 6.1: Overview of *Campylobacter* genomes with ambiguous mPCR results [n=45]. The qPCR results showed either signatures for both species (Cj/Cc mix), no result, correctly predicted *C. coli* or falsely classified the sample as *C. jejuni* (Cj)

Source	**# mPCR-based species determination**	**Cj/Cc mix**	**none**	**Cc**	**Cj (false)**
eggshells	17 (initially: 9)	5	2	8	2
chicken meat	9	9	0	0	0
duck meat	1	1	0	0	0
turkey cecum/skin	13	12	1	0	0
turkey meat	5	4	1	0	0

(4/37) isolates no *qPCR* amplification product was detected and 2 (2/37) isolates were predicted as *C. jejuni* (Table 6.1). A previously published multiplex polymerase chain reaction (mPCR) based on alternative primer targets was used as a gold standard, with internal regions of the genes *hip*O for *C. jejuni* and *gly*A for *C. coli* [242]. The PCR results identified all 37 isolates as *C. coli*. Additional 8 isolates from eggshells (Section 4.1.2) were included in further investigations since ambiguous PCR results seemed commonly associated with that particular sample origin. In total, 45 genomes were associated with isolates originating from different sources and matrices associated with poultry, i.e. poultry meat [n=9], turkey cecum or skin [n=13], turkey meat [n=5], eggs [n=17] and duck meat [n=1] (Section 4.1.2). In the following section, results of further investigations of WGS data for these 45 isolates are presented.

6.3 Genome-wide relationships of *C. jejuni* and *C. coli*

A preliminary *k-mer*-based was conducted utilizing the online tool KmerFinder v3.1 [102–104] and suggested that *Campylobacter* genomes analyzed here can be categorized into two groups (Table A.2) as part of the *C. coli* population: "Hybrid" strains that contain at least 10% genome content of *C. jejuni* and "Half hybrid" strains that contain less than 10% genome content of *C. jejuni*, but still showed ambiguous qPCR results. This initial observation was used to guide subsequent in-depth analyses.

In order to provide a reliable overview including a high-resolution analysis of phylogenetic groups, several methods to verify the *Campylobacter* species based on WGS data

have been utilized to investigate dataset 2.

The genetic relationship and taxonomic affiliation of hybrid and half hybrid strains within the *Campylobacter* population. In the first place this was assigned by an average nucleotide identity (ANI) analysis using the tool FastANI [179]. As described in Section 4.2.5, genomes with an ANI value of at least 95% belong to the same species [176, 243]. Considering the two species under investigation here, *C. jejuni* and *C. coli* showed an average ANI of 86.23% (Figure 6.1 B). Further ANI revealed that the putative hybrid strains form a separate cluster while also sharing an ANI of 96.95% with *C. coli* genomes (Figure 6.1 B). Moreover, the ANI analysis results considering the *C. jejuni* population were 97% to 100%, while 87.92% were shared with genomes of the hybrid isolates. Thus, hybrid isolates seem to regularly occur among the *C. coli* population. In contrast to *C. jejuni/C. coli* hybrid strains, half hybrid strains showed a close relationship with the majority of the *C. coli* population while sharing an ANI of 98.96% and do not form a separate cluster (Figure 6.1).

These results are mirrored by the core genome phylogeny (Figure 6.1 B) that is based on 800 core-associated genes shared by *C. coli* and *C. jejuni* with at least 80% sequence similarity. Again, half hybrid isolates are direct ancestors of the main *C. coli* population, whereas the hybrid strains form a separated clade that is still closely related to the *C. coli* group nearby (Figure 6.1 A). In general, both analyses show that the diversity/identity between *C. coli* including hybrids is similar to the diversity within the *C. jejuni* population, confirming that the hybrid strains indeed belong to the species *C. coli*. However, the *C. coli* main population is more homogeneous, whereas the *C. jejuni* population has a more diverse population structure (Figure 6.1 A). Furthermore, the clonal relationship of hybrid and half hybrid genomes was evaluated by a MLST analysis (Figure 6.2). The MLST showed a non-clonal relationship between the genomes, indicating an independent introgression of novel sequences. Besides whole-genome and population structure analysis, mass spectroscopy analysis with MALDI-TOF verified the species affiliations undoubtedly on using reference spectra on protein level of ambiguous isolates in dataset 2.

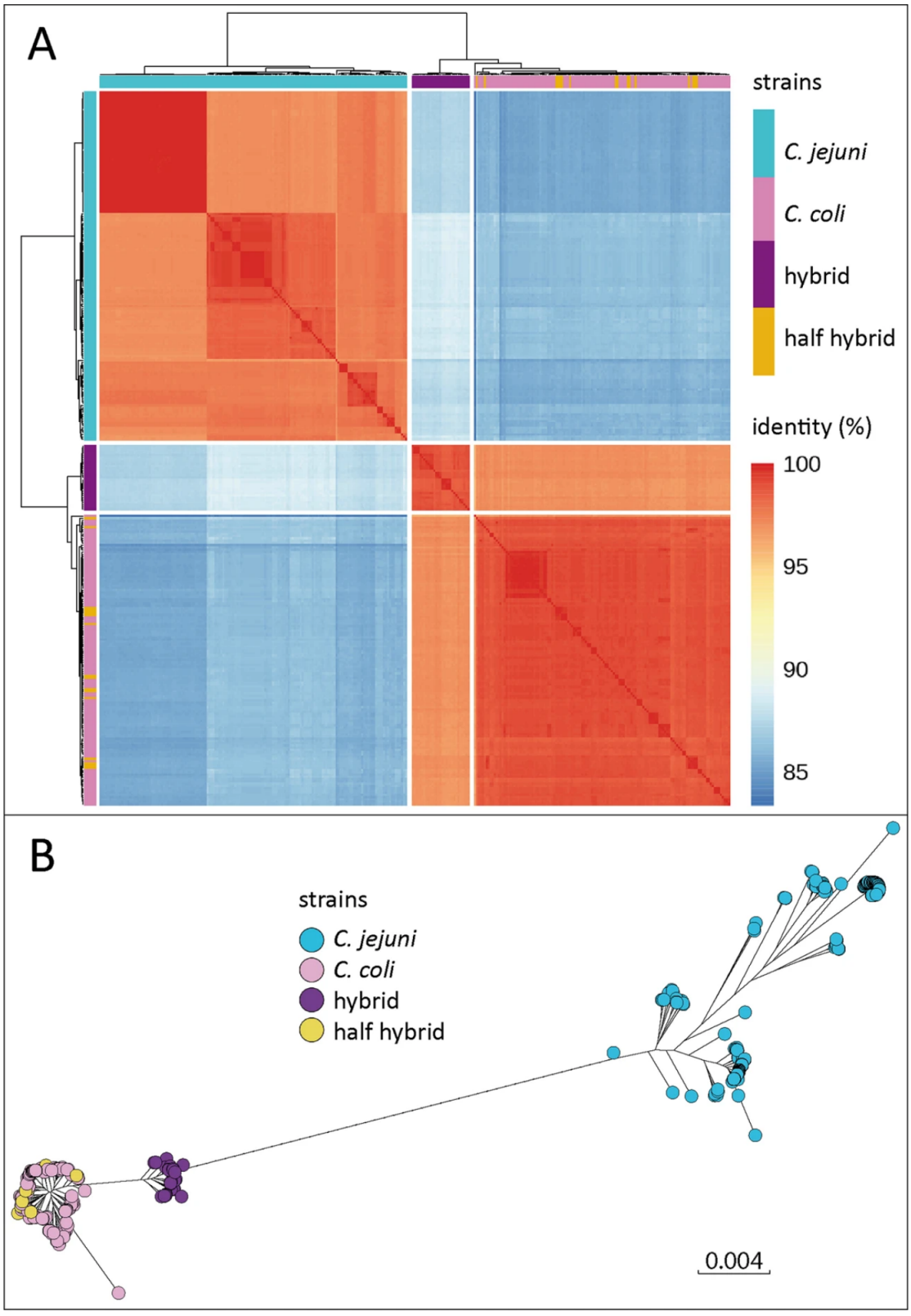

A
strains
C. jejuni
C. coli
hybrid
half hybrid
identity (%)
100
95
90
85
B
strains
C. jejuni
C. coli
hybrid
half hybrid
0.004

Figure 6.1: Phylogenetic relationship of *C. jejuni* (turquoise), *C. coli* (pink), hybrid strains (purple) and half hybrid strains (mustard) from dataset 2 based on results from ANI (A) and core genome analysis phylogeny (B). (A) ANI results are visualized in a heatmap across all isolates. Hybrid strains form a separate cluster, but still share 97% ANI with *C. coli*. Half hybrid isolates are spread across the *C. coli* population. (B) Phylogeny of the *Campylobacter* core genomes based on Roary analysis. The branch length between *C. coli*, including hybrid and half hybrid strains, and *C. jejuni* has been shortened for better visualization. This image was created and published in [39] during this work.

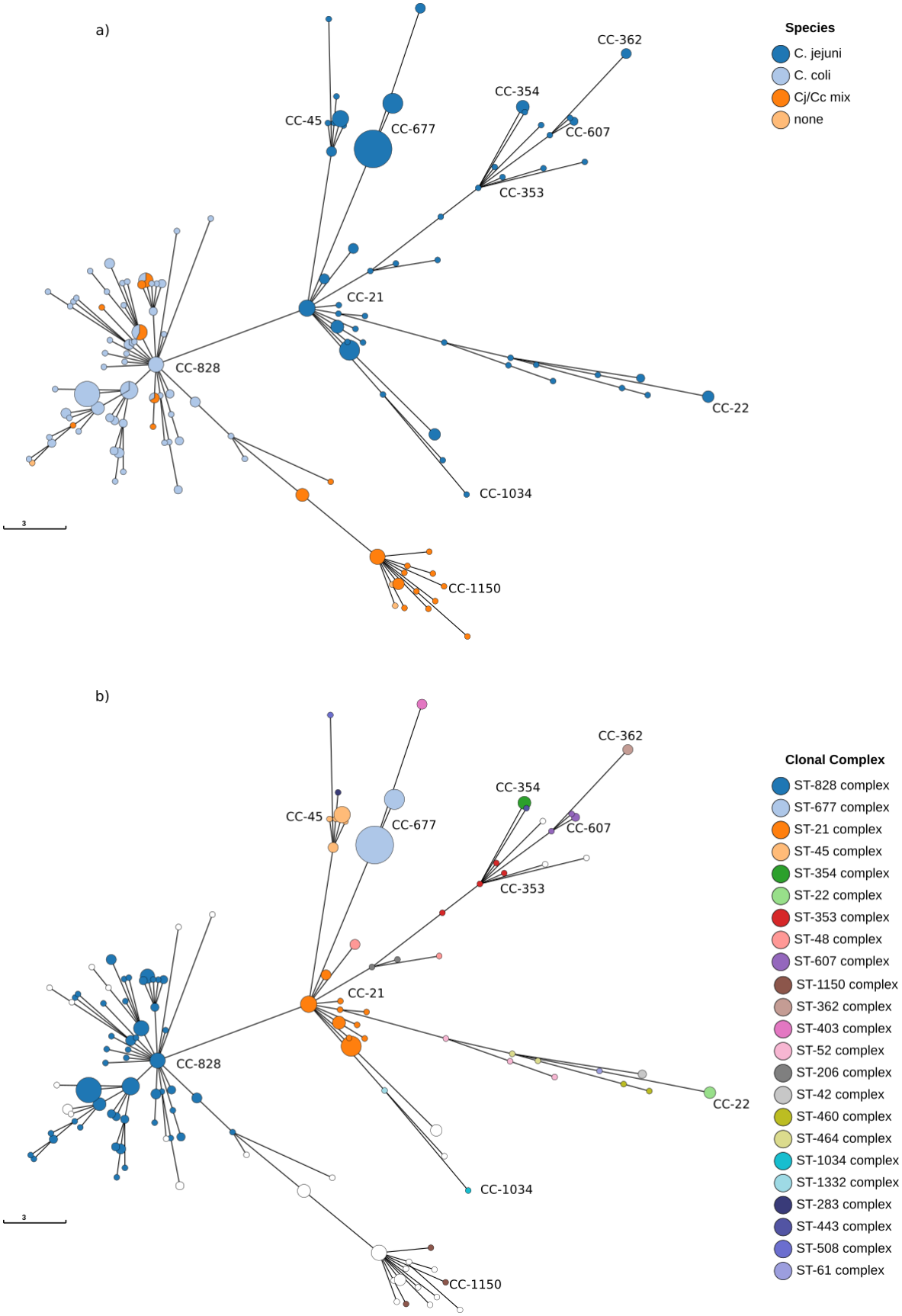

a)
Species
C. jejuni
C. coli
Cj/Cc mix
none
CC-362
CC-354
CC-607
CC-45
CC-677
CC-353
CC-21
CC-828
CC-22
CC-1034
CC-1150
3
b)
Clonal Complex
ST-828 complex
ST-677 complex
ST-21 complex
ST-45 complex
ST-354 complex
ST-22 complex
ST-353 complex
ST-48 complex
ST-607 complex
ST-1150 complex
ST-362 complex
ST-403 complex
ST-52 complex
ST-206 complex
ST-42 complex
ST-460 complex
ST-464 complex
ST-1034 complex
ST-1332 complex
ST-283 complex
ST-443 complex
ST-508 complex
ST-61 complex
CC-362
CC-354
CC-607
CC-45
CC-677
CC-353
CC-21
CC-828
CC-22
CC-1034
CC-1150
3

Figure 6.2: MSTs of *C. coli*, *C. jejuni* and hybrid strains. STs in MST a) are colored according to their *Campylobacter* species, whereas STs in MST b) are colored by their corresponding CC. The non-clonal relationship of all hybrid strains implies an independent introgression of novel genes.

6.4 Screening of recombinant regions among *C. coli*

Ambiguous *C. coli* genomes with $\geq 10\%$ introgression of *C. jejuni* sequence content were analyzed by a *k-mer*-based workflow designed and developed in scope of this work (Section 4.2.9). This approach counts *k-mers* of 16 and 31 bp length in each of the genomes and compares them against pre-calculated *C. coli* and *C. jejuni k-mer* databases. *K-mers* that were present in at least 95% of the *C. jejuni* genomes and less than 5% in *C. coli* genomes were mapped against the *C. jejuni* reference strain NCTC11168 in order to identify putative recombinant genomic regions. Example *k-mer* mappings of a hybrid ($\geq 10\%$ *C. jejuni* introgression) and a half hybrid strain ($<10\%$ *C. jejuni* introgression but with ambiguous qPCR result) against the NCTC11168 reference sequence are depicted in Figure 6.3. Loci, where recombination events lead to the incorporation of *C. jejuni* sequences in *C. coli* are scattered along the genomes. Putative recombination events involving more than 100 bp were further analyzed and *k-mers* that mapped within a distance of 100 - 500 bp were assumed to be associated with the same recombination event. As a result, putative recombination fragments of median sizes ranging from 297 to 512 bp, with most of them smaller than 1 kb were identified. However, strain BfR-CA-08318 also revealed a large recombinant site of 11.4 to 11.8 kb, demonstrating the potential of *C. coli* genomes to incorporate large sequences of *C. jejuni* origins. The total amount of recombinant bases were summarized in a cumulative plot (gap size $\leq$ 100 bp) and illustrated alongside the genomic structure of the reference sequence NCTC11168 (Figure 6.3). Recombination events in most of the *C. coli* hybrid genomes summarize to a total amount of 206.642 to 239.893 kb and mirror the results obtained from KmerFinder v3.1 [102–104] (Table A.2). Of note, the current analysis may have underestimated the number and sizes of recombination sites due to the fact that only *k-mers* with exact and unique matches to the reference *C. jejuni* NCTC11168 were considered. Furthermore, only *k-mers* represented in at least 95% of all *C. jejuni* strains included in this study were used to set up the respective database.

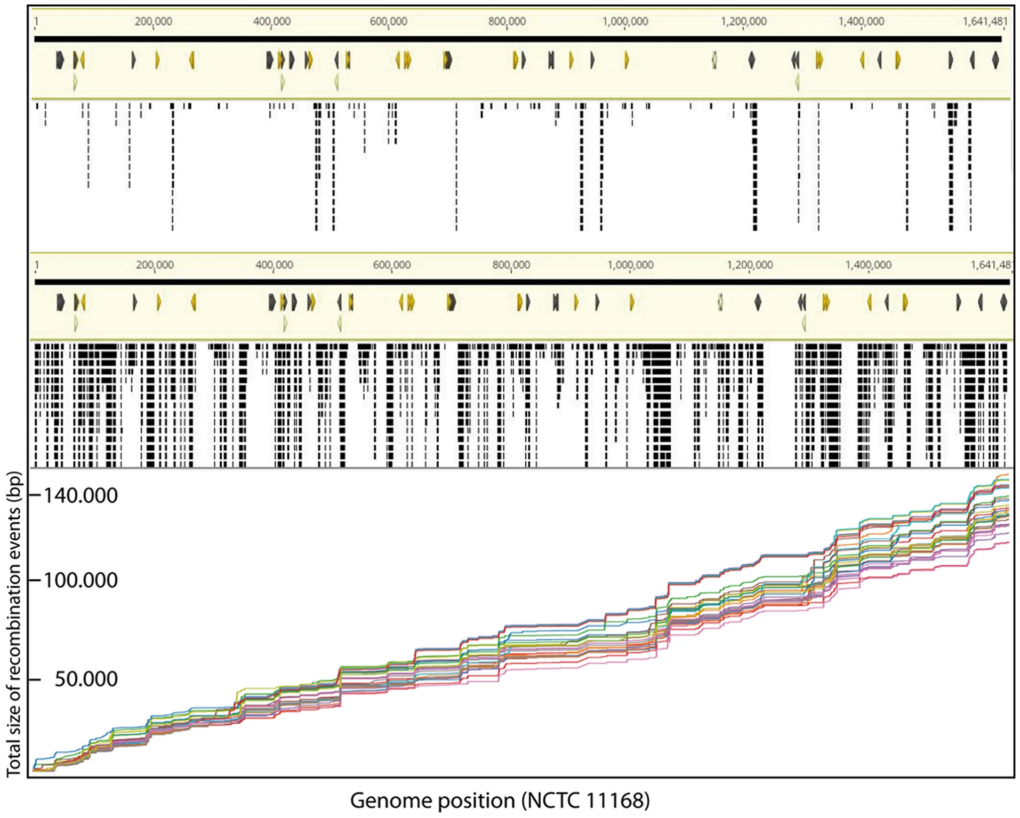


Figure 6.3: Exemplary visualization of *k-mer* mapping against the reference genome NCTC11168 with a length of 1.64 Mb. Upper and middle panel shows *k-mer* peaks for a half hybrid strain (<10% DNA introgression from *C. jejuni*) and a hybrid strain ($\geq$ 10% DNA introgression from *C. jejuni*). The genome is visualized in light yellow and black bars indicating *k-mers* mapping a specific position in the genome. The lower panel sketches the cumulative sum of predicted recombination events for the 29 hybrid genomes. Stepwise increasing curves are associated with the recombinant loci sizes. This image was created and published in [39] during this work.

Since the cumulative sums of recombination events showed a similar pattern for most of the genomes under consideration (Figure 6.3), the loci prone to recombination, likely do not occur randomly. In order to investigate this hypothesis, 300 randomly distributed recombination events within 800 core genes of *Campylobacter* were simulated (Figure 6.4). The plot shows a normal distribution with a peak at 11-12 recombination events in slightly over 100 core genes, whereas the observed recombination follows a bi-

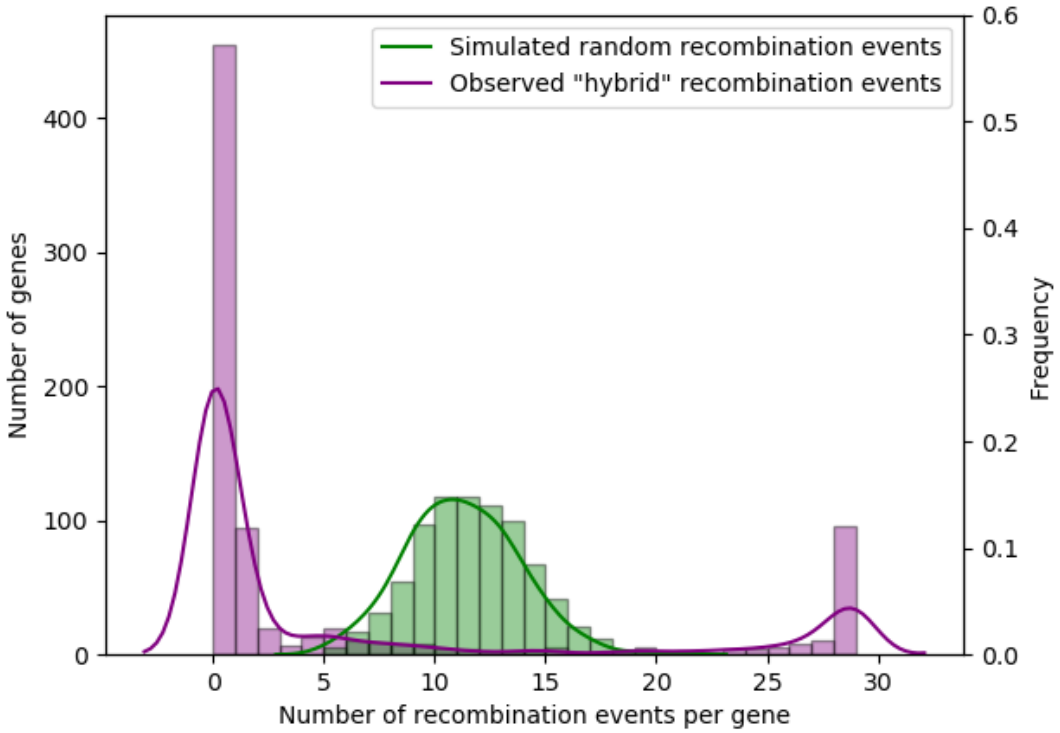


Figure 6.4: Histograms of simulated (green) and observed (purple) recombinant genes based on 800 core genes. This image was created and published in [39] during this work.

modal distribution with more than 400 genes not observed in any recombination event. Additionally, 104 genes were identified in at least 25 out of the 29 hybrid genomes. Comparing these two distributions by a χ^2 test showed a significant difference with p < 0.01. Thus, hot spots of recombination have been identified.

6.5 Functional annotation of recombinant regions between *C. jejuni* and *C. coli* hybrid strains

By the *k-mer*-based approach for recombination detection, 346 genes were identified in hybrid genomes with more than 10% introgression. Genes with at least 20% of the gene length covered by *k-mers* were identified and genes with at least 50% *k-mer* coverage in at least one hybrid genome were visualized in Figure 6.5. In total 104 genes were identified that were exchanged in at least 25 of the 29 hybrid genomes (*k-mer* size 16

and 31). Annotation of functional clusters of orthologous genes (COG) revealed that approximately 50% of these genes are involved in fitness and stress response i.e. oxidative stress response (*kat*A, Cj1386, *mrs*B, *can*B, Cj0833c *hyd*A, *hyd*A2, *nad*D, *nuo*A, *nuo*B, *nuo*C, Cj0081), stress response in general (*clp*A, *htr*B, *htr*A, *cpn*10, *cpn*60), DNA metabolism and repair (*pur*F, *pyr*G, *thy*X, *rar*A, *rec*J, *ung*, *rib*A, *gua*B, *dut*), chemotaxis and flagellar motor switch (*che*A, *che*V, *che*W, *fli*Y), signal transduction (Cj1110c, Cj1227c, Cj1258), membrane transporters (*crc*B, cj0832c, *ktr*A, *ktr*B, Cj1257c, Cj1687), cell wall and capsule biosynthesis (*kps*S, *kps*E, *kps*F, *kps*D, *kps*T, *mur*E) and a gene for S-adenosylmethionine transferase (*met*K), involved in substrate supply for methylation reactions [39].

On NCBI, one additional American *C. coli* isolate (Strain ID: RM4661), originating from a turkey carcass (NZ_CP007181.1) was identified as a *C. coli*/*C. jejuni*-hybrid strain as well. This shares 106 of the 126 *C. jejuni* introgressed genes revealed in the majority of the analyzed hybrid strains (appendix Table A.8) in this study. Therefore, the strain probably underwent a similar selection procedure.

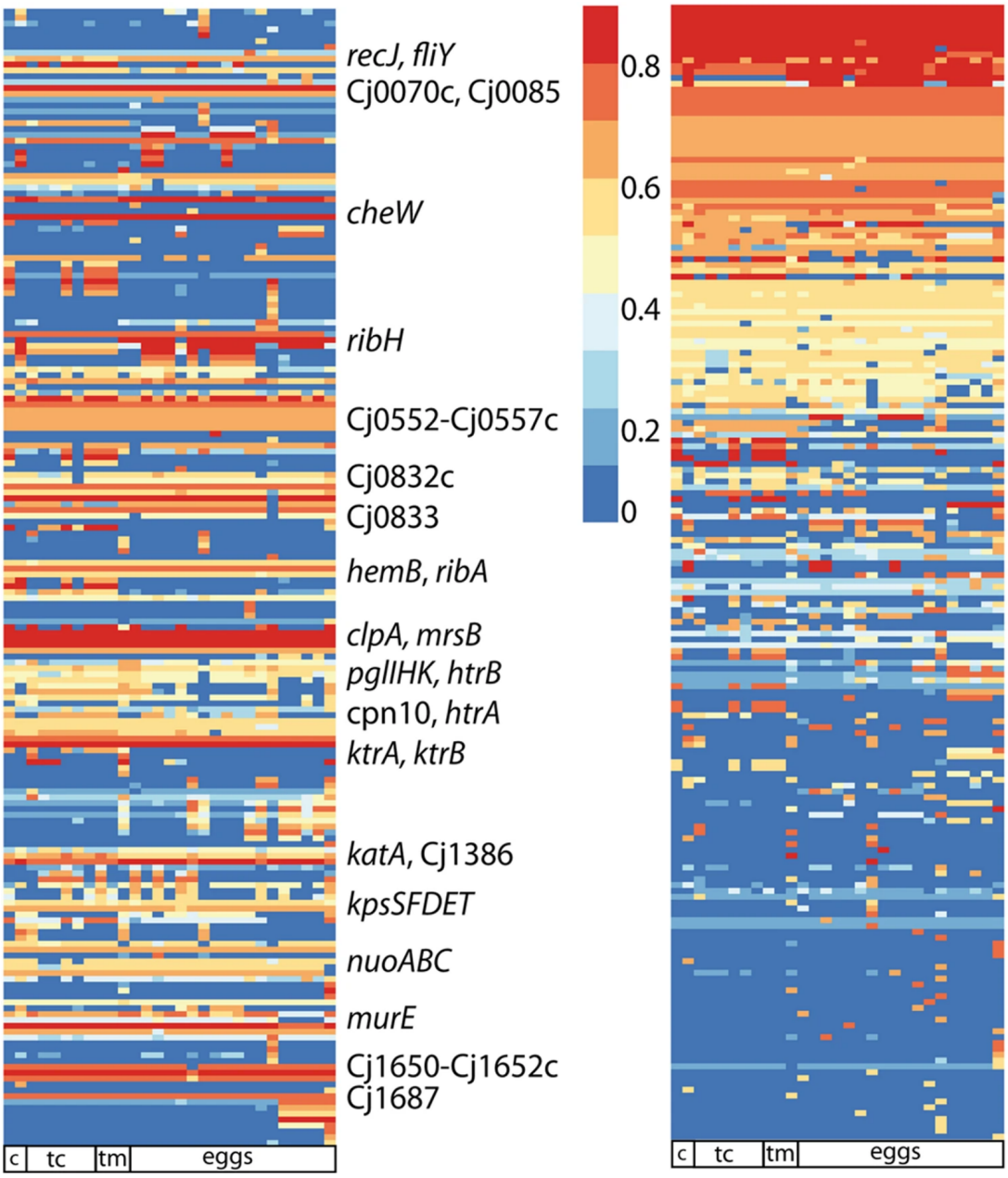


Figure 6.5: Heatmap of recombinant genes identified among hybrid genomes with at least 50% *k-mer* coverage and a *k-mer* size of 16 bp. Isolates were grouped by sample origins: Chicken meat (c), turkey cecum (tc), turkey meat (tm) and eggs. The left heatmap shows genes according to their position within the reference genome NCTC11168 and the right heatmap is sorted by *k-mer* coverage. Both maps indicate that recombinant genes are commonly shared by all hybrid genomes. This image was created and published in [39] during this work

6.6 Recombinant regions around *map*A and *ceu*E

Recombination analysis revealed that different events affecting the *map*A and *cue*E genes (Figure 6.6) limited the reliability of qPCR results initially employed for species verification. Several recombinations within two genes of the genomes with ambiguous qPCR results (Table 6.1) were identified. The ambiguous qPCR results might reflect recombination events that can be addressed in three different scenarios:

- *k-mer* mapping on genomes of isolates with positive qPCR results for both, *C. jejuni* and *C. coli* (Cj/Cc mixed samples) showed either partial substitution of the *C. coli map*A gene by sequences from *C. jejuni* (a), a very short integrated sequence exchange (c) or a larger recombination event associated with a substitution of the whole region from *mapA* to *gyr*A (b). The *cue*E gene did not seem to be affected by the *k-mer* mapping.

- *k-mer* mapping on genomes of isolates lacking qPCR results (none) revealed two different recombination events that lead to a replacement of *ceu*E together with adjacent genes by sequences of *C. jejuni* origin, while the *map*A gene was not affected (d & e).

- *k-mer* mapping on genomes of isolates yielding false positive results for *C. jejuni* (false *C. jejuni*). The *map*A gene was partially replaced by sequence variants of *C. jejuni* origin, whereas *ceu*E showed a mosaic allelic structure with the 5' start of *ceu*E displaying a typical *C. coli* sequence and the 3' end matching *C. jejuni* sequences (e & f).

Overall, introgression of genetic elements from *C. jejuni* into the *map*A locus of *C. coli* have been observed frequently in the data presented here. In contrast to a previous study [244], genetic exchange can also happen within the *cue*E locus leading to a fully replaced or a novel mosaic structure of the *cue*E allele (Figure 6.6).

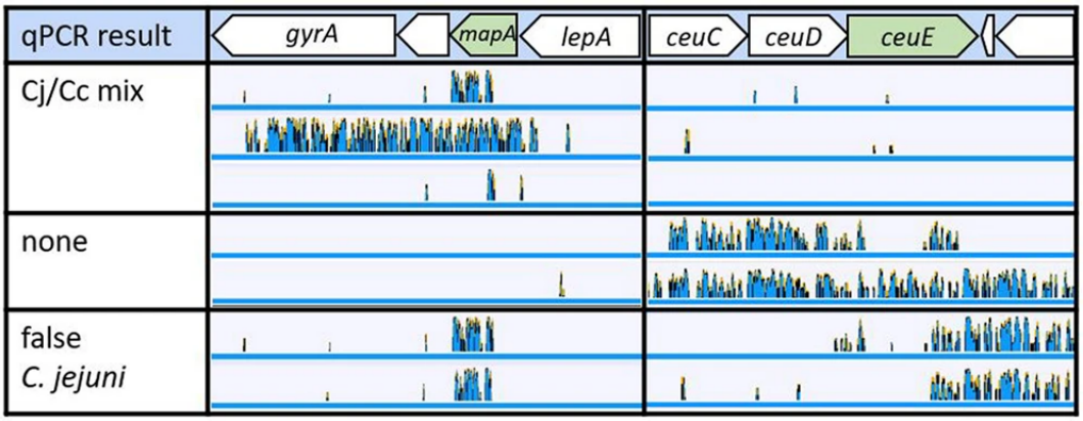


Figure 6.6: Examples of ambiguous qPCR results likely caused by recombination events at the *map*A and *ceu*E loci. *K-mer* coverage from recombinant *map*A and *cue*E regions sketch different variations within the qPCR results in reference genome NCTC11168. Isolates that lead to a *C. jejuni*/*C. coli* (Cj/Cc) mixed qPCR result showed various integration of *C. jejuni* sequences at the *map*A locus; Isolates lacking a qPCR result (none) showed integration of *C. jejuni* DNA in the *ceu*E locus and adjective genes. False positive qPCR results (*C. jejuni* predicted being *C. coli*) showed integration of *C. jejuni* sequences in *map*A as well as *ceu*E. This image was created by Julia Golz and Kerstin Stingl and published in [39] during this work.

6.7 Discussion

Recombination events between closely related bacterial species limit the resolution of accredited PCR-based methods for diagnostics purposes [245]. In this chapter, the capability and usability of a novel *k-mer*-based workflow analyzing WGS data to support diagnostic investigations were proposed. Hot spots of recombination in *C .coli* genomes, leading to ambiguous qPCR results, were detected and identified. During routine qPCR diagnostics for *Campylobacter* species determination performed at the BfR, samples from eggshells but also from poultry meat and turkey cecum show an ambiguous species differentiation in several cases. While analyzing WGS data from additional isolates taken from egg surfaces, nearly half of all *C. coli* strains displayed a non-random pattern of recombinant elements from *C. jejuni* within their genomes.

A previous study from Sheppard et al. compared agricultural-adapted *C. coli* genomes

from CC-828 and CC-1159 with non-agricultural-associated *C. coli* genomes with respect to their ability to recombine with strains from *C. jejuni* [95]. With their research they discovered 26 genes from *C. jejuni* to be present in agricultural-associated genomes which are absent in others.

In this work, several genomes of *C. coli* with a high amount of genetic elements from *C. jejuni* have been analyzed by a novel *k-mer*-based workflow. Within these genomes only 2 genes (Cj0555 and *htr*B) overlap with the 26 genes from the study of Sheppard et al.. Genomes under consideration have been shown to be non-clonal which indicates a novel and extensive development of strains within agricultural-associated *C. coli* strains.

Strains of the *C. coli* hybrids were predominantly isolated from eggshells, which is usually an uninhabitable environment for *Campylobacter*. Several studies show that *Campylobacter* from feces is not longer cultivable after 5-6 days [246, 247]. Consequently, the identified recombinant sites from *C. jejuni* might incorporate genes of functional adaptation to survival in a harsh environment. Usually, *Campylobacter* is transmitted to eggshells via fecal contamination. On the eggshell the bacterium encounters oxidative stress but also dryness and, thus, osmotic stress as well as nutrient and cold stress. This hypothesis is supported by non-random occurrences of recombination sites across the genomes of the strains analyzed here.

Adaptation to a novel environment might explain shared *C. jejuni* recombinations in *C. coli* hybrids

Functional annotation of loci that show frequent recombination in most of the *C. coli* hybrid genomes, revealed several genes included in the oxidative stress response of *Campylobacter*. These genes have either been completely exchanged with allelic variants from *C. jejuni* or formed novel *C. coli*/*C. jejuni* mosaic alleles. Genes such as katalase (katA) and Cj1386 (encoding an atypical hemin-binding protein) mediate the trafficking of hemin to katalase [248]. Katalase is one of the key protective enzymes with respect to oxidative stress due to its peroxide cleaving of water and oxygen. Another gene, *mrs*B (Cj1112c) encodes a methionine sulfoxide reductase, which protects *C. jejuni* against oxidative and nitrosative stress [249]. Additionally, genes that control the growth of *C. jejuni* at low CO_2 concentrations such as *can*B (carbonic anhydrase)

were introgressed into the hybrid strains [250]. A further oxidoreductase (Cj0833c) and genes encoding for the Ni/Fe hydrogenase small subunit *hyd*A (Cj1267c) and *hyd*A2 (Cj1399c) as well as *nad*D (Cj1404) involved in the synthesis of the redox cofactor NAD + are harbored by the *C. jejuni* sequences. Furthermore, *nuo*A, *nuo*B, *nuo*C implicated the transfer of electrons through the respiration chain and Cj0081, encoding the cyanide-resistant CioAB, which is proposed to lower oxygen levels and maintain microaerobic conditions [251], were identified to be introgressed from *C. jejuni* sequences in the hybrid strains. This might reflect adaptation towards the high oxygen tension *Campylobacter* strains have to face on the eggshells.

Besides oxidative stress, *Campylobacter* hybrid strains also have to adapt towards a dry and cooler environment with different Ph-levels in comparison to the chicken gut. As a result, genes involved in the response to harsh environmental conditions were identified. For instance, *htr*A, encoding a protease and chaperon activity with roles in virulence and oxidative stress defense [252–254] and *htr*B, encoding a lipid A acyltransferase involved in acid, heat, oxidative and osmotic stress response [255], were identified in all hybrid strains. Furthermore, a single-domain haemoglobin, encoded by the *cgb* locus was frequently discovered and associated as a response factor against nitric oxide and nitrosative stress in *Campylobacter* in a previous study [256]. Genes such as *clp*A ATPase and the chaperone genes *cpn*10 and *cpn*60 were also affected by introgression of *C. jejuni*. *Cpn*60, also known as *gro*EL, is used as a target for species differentiation by an additional qPCR assay [257] and thus would also lead to an incorrect species determination in the case of hybrid strains. Among the genes with *C. jejuni* introgression in the hybrid strains, were several associated with roles in DNA metabolism and repair, such as *pur*F, *pyr*G, *thy*X, *rar*A, *rec*J, *ung*, *rib*A, *gua*B and *dut*. Moreover, motility-associated genes, like the chemotaxis genes *che*A, *che*V, *che*W and *fli*Y, encoding a flagellar motor switch protein, represented *C. jejuni* sequences. Consequently, the non-random recombination within the gene discussed above in most of the hybrid strains, might reflect selection of survivors from harsh environments like the eggshell. Future studies should monitor if eggshells will be used by parts of the *Campylobacter* population as novel niches or if this only appears occasionally.

Practical implications for diagnostics

Many routine diagnostic and monitoring programs for *Campylobacter* spp. are based on qPCR or MLST approaches [258, 259]. A previous study reported that qPCRs based on the *map*A and *ceu*E genes for species differentiation in *Campylobacter* can lead to ambiguous results [244]. This observation was confirmed in this work by using an in-depth recombination analysis of the genomes of 37 strains (21 hybrid strains and 16 half hybrid strains) that yielded either ambiguous or false qPCR results. Of note, further qPCR tests using the gene *cpn*60 and *cad*F as targets, also failed to verify the particular species identity of all hybrid and one-half hybrid strains [257, 260, 261]. Even the MLST scheme provided only limited insights with respect to the species verification of most of the hybrid strains, since sequences of *C. jejuni* origin were detected for 6 out of the 7 housekeeping genes. A similar observation was recently reported for two *C. coli* strains isolated from turkey [262]. Hence, standardized typing methods should consider perturbations due to extended recombination activity in *Campylobacter*. Consequently, reliability of different species verification techniques should be evaluated on a regular basis, including existing norms such as ISO 10272-1/2:2017.

7 Outlook and Conclusion

Several studies already tackled questions arisen from the survival of the microaerobic species *C. jejuni* and *C. coli* exposed to different environments and host conditions. Stress tolerance, for instance, is one of the major and still enigmatic topics discussed with respect to explain the pathogen's widespread dissemination [263]. This topic highlights the need for more insights on adaptation mechanisms of zoonotic pathogens in general, since understanding is the initial and important step to develop targeted programs to prevent and combat infectious diseases. At present, most phenotype-based and molecular biological typing methods such as gene-targeting PCRs, restriction length polymorphism analysis or even fluorescence spectroscopy commonly used to identify bacterial pathogens generally lack scalability and are therefore prone to errors when recombination affects the target loci. To enhance our understanding on niche adaptation abilities of zoonotic pathogens, scalable methods provide useful additional information, including "early warning" potential during the emergence of novel high-risk variants.

In this work, novel *k-mer*-based computational methods for high-throughput analyses have been developed, improved and applied in order to study the adaptive potential of two important *Campylobacter* species using whole-genome data of two distinct cohorts. Using dataset 1, traits likely involved in niche adaption of the multi-host pathogen *C. jejuni* have been investigated. Different lineages of *C. jejuni* are either specialized for specific hosts or able to rapidly switch between hosts, known as host generalists. The main objective of the first study was to identify genomic signatures within *C. jejuni* that support or drive adaptation of this species towards particular host niches. The second study focused on ambiguous *Campylobacter* genomes that could not be assigned to a *Campylobacter* species by standard PCR protocols due to high amount of DNA introgression from *C. jejuni* to *C. coli*. The main goal of this study was to develop a *k-mer*-based workflow to identify these highly recombinant regions in WGS data.

In the following, a brief outlook is given on the future perspectives of these methods as well as practical applications for the results of this work.

Implications and perspectives of microbial GWAS

In recent years, GWAS was more and more transferred and adapted from human towards its usage in microbiological genetics. Since then, novel analysis methods based on GWAS have been applied to study direct relationships between genomic content and phenotypic traits such as pathogenicity or AMR in bacteria [136–138]. Here, host-specific genomic signatures were successfully identified by combining microbial GWAS with a bootstrapping approach. The results shown in this work emphasize the capability of GWAS approaches to analyze complex and multifactorial traits. However, the improvements and novel bioinformatic tools associated with microbial GWAS are not established for all bacterial species yet since individual properties of a population need to be taken into account during the development. It has been shown that GWAS works so far only poorly for bacterial species with highly clonal population structures such as *E. coli* or *M. tuberculosis* [134, 146]. Highly clonal population structures lead to the prediction of many non-causal variants due to the genome-wide linkage. In this work, the genomic variants detected were evaluated with respect to their putative biological impact by comparative analysis of the respective amino acid sequences. This seems a practical and easy solution to exclude non-causal variants. However, this approach requires intensive inspection of the amino acid sequences accompanied by literature research on loci of interest.

In addition to the problem of microbial GWAS described above, the false positive rate of GWAS also increases if highly imbalanced groups are compared with each other. However, in biological research it is not always possible to balance the proportion per group by the study design. In this work, highly imbalanced groups of different lifestyle preference occurred due to uneven natural appearances of various *C. jejuni* lineages in nature. In order to address this problem, a consensus GWAS was conceptualized and implemented utilizing a bootstrapping approach. As a result, mostly highly significant GWAS hits were identified and false positives were reduced. The extensions and improvements of microbial GWAS developed in the scope of this work may also help to further extend studying bacterial-host interactions in future research.

Besides technical issues, nearly all GWAS studying pathogenic bacteria were focused on genetic variations within the species itself, ignoring the probability of a combination

of genomic factors between the bacterial species and the host genome that may lead to a certain disease or colonization. A recent approach from Lees et. al 2019 introduced a double GWAS applied on genomic data from *Streptococcus pneumoniae* and the corresponding patient genomes and provided a proof of concept for such applications in the future [264]. They showed that the susceptibility to meningitis does not depend on genetic variations within *S. pneumoniae* alone, but also depends to around 29% on genetic variants within the human patients genomes. Applying a combined human-bacteria GWAS might also be relevant to study the invasive potential of *Campylobacter* or level of severeness of Campylobacteriosis of clinical relvant *Campylobacter* strains in future research. However, such complex studies require a large and well curated dataset including bacterial samples, human genome data and information about the course of the patient's disease, which can be more time consuming and costly to obtain.

Source attribution and outbreak detection of *C. jejuni*

Allelic variants of the core and the accessory genome of *C. jejuni* lineages which were identified here indicate evolutionary adaptation towards a specific host niche or even a "Jack-of-all-trades" bacterium represented by host-generalist lineages. Several genes were identified that might play important roles in transcriptional regulation, temperature adaptation or metabolic pathways. Distinct host-specific factors, including body temperature, structure and composition of the gut microbiota, the host-specific mucosal structures and the immune system have an impact on the adaptation of *C. jejuni* lineages. However, understanding of these processes leading to a certain specialization or even to an "all-rounder" with respect to niche adaptation of *C. jejuni* is still in its infancy and functional studies are clearly needed to reveal the biological impact of the results presented here. Additionally, in the chicken-related as well as the cattle-related genomes, many genes of unknown function were identified. These genes might be of significant importance as well and should be characterized throughout future experiments.

Besides academic interests to gain in-depth knowledge regarding evolutionary and adaptation processes of zoonotic pathogens, the host-specific factors identified here might also be of high relevance for practical applications such as source attribution of outbreak situations. Recently, machine learning based on WGS data of *C. jejuni* has been utilized for such purposes [265]. Host-specific genomic signatures might increase the accuracy

of source attribution even further. Additionally, incorporating markers identified here in zoonoses monitoring programs might help to track trends towards novel host niches or changes of the pathogenic potential within the *Campylobacter* population.

Detection of recombination events for microbial diagnostics

Adaptation of bacterial species or lineages towards novel environments frequently occur in nature. As shown by Baym et. al with their famous Microbial Evolution and Growth Arena (MEGA) plate experiment [266], bacterial adaption towards harsh or even hostile environments is indeed possible: They documented the adaption of *E. coli* towards exponentially increasing concentrations of antibiotic substances. Here, several *C. coli* isolates were identified on the dry surface of eggshells representing a naturally hostile environment for *Campylobacter*. The identification of hybrid strains mainly selected from a harsh environment exhibiting an extended amount of *C. jejuni* sequences in a common gene set shows the enormous potential for the *Campylobacter* of extensive genetic exchange in order to improve fitness.

The provided genetic information from this study about common gene sets among adapted *C. coli* genomes might be highly valuable for improving existing laborious workflows. Well established diagnostic methods lack resolution in order to detect these evolutionary aspects. Novel *k-mer*-based detection methods of recombination events enhanced the identification of this coherence with further practical implications for diagnostic purposes. Here, a proof of principle of novel computational approaches for the enhancement of effective monitoring of evolutionary trends in food-chain surveillance programs was shown. Detection of hot spots of recombination not only plays an important role in *Campylobacter*, but also in many other bacterial species [267]. Thus, the novel *k-mer*-based workflow developed in this work can be applied in future research projects analyzing extensive recombination sites.

A Appendix

A.1 Dataset 1: NCBI accessory numbers and metadata

Overview of sequenced genomes of the PAC-Campy dataset with additional meta data information. The table was created and published in [154] during this work.

SRR	ID	ST	CC	Origin	Association	BAPS	Source
SRR12302213	12-02934	354	ST-354	Human	Chicken	1	Germany: Federal State N
SRR12302212	12-02938	50	ST-21	Human	Generalist	2	Germany: Federal State H
SRR12302046	12-02939	10269	ST-353	Human	Chicken	1	Germany: Federal State H
SRR12302257	12-03052	48	ST-48	Human	Generalist	3	Germany: Federal State N
SRR12302223	12-03072	42	ST-42	Human	Cattle	4	Germany: Federal State J
SRR12302164	12-03499	2314	ST-1034	Human	Chicken	5	Germany: Federal State H
SRR12302153	12-03501	50	ST-21	Human	Generalist	2	Germany: Federal State G
SRR12302142	12-03536	429	ST-48	Human	Generalist	3	Germany: Federal State G
SRR12302107	12-03862	21	ST-21	Human	Generalist	6	Germany: Federal State H
SRR12302096	12-03906	19	ST-21	Human	Generalist	2	Germany: Federal State H
SRR12302211	12-03907	354	ST-354	Human	Chicken	1	Germany: Federal State H
SRR12302200	12-03908	19	ST-21	Human	Generalist	2	Germany: Federal State J
SRR12302189	12-03909	10270	ST-41	Human	unknown	4	Germany: Federal State J
SRR12302178	12-03950	464	ST-464	Human	unknown	1	Germany: Federal State N
SRR12302167	12-03951	677	ST-677	Human	unknown	7	Germany: Federal State H
SRR12302128	12-03952	21	ST-21	Human	Generalist	6	Germany: Federal State H
SRR12302117	12-03953	267	ST-283	Human	unknown	8	Germany: Federal State H
SRR12302079	12-03954	50	ST-21	Human	Generalist	2	Germany: Federal State H
SRR12302068	12-03967	677	ST-677	Human	unknown	7	Germany: Federal State P
SRR12302057	13-00141	53	ST-21	Human	Generalist	6	Germany: Federal State N
SRR12302045	13-00142	354	ST-354	Human	Chicken	1	Germany: Federal State P
SRR12302033	13-00156	21	ST-21	Human	Generalist	6	Germany: Federal State J
SRR12302022	13-00216	918	ST-48	Human	Generalist	3	Germany
SRR12302011	13-00388	1326	ST-45	Human	Generalist	8	Germany: Federal State H
SRR12302000	13-00578	572	ST-206	Human	Generalist	3	Germany: Federal State H
SRR12301989	13-00579	53	ST-21	Human	Generalist	6	Germany: Federal State J
SRR12301978	13-00893	21	ST-21	Human	Generalist	6	Germany: Federal State H
SRR12301967	13-01164	50	ST-21	Human	Generalist	2	Germany: Federal State P
SRR12302279	13-01279	21	ST-21	Human	Generalist	6	Germany: Federal State H
SRR12302268	13-01855	1519	ST-21	Human	Generalist	2	Germany: Federal State N
SRR12302256	13-01999	206	ST-206	Human	Generalist	3	Germany: Federal State H
SRR12302245	13-02263	50	ST-21	Human	Generalist	2	Germany
SRR12302234	13-02264	50	ST-21	Human	Generalist	2	Germany
SRR12302230	13-02292	354	ST-354	Human	Chicken	1	Germany: Federal State G
SRR12302229	13-02407	19	ST-21	Human	Generalist	2	Germany: Federal State H
SRR12302228	13-02841	45	ST-45	Human	Generalist	8	Germany
SRR12302227	13-02842	45	ST-45	Human	Generalist	8	Germany
SRR12302226	14-01211	46	ST-206	Human	Generalist	3	Germany: Federal State H
SRR12302225	14-01213	21	ST-21	Human	Generalist	6	Germany: Federal State H
SRR12302224	14-01252	990	ST-257	Human	Chicken	9	Germany: Federal State H
SRR12302222	14-01255	824	ST-257	Human	Chicken	1	Germany: Federal State H
SRR12302221	14-01256	46	ST-206	Human	Generalist	3	Germany: Federal State H
SRR12302220	14-01401	7231	unknown	Human	unknown	3	Germany: Federal State H
SRR12302219	14-01734	46	ST-206	Human	Generalist	3	Germany: Federal State B
SRR12302218	14-01866	50	ST-21	Human	Generalist	2	Germany: Federal State H
SRR12302217	14-02206	2036	ST-353	Human	Chicken	1	Germany: Federal State P
SRR12302216	14-02234	1519	ST-21	Human	Generalist	2	Germany: Federal State J
SRR12302215	14-02237	50	ST-21	Human	Generalist	2	Germany: Federal State H
SRR12302214	14-02514	354	ST-354	Human	Chicken	1	Germany: Federal State H

Continued on next page

Table A.1 Continued from previous page

SRR	ID	ST	CC	Origin	Association	BAPS	Source
SRR12302165	14-02648	50	ST-21	Human	Generalist	2	Germany: Federal State P
SRR12302163	14-02649	61	ST-61	Human	Cattle	10	Germany: Federal State H
SRR12302162	14-02926	19	ST-21	Human	Generalist	2	Germany: Federal State K
SRR12302161	14-02989	44	ST-21	Human	Generalist	2	Germany: Federal State K
SRR12302160	14-02992	45	ST-45	Human	Generalist	8	Germany: Federal State P
SRR12302159	14-03375	262	ST-21	Human	Generalist	6	Germany: Federal State P
SRR12302158	15-00051	50	ST-21	Human	Generalist	2	Germany
SRR12302157	15-00397	1044	ST-658	Human	unknown	1	Germany: Federal State J
SRR12302156	15-00432	10271	ST-464	Human	unknown	1	Germany: Federal State H
SRR12302155	15-00440	475	ST-48	Human	Generalist	3	Germany: Federal State H
SRR12302154	15-00441	475	ST-48	Human	Generalist	3	Germany: Federal State A
SRR12302152	15-01126	354	ST-354	Human	Chicken	1	Germany: Federal State N
SRR12302151	15-01144	1519	ST-21	Human	Generalist	2	Germany: Federal State H
SRR12302150	15-01231	824	ST-257	Human	Chicken	1	Germany: Federal State H
SRR12302149	15-01311	2036	ST-353	Human	Chicken	1	Germany: Federal State N
SRR12302148	15-01320	2274	unknown	Human	unknown	1	Germany: Federal State H
SRR12302147	15-01321	1519	ST-21	Human	Generalist	2	Germany: Federal State H
SRR12302146	15-01399	50	ST-21	Human	Generalist	2	Germany: Federal State G
SRR12302145	15-01400	1003	ST-45	Human	Generalist	8	Germany: Federal State H
SRR12302144	15-01539	61	ST-61	Human	Cattle	10	Germany
SRR12302143	15-01638	61	ST-61	Human	Cattle	10	Germany
SRR12302141	15-01641	61	ST-61	Human	Cattle	10	Germany: Federal State J
SRR12302140	15-01705	257	ST-257	Human	Chicken	9	Germany
SRR12302139	15-01779	137	ST-45	Human	Generalist	8	Germany
SRR12302138	15-01780	1519	ST-21	Human	Generalist	2	Germany
SRR12302113	15-00398	50	ST-21	Human	Generalist	2	Germany: Federal State P
SRR12302112	16-00049	46	ST-206	Human	Generalist	3	Germany: Federal State H
SRR12302111	16-00112	122	ST-206	Human	Generalist	3	Germany: Federal State H
SRR12302110	16-00241	48	ST-48	Human	Generalist	3	Germany: Federal State H
SRR12302109	16-00242	6461	ST-353	Human	Chicken	1	Germany: Federal State N
SRR12302108	16-00263	658	ST-658	Human	unknown	1	Germany: Federal State H
SRR12302106	16-00264	1519	ST-21	Human	Generalist	2	Germany: Federal State H
SRR12302105	16-00268	48	ST-48	Human	Generalist	3	Germany: Federal State J
SRR12302104	16-00290	6175	ST-21	Human	Generalist	3	Germany: Federal State N
SRR12302103	16-00409	2274	unknown	Human	unknown	1	Germany: Federal State P
SRR12302102	16-00431	22	ST-22	Human	unknown	4	Germany: Federal State G
SRR12302101	16-00576	1519	ST-21	Human	Generalist	2	Germany: Federal State H
SRR12302100	16-00632	50	ST-21	Human	Generalist	2	Germany: Federal State N
SRR12302099	16-00663	21	ST-21	Human	Generalist	6	Germany: Federal State H
SRR12302098	16-01339	1775	ST-403	Human	Pig	11	Germany: Federal State H
SRR12302097	16-01340	21	ST-21	Human	Generalist	6	Germany: Federal State H
SRR12302095	16-01341	50	ST-21	Human	Generalist	2	Germany: Federal State J
SRR12302094	19-00788	658	ST-177	Human	unknown	1	Germany
SRR12302093	19-00791	21	ST-21	Human	Generalist	6	Germany
SRR12302092	19-00792	22	ST-22	Human	unknown	4	Germany
SRR12302091	19-00811	61	ST-61	Human	Cattle	10	Germany
SRR12302090	19-00812	61	ST-61	Human	Cattle	10	Germany
SRR12302089	BfRCA02030	435	ST-403	Pig	Pig	11	Germany
SRR12302088	BfRCA05380	257	ST-257	Pig	Chicken	9	Germany
SRR12302087	BfRCA05532NCTC11168		ST-21	Chicken	Generalist	6	Germany
SRR12302038	BfRCA05915	257	ST-257	Pig	Chicken	9	Germany
SRR12302210	BfRCA05925	9340	ST-403	Pig	Pig	11	Germany
SRR12302209	BfRCA06044	21	ST-21	Pig	Generalist	6	Germany
SRR12302208	BfRCA06058	10272	ST-403	Pig	Pig	11	Germany
SRR12302207	BfRCA06059	435	ST-403	Pig	Pig	11	Germany
SRR12302206	BfRCA06089	435	ST-403	Pig	Pig	11	Germany

Continued on next page

Table A.1 Continued from previous page

SRR	ID	ST	CC	Origin	Association	BAPS	Source
SRR12302205	BfRCA06331	21	ST-21	Pig	Generalist	6	Germany
SRR12302204	BfRCA07435	1775	ST-403	Pig	Pig	11	Germany
SRR12302203	BfRCA07767	1777	ST-403	Pig	Pig	11	Germany
SRR12302202	BfRCA08326	1777	ST-403	Pig	Pig	11	Germany: Federal State L
SRR12302201	BfRCA08665	583	ST-45	Chicken	Generalist	8	Germany: Federal State J
SRR12302199	BfRCA10009	21	ST-21	Chicken	Generalist	6	Germany: Federal State B
SRR12302198	BfRCA10050	51	ST-443	Chicken	Chicken	1	Germany: Federal State K
SRR12302197	BfRCA10051	48	ST-48	Cattle	Generalist	3	Germany: Federal State H
SRR12302196	BfRCA10084	1073	ST-354	Chicken	Chicken	1	Germany: Federal State J
SRR12302195	BfRCA10129	53	ST-21	Chicken	Generalist	6	Germany: Federal State D
SRR12302194	BfRCA10142	6409	ST-1034	Chicken	Chicken	5	Germany: Federal State K
SRR12302193	BfRCA10153	21	ST-21	Chicken	Generalist	6	Germany: Federal State A
SRR12302192	BfRCA10170	257	ST-257	Chicken	Chicken	9	Germany: Federal State A
SRR12302191	BfRCA10173	48	ST-48	Chicken	Generalist	3	Germany: Federal State A
SRR12302190	BfRCA10176	4754	unknown	Chicken	Chicken	5	Germany: Federal State C
SRR12302188	BfRCA10204	61	ST-61	Cattle	Cattle	10	Germany: Federal State G
SRR12302187	BfRCA10221	48	ST-48	Chicken	Generalist	3	Germany: Federal State N
SRR12302186	BfRCA10231	429	ST-48	Chicken	Generalist	3	Germany: Federal State E
SRR12302185	BfRCA10238	51	ST-443	Chicken	Chicken	1	Germany: Federal State J
SRR12302184	BfRCA10257	354	ST-354	Chicken	Chicken	1	Germany: Federal State H
SRR12302183	BfRCA10272	3100	unknown	Cattle	unknown	3	Germany: Federal State H
SRR12302182	BfRCA10285	19	ST-21	Cattle	Generalist	2	Germany: Federal State H
SRR12302181	BfRCA10287	273	ST-206	Cattle	Generalist	3	Germany: Federal State H
SRR12302180	BfRCA10303	38	ST-48	Cattle	Generalist	3	Germany: Federal State O
SRR12302179	BfRCA10338	45	ST-45	Chicken	Generalist	8	Germany: Federal State A
SRR12302177	BfRCA10393	2274	unknown	Chicken	unknown	1	Germany: Federal State D
SRR12302176	BfRCA10394	46	ST-206	Chicken	Generalist	3	Germany: Federal State M
SRR12302175	BfRCA10443	21	ST-21	Cattle	Generalist	6	Germany: Federal State O
SRR12302174	BfRCA10483	38	ST-48	Cattle	Generalist	3	Germany: Federal State M
SRR12302173	BfRCA10491	658	ST-658	Chicken	unknown	1	Germany: Federal State M
SRR12302172	BfRCA10492	50	ST-21	Chicken	Generalist	2	Germany: Federal State J
SRR12302171	BfRCA10572	2314	ST-1034	Chicken	Chicken	5	Germany: Federal State D
SRR12302170	BfRCA10615	19	ST-21	Cattle	Generalist	2	Germany: Federal State J
SRR12302169	BfRCA10644	1911	unknown	Chicken	unknown	1	Germany: Federal State J
SRR12302168	BfRCA10649	45	ST-45	Chicken	Generalist	8	Germany: Federal State J
SRR12302166	BfRCA10651	7958	unknown	Chicken	unknown	5	Germany: Federal State C
SRR12302137	BfRCA10667	21	ST-21	Cattle	Generalist	6	Germany: Federal State H
SRR12302136	BfRCA10675	3100	unknown	Cattle	unknown	3	Germany: Federal State H
SRR12302135	BfRCA10677	7041	ST-21	Cattle	Generalist	3	Germany: Federal State H
SRR12302134	BfRCA10717	257	ST-257	Cattle	Chicken	9	Germany: Federal State H
SRR12302133	BfRCA10738	10273	ST-61	Cattle	Cattle	10	Germany: Federal State H
SRR12302132	BfRCA10753	21	ST-21	Cattle	Generalist	6	Germany: Federal State H
SRR12302131	BfRCA10776b	403	ST-403	Cattle	Pig	11	Germany: Federal State H
SRR12302130	BfRCA10820	51	ST-443	Cattle	Chicken	1	Germany: Federal State A
SRR12302129	BfRCA10827	21	ST-21	Cattle	Generalist	6	Germany: Federal State M
SRR12302127	BfRCA10834	6728	ST-353	Chicken	Chicken	1	Germany: Federal State D
SRR12302126	BfRCA10850	233	ST-45	Chicken	Generalist	8	Germany: Federal State J
SRR12302125	BfRCA10877	48	ST-48	Cattle	Generalist	3	Germany: Federal State M
SRR12302124	BfRCA10905	48	ST-48	Cattle	Generalist	3	Germany: Federal State A
SRR12302123	BfRCA10926	21	ST-21	Cattle	Generalist	6	Germany: Federal State A
SRR12302122	BfRCA10942	206	ST-206	Cattle	Generalist	3	Germany: Federal State J
SRR12302121	BfRCA10958	5970	unknown	Chicken	unknown	5	Germany: Federal State H
SRR12302120	BfRCA10959	19	ST-21	Cattle	Generalist	2	Germany: Federal State O
SRR12302119	BfRCA10961	233	ST-45	Chicken	Generalist	8	Germany: Federal State D
SRR12302118	BfRCA10962	50	ST-21	Chicken	Generalist	2	Germany: Federal State D
SRR12302116	BfRCA10963	61	ST-61	Cattle	Cattle	10	Germany: Federal State A

Continued on next page

Table A.1 Continued from previous page

SRR	ID	ST	CC	Origin	Association	BAPS	Source
SRR12302115	BfRCA10977	2156	unknown	Cattle	unknown	3	Germany: Federal State B
SRR12302114	BfRCA10984	21	ST-21	Cattle	Generalist	6	Germany: Federal State O
SRR12302086	BfRCA10988	50	ST-21	Chicken	Generalist	2	Germany: Federal State P
SRR12302085	BfRCA11042	61	ST-61	Cattle	Cattle	10	Germany: Federal State P
SRR12302084	BfRCA11044	2274	unknown	Chicken	unknown	1	Germany: Federal State J
SRR12302083	BfRCA11054	2274	unknown	Chicken	unknown	1	Germany: Federal State J
SRR12302082	BfRCA11065	3098	unknown	Cattle	unknown	12	Germany: Federal State D
SRR12302081	BfRCA11066	10275	ST-21	Cattle	Generalist	6	Germany: Federal State H
SRR12302080	BfRCA11083	257	ST-257	Cattle	Chicken	9	Germany: Federal State H
SRR12302078	BfRCA11096	21	ST-21	Cattle	Generalist	6	Germany: Federal State H
SRR12302077	BfRCA11177	21	ST-21	Cattle	Generalist	6	Germany: Federal State A
SRR12302076	BfRCA11178	21	ST-21	Cattle	Generalist	6	Germany: Federal State A
SRR12302075	BfRCA11192	10276	unknown	Cattle	unknown	12	Germany: Federal State J
SRR12302074	BfRCA11199	2274	unknown	Chicken	unknown	1	Germany: Federal State B
SRR12302073	BfRCA11202	21	ST-21	Chicken	Generalist	6	Germany: Federal State O
SRR12302072	BfRCA11209	42	ST-42	Cattle	Cattle	4	Germany: Federal State A
SRR12302071	BfRCA11214	61	ST-61	Cattle	Cattle	10	Germany: Federal State J
SRR12302070	BfRCA11215	19	ST-21	Cattle	Generalist	2	Germany: Federal State J
SRR12302069	BfRCA11219	48	ST-48	Chicken	Generalist	3	Germany: Federal State B
SRR12302067	BfRCA11234	55	ST-403	Cattle	Pig	11	Germany: Federal State M
SRR12302066	BfRCA11258	10277	unknown	Chicken	unknown	1	Germany: Federal State A
SRR12302065	BfRCA11315	21	ST-21	Cattle	Generalist	6	Germany: Federal State J
SRR12302064	BfRCA11319	61	ST-61	Cattle	Cattle	10	Germany: Federal State J
SRR12302063	BfRCA11330	45	ST-45	Cattle	Generalist	8	Germany: Federal State J
SRR12302062	BfRCA11331	21	ST-21	Cattle	Generalist	6	Germany: Federal State J
SRR12302061	BfRCA11344	61	ST-61	Cattle	Cattle	10	Germany: Federal State H
SRR12302060	BfRCA11346	61	ST-61	Cattle	Cattle	10	Germany: Federal State H
SRR12302059	BfRCA11347	21	ST-21	Cattle	Generalist	6	Germany: Federal State H
SRR12302058	BfRCA11352	38	ST-48	Cattle	Generalist	3	Germany: Federal State H
SRR12302056	BfRCA11375	50	ST-21	Cattle	Generalist	2	Germany: Federal State B
SRR12302055	BfRCA11386	7515	ST-42	Cattle	Cattle	4	Germany: Federal State A
SRR12302054	BfRCA11387	10278	ST-42	Cattle	Cattle	4	Germany: Federal State A
SRR12302053	BfRCA11388	61	ST-61	Cattle	Cattle	10	Germany: Federal State A
SRR12302052	BfRCA11390	50	ST-21	Chicken	Generalist	2	Germany: Federal State A
SRR12302051	BfRCA11392	21	ST-21	Chicken	Generalist	6	Germany: Federal State A
SRR12302050	BfRCA11421	45	ST-45	Chicken	Generalist	8	Germany: Federal State B
SRR12302049	BfRCA11438	904	ST-607	Chicken	unknown	1	Germany: Federal State G
SRR12302048	BfRCA11498	21	ST-21	Cattle	Generalist	6	Germany: Federal State J
SRR12302047	BfRCA11566	2153	unknown	Chicken	unknown	1	Germany: Federal State J
SRR12302044	BfRCA11567	3766	unknown	Chicken	unknown	5	Germany: Federal State J
SRR12302043	BfRCA11573	464	ST-464	Cattle	unknown	1	Germany: Federal State J
SRR12302042	BfRCA11581	45	ST-45	Cattle	Generalist	8	Germany: Federal State D
SRR12302041	BfRCA11590	583	ST-45	Cattle	Generalist	8	Germany: Federal State K
SRR12302040	BfRCA11610	933	ST-403	Cattle	Pig	11	Germany: Federal State P
SRR12302039	BfRCA11627	21	ST-21	Cattle	Generalist	6	Germany: Federal State B
SRR12302037	BfRCA11629	6461	ST-353	Chicken	Chicken	1	Germany: Federal State B
SRR12302036	BfRCA11633	19	ST-21	Chicken	Generalist	2	Germany: Federal State K
SRR12302035	BfRCA11654	50	ST-21	Cattle	Generalist	2	Germany: Federal State B
SRR12302034	BfRCA11663	586	unknown	Cattle	unknown	4	Germany: Federal State B
SRR12302032	BfRCA11664	10280	unknown	Chicken	unknown	5	Germany: Federal State K
SRR12302031	BfRCA11665	21	ST-21	Cattle	Generalist	6	Germany: Federal State K
SRR12302030	BfRCA11667	21	ST-21	Cattle	Generalist	6	Germany: Federal State K
SRR12302029	BfRCA11700	1301	ST-692	Chicken	unknown	5	Germany: Federal State P
SRR12302028	BfRCA11706	10281	ST-21	Cattle	Generalist	6	Germany: Federal State J
SRR12302027	BfRCA11713	10282	unknown	Cattle	unknown	12	Germany: Federal State J
SRR12302026	BfRCA11722	21	ST-21	Cattle	Generalist	6	Germany: Federal State A

Continued on next page

Table A.1 Continued from previous page

SRR	ID	ST	CC	Origin	Association	BAPS	Source
SRR12302025	BfRCA11723	19	ST-21	Pig	Generalist	2	Germany: Federal State O
SRR12302024	BfRCA11724	61	ST-61	Cattle	Cattle	10	Germany: Federal State O
SRR12302023	BfRCA11725	432	ST-61	Cattle	Cattle	10	Germany: Federal State O
SRR12302021	BfRCA11846	44	ST-21	Chicken	Generalist	2	Germany: Federal State I
SRR12302020	BfRCA11848b	51	ST-443	Cattle	Chicken	1	Germany: Federal State A
SRR12302019	BfRCA11849	356	ST-353	Chicken	Chicken	1	Germany: Federal State A
SRR12302018	BfRCA11850	19	ST-21	Cattle	Generalist	2	Germany: Federal State A
SRR12302017	BfRCA11851	2274	unknown	Chicken	unknown	1	Germany: Federal State A
SRR12302016	BfRCA11852	21	ST-21	Cattle	Generalist	6	Germany: Federal State H
SRR12302015	BfRCA11853	50	ST-21	Cattle	Generalist	2	Germany: Federal State A
SRR12302014	BfRCA11884	583	ST-45	Chicken	Generalist	8	Germany: Federal State J
SRR12302013	BfRCA11888	2066	ST-52	Chicken	unknown	1	Germany: Federal State B
SRR12302012	BfRCA11917	42	ST-42	Chicken	Cattle	4	Germany: Federal State J
SRR12302010	BfRCA11926	257	ST-257	Chicken	Chicken	9	Germany: Federal State J
SRR12302009	BfRCA11946	50	ST-21	Chicken	Generalist	2	Germany: Federal State A
SRR12302008	BfRCA12057	21	ST-21	Chicken	Generalist	6	Germany: Federal State B
SRR12302007	BfRCA12154	5798	unknown	Chicken	unknown	5	Germany: Federal State B
SRR12302006	BfRCA12659	21	ST-21	Chicken	Generalist	6	Germany: Federal State B
SRR12302005	BfRCA12662	5840	ST-353	Chicken	Chicken	1	Germany: Federal State B
SRR12302004	BfRCA12663	50	ST-21	Chicken	Generalist	2	Germany: Federal State B
SRR12302003	BfRCA12891	464	ST-464	Chicken	unknown	1	Germany: Federal State B
SRR12302002	BfRCA12978	2254	ST-257	Chicken	Chicken	9	Germany: Federal State N
SRR12302001	BfRCA13157	21	ST-21	Chicken	Generalist	6	Germany: Federal State B
SRR12301999	BfRCA13162	6175	ST-21	Chicken	Generalist	3	Germany: Federal State B
SRR12301998	BfRCA13163	46	ST-206	Chicken	Generalist	3	Germany: Federal State A
SRR12301997	BfRCA13168	3628	ST-443	Chicken	Chicken	1	Germany: Federal State A
SRR12301996	BfRCA13169	10283	ST-443	Chicken	Chicken	1	Germany: Federal State A
SRR12301995	BfRCA13171	1519	ST-21	Chicken	Generalist	2	Germany: Federal State J
SRR12301994	BfRCA13189	21	ST-21	Cattle	Generalist	6	Germany: Federal State M
SRR12301993	BfRCA13199	861	ST-21	Cattle	Generalist	6	Germany: Federal State O
SRR12301992	BfRCA13206	5103	ST-22	Cattle	unknown	4	Germany: Federal State A
SRR12301991	BfRCA13207	42	ST-42	Cattle	Cattle	4	Germany: Federal State A
SRR12301990	BfRCA13233b	403	ST-403	Pig	Pig	11	Germany
SRR12301988	BfRCA13265	354	ST-354	Pig	Chicken	1	Germany
SRR12301987	BfRCA13281	403	ST-403	Pig	Pig	11	Germany: Federal State J
SRR12301986	BfRCA13282	403	ST-403	Pig	Pig	11	Germany: Federal State J
SRR12301985	BfRCA13292	22	ST-22	Cattle	unknown	4	Germany: Federal State J
SRR12301984	BfRCA13298	22	ST-22	Cattle	unknown	4	Germany: Federal State J
SRR12301983	BfRCA13324	10284	unknown	Cattle	unknown	3	Germany: Federal State J
SRR12301982	BfRCA13330	441	unknown	Cattle	unknown	13	Germany: Federal State J
SRR12301981	BfRCA13394	61	ST-61	Cattle	Cattle	10	Germany: Federal State A
SRR12301980	BfRCA13398	2274	unknown	Chicken	unknown	1	Germany: Federal State H
SRR12301979	BfRCA13453	38	ST-48	Cattle	Generalist	3	Germany: Federal State J
SRR12301977	BfRCA13463	50	ST-21	Chicken	Generalist	2	Germany: Federal State J
SRR12301976	BfRCA13512	21	ST-21	Cattle	Generalist	6	Germany: Federal State J
SRR12301975	BfRCA13514	21	ST-21	Cattle	Generalist	6	Germany: Federal State J
SRR12301974	BfRCA13527	1709	ST-1034	Chicken	Chicken	5	Germany: Federal State N
SRR12301973	BfRCA13538	10285	ST-61	Cattle	unknown	10	Germany: Federal State I
SRR12301972	BfRCA13539	206	ST-206	Cattle	Generalist	3	Germany: Federal State I
SRR12301971	BfRCA13541	21	ST-21	Cattle	Generalist	6	Germany: Federal State I
SRR12301970	BfRCA13564	61	ST-61	Cattle	Cattle	10	Germany: Federal State J
SRR12301969	BfRCA13582	38	ST-48	Cattle	Generalist	3	Germany: Federal State I
SRR12301968	BfRCA13613	42	ST-42	Cattle	Cattle	4	Germany: Federal State J
SRR12301966	BfRCA13729	3155	ST-354	Pig	Chicken	1	Germany
SRR12301965	BfRCA13756	403	ST-403	Cattle	Pig	11	Germany: Federal State J
SRR12301964	BfRCA13758	21	ST-21	Cattle	Generalist	6	Germany: Federal State J

Continued on next page

Table A.1 Continued from previous page

SRR	ID	ST	CC	Origin	Association	BAPS	Source
SRR12301963	BfRCA13795	19	ST-21	Chicken	Generalist	2	Germany: Federal State H
SRR12301962	BfRCA13811	1459	ST-21	Cattle	Generalist	6	Germany: Federal State J
SRR12301961	BfRCA13816	257	ST-257	Cattle	Chicken	9	Germany: Federal State J
SRR12301960	BfRCA13821	42	ST-42	Cattle	Cattle	4	Germany: Federal State J
SRR12301959	BfRCA13824	10286	ST-61	Cattle	unknown	10	Germany: Federal State J
SRR12302281	BfRCA13826	19	ST-21	Cattle	Generalist	2	Germany: Federal State J
SRR12302280	BfRCA13835	354	ST-354	Cattle	Chicken	1	Germany: Federal State J
SRR12302278	BfRCA13836	21	ST-21	Cattle	Generalist	6	Germany: Federal State J
SRR12302277	BfRCA13838	21	ST-21	Cattle	Generalist	6	Germany: Federal State J
SRR12302276	BfRCA13918	1003	ST-45	Chicken	Generalist	8	Germany: Federal State B
SRR12302275	BfRCA13937	1519	ST-21	Chicken	Generalist	2	Germany: Federal State K
SRR12302274	BfRCA13939	607	ST-607	Chicken	unknown	1	Germany: Federal State A
SRR12302273	BfRCA14088	48	ST-48	Chicken	Generalist	3	Germany: Federal State B
SRR12302272	BfRCA14109	45	ST-45	Chicken	Generalist	8	Germany: Federal State M
SRR12302271	BfRCA14180	354	ST-354	Chicken	Chicken	1	Germany: Federal State B
SRR12302270	BfRCA14181	2066	ST-52	Chicken	unknown	1	Germany: Federal State B
SRR12302269	BfRCA14304	2304	unknown	Chicken	unknown	3	Germany: Federal State I
SRR12302267	BfRCA14323	1519	ST-21	Chicken	Generalist	2	Germany: Federal State B
SRR10103069, SRR10103068	BfRCA14430	44	ST-21	Chicken	Generalist	2	Germany
SRR12302266	BfRCA14435	1519	ST-21	Chicken	Generalist	2	Germany: Federal State B
SRR12302265	BfRCA14444	2274	unknown	Chicken	unknown	1	Germany: Federal State J
SRR12302264	BfRCA14553	44	ST-21	Chicken	Generalist	2	Germany: Federal State P
SRR12302263	BfRCA14579	977	ST-1034	Chicken	Chicken	5	Germany: Federal State J
SRR12302262	BfRCA14704	538	ST-45	Chicken	Generalist	8	Germany: Federal State P
SRR12302261	BfRCA14734	977	ST-1034	Chicken	Chicken	5	Germany: Federal State J
SRR12302260	BfRCA14814	2275	ST-52	Chicken	unknown	1	Germany: Federal State H
SRR12302259	BfRCA14836	464	ST-464	Chicken	unknown	1	Germany: Federal State J
SRR12302258	BfRCA14940	267	ST-283	Chicken	unknown	8	Germany: Federal State F
SRR12302255	BfRCA14957	50	ST-21	Chicken	Generalist	2	Germany: Federal State O
SRR12302254	BfRCA14962	2275	ST-52	Chicken	unknown	1	Germany: Federal State J
SRR12302253	BfRCA14988	8334	ST-353	Chicken	Chicken	1	Germany: Federal State B
SRR12302252	BfRCA14993	2275	ST-52	Chicken	unknown	1	Germany: Federal State J
SRR12302251	BfRCA15004	2254	ST-257	Chicken	Chicken	9	Germany: Federal State P
SRR12302250	BfRCA15023b	1846	ST-403	Pig	Pig	11	Germany
SRR12302249	BfRCA15024	1846	ST-403	Pig	Pig	11	Germany
SRR12302248	BfRCA15054	1846	ST-403	Pig	Pig	11	Germany
SRR12302247	BfRCA15085	9351	unknown	Chicken	unknown	5	Germany: Federal State P
SRR12302246	BfRCA15095	3335	ST-206	Pig	Generalist	3	Germany
SRR12302244	BfRCA15119	45	ST-45	Chicken	Generalist	8	Germany: Federal State H
SRR12302243	BfRCA15155	400	ST-353	Chicken	Chicken	1	Germany: Federal State M
SRR12302242	BfRCA15166	9366	ST-403	Pig	Pig	11	Germany: Federal State E
SRR12302241	BfRCA15240	257	ST-257	Chicken	Chicken	9	Germany: Federal State E
SRR12302240	BfRCA15256	1942	unknown	Pig	unknown	11	Germany
SRR12302239	BfRCA15265	3628	ST-443	Chicken	Chicken	1	Germany: Federal State A
SRR12302238	BfRCA15279	21	ST-21	Pig	Generalist	6	Germany
SRR12302237	BfRCA15284	400	ST-353	Chicken	Chicken	1	Germany: Federal State M
SRR12302236	BfRCA15332	1775	ST-403	Pig	Pig	11	Germany: Federal State H
SRR12302235	BfRCA15395	464	ST-464	Chicken	unknown	1	Germany: Federal State P
SRR12302233	IMT468	10287	ST-403	Pig	Pig	11	Germany
SRR12302232	IMT538	1775	ST-403	Pig	Pig	11	Germany
SRR12302231	IMT541	403	ST-403	Pig	Pig	11	Germany
SRR5209454	SRR5209454	61	ST-61	Horse	Cattle	10	Canada: Alberta
SRR5209455	SRR5209455	61	ST-61	Horse	Cattle	10	Canada: Alberta
SRR5209456	SRR5209456	1244	ST-61	Cattle	Cattle	10	Canada: Alberta
SRR5209457	SRR5209457	1244	ST-61	Cattle	Cattle	10	Canada: Alberta

Continued on next page

Table A.1 Continued from previous page

SRR	ID	ST	CC	Origin	Association	BAPS	Source
SRR5209458	SRR5209458	45	ST-45	Water	Generalist	8	Canada: South Nation Watershed Ontario
SRR5209459	SRR5209459	2539	ST-177	Raccoon	unknown	7	Canada: Ontario
SRR5209460	SRR5209460	137	ST-45	Raccoon	Generalist	8	Canada: Ontario
SRR5209461	SRR5209461	137	ST-45	Raccoon	Generalist	8	Canada: Ontario
SRR5209462	SRR5209462	137	ST-45	Raccoon	Generalist	8	Canada: Ontario
SRR5209463	SRR5209463	45	ST-45	Water	Generalist	8	Canada: Ontario
SRR5209464	SRR5209464	61	ST-61	Cattle	Cattle	10	Canada: Alberta
SRR5209465	SRR5209465	1244	ST-61	Cattle	Cattle	10	Canada: Alberta
SRR5209466	SRR5209466	45	ST-45	Water	Generalist	8	Canada: Ontario
SRR5209467	SRR5209467	682	ST-682	Water	unknown	7	Canada: Alberta
SRR5209468	SRR5209468	45	ST-45	Water	Generalist	8	Canada: Ontario
SRR5209469	SRR5209469	10289	ST-177	Water	unknown	7	Canada: Quebec
SRR5209470	SRR5209470	682	ST-682	Water	unknown	7	Canada: Ontario
SRR5209471	SRR5209471	45	ST-45	Water	Generalist	8	Canada: Ontario
SRR5209472	SRR5209472	10290	ST-42	Water	Cattle	4	Canada: Ontario
SRR5209473	SRR5209473	132	ST-508	Cattle	unknown	14	Canada: Alberta
SRR5209474	SRR5209474	132	ST-508	Cattle	unknown	14	Canada: Alberta
SRR5209475	SRR5209475	42	ST-42	Water	Cattle	4	Canada: Ontario
SRR5209476	SRR5209476	45	ST-45	Water	Generalist	8	Canada: Alberta
SRR5209477	SRR5209477	45	ST-45	Water	Generalist	8	Canada: Ontario
SRR5209478	SRR5209478	137	ST-45	Water	Generalist	8	Canada: Alberta
SRR5209479	SRR5209479	45	ST-45	Water	Generalist	8	Canada: Ontario
SRR5209480	SRR5209480	682	ST-682	Water	unknown	7	Canada: New Brunswick
SRR5209481	SRR5209481	45	ST-45	Chicken	Generalist	8	Canada: Alberta
SRR5209482	SRR5209482	45	ST-45	Duck	Generalist	8	Canada: Alberta
SRR5209483	SRR5209483	45	ST-45	Dog	Generalist	8	Canada: Alberta
SRR5209484	SRR5209484	45	ST-45	Goose	Generalist	8	Canada: Alberta
SRR5209485	SRR5209485	45	ST-45	Water	Generalist	8	Canada: Alberta
SRR5209486	SRR5209486	1244	ST-61	Cattle	Cattle	10	Canada: Alberta
SRR5209487	SRR5209487	682	ST-682	Water	unknown	7	Canada: Ontario
SRR5209488	SRR5209488	5128	ST-682	Water	unknown	7	Canada: New Brunswick
SRR5209489	SRR5209489	45	ST-45	Sheep	Generalist	8	Canada: Alberta
SRR5209490	SRR5209490	45	ST-45	Cattle	Generalist	8	Canada: Alberta
SRR5209491	SRR5209491	45	ST-45	Sheep	Generalist	8	Canada: Alberta
SRR5209492	SRR5209492	61	ST-61	Cattle	Cattle	10	Canada: Alberta
SRR5209493	SRR5209493	459	ST-42	Cattle	Cattle	4	Canada: Alberta
SRR5209494	SRR5209494	48	ST-48	Chicken	Generalist	3	Canada: Alberta
SRR5209495	SRR5209495	48	ST-48	Chicken	Generalist	3	Canada: Alberta
SRR5209496	SRR5209496	45	ST-45	Sewage	Generalist	8	Canada: Alberta
SRR5209497	SRR5209497	137	ST-45	Water	Generalist	8	Canada: Alberta
SRR5209498	SRR5209498	459	ST-42	Cat	Cattle	4	Canada: Alberta
SRR5209499	SRR5209499	132	ST-508	Sewage	unknown	14	Canada: Alberta
SRR5209500	SRR5209500	806	ST-21	Sheep	Generalist	6	Canada: Alberta
SRR5209501	SRR5209501	61	ST-61	Cattle	Cattle	10	Canada: Alberta
SRR5209502	SRR5209502	45	ST-45	Cattle	Generalist	8	Canada: Alberta
SRR5209503	SRR5209503	61	ST-61	Water	Cattle	10	Canada: Alberta
SRR5209504	SRR5209504	132	ST-508	Cattle	unknown	14	Canada: Alberta
SRR5209505	SRR5209505	132	ST-508	Cattle	unknown	14	Canada: Alberta
SRR5209506	SRR5209506	459	ST-42	Water	Cattle	4	Canada: Alberta
SRR5209507	SRR5209507	262	ST-21	Human	Generalist	6	Canada: Alberta
SRR5209508	SRR5209508	61	ST-61	Human	Cattle	10	Canada: Alberta
SRR5209509	SRR5209509	10291	ST-48	Human	unknown	3	Canada: Alberta
SRR5209510	SRR5209510	45	ST-45	Human	Generalist	8	Canada: Alberta
SRR5209511	SRR5209511	19	ST-21	Human	Generalist	2	Canada: Alberta
SRR5209512	SRR5209512	45	ST-45	Human	Generalist	8	Canada: Alberta

Continued on next page

Table A.1 Continued from previous page

SRR	ID	ST	CC	Origin	Association	BAPS	Source
SRR5209513	SRR5209513	45	ST-45	Human	Generalist	8	Canada: Alberta
SRR5209514	SRR5209514	48	ST-48	Human	Generalist	3	Canada: Alberta
SRR5209515	SRR5209515	10291	ST-48	Human	unknown	3	Canada: Ontario
SRR5209516	SRR5209516	48	ST-48	Human	Generalist	3	Canada: Alberta
SRR5209517	SRR5209517	61	ST-61	Human	Cattle	10	Canada: Alberta
SRR5209518	SRR5209518	45	ST-45	Human	Generalist	8	Canada: Alberta
SRR5209519	SRR5209519	61	ST-61	Human	Cattle	10	Canada: Alberta
SRR5209520	SRR5209520	50	ST-21	Human	Generalist	2	Canada: Alberta
SRR5209521	SRR5209521	61	ST-61	Human	Cattle	10	Canada: Alberta
SRR5209522	SRR5209522	45	ST-45	Chicken	Generalist	8	Canada: British Columbia
SRR5209523	SRR5209523	267	ST-283	Chicken	unknown	8	Canada: British Columbia
SRR5209524	SRR5209524	267	ST-283	Turkey	unknown	8	Canada: Ontario
SRR5209525	SRR5209525	45	ST-45	Chicken	Generalist	8	Canada: Ontario
SRR5209526	SRR5209526	21	ST-21	Chicken	Generalist	6	Canada: Ontario
SRR5209527	SRR5209527	45	ST-45	Chicken	Generalist	8	Canada: Ontario
SRR5209528	SRR5209528	45	ST-45	Cattle	Generalist	8	Canada: Ontario
SRR5209529	SRR5209529	933	ST-403	Human	Pig	11	Canada: Ontario
SRR5209530	SRR5209530	42	ST-42	Human	Cattle	4	Canada: Ontario
SRR5209531	SRR5209531	21	ST-21	Human	Generalist	6	Canada: Ontario
SRR5209532	SRR5209532	45	ST-45	Human	Generalist	8	Canada: Ontario
SRR5209533	SRR5209533	45	ST-45	Chicken	Generalist	8	Canada: Ontario
SRR5209534	SRR5209534	4080	unknown	Water	unknown	15	Canada: Ontario
SRR5209535	SRR5209535	1030	unknown	Water	unknown	15	Canada: Quebec
SRR5209536	SRR5209536	3112	unknown	Water	unknown	15	Canada: Ontario
SRR5209537	SRR5209537	1294	unknown	Water	unknown	15	Canada: Alberta
SRR5209538	SRR5209538	1030	unknown	Water	unknown	15	Canada: Ontario
SRR5209539	SRR5209539	3495	unknown	Water	unknown	15	Canada: Ontario
SRR5209540	SRR5209540	996	unknown	Water	unknown	15	Canada: British Columbia
SRR5209541	SRR5209541	996	unknown	Water	unknown	15	Canada: Alberta
SRR5209542	SRR5209542	693	unknown	Water	unknown	15	Canada: Alberta
SRR5209543	SRR5209543	699	ST-692	Water	unknown	5	Canada: Alberta
SRR5209544	SRR5209544	991	ST-692	Water	unknown	5	Canada: British Columbia
SRR5209545	SRR5209545	991	ST-692	Water	unknown	5	Canada: British Columbia
SRR5209546	SRR5209546	6516	unknown	Water	unknown	15	Canada: Alberta
SRR5209547	SRR5209547	9353	ST-1034	Water	Chicken	15	Canada: British Columbia
SRR5209548	SRR5209548	4071	ST-1034	Water	Chicken	15	Canada: Alberta
SRR5209549	SRR5209549	693	unknown	Water	unknown	15	Canada: Ontario
SRR5209550	SRR5209550	991	ST-692	Water	unknown	5	Canada: British Columbia
SRR5209551	SRR5209551	693	unknown	Water	unknown	15	Canada: Alberta
SRR5209552	SRR5209552	4071	ST-1034	Water	Chicken	15	Canada: Alberta
SRR5209553	SRR5209553	5452	unknown	Water	unknown	15	Canada: British Columbia
SRR5209554	SRR5209554	991	ST-692	Water	unknown	5	Canada: British Columbia
SRR5209555	SRR5209555	10293	unknown	Water	unknown	15	Canada: Ontario
SRR5209556	SRR5209556	5705	unknown	Duck	unknown	15	Canada: Ontario
SRR5209557	SRR5209557	5705	unknown	Goose	unknown	15	Canada: Ontario
SRR5209558	SRR5209558	995	unknown	Water	unknown	15	Canada: British Columbia
SRR5209559	SRR5209559	1030	unknown	Water	unknown	15	Canada: Alberta
SRR5209560	SRR5209560	3495	unknown	Water	unknown	15	Canada: British Columbia
SRR5209561	SRR5209561	710	unknown	Goose	unknown	15	Canada: Alberta
SRR5209562	SRR5209562	996	unknown	Water	unknown	15	Canada: Alberta
SRR5209563	SRR5209563	3112	unknown	Water	unknown	15	Canada: Alberta
SRR5209564	SRR5209564	10296	unknown	Water	unknown	15	Canada: Alberta
SRR5209565	SRR5209565	693	unknown	Water	unknown	15	Canada: Alberta
SRR5209566	SRR5209566	1206	unknown	Goose	unknown	15	Canada: Alberta
SRR5209567	SRR5209567	1206	unknown	Goose	unknown	15	Canada: Alberta
SRR5209568	SRR5209568	1206	unknown	Goose	unknown	15	Canada: Alberta

Continued on next page

Table A.1 Continued from previous page

SRR	ID	ST	CC	Origin	Association	BAPS	Source
SRR5209569	SRR5209569	929	ST-257	Water	Chicken	9	Canada: Alberta
SRR5209570	SRR5209570	929	ST-257	Water	Chicken	9	Canada: Alberta
SRR5209571	SRR5209571	929	ST-257	Cattle	Chicken	9	Canada: Alberta
SRR5209572	SRR5209572	929	ST-257	Cattle	Chicken	9	Canada: Alberta
SRR5209573	SRR5209573	929	ST-257	Human	Chicken	9	Canada: Alberta
SRR5209574	SRR5209574	929	ST-257	Human	Chicken	9	Canada: Alberta
SRR5209575	SRR5209575	982	ST-21	Cattle	Generalist	6	Canada: Alberta
SRR5209576	SRR5209576	21	ST-21	Cattle	Generalist	6	Canada: Alberta
SRR5209577	SRR5209577	21	ST-21	Cattle	Generalist	6	Canada: Alberta
SRR5209578	SRR5209578	21	ST-21	Cattle	Generalist	6	Canada: Alberta
SRR5209579	SRR5209579	8	ST-21	Cattle	Generalist	6	Canada: Alberta
SRR5209580	SRR5209580	8	ST-21	Cattle	Generalist	6	Canada: Alberta
SRR5209581	SRR5209581	8	ST-21	Cattle	Generalist	6	Canada: Alberta
SRR5209582	SRR5209582	8	ST-21	Cattle	Generalist	6	Canada: Alberta
SRR5209583	SRR5209583	8	ST-21	Cattle	Generalist	6	Canada: Alberta
SRR5209584	SRR5209584	3391	ST-21	Cattle	Generalist	6	Canada: Alberta
SRR5209585	SRR5209585	3391	ST-21	Cattle	Generalist	6	Canada: Alberta
SRR5209586	SRR5209586	3391	ST-21	Cattle	Generalist	6	Canada: Alberta
SRR5209587	SRR5209587	8	ST-21	Cattle	Generalist	6	Canada: Alberta
SRR5209588	SRR5209588	8	ST-21	Cattle	Generalist	6	Canada: Alberta
SRR5209589	SRR5209589	982	ST-21	Chicken	Generalist	6	Canada: Alberta
SRR5209590	SRR5209590	982	ST-21	Chicken	Generalist	6	Canada: Alberta
SRR5209591	SRR5209591	922	unknown	Cattle	unknown	13	Canada: Alberta
SRR5209592	SRR5209592	45	ST-45	Sheep	Generalist	8	Canada: Alberta
SRR5209593	SRR5209593	21	ST-21	Cattle	Generalist	6	Canada: Alberta
SRR5209594	SRR5209594	922	unknown	Cattle	unknown	13	Canada: Alberta
SRR5209595	SRR5209595	8	ST-21	Cattle	Generalist	6	Canada: Alberta
SRR5209596	SRR5209596	8	ST-21	Cattle	Generalist	6	Canada: Alberta
SRR5209597	SRR5209597	922	unknown	Cattle	unknown	13	Canada: Alberta
SRR5209598	SRR5209598	8	ST-21	Cattle	Generalist	6	Canada: Alberta
SRR5209599	SRR5209599	679	ST-45	Human	Generalist	8	Canada: Alberta
SRR5209600	SRR5209600	982	ST-21	Human	Generalist	6	Canada: Alberta
SRR5209601	SRR5209601	922	unknown	Human	unknown	13	Canada: Alberta
SRR5209602	SRR5209602	8	ST-21	Human	Generalist	6	Canada: Alberta
SRR5209603	SRR5209603	21	ST-21	Human	Generalist	6	Canada: Alberta
SRR5209604	SRR5209604	922	unknown	Human	unknown	13	Canada: Alberta
SRR5209605	SRR5209605	8	ST-21	Human	Generalist	6	Canada: Alberta
SRR5209606	SRR5209606	8	ST-21	Human	Generalist	6	Canada: Alberta
SRR5209607	SRR5209607	982	ST-21	Human	Generalist	6	Canada: Alberta
SRR5209608	SRR5209608	982	ST-21	Human	Generalist	6	Canada: Alberta
SRR5209609	SRR5209609	982	ST-21	Human	Generalist	6	Canada: Alberta
SRR5209610	SRR5209610	982	ST-21	Human	Generalist	6	Canada: Alberta
SRR5209611	SRR5209611	922	unknown	Human	unknown	13	Canada: Alberta
SRR5209612	SRR5209612	679	ST-45	Human	Generalist	8	Canada: Alberta
SRR5209613	SRR5209613	2306	ST-21	Human	Generalist	2	Canada: Alberta
SRR5209614	SRR5209614	46	ST-206	Human	Generalist	3	Canada: Alberta
SRR5209615	SRR5209615	46	ST-206	Human	Generalist	3	Canada: Alberta
SRR5209616	SRR5209616	19	ST-21	Human	Generalist	2	Canada: Alberta
SRR5209617	SRR5209617	52	ST-52	Cattle	unknown	1	Canada: Ontario
SRR5209618	SRR5209618	429	ST-48	Chicken	Generalist	3	Canada: Ontario
SRR5209619	SRR5209619	918	ST-48	Human	Generalist	3	Canada: Ontario

End of table

A.2 Dataset 2 NCBI accessory numbers and metadata

Overview of the Campylobacter Hybrid Strains used from the zoonoses monitoring. The table was created and published in [39] during this work.

Strain No.	Species	Matrix catagory	qPCR result (Best/-Mayr)	result (Denis)	result (cpn60)	result (cadF)	WGS result	%Cj (CGE)	MLST ST	MLST CC	Accession-No.
BfR-CA-08175	C. coli	eggs	coli	coli	jejuni	coli	hybrid	13,36	**10180**	?	SAMN13577876
BfR-CA-08176	C. coli	eggs	coli	coli	jejuni	coli	hybrid	13,13	7018	?	SAMN13577877
BfR-CA-08318	C. coli	eggs	none	none	jejuni	coli	hybrid	13,76	9102	?	SAMN13577878
BfR-CA-08393	C. coli	eggs	coli	none	jejuni	coli	hybrid	13,2	**10181**	?	SAMN13577879
BfR-CA-08683	C. coli	eggs	coli	coli	jejuni	coli	hybrid	13,33	4148	?	SAMN13577880
BfR-CA-08836	C. coli	eggs	coli	coli	jejuni	coli	hybrid	12,77	4148	?	SAMN13577881
BfR-CA-08928	C. coli	eggs	false jejuni	false jejuni	jejuni	coli	hybrid	15,54	1487	ST-1150 complex	SAMN13577882
BfR-CA-09211	C. coli	eggs	coli	coli	jejuni	coli	hybrid	13,44	4148	?	SAMN13577883
BfR-CA-11586	C. coli	chicken meat	Cj/Cc mix	Cj/Cc mix	coli	coli	half hybrid	1,46	1595	ST-828 complex	SAMN13577884
BfR-CA-13047	C. coli	eggs	coli	coli	jejuni	coli	hybrid	13,85	4148	?	SAMN13577885
BfR-CA-13120	C. coli	chicken meat	Cj/Cc mix	Cj/Cc mix	jejuni	coli	hybrid	13,87	**10182**	?	SAMN13577886
BfR-CA-13188	C. coli	eggs	Cj/Cc mix	coli	jejuni	coli	hybrid	12,19	5439	?	SAMN13577887
BfR-CA-13619	C. coli	eggs	Cj/Cc mix	coli	jejuni	coli	hybrid	12,3	5439	?	SAMN13577888
BfR-CA-13895	C. coli	eggs	Cj/Cc mix	coli	jejuni	coli	hybrid	13,06	5439	?	SAMN13577889

Continued on next page

Table A.2 Continued from previous page

Strain No.	Species	Matrix catagory	qPCR result (Best/-Mayr)	result (Denis)	result (cpn60)	result (cadF)	WGS result	%Cj (CGE)	MLST ST	MLST CC	Accession-No.
BfR-CA-13919	C. coli	turkey cecum	Cj/Cc mix	Cj/Cc mix	jejuni	coli	hybrid	12,32	4148	?	SAMN13577890
BfR-CA-13953	C. coli	turkey cecum	Cj/Cc mix	Cj/Cc mix	jejuni	coli	hybrid	14,31	**10183**	?	SAMN13577891
BfR-CA-14226	C. coli	turkey cecum	Cj/Cc mix	coli	jejuni	coli	hybrid	13,96	**10184**	?	SAMN13577892
BfR-CA-14582	C. coli	chicken meat	Cj/Cc mix	Cj/Cc mix	coli	coli	half hybrid	1,99	830	ST-828 complex	SAMN13577893
BfR-CA-14610	C. coli	turkey cecum	Cj/Cc mix	Cj/Cc mix	coli	coli	half hybrid	1,55	1586	ST-828 complex	SAMN13577894
BfR-CA-14731	C. coli	turkey meat	Cj/Cc mix	coli	jejuni	coli	hybrid	12,69	**10185**	?	SAMN13577895
BfR-CA-14810	C. coli	eggs	Cj/Cc mix	coli	jejuni	coli	hybrid	13,66	**10186**	?	SAMN13577896
BfR-CA-14825	C. coli	turkey cecum	Cj/Cc mix	Cj/Cc mix	jejuni	coli	hybrid	12,13	4148	?	SAMN13577897
BfR-CA-14833	C. coli	eggs	Cj/Cc mix	coli	jejuni	coli	hybrid	13,1	5439	?	SAMN13577898
BfR-CA-14943	C. coli	turkey cecum	Cj/Cc mix	Cj/Cc mix	jejuni	coli	hybrid	13,85	4148	?	SAMN13577899
BfR-CA-14973	C. coli	eggs	none	none	jejuni	coli	hybrid	12,72	**10187**	?	SAMN13577900
BfR-CA-15005	C. coli	turkey skin	Cj/Cc mix	Cj/Cc mix	coli	coli	half hybrid	1,17	1586	ST-828 complex	SAMN13577901
BfR-CA-15124	C. coli	chicken meat	Cj/Cc mix	coli	coli	coli	half hybrid	2,56	832	ST-828 complex	SAMN13577902
BfR-CA-15267	C. coli	eggs	false jejuni	false jejuni	jejuni	coli	hybrid	14,93	5903	ST-1150 complex	SAMN13577903
BfR-CA-15268	C. coli	chicken meat	Cj/Cc mix	Cj/Cc mix	coli	coli	half hybrid	1,54	1595	ST-828 complex	SAMN13577904

Continued on next page

Table A.2 Continued from previous page

Strain No.	Species	Matrix catagory	qPCR result (Best/-Mayr)	result (Denis)	result (cpn60)	result (cadF)	WGS result	%Cj (CGE)	MLST ST	MLST CC	Accession-No.
BfR-CA-15281	C. coli	turkey meat	Cj/Cc mix	coli	coli	coli	half hybrid	2,2	832	ST-828 complex	SAMN13577905
BfR-CA-15286	C. coli	chicken meat	Cj/Cc mix	Cj/Cc mix	coli	coli	half hybrid	0	9168	ST-828 complex	SAMN13577906
BfR-CA-15287	C. coli	chicken meat	Cj/Cc mix	Cj/Cc mix	coli	coli	half hybrid	1,16	9168	ST-828 complex	SAMN13577907
BfR-CA-15301	C. coli	eggs	coli	coli	jejuni	coli	hybrid	13,69	**10188**	?	SAMN13577908
BfR-CA-15396	C. coli	chicken meat	Cj/Cc mix	Cj/Cc mix	jejuni	coli	hybrid	14,31	**10183**	?	SAMN13577909
BfR-CA-15426	C. coli	chicken meat	Cj/Cc mix	Cj/Cc mix	coli	coli	half hybrid	1,01	1595	ST-828 complex	SAMN13577910
BfR-CA-15489	C. coli	turkey cecum	none	none	coli	coli	half hybrid	6,49	**10190**	?	SAMN13577911
BfR-CA-15533	C. coli	duck meat	Cj/Cc mix	Cj/Cc mix	coli	coli	half hybrid	1,11	1595	ST-828 complex	SAMN13577912
BfR-CA-15630	C. coli	turkey meat	Cj/Cc mix	Cj/Cc mix	jejuni	coli	hybrid	15,46	**10183**	?	SAMN13577913
BfR-CA-15892	C. coli	turkey cecum	Cj/Cc mix	Cj/Cc mix	jejuni	coli	hybrid	14,35	**10183**	?	SAMN13577914
BfR-CA-16767	C. coli	turkey cecum	Cj/Cc mix	Cj/Cc mix	jejuni	coli	hybrid	14,39	**10183**	?	SAMN13577915
BfR-CA-16822	C. coli	turkey cecum	Cj/Cc mix	coli	coli	coli	half hybrid	2,22	832	ST-828 complex	SAMN13577916
BfR-CA-16834	C. coli	turkey cecum	Cj/Cc mix	coli	coli	coli	half hybrid	3,48	832	ST-828 complex	SAMN13577917
BfR-CA-16942	C. coli	turkey meat	none	coli	jejuni	coli	hybrid	15,17	**10194**	?	SAMN13577918
BfR-CA-17078	C. coli	turkey cecum	Cj/Cc mix	Cj/Cc mix	coli	coli	half hybrid	4,78	**10195**	?	SAMN13577919
BfR-CA-17110	C. coli	turkey meat	Cj/Cc mix	Cj/Cc mix	coli	jejuni	half hybrid	1,53	1769	ST-828 complex	SAMN13577920

End of table

A.3 Wet-lab workflows for species determination

Wet-lab workflows were performed by Julia Golz and Kerstin Stingl at the BfR.

A.3.1 Discrimination among *Campylobacter* spp. using qPCR

Resuspended cell pellets were vortexted with 5% Chelex 100 resin (Bio-Rad Laboratories GmbH, Germany) and incubated for 15 min at 95 ℃. Subsequent centrifugation was performed and the supernatant was used as PCR target. To distinguish between *C. coli* and *C. jejuni* a real-time qPCR-based on the genes *map*A, *gyr*A and *ceu*E was used [157]. The *map*A gene encodes an outer-membrane gene of *C. jejuni*, whereas *ceu*E encodes the enterochelin uptake substrate-binding protein for iron acquisition in *C. coli* [158]. In case of ambiguous results, an additional gel-based multiplex-PCR was applied. For this purpose, different species-specific fragments were targeted by corresponding primer pairs: internal regions of the *hip*O gene for *C. jejuni*, the *gly*A gene for *C. coli* and *C. upsaliensis*, the *cpn*60 for *C. lari*, the *sap*B2 gene of *C. fetus* and a *Campylobacterales* specific fragment of the 23S rRNA gene [242].

A.3.2 Matrix-assisted laser desorption/ionization-time-of-flight (MALDI-TOF) analysis

Campylobacter colonies incubated overnight on columbia agar plates supplemented with 5% sheep blood were spotted onto a polished steel plate with 96 targets (MicroScout Target plate; Bruker Daltonik, Germany). The colonoies were overlaid with 1 μl of saturated *alpha*-cyano-4-hydroxy-cinnamic acid matrix solution (200 mg in 2.5% trifluoroacetic acid/ 50% acetonitrile) and dried completely. MALDI-TOF mass spectrometry analysis was performed using a MALDI-TOF Microflex LT (Bruker Daltonics, Germany) within a range of 2,000-20,000 $\frac{m}{z}$ (mass to charge ratio) as suggested by Bruker Bacterial Test Standards (Bruker Daltonics, Germany). Each of the 240 laser shot spectra were summed up in 40 steps, covering at least 80 shots per raster spot from different positions within the sample by the AutoXecute method using the software FlexAnalysis 3.4. The spectra were compared with the MBT Compass Library, Revision F (Bruker Daltonics, Germany). Each identification obtains a score value, for *Campylobacter* a score value $\geq$ 2.000 was considered to be correct identification at species level [268–270].

A.4 NCBI accessory numbers of strains used to build a k-mer database

NCBI accessory numbers/ BfR ID and species classification of strains used to build a k-mer database. The table was created and published in [39] during this work.

ID	species
GCA_000146835.1_ASM14683v1	Campylobacter coli
GCA_000167415.1_ASM16741v1	Campylobacter coli
GCA_000505605.1_K3	Campylobacter coli
GCA_000505625.1_K7	Campylobacter coli
GCA_000531565.1_IPSID-1	Campylobacter coli
GCA_001225145.1_7092_1_57	Campylobacter coli
GCA_001228685.1_7092_1_27	Campylobacter coli
GCA_001228905.1_7038_3_46	Campylobacter coli
GCA_001228985.1_7213_3_54	Campylobacter coli
GCA_001234385.1_7092_1_13	Campylobacter coli
GCA_001235265.1_7038_3_62	Campylobacter coli
GCA_001235285.1_7092_1_29	Campylobacter coli
GCA_001236625.1_7065_7_40	Campylobacter coli
GCA_001236965.1_7213_3_16	Campylobacter coli
GCA_001291485.1_RC282_S32contigs.fa	Campylobacter coli
GCA_001291525.1_RC105_S13contigs.fa	Campylobacter coli
GCA_001291545.1_RC382_S38contigs.fa	Campylobacter coli
GCA_001291605.1_RC148_S18contigs.fa	Campylobacter coli
GCA_001291665.1_RC264_S27contigs.fa	Campylobacter coli
GCA_001291725.1_RC383_S39contigs.fa	Campylobacter coli
GCA_001291745.1_RC285_S34contigs.fa	Campylobacter coli
GCA_001291765.1_RC126_S16contigs.fa	Campylobacter coli
GCA_001291785.1_RC182_S22contigs.fa	Campylobacter coli
GCA_001291845.1_RC387_S41contigs.fa	Campylobacter coli
GCA_001291865.1_RC023_S5contigs.fa	Campylobacter coli
GCA_001291885.1_RC127_S17contigs.fa	Campylobacter coli
GCA_001291905.1_RC269_S28contigs.fa	Campylobacter coli
GCA_001291985.1_RC037_S7contigs.fa	Campylobacter coli
GCA_001292005.1_RC026_S6contigs.fa	Campylobacter coli
GCA_001292065.1_RC106_S14contigs.fa	Campylobacter coli
GCA_001292085.1_RC430_S42contigs.fa	Campylobacter coli
GCA_001292105.1_RC096_S11contigs.fa	Campylobacter coli
GCA_001292125.1_RC289_S35contigs.fa	Campylobacter coli
GCA_001292165.1_RC415_S43contigs.fa	Campylobacter coli
GCA_001292225.1_RC116_S15contigs.fa	Campylobacter coli
GCA_001292305.1_RC038_S8contigs.fa	Campylobacter coli
GCA_001292325.1_RC428_S45contigs.fa	Campylobacter coli
GCA_001292345.1_RC043_S10contigs.fa	Campylobacter coli
GCA_001292365.1_RC281_S31contigs.fa	Campylobacter coli
GCA_001292405.1_RC018_S4contigs.fa	Campylobacter coli
GCA_001292425.1_RC386_S40contigs.fa	Campylobacter coli
GCA_001292445.1_RC284_S33contigs.fa	Campylobacter coli
GCA_001292505.1_RC008_S1contigs.fa	Campylobacter coli
GCA_001305715.1_ASM130571v1	Campylobacter coli
GCA_001419355.1_Campylobacter_coli_CVM_41957_v1.0	Campylobacter coli
GCA_001491115.1_SWAN361	Campylobacter coli
GCA_001498315.1_H074360315	Campylobacter coli
GCA_001545275.1_ASM154527v1	Campylobacter coli
GCA_001545285.1_ASM154528v1	Campylobacter coli
GCA_001545295.1_ASM154529v1	Campylobacter coli
GCA_001545335.1_ASM154533v1	Campylobacter coli
GCA_001545355.1_ASM154535v1	Campylobacter coli

Continued on next page

Table A.3 Continued from previous page

ID	species
GCA_001545365.1_ASM154536v1	Campylobacter coli
GCA_001545415.1_ASM154541v1	Campylobacter coli
GCA_001545425.1_ASM154542v1	Campylobacter coli
GCA_001545435.1_ASM154543v1	Campylobacter coli
GCA_001545445.1_ASM154544v1	Campylobacter coli
GCA_001761965.1_ASM176196v1	Campylobacter coli
GCA_001761985.1_ASM176198v1	Campylobacter coli
GCA_001761995.1_ASM176199v1	Campylobacter coli
GCA_001762015.1_ASM176201v1	Campylobacter coli
GCA_001762045.1_ASM176204v1	Campylobacter coli
GCA_001762065.1_ASM176206v1	Campylobacter coli
GCA_001762085.1_ASM176208v1	Campylobacter coli
GCA_001762095.1_ASM176209v1	Campylobacter coli
GCA_001762225.1_BCW_6447	Campylobacter coli
GCA_001762245.1_BCW_6448	Campylobacter coli
GCA_001763285.1_ASM176328v1	Campylobacter coli
GCA_001763685.1_ASM176368v1	Campylobacter coli
GCA_001764015.1_ASM176401v1	Campylobacter coli
GCA_001765195.1_ASM176519v1	Campylobacter coli
GCA_002000205.1_ASM200020v1	Campylobacter coli
GCA_002178015.1_ASM217801v1	Campylobacter coli
GCA_002207945.1_ASM220794v1	Campylobacter coli
GCA_002207965.1_ASM220796v1	Campylobacter coli
GCF_000465235.1_ASM46523v1	Campylobacter coli
GCF_000494775.1_ASM49477v1	Campylobacter coli
GCF_000583755.1_ASM58375v1	Campylobacter coli
GCF_000583795.1_ASM58379v1	Campylobacter coli
GCF_000954195.1_ASM95419v1	Campylobacter coli
GCF_001417635.1_ASM141763v1	Campylobacter coli
GCF_001483845.1_ASM148384v1	Campylobacter coli
GCF_001639125.1_ASM163912v1	Campylobacter coli
GCF_001717605.1_ASM171760v1	Campylobacter coli
GCF_001865455.1_ASM186545v1	Campylobacter coli
GCF_001865475.1_ASM186547v1	Campylobacter coli
GCF_001865495.1_ASM186549v1	Campylobacter coli
GCF_001865515.1_ASM186551v1	Campylobacter coli
GCF_001865535.1_ASM186553v1	Campylobacter coli
GCF_001865555.1_ASM186555v1	Campylobacter coli
GCF_001936355.1_ASM193635v1	Campylobacter coli
GCF_002024185.1_ASM202418v1	Campylobacter coli
GCF_002407145.1_ASM240714v1	Campylobacter coli
GCF_002843985.1_ASM284398v1	Campylobacter coli
GCF_003030205.1_ASM303020v1	Campylobacter coli
GCF_000466065.2_ASM46606v2	Campylobacter jejuni
GCF_000466075.2_ASM46607v2	Campylobacter jejuni
GCF_002407125.1_ASM240712v1	Campylobacter jejuni
GCF_000009085.1_ASM908v1	Campylobacter jejuni
GCF_000011865.1_ASM1186v1	Campylobacter jejuni
GCF_000015525.1_ASM1552v1	Campylobacter jejuni
GCF_000017905.1_ASM1790v1	Campylobacter jejuni
GCF_000025425.1_ASM2542v1	Campylobacter jejuni
GCF_000148705.1_ASM14870v1	Campylobacter jejuni
GCF_000171795.2_ASM17179v2	Campylobacter jejuni
GCF_000184205.1_ASM18420v1	Campylobacter jejuni
GCF_000302555.5_ASM30255v4	Campylobacter jejuni
GCF_000304375.1_AINO	Campylobacter jejuni

Continued on next page

Table A.3 Continued from previous page

ID	species
GCF_000430385.1_ASM43038v1	Campylobacter jejuni
GCF_000466105.2_ASM46610v2	Campylobacter jejuni
GCF_000468915.2_ASM46891v2	Campylobacter jejuni
GCF_000493495.1_TS	Campylobacter jejuni
GCF_000737085.1_ASM73708v1	Campylobacter jejuni
GCF_000772225.1_ASM77222v1	Campylobacter jejuni
GCF_000830775.1_ASM83077v1	Campylobacter jejuni
GCF_000830805.1_ASM83080v1	Campylobacter jejuni
GCF_000830825.1_ASM83082v1	Campylobacter jejuni
GCF_000830845.1_ASM83084v1	Campylobacter jejuni
GCF_000830865.1_ASM83086v1	Campylobacter jejuni
GCF_000835285.1_ASM83528v1	Campylobacter jejuni
GCF_000835305.1_ASM83530v1	Campylobacter jejuni
GCF_000835345.1_ASM83534v1	Campylobacter jejuni
GCF_000835365.1_ASM83536v1	Campylobacter jejuni
GCF_000934305.1_ASM93430v1	Campylobacter jejuni
GCF_001299565.1_ASM129956v1	Campylobacter jejuni
GCF_001299595.1_ASM129959v1	Campylobacter jejuni
GCF_001314285.1_ASM131428v1	Campylobacter jejuni
GCF_001412295.1_ASM141229v1	Campylobacter jejuni
GCF_001457695.1_NCTC11351	Campylobacter jejuni
GCF_001506185.1_ASM150618v1	Campylobacter jejuni
GCF_001506205.1_ASM150620v1	Campylobacter jejuni
GCF_001506225.1_ASM150622v1	Campylobacter jejuni
GCF_001506245.1_ASM150624v1	Campylobacter jejuni
GCF_001506265.1_ASM150626v1	Campylobacter jejuni
GCF_001506285.1_ASM150628v1	Campylobacter jejuni
GCF_001506305.1_ASM150630v1	Campylobacter jejuni
GCF_001506345.1_ASM150634v1	Campylobacter jejuni
GCF_001506365.1_ASM150636v1	Campylobacter jejuni
GCF_001506385.1_ASM150638v1	Campylobacter jejuni
GCF_001506405.1_ASM150640v1	Campylobacter jejuni
GCF_001506425.1_ASM150642v1	Campylobacter jejuni
GCF_001506445.1_ASM150644v1	Campylobacter jejuni
GCF_001506465.1_ASM150646v1	Campylobacter jejuni
GCF_001506485.1_ASM150648v1	Campylobacter jejuni
GCF_001506505.1_ASM150650v1	Campylobacter jejuni
GCF_001506525.1_ASM150652v1	Campylobacter jejuni
GCF_001506545.1_ASM150654v1	Campylobacter jejuni
GCF_001506565.1_ASM150656v1	Campylobacter jejuni
GCF_001506585.1_ASM150658v1	Campylobacter jejuni
GCF_001506605.1_ASM150660v1	Campylobacter jejuni
GCF_001506625.1_ASM150662v1	Campylobacter jejuni
GCF_001506645.1_ASM150664v1	Campylobacter jejuni
GCF_001506665.1_ASM150666v1	Campylobacter jejuni
GCF_001506685.1_ASM150668v1	Campylobacter jejuni
GCF_001506705.1_ASM150670v1	Campylobacter jejuni
GCF_001506725.1_ASM150672v1	Campylobacter jejuni
GCF_001506745.1_ASM150674v1	Campylobacter jejuni
GCF_001506765.1_ASM150676v1	Campylobacter jejuni
GCF_001506785.1_ASM150678v1	Campylobacter jejuni
GCF_001506805.1_ASM150680v1	Campylobacter jejuni
GCF_001506825.1_ASM150682v1	Campylobacter jejuni
GCF_001506845.1_ASM150684v1	Campylobacter jejuni
GCF_001506865.1_ASM150686v1	Campylobacter jejuni
GCF_001506885.1_ASM150688v1	Campylobacter jejuni

Continued on next page

Table A.3 Continued from previous page

ID	species
GCF_001506905.1_ASM150690v1	Campylobacter jejuni
GCF_001506925.1_ASM150692v1	Campylobacter jejuni
GCF_001506945.1_ASM150694v1	Campylobacter jejuni
GCF_001506965.1_ASM150696v1	Campylobacter jejuni
GCF_001506985.1_ASM150698v1	Campylobacter jejuni
GCF_001507005.1_ASM150700v1	Campylobacter jejuni
GCF_001507025.1_ASM150702v1	Campylobacter jejuni
GCF_001507045.1_ASM150704v1	Campylobacter jejuni
GCF_001507065.1_ASM150706v1	Campylobacter jejuni
GCF_001507085.1_ASM150708v1	Campylobacter jejuni
GCF_001507105.1_ASM150710v1	Campylobacter jejuni
GCF_001507125.1_ASM150712v1	Campylobacter jejuni
GCF_001507145.1_ASM150714v1	Campylobacter jejuni
GCF_001507165.1_ASM150716v1	Campylobacter jejuni
GCF_001507185.1_ASM150718v1	Campylobacter jejuni
GCF_001507205.1_ASM150720v1	Campylobacter jejuni
GCF_001507225.1_ASM150722v1	Campylobacter jejuni
GCF_001507245.1_ASM150724v1	Campylobacter jejuni
GCF_001507265.1_ASM150726v1	Campylobacter jejuni
GCF_001563565.1_ASM156356v1	Campylobacter jejuni
GCF_001587015.1_ASM158701v1	Campylobacter jejuni
GCF_001587035.1_ASM158703v1	Campylobacter jejuni
GCF_001686905.1_ASM168690v1	Campylobacter jejuni
GCF_001717625.1_ASM171762v1	Campylobacter jejuni
GCF_001721945.1_ASM172194v1	Campylobacter jejuni
GCF_001721965.1_ASM172196v1	Campylobacter jejuni
GCF_001721985.1_ASM172198v1	Campylobacter jejuni
GCF_001767215.1_ASM176721v1	Campylobacter jejuni
GCF_001865395.1_ASM186539v1	Campylobacter jejuni
GCF_001865415.1_ASM186541v1	Campylobacter jejuni
GCF_001865435.1_ASM186543v1	Campylobacter jejuni
GCF_001865595.1_ASM186559v1	Campylobacter jejuni
GCF_001870085.1_ASM187008v1	Campylobacter jejuni
GCF_001870105.1_ASM187010v1	Campylobacter jejuni
GCF_001951235.1_ASM195123v1	Campylobacter jejuni
GCF_001951255.1_ASM195125v1	Campylobacter jejuni
GCF_001951275.1_ASM195127v1	Campylobacter jejuni
GCF_001951295.1_ASM195129v1	Campylobacter jejuni
GCF_001951315.1_ASM195131v1	Campylobacter jejuni
GCF_001951335.1_ASM195133v1	Campylobacter jejuni
GCF_002024325.1_ASM202432v1	Campylobacter jejuni
GCF_002028305.1_ASM202830v1	Campylobacter jejuni
GCF_002101355.1_ASM210135v1	Campylobacter jejuni
GCF_002209005.1_ASM220900v1	Campylobacter jejuni
GCF_002209025.1_ASM220902v1	Campylobacter jejuni
GCF_002209045.1_ASM220904v1	Campylobacter jejuni
GCF_002209065.1_ASM220906v1	Campylobacter jejuni
GCF_002214785.1_ASM221478v1	Campylobacter jejuni
GCF_002224325.1_ASM222432v1	Campylobacter jejuni
GCF_002224385.1_ASM222438v1	Campylobacter jejuni
GCF_002234455.1_ASM223445v1	Campylobacter jejuni
GCF_002238375.1_ASM223837v1	Campylobacter jejuni
GCF_002587105.1_ASM258710v1	Campylobacter jejuni
GCF_002587225.1_ASM258722v1	Campylobacter jejuni
GCF_003030185.1_ASM303018v1	Campylobacter jejuni
GCF_003060725.1_ASM306072v1	Campylobacter jejuni

Continued on next page

Table A.3 Continued from previous page

ID	species
GCF_003060745.1_ASM306074v1	Campylobacter jejuni
GCF_003060765.1_ASM306076v1	Campylobacter jejuni
GCF_003060785.1_ASM306078v1	Campylobacter jejuni
GCF_003368045.1_ASM336804v1	Campylobacter jejuni
GCF_003368065.1_ASM336806v1	Campylobacter jejuni
GCF_003368085.1_ASM336808v1	Campylobacter jejuni
GCF_003368105.1_ASM336810v1	Campylobacter jejuni
GCF_003368125.1_ASM336812v1	Campylobacter jejuni
GCF_003368145.1_ASM336814v1	Campylobacter jejuni
GCF_003368165.1_ASM336816v1	Campylobacter jejuni
GCF_003368185.1_ASM336818v1	Campylobacter jejuni
GCF_003368205.1_ASM336820v1	Campylobacter jejuni
GCF_003368225.1_ASM336822v1	Campylobacter jejuni
GCF_003368245.1_ASM336824v1	Campylobacter jejuni
GCF_003574945.1_ASM357494v1	Campylobacter jejuni
GCF_003950275.1_ASM395027v1	Campylobacter jejuni
GCF_003971585.1_ASM397158v1	Campylobacter jejuni
GCF_900475265.1_43024_F01	Campylobacter jejuni
GCF_900638165.1_56527_E01	Campylobacter jejuni
GCF_900638175.1_56527_F01	Campylobacter jejuni
GCF_900638185.1_56527_G01	Campylobacter jejuni
GCF_900638195.1_56553_F01	Campylobacter jejuni
GCF_900638205.1_56553_D01	Campylobacter jejuni
GCF_900638225.1_56553_E01	Campylobacter jejuni
GCF_900638235.1_56772_E02	Campylobacter jejuni
GCF_900638285.1_57043_B01	Campylobacter jejuni
GCF_900638365.1_57428_D01	Campylobacter jejuni
BfR-CA-11057	Campylobacter coli
BfR-CA-13264	Campylobacter coli
BfR-CA-13971	Campylobacter coli
BfR-CA-14216	Campylobacter coli
BfR-CA-14583	Campylobacter coli
BfR-CA-14709	Campylobacter coli
BfR-CA-14751	Campylobacter coli
BfR-CA-14815	Campylobacter coli
BfR-CA-15034	Campylobacter coli
BfR-CA-15062	Campylobacter coli
BfR-CA-15077	Campylobacter coli
BfR-CA-15371	Campylobacter coli
BfR-CA-15403	Campylobacter coli
BfR-CA-15629	Campylobacter coli
BfR-CA-15913	Campylobacter coli
BfR-CA-15969	Campylobacter coli
BfR-CA-15978	Campylobacter coli
DSM 4689	Campylobacter coli
BfRCA15282	Campylobacter jejuni
BfRCA15395	Campylobacter jejuni
BfRCA16737	Campylobacter jejuni

End of table

A.5 Identified genes by the host-specific GWAS

A.5.1 Overview of cattle associated accessory genes and core gene variants identified by k-mer mapping.

Overview of cattle associated accessory genes and core gene variants identified by *k-mer* mapping. The table was created and published in [154] during this work.

Reference Gene Name	Alias	Predicted Function	COG Family	# Cattle	# Chicken	# Pig	# Generalist	# Others	Total	Core/ Accessory
NCTC13261_01393	-	plasmid stabilization system protein, RelE/ParE family (HicA/HicB)	S	55	1	0	59	7	122	Accessory
NCTC13261_01392	-	hypothetical protein	-	55	2	0	61	7	125	Accessory
Cj1355	*ceuE*	Enterochelin uptake periplasmic binding protein	P	55	90	26	255	64	490	Core
Cj1278c	*trmB*	putative tRNA (guanine-N(7)-)-methyltransferase	J	55	90	26	255	64	490	Core
Cj1250	*purD*	phosphoribosylamine–glycine ligase	F	55	90	26	255	64	490	Core
Cj1248	*guaA*	GMP synthase (glutamine-hydrolyzing)	F	55	90	26	255	64	490	Core
Cj1246c	*uvrC*	excinuclease ABC subunit C	L	53	84	26	247	63	473	Accessory
Cj1240c	-	Hypothetical protein	-	55	86	21	249	64	475	Accessory
Cj1234	*glyS*	glycyl-tRNA synthetase beta chain	J	55	90	26	255	64	490	Core
Cj1233	*ppaX*	putative HAD-superfamily hydrolase	S	55	87	25	255	63	485	Accessory
Cj1228c	*htrA*	serine protease (protease DO)	M	55	90	26	255	64	490	Core
Cj1163c	*czcD*	putative cation transport protein / putative heavy-metal-associated domain protein	P	55	90	26	255	64	490	Core
Cj1162c	*copZ*	putative cation transport protein / putative heavy-metal-associated domain protein	P	55	90	26	255	64	490	Core
Cj1048c	*dapE*	succinyl-diaminopimelate desuccinylase	E	55	90	26	255	64	490	Core
NCTC13261_01099	-	dna methylase-type I restriction-modification system	V	30	34	0	46	27	137	Accessory
NCTC13261_01098	*rlfA*	RlfA	S	29	8	13	2	15	67	Accessory
NCTC13261_01097	-	hypothetical protein	-	36	35	14	50	34	169	Accessory
NCTC13261_01096	-	Putative uncharacterized protein	V	36	30	0	47	21	134	Accessory
Cj1047c	*thiS*	thiamine biosynthesis protein	H	55	90	14	254	51	464	Accessory
Cj1008c	*aroB*	3-dehydroquinate synthase	E	55	90	26	255	63	489	Core
Cj1002c	*sixA*	putative phosphoglycerate/bisphosphoglycerate mutase	T	42	88	23	216	32	401	Accessory

Continued on next page

Table A.4 Continued from previous page

Reference Gene Name	Alias	Predicted Function	COG Family	# Cattle	# Chicken	# Pig	# Generalist	# Others	Total	Core/ Accessory
Cj1001	*rpoD*	RNA polymerase sigma factor (sigma-70)	K	55	90	26	255	64	490	Core
Cj0999c	*yeiH*	putative integral membrane protein	S	55	90	26	255	64	490	Core
Cj0995c	*hemB*	delta-aminolevulinic acid dehydratase	H	55	90	26	255	64	490	Core
Cj0994c	*argF*	ornithine carbamoyltransferase	E	55	90	26	255	64	490	Core
Cj0991c	*glpC*	putative oxidoreductase ferredoxin-type electron transport protein	C	55	90	26	255	64	490	Core
Cj0967	-	putative periplasmic protein	-	37	79	26	192	45	379	Accessory
Cj0967	-	periplasmic protein	-	37	54	26	60	34	211	Accessory
Cj0964	-	putative periplasmic protein	E	55	90	26	255	64	490	Core
Cj0959c	*yidD*	hypothetical protein	S	55	90	26	255	63	489	Core
Cj0946	*spr7*	putative lipoprotein	M	55	90	26	255	64	490	Core
Cj0944c	-	putative periplasmic protein	-	55	90	25	255	62	487	Core
Cj0940c	*glnM*	putative glutamine transport system permease	P	55	90	26	255	64	490	Core
Cj0939c	-	hypothetical protein	-	53	90	26	255	63	487	Core
Cj0938c	*aas*	putative 2-acylglycerophosphoethanolamine acyltransferase / acyl-[acp] synthetase	EGP	45	82	25	215	34	401	Accessory
Cj0934c	IV02 _29000	putative sodium:amino-acid symporter family protein	P	32	81	0	175	15	303	Accessory
Cj0933c	*pycB*	putative pyruvate carboxylase B subunit	C	55	90	26	255	64	490	Core
Cj0932c	*pckA*	phosphoenolpyruvate carboxykinase (ATP)	H	55	90	26	255	64	490	Core
Cj0920c	*gltJ*	putative ABC-type amino-acid transporter permease protein	P	55	90	26	255	64	490	Core
Cj0919c	*tcyB_2*	Amino acid ABC transporter, permease protein PEB1	P	55	90	26	255	64	490	Core
Cj0917c	*cstA*	putative integral membrane protein	T	52	88	26	245	63	474	Accessory
Cj0916c	*ybdD*	hypothetical protein	S	55	90	26	254	64	489	Core
Cj0915	*yciA*	putative hydrolase	I	55	90	26	255	64	490	Core
Cj0912c	*cysK*	cysteine synthase	E	52	88	26	249	64	479	Accessory
Cj0911	hyaE	putative periplasmic protein	S	55	90	26	255	64	490	Core

Continued on next page

Table A.4 Continued from previous page

Reference Gene Name	Alias	Predicted Function	COG Family	# Cattle	# Chicken	# Pig	# Generalist	# Others	Total	Core/ Accessory
Cj0909	VY92 _09940	putative periplasmic protein	S	55	89	26	252	64	486	Core
Cj0908	-	putative periplasmic protein	-	46	80	24	211	35	396	Accessory
Cj0905c	*alr*	alanine racemase	E	55	90	26	255	64	490	Core
Cj0903c	*agcS*	putative amino-acid transport protein	E	55	90	25	254	64	488	Core
Cj0902	*artM*	putative glutamine transport ATP-binding protein	E	55	90	26	255	64	490	Core
Cj0886c	*ftsK*	putative cell division protein	D	55	90	26	255	64	490	Core
Cj0879c	-	putative periplasmic protein	-	38	57	4	140	57	296	Accessory
NCTC13261_00864	-	Arylsulfotransferase	M	33	36	1	74	25	169	Accessory
NCTC13261_00861	-	Arylsulfotransferase	M	48	65	25	118	26	282	Accessory
Cj0865	*dsbB*	disulfide oxidoreductase	C	55	90	26	254	64	489	Core
NCTC13261_00859	*dsbA*	thiol:disulfide interchange protein	O	55	78	26	255	27	441	Accessory
Cj0863c	*xerH*	DNA recombinase	L	55	90	26	255	64	490	Core
Cj0821	*glmU*	UDP-N-acetylglucosamine pyrophosphorylase	M	55	90	26	255	64	490	Core
Cj0794	-	hypothetical protein	S	43	81	25	20	44	213	Accessory
Cj0814	-	hypothetical protein	S	42	18	0	102	18	180	Accessory
Cj0812	*thrC*	threonine synthase	E	55	80	26	255	64	480	Accessory
Cj0800c	-	putative ATPase	L	55	90	26	255	64	490	Core
Cj0799c	*ruvA*	putative Holliday junction ATP-dependent DNA helicase	L	55	90	26	255	64	490	Core
Cj0798c	*ddl*	D-alanine–D-alanine ligase	F	55	90	26	255	64	490	Core
Cj0794	-	hypothetical protein	S	48	76	19	174	22	339	Accessory
Cj0793	*flgS*	signal transduction histidine kinase	T	55	90	26	237	61	469	Accessory
Cj0792	-	hypothetical protein	S	55	90	26	250	62	483	Accessory
Cj0791c	*csdA*	putative aminotransferase	E	55	90	26	255	63	489	Core
Cj0783	*napB*	periplasmic nitrate reductase small subunit (cytochrome C-type protein)	C	55	90	26	255	64	490	Core
Cj0780	*napA*	periplasmic nitrate reductase	C	55	90	26	254	64	489	Core
Cj0777	*rep*	putative ATP-dependent DNA helicase	L	55	90	26	255	64	490	Core
Cj0776c	-	putative periplasmic protein	-	55	90	26	255	64	490	Core
Cj0775c	*valS*	valyl-tRNA synthetase	J	55	90	26	255	64	490	Core

Continued on next page

Table A.4 Continued from previous page

Reference Gene Name	Alias	Predicted Function	COG Family	# Cattle	# Chicken	# Pig	# Generalist	# Others	Total	Core/ Accessory
Cj0772c	*metQ*	putative NLPA family lipoprotein	P	55	90	26	255	64	490	Core
Cj0771c	*metQ*	putative NLPA family lipoprotein	M	55	90	26	255	64	490	Core
Cj0763c	*cysE*	serine acetyltransferase	E	55	90	26	255	64	490	Core
Cj0762c	*aspC*	aspartate aminotransferase	E	55	90	26	255	64	490	Core
Cj0757	*hrcA*	putative heat shock regulator	K	36	74	0	195	15	320	Accessory
Cj0755	*cfrA*	ferric receptor CfrA	P	26	55	0	144	1	226	Accessory
Cj0753c	*tonB*	energy transducer TonB	M	36	73	0	192	12	313	Accessory
Cj0752	-	ISCco1, transposase orfB	-	29	65	0	176	12	282	Accessory
Cj0718	*dnaE*	DNA polymerase III, alpha chain	L	55	90	26	255	64	490	Core
Cj0717	*spxA*	putative ArsC family protein	P	55	90	26	255	64	490	Core
Cj0716	*aroF*	putative phospho-2-dehydro-3-deoxyheptonate aldolase	E	55	90	26	255	64	490	Core
Cj0715	*uraH*	transthyretin-like periplasmic protein	S	55	90	26	255	64	490	Core
Cj0714	*rplS*	50S ribosomal protein L19	J	55	90	26	255	64	490	Core
Cj0713	*trmD*	tRNA (guanine-N1)-methyltransferase	J	55	90	26	255	64	490	Core
Cj0712	*rimM*	putative 16S rRNA processing protein	J	55	89	26	254	63	487	Core
Cj0710	*rpsP*	30S ribosomal protein S16	J	55	90	26	255	64	490	Core
Cj0709	*ffh*	signal recognition particle protein	U	55	90	26	255	64	490	Core
Cj0535	*oorD*	OORD subunit of 2-oxoglutarate:acceptor oxidoreductase	C	55	90	26	255	64	490	Core
Cj0534	*sucD*	succinyl-coA synthetase alpha chain	C	55	90	26	255	64	490	Core
Cj0532	*mdh*	malate dehydrogenase	C	55	90	26	255	64	490	Core
Cj0495	-	putative methyltransferase domain protein	S	55	90	26	255	64	490	Core
Cj0494	-	putative exporting protein	-	46	53	25	72	25	221	Accessory
Cj0493	*fusA*	elongation factor G	J	55	90	26	255	64	490	Core
Cj0492	*rpsG*	30S ribosomal protein S7	J	55	90	26	255	64	490	Core
Cj0491	*rpsL*	30S ribosomal protein S12	J	55	90	26	255	64	490	Core
Cj0479	*rpoC*	DNA-directed RNA polymerase beta' chain	K	55	90	26	255	64	490	Core
Cj0478	*rpoB*	DNA-directed RNA polymerase beta chain	K	55	90	26	255	64	490	Core
Cj0464	*recG*	ATP-dependent DNA helicase	L	55	89	26	250	63	483	Accessory
Cj0463	*ymxG*	zinc protease-like protein	S	55	90	26	255	64	490	Core

Continued on next page

Table A.4 Continued from previous page

Reference Gene Name	Alias	Predicted Function	COG Family	# Cattle	# Chicken	# Pig	# Generalist	# Others	Total	Core/ Accessory
Cj0462	*mqnC*	putative radical SAM domain protein	H	55	90	26	255	64	490	Core
Cj0461c	*bacE*	putative MFS (Major Facilitator Superfamily) transport protein	EGP	55	90	26	255	64	490	Core
Cj0460	*nusA*	transcription termination factor	K	55	90	26	255	64	490	Core
Cj0459c	-	hypothetical protein	-	55	90	26	255	64	490	Core
Cj0458c	*miaB*	putative tRNA 2-methylthioadenosine synthase	J	55	90	25	255	64	489	Core
Cj0456c	-	hypothetical protein	-	55	90	26	255	64	490	Core
Cj0455c	*pilN*	hypothetical protein	NU	55	90	26	255	64	490	Core
Cj0454c	-	putative membrane protein	-	55	90	26	255	64	490	Core
Cj0453	*thiC*	thiamin biosynthesis protein ThiC	H	3	0	0	3	0	6	Accessory
Cj0451	*rpe*	Ribulose-phosphate 3-epimerase	G	55	90	26	255	64	490	Core
Cj0450c	*rpmB*	50S ribosomal protein L28	J	55	90	26	255	64	490	Core
Cj0449c	*ydcH*	hypothetical protein	S	55	90	26	255	64	490	Core
Cj0448c	-	putative MCP-type signal transduction protein	NT	21	24	26	81	58	210	Accessory
Cj0444	*cirA_3*	TonB-dependent receptor	P	35	66	0	174	6	281	Accessory
Cj0444	*cirA_3*	Ferric receptor CfrA	P	35	53	0	140	0	228	Accessory
Cj0437	*frdA*	succinate dehydrogenase flavoprotein subunit	C	55	78	26	249	62	470	Accessory
Cj0435	-	3-oxoacyl-[acyl-carrier protein] reductase	IQ	55	90	26	255	64	490	Core
Cj0434	*gpmI*	2,3-bisphosphoglycerate-independent phosphoglycerate mutase	G	55	90	26	255	64	490	Core
Cj0432c	*murD*	UDP-N-acetylmuramoylalanine–D-glutamate ligase	M	55	90	26	255	64	490	Core
NCTC13261_01757	*rplV*	ribosomal protein L22	J	55	90	26	255	64	490	Core
Cj1701c	*rpsC*	30S ribosomal protein S3	J	55	90	26	255	64	490	Core
Cj1688c	*secY*	preprotein translocase subunit	U	55	90	26	255	64	490	Core
Cj1687	-	putative efflux protein	EGP	37	77	4	196	50	364	Accessory
Cj1687	-	Major Facilitator Superfamily protein	EGP	54	14	25	82	19	194	Accessory
Cj1686c	*topA*	DNA topoisomerase I	L	55	90	26	255	64	490	Core
Cj1685c	*bioB*	putative biotin synthase	H	55	90	26	255	64	490	Core
Cj1681c	*cysQ*	CysQ protein	P	55	90	26	255	64	490	Core

Continued on next page

Table A.4 Continued from previous page

Reference Gene Name	Alias	Predicted Function	COG Family	# Cattle	# Chicken	# Pig	# Generalist	# Others	Total	Core/ Accessory
Cj1673c	*recA*	recA protein	L	55	90	26	255	64	490	Core
Cj1672c	*eno*	enolase	G	55	90	26	255	64	490	Core
Cj1670c	*cgpA*	putative periplasmic protein	S	55	90	26	255	64	490	Core
Cj1669c	*lig*	putative ATP-dependent DNA ligase	L	55	90	26	255	64	490	Core
NCTC13261_01720	-	integrase	L	35	20	0	1	4	60	Accessory
NCTC13261_01719	*HicA*	HicA	-	35	20	0	0	4	59	Accessory
NCTC13261_01718	-	hypothetical protein	N	35	20	0	0	4	59	Accessory
NCTC13261_01717	*HicB*	HicB	S	34	14	0	1	4	53	Accessory
NCTC13261_01716	-	putative protein	-	35	0	0	0	0	35	Accessory
NCTC13261_01715	*yafQ*	putative RelE/StbE family addiction module toxin	-	35	0	0	0	0	35	Accessory
NCTC13261_01714	-	helix-turn-helix domain-containing protein	-	35	19	0	1	4	59	Accessory
NCTC13261_01713	-	hypothetical protein	-	35	0	0	1	0	36	Accessory
NCTC13261_01712	-	hypothetical protein	-	34	7	0	1	4	46	Accessory
NCTC13261_01711	*dnaG*	DnaB-like protein helicase-like protein	L	30	19	0	0	4	53	Accessory
NCTC13261_01710	-	hypothetical protein	-	35	0	0	0	0	35	Accessory
NCTC13261_01709	-	acyl carrier protein	K	34	0	0	0	0	34	Accessory
NCTC13261_01708	-	hypothetical protein	-	35	0	0	0	0	35	Accessory
NCTC13261_01707	-	hypothetical protein	-	35	0	0	0	0	35	Accessory
NCTC13261_01706	-	RelE/ParE family plasmid stabilization system protein	S	35	20	0	0	4	59	Accessory
NCTC13261_01705	-	putative periplasmic protein	-	35	38	0	193	43	309	Accessory
Cj0313	CP _0860	putative integral membrane protein	S	55	90	26	255	64	490	Core
Cj0314	*lysA*	diaminopimelate decarboxylase	E	55	90	26	255	64	490	Core
Cj0316	*pheA*	chorismate mutase/prephenate dehydratase	E	55	90	26	255	64	490	Core
Cj0317	*hisC*	histidinol-phosphate aminotransferase	E	55	90	26	255	64	490	Core
Cj0318	*fliF*	flagellar M-ring protein	N	55	90	26	255	64	490	Core
Cj0360	*glmM*	phosphoglucosamine mutase	G	55	90	26	255	64	490	Core
Cj0392c	*pyk*	pyruvate kinase I	G	55	90	25	255	64	489	Core
Cj0393c	*mqo*	putative malate:quinone oxidoreductase	C	55	90	26	255	64	490	Core
Cj0394c	*coaX*	putative transcriptional activator	F	55	90	26	255	64	490	Core

Continued on next page

Table A.4 Continued from previous page

Reference Gene Name	Alias	Predicted Function	COG Family	# Cattle	# Chicken	# Pig	# Generalist	# Others	Total	Core/ Accessory
Cj0396c	*pgbB*	putative lipoprotein	-	55	90	26	255	63	489	Core
Cj0397c	-	hypothetical protein	-	55	89	26	252	64	486	Core
Cj0398	*gatC*	Glu-tRNAGln amidotransferase subunit C	J	55	90	26	255	64	490	Core
Cj0399	*cvpA*	cvpA family protein	S	55	90	26	255	64	490	Core
Cj0401	*lysS*	lysyl-tRNA synthetase	J	55	90	26	255	64	490	Core
Cj0404	-	putative transmembrane protein	S	55	90	26	255	64	490	Core
Cj0405	*aroE*	shikimate 5-dehydrogenase	E	55	90	26	255	64	490	Core
Cj0406c	-	lipoprotein, putative	-	55	90	25	255	59	484	Accessory
Cj0418c	Cj0418c	hypothetical protein	M	54	34	26	93	41	248	Accessory
Cj0420	Cj0420	putative periplasmic protein	S	55	90	26	255	64	490	Core
Cj0421c	Cj0421c	putative integral membrane protein	-	55	90	26	255	64	490	Core
Cj0422c	-	putative H-T-H containing protein	-	36	85	26	255	49	451	Accessory
NCTC13261_00426	-	integral membrane protein	-	35	0	23	44	14	116	Accessory
NCTC13261_00427	-	lipoprotein, putative	-	35	0	23	44	15	117	Accessory
NCTC13261_00428	-	Integral membrane protein	-	26	0	23	30	11	90	Accessory
Cj0426	*ybiT*	putative ABC transporter ATP-binding protein	S	55	90	26	255	64	490	Core
Cj0427	-	hypothetical protein	-	55	90	26	255	64	490	Core
Cj0428	-	hypothetical protein	-	55	90	26	255	64	490	Core
Cj0429c	*yigZ*	hypothetical protein	S	55	90	26	255	64	490	Core
Cj0430	-	putative protein, PMT family	M	55	90	26	255	64	490	Core
Cj0431	-	hypothetical protein	NU	55	90	26	255	64	490	Core
Cj0129c	*bamA*	outer membrane protein assembly complex, YaeT protein	M	55	90	26	255	64	490	Core
Cj0127c	*accD*	acetyl-coenzyme A carboxylase carboxyl transferase subunit beta	I	55	90	26	255	64	490	Core
Cj0105	*atpA*	ATP synthase F1 sector alpha subunit	C	55	90	26	255	64	490	Core
Cj0100	*parA*	parA family protein	D	55	90	26	255	64	490	Core
Cj0099	*birA*	putative biotin–[acetyl-CoA-carboxylase] synthetase	H	55	90	26	255	64	490	Core
Cj0098	*fmt*	methionyl-tRNA formyltransferase	J	55	90	26	255	64	490	Core
Cj0096	*obg*	GTP-binding protein, GTP1/Obg family	S	55	90	26	255	64	490	Core
Cj0093	-	putative periplasmic protein	M	55	88	26	214	42	425	Accessory

Continued on next page

Table A.4 Continued from previous page

Reference Gene Name	Alias	Predicted Function	COG Family	# Cattle	# Chicken	# Pig	# Generalist	# Others	Total	Core/ Accessory
Cj0089	-	putative lipoprotein	S	55	90	26	255	64	490	Core
Cj0088	*dcuA*	anaerobic C4-dicarboxylate transporter	U	55	88	26	221	62	452	Accessory
Cj0087	*aspA*	aspartate ammonia-lyase	E	55	90	26	255	64	490	Core
Cj0086c	*ung*	uracil-DNA glycosylase	L	55	90	26	255	64	490	Core
Cj0085c	*racD*	putative amino acid recemase	M	36	90	26	187	28	367	Accessory
Cj0082	*cydB*	cytochrome bd oxidase subunit II	C	55	90	26	255	64	490	Core
Cj0076c	*lldP*	L-lactate permease	C	54	87	1	254	64	460	Accessory
Cj0197c	*dapB*	dihydrodipicolinate reductase	E	55	90	26	255	64	490	Core
Cj0196c	*purF*	amidophosphoribosyltransferase	F	55	90	26	255	64	490	Core
Cj0195	*fliI*	flagellum-specific ATP synthase	NU	55	90	26	255	64	490	Core
Cj0193c	*tig*	trigger factor (peptidyl-prolyl cis /trans isomerase, chaperone)	D	55	90	26	255	64	490	Core
Cj0192c	*clpP*	ATP-dependent Clp protease proteolytic subunit	O	55	90	26	255	64	490	Core
Cj0191c	*def*	polypeptide deformylase	J	55	90	26	255	64	490	Core
Cj0189c	-	hypothetical protein	S	55	90	26	255	64	490	Core
Cj0188c	*nnrD*	putative kinase	-	55	90	26	255	64	490	Core
Cj0184c	-	Ser/Thr protein phosphatase family protein	T	55	90	26	255	64	490	Core
Cj0183	*corC*	putative integral membrane protein with haemolysin domain protein	S	55	90	26	255	64	490	Core
Cj0182	*sbmA*	putative transmembrane transport protein	I	55	76	26	255	29	441	Accessory
Cj0179	*exbB*	biopolymer transport protein	U	36	64	0	194	14	308	Accessory
Cj0178	-	putative TonB-denpendent outer membrane receptor	P	34	9	0	58	6	107	Accessory
Cj0178	-	putative TonB-denpendent outer membrane receptor	P	23	3	0	24	3	53	Accessory

End of table

A.5.2 Overview of chicken associated accessory genes and core gene variants identified by k-mer mapping.

Overview of chicken associated accessory genes and core gene variants identified by *k-mer* mapping. The table was created and published in [154] during this work.

Reference Gene Name	Alias	Predicted Function	COG Family	# Cattle	# Chicken	# Pig	# Generalist	# Others	Total	Core/ Accessory
Cj1033	*cmeF*	integral membrane component of efflux system (multidrug efflux system CmeDEF)	V	54	85	26	220	60	445	Accessory
NCTC13265_01618	*traG*	conjugal transfer protein TraG	U	1	59	1	1	13	75	Accessory
NCTC13265_01619	-	TraG-like protein	-	0	16	0	0	0	16	Accessory
NCTC13265_01620	-	TraG-like protein	-	0	17	0	0	0	17	Accessory
NCTC13265_01623	-	putative protein	-	0	35	0	0	5	40	Accessory
NCTC13265_01624	-	death-on-curing family protein	-	0	35	0	0	5	40	Accessory
LR59_01905	*doc*	death-on-curing family protein	S	0	30	0	0	5	35	Accessory
NCTC13265_01627	-	putative protein	-	0	66	3	1	7	77	Accessory
NCTC13265_01633	-	putative protein	-	0	68	3	0	0	71	Accessory
Cj0976	*cmoB*	putative methyltransferase	J	20	63	24	109	37	253	Accessory
NCTC13265_01672	-	putative periplasmic protein	U	19	20	0	21	14	74	Accessory
Cj0933c	*pycB*	putative pyruvate carboxylase B subunit	C	55	90	26	255	64	490	Core
Cj0912c	*cysK*	cysteine synthase	E	52	88	26	249	64	479	Accessory
Cj0861c	*trpG*	glutamine amidotransferase	EH	55	90	26	242	57	470	Accessory
Cj0849c	-	FIG00469420: hypothetical protein	-	0	65	26	14	9	114	Accessory
Cj0780	*napA*	periplasmic nitrate reductase	C	55	90	26	254	64	489	Core
Cj0652	*pbpC*	penicillin-binding protein 2	M	55	90	26	255	64	490	Core
Cj0651	-	putative integral membrane protein	-	55	90	26	255	64	490	Core
Cj0577c	*queA*	S-adenosylmethionine:tRNA ribosyltransferase-isomerase	F	55	90	26	255	64	490	Core
Cj0528c	*flgB*	flagellar basal-body rod protein	N	55	90	26	255	64	490	Core
Cj0509c	*clpB*	ATP-dependent chaperone protein ClpB	O	55	90	26	255	64	490	Core
Cj0508	*pbpA*	penicillin-binding protein	M	55	90	26	255	64	490	Core
Cj0507	*maf*	Maf-like protein	D	55	90	26	255	64	490	Core
Cj0497	-	ATP-dependent nuclease subunit B	S	55	90	26	255	64	490	Core
Cj0496	-	hypothetical protein	S	55	90	26	255	64	490	Core
Cj0495	-	putative methyltransferase domain protein	S	55	90	26	255	64	490	Core
Cj0493	*fusA*	elongation factor G	J	55	90	26	255	64	490	Core
Cj0492	*rpsG*	30S ribosomal protein S7	J	55	90	26	255	64	490	Core

Continued on next page

Table A.5 Continued from previous page

Reference Gene Name	Alias	Predicted Function	COG Family	# Cattle	# Chicken	# Pig	# Generalist	# Others	Total	Core/ Accessory
Cj0479	*rpoC*	DNA-directed RNA polymerase beta' chain	K	55	90	26	255	64	490	Core
Cj0478	*rpoB*	DNA-directed RNA polymerase beta chain	K	55	90	26	255	64	490	Core
Cj0444	*cirA_3*	TonB-dependent receptor, putative, degenerate	P	35	53	0	140	0	228	Accessory
Cj1623	-	putative membrane protein	-	55	90	26	255	64	490	Core
Cj1633	*tilS*	thiamine biosynthesis protein:ExsB	D	55	90	26	255	64	490	Core
Cj0391c	-	hypothetical protein	-	55	90	25	255	64	489	Core
NCTC13265_00510	*bioC*	hypothetical protein	S	14	31	0	0	0	45	Accessory
Cj0131	-	putative peptidase M23 family protein	M	55	90	26	255	64	490	Core
Cj0662c	*hslU*	ATP-dependent Hsl protease ATP-binding subunit	O	55	90	26	255	64	490	Core
Cj0718	*dnaE*	DNA polymerase III, alpha chain	L	55	90	26	255	64	490	Core
Cj1481c	*addA*	putative helicase	-	55	90	26	255	64	490	Core
Cj1478c	*oprF*	outer membrane fibronectin-binding protein	M	55	90	26	255	64	490	Core
Cj1477c	*ppaX*	putative hydrolase	S	55	90	26	255	64	490	Core
NCTC13265_01596	-	hypothetical protein	I	0	20	0	0	0	20	Accessory

End of table

A.5.3 Overview of pig associated accessory genes and core gene variants identified by *k-mer* mapping.

Overview of pig associated accessory genes and core gene variants identified by k-mer mapping. The table was created and published in [154] during this work.

Reference Gene Name	Alias	Predicted Function	COG Family	# Cattle	# Chicken	# Pig	# Generalist	# Others	Total	Core/ Accessory
Cj1019c	*livJ*	branched-chain amino-acid ABC transport system periplasmic binding protein	E	55	90	26	255	64	490	Core
Cj1033	*cmeF*	integral membrane component of efflux system (multidrug efflux system CmeDEF)	V	54	85	26	220	60	445	Accessory
Cj1034c	-	putative DnaJ-like protein	O	55	90	26	255	64	490	Core
Cj1037c	*pycA*	acetyl-CoA carboxylase, biotin carboxylase	I	55	90	26	255	64	490	Core
Cj1039	*murG*	putative undecaprenyldiphospho-muramoylpentapeptide b-N-acetylglucosaminyltransferase	M	53	90	26	255	64	488	Core
Cj1040c	*yeaN*	membrane protein, putative	P	0	0	25	0	10	35	Accessory
Cj1040c	*yeaN*	putative transmembrane transport protein	P	1	0	26	14	0	41	Accessory
Cj1040c	*yeaN*	Putative transmembrane transport protein	P	0	0	26	0	4	30	Accessory
Cj1043c	*tenI*	putative thiamine-phosphate pyrophosphorylase	H	55	90	26	255	64	490	Core
Cj1044c	*thiH*	thiazole biosynthesis protein ThiH	C	55	90	26	255	64	490	Core
A6J90_06670	Z012 _06150	Type II restriction endonuclease	L	0	0	26	0	1	27	Accessory
A6J90_06675	*dcm*	cytosine-specific methyltransferase NlaX	H	0	0	26	0	1	27	Accessory
Cj1045c	*thiG*	Thiazole biosynthesis protein ThiG	H	55	90	26	255	64	490	Core
Cj1052c	*mutS2*	putative mismatch repair protein	L	55	77	26	255	62	475	Accessory
Cj1097	*sstT*	putative transmembrane transport protein	E	49	88	25	236	48	446	Accessory
Cj1124c	*pglA*	GalNAc transferase	M	55	86	26	253	64	484	Accessory
Cj1133	*waaC*	heptosyltransferase I	M	55	90	26	255	64	490	Core
Cj1210	*yohD*	putative integral membrane protein	S	55	90	26	252	64	487	Core
Cj1211	*comEC*	putative competence family protein	S	54	87	25	231	60	457	Accessory
Cj1213c	*glcD*	putative glycolate oxidase subunit D	C	55	90	26	255	62	488	Core
Cj1218c	*ribE*	putative riboflavin synthase alpha chain	H	55	90	26	255	64	490	Core
Cj1219c	*ytfN*	putative periplasmic protein	S	55	90	26	255	64	490	Core

Continued on next page

Table A.6 Continued from previous page

Reference Gene Name	Alias	Predicted Function	COG Family	# Cattle	# Chicken	# Pig	# Generalist	# Others	Total	Core/ Accessory
Cj1221	*groL*	60 kD chaperonin (cpn60)	O	55	90	26	255	64	490	Core
Cj1226c	*cprS*	putative two-component sensor (histidine kinase)	T	55	90	26	254	64	489	Core
Cj1228c	*htrA*	serine protease (protease DO)	M	55	90	26	255	64	490	Core
Cj1231	*kefB*	putative glutathione-regulated potassium-efflux system protein	P	55	89	26	254	64	488	Core
Cj1253	*pnp*	polynucleotide phosphorylase	J	55	90	25	253	64	487	Core
Cj1257c	*mdtG*	putative efflux pump	EGP	54	88	26	245	57	470	Accessory
Cj1283	*ktrB*	putative K+ uptake protein	P	55	90	26	255	64	490	Core
Cj1290c	*accC*	biotin carboxylase	I	55	90	26	255	64	490	Core
Cj1311	*pseF*	putative acylneuraminate cytidylyltransferase	M	55	90	26	254	62	487	Core
Cj1312	*pseG*	UDP-2,4-diacetamido-2,4,6-trideoxy-beta-L- altropyranose hydrolase	M	55	72	26	251	64	468	Accessory
Cj0719c	*yggS*	hypothetical protein	S	55	90	26	255	64	490	Core
Cj0718	*dnaE*	DNA polymerase III, alpha chain	L	55	90	26	255	64	490	Core
Cj0709	*ffh*	signal recognition particle protein	U	55	90	26	255	64	490	Core
Cj0686	*ispG*	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase	I	55	90	26	255	64	490	Core
Cj0685c	-	invasion phenotype protein	S	15	39	10	58	13	135	Accessory
A6J90_04300	-	invasion phenotype protein	S	55	85	26	199	59	424	Accessory
Cj0684	*priA*	putative primosomal protein N'	L	55	90	26	255	64	490	Core
Cj0608	*cusC*	putative outer membrane efflux protein	MU	55	90	26	255	64	490	Core
Cj0607	*macB*	ABC-type transmembrane transport protein	V	35	55	21	165	28	304	Accessory
Cj0606	*macA*	amidohydrolase	M	35	55	21	163	24	298	Accessory
Cj0605	-	putative amidohydrolase	E	55	90	22	254	64	485	Accessory
Cj0597	*fbaA*	Fructose-bisphosphate aldolase class II	G	55	90	26	255	64	490	Core
Cj0586	*ligA*	DNA ligase	L	55	90	26	255	64	490	Core
Cj0574	*ilvI*	acetolactate synthase large subunit	H	55	90	26	255	64	490	Core
Cj0557c	-	putative integral membrane protein	S	55	85	24	254	59	477	Accessory
Cj0551	*efp*	elongation factor P	J	55	90	26	255	64	490	Core

Continued on next page

Table A.6 Continued from previous page

Reference Gene Name	Alias	Predicted Function	COG Family	# Cattle	# Chicken	# Pig	# Generalist	# Others	Total	Core/ Accessory
Cj0545	*ubiD*	putative 3-octaprenyl-4-hydroxybenzoate carboxy-lyase	H	55	90	26	255	64	490	Core
Cj0531	*icd*	isocitrate dehydrogenase, NADP-dependent	C	55	90	26	255	64	490	Core
Cj0530	-	putative periplasmic protein	M	55	90	26	254	63	488	Core
Cj0525c	*ftsI*	putative penicillin-binding protein	M	55	90	26	255	64	490	Core
Cj0518	*htpG*	hsp90 family heat shock protein	O	55	90	26	255	64	490	Core
Cj0506	*alaS*	alanyl-tRNA synthetase	J	55	90	26	255	64	490	Core
Cj0503c	*hemH*	ferrochelatase	H	55	90	26	255	64	490	Core
Cj0500	*selU*	tRNA 2-selenouridine synthase	H	31	24	18	83	59	215	Accessory
Cj0499	*hit*	putative histidine triad (HIT) family protein	FG	31	25	18	84	59	217	Accessory
Cj0479	*rpoC*	DNA-directed RNA polymerase beta' chain	K	55	90	26	255	64	490	Core
Cj0465c	*ctb*	group III truncated haemoglobin	S	55	90	26	255	64	490	Core
Cj0464	*recG*	ATP-dependent DNA helicase	L	55	89	26	250	63	483	Accessory
Cj0463	*ymxG*	zinc protease-like protein	S	55	90	26	255	64	490	Core
Cj0462	*mqnC*	putative radical SAM domain protein	H	55	90	26	255	64	490	Core
Cj0444	*cirA_3*	TonB-dependent receptor, putative, degenerate	P	55	89	26	254	63	487	Core
Cj1576c	*nuoD*	NADH dehydrogenase I chain D	C	55	90	26	255	64	490	Core
A6J90_00190	-	putative protein	-	0	0	25	0	0	25	Accessory
A6J90_00195	-	FIG00470712: hypothetical protein	S	0	0	26	0	0	26	Accessory
A6J90_00200	-	FIG00471113: hypothetical protein	-	1	0	26	0	0	27	Accessory
Cj0571	-	transcriptional regulator	K	0	0	18	0	0	18	Accessory
A6J90_00270	-	putative protein	-	0	0	26	0	0	26	Accessory
A6J90_00275	*ccrM*	DNA methylase	L	0	0	26	0	0	26	Accessory
Cj1612	*prfA*	peptide chain release factor 1	J	55	90	26	255	64	490	Core
Cj1630	*tonB*	putative TonB transport protein	U	17	19	25	168	13	242	Accessory
Cj1631c	-	FIG00469530: hypothetical protein	-	53	90	26	255	54	478	Accessory
Cj1633	*tilS*	thiamine biosynthesis protein:ExsB	D	55	90	26	255	64	490	Core
Cj1634c	*aroC*	chorismate synthase	E	55	90	26	255	64	490	Core

Continued on next page

Table A.6 Continued from previous page

Reference Gene Name	Alias	Predicted Function	COG Family	# Cattle	# Chicken	# Pig	# Generalist	# Others	Total	Core/ Accessory
Cj0274	*lpxA*	acyl-[acyl-carrier-protein]–UDP-N-acetylglucosamine O-acyltransferase	M	55	90	26	255	64	490	Core
Cj0279	*carB*	carbamoyl-phosphate synthase large chain	F	55	90	26	255	64	490	Core
A6J90_02340	-	putative protein	-	0	0	25	0	0	25	Accessory
A6J90_02350	-	R Pab1 restriction endonuclease	L	0	0	25	0	0	25	Accessory
A6J90_02350	*sua5*	hypothetical protein	J	0	0	26	0	0	26	Accessory
A6J90_02420	*pgtP*	MFS transporter%2C OPA family%2C phosphoglycerate transporter protein	G	33	29	26	104	52	244	Accessory
Cj0293	*surE*	multifunctional protein SurE	S	55	90	26	255	64	490	Core
Cj0300c	*modC*	putative molybdenum transport ATP-binding protein	P	40	61	26	255	54	436	Accessory
Cj0321	*dxs*	L-deoxy-D-xylulose-5-phosphate synthase	H	55	90	26	254	64	489	Core
Cj0328c	*fabH*	3-oxoacyl-[acyl-carrier-protein] synthase	I	55	90	26	255	64	490	Core
Cj0329c	*plsX*	putative fatty acid/phospholipid synthesis protein	I	55	90	26	255	64	490	Core
Cj0352	*fliZ*	putative transmembrane protein	N	55	90	26	255	64	490	Core
Cj0356c	*folB*	putative dihydroneopterin aldolase	H	55	90	26	255	64	490	Core
Cj0428	-	hypothetical protein	-	55	90	26	255	64	490	Core
A6J90_00035	*mloA*	MloA protein, putative	S	2	18	12	8	0	40	Accessory
Cj1543	*kipA*	hypothetical protein	E	52	86	20	235	52	445	Accessory
A6J90_08990	*hsdR*	type I restriction enzyme EcoR124II R protein	V	0	1	26	0	0	27	Accessory
Cj1500	*yedE*	putative integral membrane protein	S	55	90	26	255	64	490	Core
Cj1484c	-	putative membrane protein	-	55	90	25	255	64	489	Core
Cj1482c	*addB*	hypothetical protein	L	55	90	26	255	64	490	Core
Cj1481c	*addA*	putative helicase	L	55	90	26	255	64	490	Core
Cj1476c	*nifJ*	pyruvate-flavodoxin oxidoreductase	C	55	90	26	255	64	490	Core
Cj1457c	*truD*	tRNA pseudouridine synthase D	J	55	90	26	255	64	490	Core
Cj0574	*ilvA*	threonine dehydratase biosynthetic	E	55	90	26	255	64	490	Core
Cj0886c	*ftsK*	putative cell division protein	D	55	90	26	255	64	490	Core
Cj0887c	*flgL*	putative flagellin	N	55	90	26	255	57	483	Accessory
Cj0891c	*serA*	D-3-phosphoglycerate dehydrogenase	E	55	90	26	255	64	490	Core

Continued on next page

Table A.6 Continued from previous page

Reference Gene Name	Alias	Predicted Function	COG Family	# Cattle	# Chicken	# Pig	# Generalist	# Others	Total	Core/ Accessory
Cj0924c	*cheB*	putative MCP protein-glutamate methylesterase	NT	55	90	26	255	64	490	Core
Cj0929	*pepA*	aminopeptidase	E	55	90	26	255	63	489	Core
Cj0930	*ychF*	putative GTP-binding protein	J	55	90	26	255	64	490	Core
Cj0931c	*argH*	argininosuccinate lyase	E	53	86	26	229	53	447	Accessory
A6J90_01640	-	hypothetical protein	-	0	0	26	0	0	26	Accessory
Cj0107	*atpD*	ATP synthase F1, beta subunit	F	55	90	26	255	64	490	Core
Cj0105	*atpA*	ATP synthase F1 sector alpha subunit	C	55	90	26	255	64	490	Core
Cj0101	*parB*	parB family protein	K	55	90	26	254	64	489	Core
A6J90_01500 A6J90_01505	-	putative protein	V	0	0	25	0	0	25	Accessory
A6J90_01490	-	putative protein	-	0	0	26	0	0	26	Accessory
Cj0077c	*cdtC*	cytolethal distending toxin%2C subunit C	S	0	0	26	4	0	30	Accessory
Cj0076c	*lldP*	L-lactate permease	C	55	90	26	255	64	490	Core
A6J90_01375	-	Non-heme iron protein, hemerythrin family	P	0	0	12	0	0	12	Accessory
A6J90_01275	-	anion transporter	P	12	77	23	113	17	242	Accessory
Cj0038c	-	putative poly(A) polymerase family protein	-	49	81	25	241	45	441	Accessory
Cj1414c	*kpsC*	capsule polysaccharide export protein KpsC	M	4	0	26	2	1	33	Accessory
Cj0943	*lolA*	putative outer-membrane lipoprotein carrier protein precursor	M	55	90	26	255	64	490	Core
Cj0944c	-	putative periplasmic protein	-	55	90	25	255	62	487	Core
Cj0945c	-	putative helicase	L	55	90	23	252	64	484	Accessory
Cj0946	*spr7*	putative lipoprotein	M	55	90	26	255	64	490	Core
Cj0995c	*hemB*	delta-aminolevulinic acid dehydratase	H	55	90	26	255	64	490	Core
Cj1011	*corA*	putative CorA-like Mg2+ transporter protein	P	55	90	26	255	64	490	Core
Cj0773c	*metI*	putative ABC transport system permease protein	P	55	90	26	255	64	490	Core
A6J90_04615	-	hypothetical protein	-	17	14	24	57	49	161	Accessory
Cj0796c	*mhpC*	putative hydrolase	S	55	90	26	255	64	490	Core

End of table

A.5.4 Overview of host-generalist associated accessory genes and core gene variants identified by *k-mer* mapping.

Overview of host-generalist associated accessory genes and core gene variants identified by k-mer mapping. The table was created and published in [154] during this work.

Reference Gene Name	Alias	Predicted Function	COG Family	# Cattle	# Chicken	# Pig	# Generalist	# Others	Total	Core/ Accessory
Cj1342c	-	A member of the 617 family of C.j. proteins containing homopolymeric tracts	E	12	39	25	167	30	273	Accessory
Cj1341c	*pseE*	Protein of unknown function DUF115	S	1	33	2	146	17	199	Accessory
Cj1276c	*ftsX*	Cell division protein FtsX	D	55	90	26	255	64	490	Core
Cj1266c	*hydB*	Belongs to the NiFe NiFeSe hydrogenase large subunit family	C	55	90	26	255	64	490	Core
Cj1252	*lptD*	LPS-assembly protein of lipopolysaccharide (LPS) at the surface of the outer membrane	M	55	90	25	255	64	489	Core
Cj1250	*purD*	Belongs to the GARS family	F	55	90	26	255	64	490	Core
Cj1240c	-	-	-	55	86	21	249	64	475	Accessory
Cj1019c	*livJ*	amino acid abc transporter	E	55	90	26	255	64	490	Core
Cj1008c	*aroB*	3-dehydroquinate synthase	E	55	90	26	255	63	489	Core
Cj1001	*rpoD*	RNA polymerase sigma factor RpoD	K	55	90	26	255	64	490	Core
Cj0975	*hxuB*	Haemolysin secretion/activation protein ShlB/FhaC/HecB	U	37	90	25	195	43	390	Accessory
Cj0964	-	leucine binding	E	55	90	26	255	64	490	Core
Cj0961c	*rpmH*	Belongs to the bacterial ribosomal protein bL34 family	J	55	90	26	255	64	490	Core
Cj0960c	*rnpA*	Ribonuclease P protein component	J	55	90	26	255	64	490	Core
Cj0958c	*yidC*	membrane protein insertase	U	55	90	26	255	64	490	Core
Cj0956c	*mnmE*	tRNA modification GTPase MnmE	J	55	90	26	255	64	490	Core
Cj0945c	-	COG0507 ATP-dependent exoDNAse (exonuclease V) alpha subunit - helicase superfamily I member	L	55	90	23	252	64	484	Accessory
Cj0943	*lolA*	Participates in the translocation of lipoproteins from the inner membrane to the outer membrane	M	55	90	26	255	64	490	Core
Cj0934c	IV02 _29000	Belongs to the sodium neurotransmitter symporter (SNF) (TC 2.A.22) family	P	32	81	0	175	15	303	Accessory

Continued on next page

Table A.7 Continued from previous page

Reference Gene Name	Alias	Predicted Function	COG Family	# Cattle	# Chicken	# Pig	# Generalist	# Others	Total	Core/ Accessory
Cj0932c	*pckA*	Phosphoenolpyruvate carboxykinase (ATP)	H	55	90	26	255	64	490	Core
Cj0917c	*cstA*	Carbon starvation protein	T	52	88	26	245	63	474	Accessory
Cj0915	*yciA*	putative hydrolase	I	55	90	26	255	64	490	Core
Cj0912c	*cysM*	Belongs to the cysteine synthase cystathionine beta- synthase family	E	52	88	26	249	64	479	Accessory
Cj0911	*hyaE*	SCO1 SenC	S	55	90	26	255	64	490	Core
Cj0909	VY92 _09940	Putative periplasmic protein	S	55	89	26	252	64	486	Core
Cj0905c	*alr*	Alanine racemase	E	55	90	26	255	64	490	Core
Cj0898	*hinT*	Putative histidine triad (HIT) family protein	FG	55	90	26	254	64	489	Core
Cj0886c	*ftsK*	DNA translocase	D	55	90	26	255	64	490	Core
Cj0879c	-	-	-	38	57	4	140	57	296	Accessory
Cj0874c	*petA*	cytochrome C	C	52	73	26	239	54	444	Accessory
Cj0874c	*petA*	cytochrome C	C	0	24	0	87	7	118	Accessory
Cj0866	-	hmm pf05935	M	0	37	0	97	0	134	Accessory
Cj0865	*dsbB*	Putative protein-disulfide oxidoreductase	C	55	90	26	254	64	489	Core
Cj0835c	*acnB*	Belongs to the aconitase IPM isomerase family	C	55	90	26	255	64	490	Core
Cj0812	*thrC*	threonine synthase	E	55	80	26	255	64	480	Accessory
Cj0800c	-	Flagellar Assembly Protein A	L	55	90	26	255	64	490	Core
Cj0799c	*ruvA*	Holliday junction ATP-dependent DNA helicase	L	55	90	26	255	64	490	Core
Cj0798c	*ddl*	Belongs to the D-alanine–D-alanine ligase family	F	55	90	26	255	64	490	Core
Cj0794	-	Annotation was generated automatically without manual curation	S	48	76	19	174	22	339	Accessory
Cj0791c	*csdA*	Aminotransferase	E	55	90	26	255	63	489	Core
Cj0780	*napA*	Periplasmic nitrate reductase	C	55	90	26	254	64	489	Core
Cj0776c	-	-	-	55	90	26	255	64	490	Core
Cj0775c	*valS*	Valine–tRNA ligase	J	55	90	26	255	64	490	Core
Cj0737	-	haemagglutination activity domain	U	51	77	15	195	34	372	Accessory

Continued on next page

Table A.7 Continued from previous page

Reference Gene Name	Alias	Predicted Function	COG Family	# Cattle	# Chicken	# Pig	# Generalist	# Others	Total	Core/ Accessory
Cj0718	*dnaE*	DNA polymerase	L	55	90	26	255	64	490	Core
Cj0717	*spxA*	Belongs to the ArsC family	P	55	90	26	255	64	490	Core
Cj0710	*rpsP*	Belongs to the bacterial ribosomal protein bS16 family	J	55	90	26	255	64	490	Core
Cj0709	ffh	signal recognition particle protein	U	55	90	26	255	64	490	Core
Cj0703	-	Protein of unknown function (DUF3972)	S	55	90	26	255	64	490	Core
Cj0700	-	-	-	55	90	26	255	64	490	Core
Cj0699c	*glnA*	glutamine synthetase	E	55	90	26	255	64	490	Core
Cj0684	*priA*	Primosomal protein N', involved in the restart of stalled replication forks.	L	55	90	26	255	64	490	Core
Cj0680c	*uvrB*	The UvrABC repair system catalyzes the recognition and processing of DNA lesions	L	55	90	26	255	64	490	Core
Cj0628	-	Autotransporter beta-domain	S	0	21	0	13	0	34	Accessory
Cj0506	*alaS*	Alanine–tRNA ligase	J	55	90	26	255	64	490	Core
Cj0496	-	Uncharacterised protein family (UPF0153)	S	55	90	26	255	64	490	Core
Cj0493	*fusA*	Elongation factor G	J	55	90	26	255	64	490	Core
Cj0492	*rpsG*	30S ribosomal protein S7	J	55	90	26	255	64	490	Core
Cj0491	*rpsL*	30S ribosomal protein S12	J	55	90	26	255	64	490	Core
Cj0490	*aldA*	Belongs to the aldehyde dehydrogenase family	C	0	36	0	177	0	213	Accessory
Cj0483	*uxaA*	Altronate hydrolase	G	0	57	0	195	2	254	Accessory
Cj0480c	-	Transcriptional regulator	K	0	57	0	194	0	251	Accessory
Cj0479	*rpoC*	DNA-directed RNA polymerase subunit beta'	K	55	90	26	255	64	490	Core
Cj0478	*rpoB*	DNA-directed RNA polymerase subunit beta	K	55	90	26	255	64	490	Core
Cj0477	*rplL*	50S ribosomal protein L7/L12	J	55	90	26	255	64	490	Core
Cj0475	*rplA*	50S ribosomal protein L1	J	55	90	26	255	64	490	Core
Cj0470	*tuf*	Elongation factor	J	55	90	26	255	64	490	Core
Cj0464	*recG*	ATP-dependent DNA helicase	L	55	89	26	250	63	483	Accessory
Cj0463	*ymxG*	Peptidase, M16	S	55	90	26	255	64	490	Core
Cj0461c	*bacE*	Major facilitator Superfamily	EGP	55	90	26	255	64	490	Core

Continued on next page

Table A.7 Continued from previous page

Reference Gene Name	Alias	Predicted Function	COG Family	# Cattle	# Chicken	# Pig	# Generalist	# Others	Total	Core/ Accessory
Cj0460	*nusA*	tRNA-2-methylthio-N(6)-dimethylallyladenosine synthase	K	55	90	26	255	64	490	Core
Cj0459c	-	-	-	55	90	26	255	64	490	Core
Cj0458c	*miaB*	Catalyzes the methylthiolation of N6-(dimethylallyl)adenosine (i(6)A), leading to the formation of 2-methylthio-N6-(dimethylallyl)adenosine (ms(2)i(6)A) at position 37 in tRNAs that read codons beginning with uridine	J	55	90	25	255	64	489	Core
Cj0457c	MA20_05800	protein conserved in bacteria	S	55	90	26	255	64	490	Core
Cj0453	*thiC*	Phosphomethylpyrimidine synthase	H	54	90	26	252	64	486	Core
Cj0452	*dnaQ*	dna polymerase iii	L	53	90	26	255	64	488	Core
Cj0451	*rpe*	Belongs to the ribulose-phosphate 3-epimerase family	G	55	90	26	255	64	490	Core
Cj0444	*cirA_3*	receptor	P	35	66	0	174	6	281	Accessory
Cj0434	*gpmI*	Catalyzes the interconversion of 2-phosphoglycerate and 3-phosphoglycerate	G	55	90	26	255	64	490	Core
Cj0432c	*murD*	UDP-N-acetylmuramoylalanine–D-glutamate ligase	M	55	90	26	255	64	490	Core
Cj0431	-	general secretion pathway protein	NU	55	90	26	255	64	490	Core
Cj0431	-	integral membrane protein	M	55	90	26	255	64	490	Core
Cj0429c	*yigZ*	hmm pf01205	S	55	90	26	255	64	490	Core
Cj0428	-	-	-	55	90	26	255	64	490	Core
Cj0426	*ybiT*	abc transporter atp-binding protein	S	55	90	26	255	64	490	Core
Cj0422c	-	-	-	36	85	26	255	49	451	Accessory
Cj0404	-	Sporulation related domain	S	55	90	26	255	64	490	Core
Cj0396c	*pgbB*	-	-	55	90	26	255	63	489	Core
Cj0393c	*mqo*	Malate quinone- oxidoreductase	C	55	90	26	255	64	490	Core
Cj0362	-	membrane	S	55	90	26	255	64	490	Core
Cj0321	*dxs*	1-deoxy-D-xylulose-5-phosphate synthase	H	55	90	26	254	64	489	Core

Continued on next page

Table A.7 Continued from previous page

Reference Gene Name	Alias	Predicted Function	COG Family	# Cattle	# Chicken	# Pig	# Generalist	# Others	Total	Core/ Accessory
Cj0318	*fliF*	The M ring may be actively involved in energy transduction	N	55	90	26	255	64	490	Core
Cj0292c	*pgtP*	Catalyzes the uptake of glycerol-3-phosphate with the simultaneous export of inorganic phosphate from the cell	G	38	20	0	88	6	152	Accessory
Cj0248	-	Signal transduction protein	T	55	90	26	255	64	490	Core
Cj0247c	-	FIST N domain	NT	54	90	26	240	55	465	Accessory
Cj0196c	*purF*	Amidophosphoribosyltransferase	F	55	90	26	255	64	490	Core
Cj0192c	*clpP*	ATP-dependent Clp protease proteolytic subunit	O	55	90	26	255	64	490	Core
Cj0186c	*TerC*	Membrane protein	P	54	88	25	253	64	484	Accessory
Cj0185c	*phnA*	Zn-ribbon-containing protein involved in phosphonate metabolism	P	3	57	1	124	5	190	Accessory
Cj0183	*corC*	COG1253 Hemolysins and related proteins containing CBS domains	S	55	90	26	255	64	490	Core
Cj0182	*sbmA*	ABC transporter transmembrane region 2	I	55	76	26	255	29	441	Accessory
Cj0108	*atpC*	Produces ATP from ADP in the presence of a proton gradient across the membrane	C	55	90	26	255	64	490	Core
Cj0105	*atpA*	ATP synthase subunit alpha	C	55	90	26	255	64	490	Core
Cj0100	*parA*	involved in chromosome partitioning	D	55	90	26	255	64	490	Core
Cj0093	-	curli production assembly transport component CsgG	M	55	88	26	214	42	425	Accessory
Cj0089	-	protein conserved in bacteria	S	55	90	26	255	64	490	Core
Cj0087	*aspA*	Aspartate ammonia-lyase	E	55	90	26	255	64	490	Core
Cj0086c	*ung*	Uracil-DNA glycosylase	L	55	90	26	255	64	490	Core
Cj0082	*cydB*	cytochrome d ubiquinol oxidase, subunit II	C	55	90	26	255	64	490	Core
Cj0076c	*lldP*	L-lactate permease	C	54	87	1	254	64	460	Accessory
Cj0036	-	protein conserved in bacteria	S	53	90	26	252	62	483	Accessory
Cj1713	*rlmN*	Dual-specificity RNA methyltransferase	J	55	90	26	255	64	490	Core
Cj1703c	*rpsS*	ribosomal protein S19	J	55	90	26	255	64	490	Core
Cj1702c	*rplV*	ribosomal protein L22	J	55	90	26	255	64	490	Core

Continued on next page

Table A.7 Continued from previous page

Reference Gene Name	Alias	Predicted Function	COG Family	# Cattle	# Chicken	# Pig	# Generalist	# Others	Total	Core/ Accessory
Cj1701c	*rpsC*	30S ribosomal protein S3	J	55	90	26	255	64	490	Core
Cj1690c	*rpsE*	L30S ribosomal protein S5	J	55	90	26	255	64	490	Core
Cj0628	-	Autotransporter beta-domain	S	0	0	0	5	0	5	Accessory
Cj1669c	*lig*	DNA ligase	L	55	90	26	255	64	490	Core
Cj1474c	*ctsD*	Type II and III secretion system protein	NU	51	83	24	233	45	436	Accessory
GRN82_06795	-	Protein of unknown function (DUF2972)	S	0	3	0	20	0	23	Accessory

End of table

A.6 Identified genes interogressed from *C. jejuni* to *C. coli*

Genes with at least 10 % *C. jejuni* introgression in at least 25 of the 29 *C. coli* hybrid strains and their potential function. The table was created and published in [39] during this work.

	16-mer coverage										32-mer coverage										
Gene	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**Gene**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**Gene name (putative function)**	**Functional categories**
Cj0026c	29	29	2	2	2	1	1	0	0											*thy*X (flavin dependent thymidylate synthase; catalyzes the formation of dTMP and tetrahydrofolate from dUMP and methylenetetrahydrofolate)	DNA metabolism and repair
Cj0027	29	29	13	0	0	0	0	0	0	Cj0027	29	29	29	10	0	0	0	0	0	*pyr*G; cytidine triphosphate synthetase; catalyzes the ATP-dependent amination of UTP to CTP with either L-glutamine or ammonia as the source of nitrogen	DNA metabolism and repair
Cj0028	29	29	29	29	29	24	0	0	0	Cj0028	29	29	29	29	29	24	0	0	0	*rec*J (putative single-stranded-DNA-specific exonuclease)	DNA metabolism and repair
Cj0029	29	29	29	0	0	0	0	0	0	Cj0029	29	29	7	0	0	0	0	0	0	*ans*A (cytoplasmic L-asparaginase)	Protein synthesis, Amino Acid metabolism
Cj0059c	29	28	28	27	27	1	0	0	0	Cj0059c	29	28	27	26	1	0	0	0	0	*fli*Y (one of three proteins for switching the direction of the flagellar motor)	cell motility
Cj0063c	28	28	1	0	0	0	0	0	0	Cj0063c	28	28	1	0	0	0	0	0	0	ATP-binding protein	signal transduction
Cj0069	29	29	2	2	2	2	0	0	0											hypothetical protein	unknown
Cj0070c	29	29	29	29	29	29	29	29	0	Cj0070c	29	29	29	29	29	29	29	0	0	hypothetical protein	unknown
Cj0081	29	29	29	0	0	0	0	0	0	Cj0081	29	29	4	0	0	0	0	0	0	*cyd*A (better cioA, cyanide-insensitive oxidase)	stress response (oxidative)
Cj0085c	29	29	29	29	29	29	0	0	0	Cj0085c	29	29	29	29	29	28	0	0	0	putative amino acid racemase	Cell wall/membrane/-capsule
Cj0086c	29	28	28	3	0	0	0	0	0	Cj0086c	29	28	28	0	0	0	0	0	0	*ung* (uracil-DNA glycosylase, excises uracil residues from the DNA which can arise as a result of misincorporation of dUMP residues by DNA polymerase or due to deamination of cytosine)	DNA metabolism and repair

Continued on next page

Table A.8 Continued from previous page

Gene	16-mer coverage									Gene	32-mer coverage									Gene name (putative function)	Functional categories
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9		
Cj0087	29	29	29	2	0	0	0	0	0	Cj0087	29	29	29	28	2	0	0	0	0	*asp*A (aspartate ammonia-lyase)	Protein synthesis, Amino Acid metabolism
Cj0095	29	29	29	28	0	0	0	0	0	Cj0095	29	29	29	29	0	0	0	0	0	*rpm*A (50S ribosomal protein L27, involved in the peptidyltransferase reaction during translation)	Protein synthesis, Amino Acid metabolism
Cj0131	29	29	29	29	29	26	26	0	0	Cj0131	29	29	29	29	29	26	17	0	0	putative periplasmic protein	unknown
Cj0194	29	29	29	29	29	27	0	0	0	Cj0194	29	29	28	0	0	0	0	0	0	*fol*E (GTP cyclohydrolase I, involved in the first step of tetrahydrofolate biosynthesis)	Metabolism of cofactors and vitamins
Cj0196c	28	28	28	11	0	0	0	0	0											*pur*F (amidophosphoribosyltransferase, catalyzes first step of the de novo purine nucleotide biosynthetic pathway)	DNA metabolism and repair
Cj0197c	28	28	28	28	28	0	0	0	0	Cj0197c	28	28	28	28	18	0	0	0	0	*dap*B (4-hydroxytetrahydrodipicolinate reductase)	Protein synthesis, Amino Acid metabolism
Cj0198c	28	28	8	4	3	3	0	0	0	Cj0198c	28	28	4	3	3	0	0	0	0	recombination factor protein RarA (maintainance of genome stability)	DNA metabolism and repair
Cj0203	29	29	29	29	29	29	29	19	0	Cj0203	29	29	29	29	29	20	2	0	0	putative transmembrane transport protein	membrane transport
Cj0237	29	29	29	0	0	0	0	0	0	Cj0237	29	29	28	0	0	0	0	0	0	*can*B (carbonic anhydrase); enables growth at low CO2	stress response (oxidative)
Cj0254	28	26	25	0	0	0	0	0	0	Cj0254	29	26	25	5	0	0	0	0	0	hypothetical protein	unknown
Cj0266c	28	28	28	0	0	0	0	0	0											integral membrane protein	membrane transport
Cj0283c	29	29	29	29	29	29	29	29	29	Cj0283c	29	29	29	29	29	29	29	29	29	*che*W (chemotaxis protein)	cell motility
Cj0284c	29	29	0	0	0	0	0	0	0											*che*A (chemotaxis histidine kinase)	cell motility
Cj0285c	29	29	0	0	0	0	0	0	0	Cj0285c	29	29	0	0	0	0	0	0	0	*che*V (chemotaxis protein)	cell motility
Cj0286c	29	29	0	0	0	0	0	0	0	Cj0286c	29	29	0	0	0	0	0	0	0	hypothetical protein	unknown
Cj0379c	29	28	28	28	28	28	28	0	0	Cj0379c	28	28	28	28	28	28	7	0	0	hypothetical protein	unknown
Cj0383c	29	27	27	24	24	14	14	14	0	Cj0383c	29	27	27	24	24	24	14	14	9	*rib*H (6,7-dimethyl-8-ribityllumazine synthase), biosynthesis of riboflavin	Metabolism of cofactors and vitamins

Continued on next page

Table A.8 Continued from previous page

Gene	16-mer coverage									Gene	32-mer coverage									Gene name (putative function)	Functional categories
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9		
										Cj0387	29	29	9	9	0	0	0	0	0	*aro*K (shikimate kinase, catalyzes the formation of shikimate 3-phosphate from shikimate in aromatic amino acid biosynthesis)	Protein synthesis, Amino Acid metabolism
Cj0435	28	28	24	24	7	0	0	0	0											*fab*G (3-oxoacyl-ACP reductase, catalyzes the first of the two reduction steps in the elongation cycle of fatty acid synthesis)	Lipid metabolism
Cj0441	29	29	13	0	0	0	0	0	0	Cj0441	29	26	0	0	0	0	0	0	0	*acp*P (acyl carrier protein, carries the fatty acid chain in fatty acid biosynthesis)	Lipid metabolism
Cj0442	29	29	29	29	29	25	5	0	0	Cj0442	29	29	29	29	29	23	0	0	0	*fab*F (3-oxoacyl-[acyl-carrier-protein] synthase) (fatty acid biosynthesis)	Lipid metabolism
Cj0511	29	29	29	0	0	0	0	0	0	Cj0511	29	29	0	0	0	0	0	0	0	protease	Protein synthesis, Amino Acid metabolism
Cj0517	28	28	28	0	0	0	0	0	0	Cj0517	28	26	0	0	0	0	0	0	0	*crc*B (fluoride ion transporter, campher resistance)	membrane transport
Cj0552	29	29	29	29	29	29	19	19	0	Cj0552	29	29	29	29	29	20	19	18	0	putative membrane protein	membrane transport
Cj0553	29	29	29	29	29	29	29	0	0	Cj0553	29	29	29	29	29	19	0	0	0	integral membrane protein	membrane transport
Cj0554	29	29	29	29	29	29	0	0	0	Cj0554	29	29	29	29	29	0	0	0	0	hypothetical protein	unknown
Cj0555	29	29	29	29	29	29	0	0	0	Cj0555	29	29	29	29	29	29	0	0	0	putative integral membrane protein	membrane transport
Cj0556	29	29	29	29	29	29	0	0	0	Cj0556	29	29	29	29	0	0	0	0	0	amidohydrolase family protein	unknown
Cj0557c	29	29	29	29	29	29	0	0	0	Cj0557c	29	29	29	29	0	0	0	0	0	integral membrane protein	membrane transport
Cj0630c	25	25	25	25	0	0	0	0	0	Cj0630c	25	25	25	24	0	0	0	0	0	*hol*A (DNA polymerase III subunit delta)	DNA metabolism and repair
Cj0632	29	29	29	0	0	0	0	0	0	Cj0632	29	29	0	0	0	0	0	0	0	*ilvC (ketol-acid reductoisomerase, valine and isoleucine biosynthesis)*	Protein synthesis, Amino Acid metabolism
Cj0685c	27	27	27	27	27	27	1	0	0	Cj0685c	27	27	27	27	27	27	5	0	0	*cipA (invasion protein)*	other
Cj0686	26	26	10	10	10	9	6	0	0											*isp*G (4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (flavodoxin); involved in isoprenoid synthesis)	Lipid metabolism

Continued on next page

Table A.8 Continued from previous page

Gene	16-mer coverage									Gene	32-mer coverage									Gene name (putative function)	Functional categories
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9		
Cj0763c	27	27	27	17	0	0	0	0	0	Cj0763c	27	27	17	0	0	0	0	0	0	*cys*E (serine acetyltransferase)	Protein synthesis, Amino Acid metabolism
Cj0764c	27	27	25	25	25	2	0	0	0	Cj0764c	27	27	25	18	0	0	0	0	0	*spe*A (arginine decarboxylase)	Protein synthesis, Amino Acid metabolism
Cj0765c	28	26	25	0	0	0	0	0	0	Cj0765c	28	26	25	0	0	0	0	0	0	*his*S (histidine-tRNA ligase)	Protein synthesis, Amino Acid metabolism
Cj0766c	28	28	28	28	1	0	0	0	0	Cj0766c	28	28	28	28	4	0	0	0	0	*tmk* (thymidylate kinase)	DNA metabolism and repair
Cj0776c	29	29	29	29	29	29	29	0	0	Cj0776c	29	29	29	29	29	29	0	0	0	putative periplasmic protein	unknown
Cj0832c	29	28	28	28	28	2	0	0	0	Cj0832c	28	28	28	2	0	0	0	0	0	Na+/H+ antiporter family protein	membrane transport
Cj0833c	28	28	28	28	28	28	28	28	0	Cj0833c	28	28	28	28	28	28	26	0	0	oxidoreductase	stress response (oxidative)
Cj0839c	28	28	28	28	28	28	28	0	0	Cj0839c	28	28	28	28	28	28	28	28	0	hypothetical protein	unknown
Cj0840c	27	27	8	8	0	0	0	0	0											*fbp* (fructose 1,6 bisphosphatase)	Carbohydrate metabolism
Cj0841c	27	27	27	27	0	0	0	0	0	Cj0841c	27	27	27	27	1	0	0	0	0	*mob*B (molybdopterin-guanine dinucleotide biosynthesis protein)	Metabolism of cofactors and vitamins
Cj0995c	29	29	29	29	28	0	0	0	0	Cj0995c	29	29	29	0	0	0	0	0	0	*hem*B (delta-aminolevulinic acid dehydratase)	Metabolism of cofactors and vitamins
Cj0996	29	29	29	29	29	29	29	0	0	Cj0996	29	29	29	29	29	0	0	0	0	*rib*A (GTP cyclohydrolase II)	DNA metabolism and repair
Cj0997	29	29	29	28	0	0	0	0	0	Cj0997	29	29	29	0	0	0	0	0	0	rRNA small subunit methyltransferase G	Protein synthesis, Amino Acid metabolism
Cj0998c	29	29	29	29	28	0	0	0	0	Cj0998c	29	29	29	29	28	0	0	0	0	periplasmic protein	unknown
Cj1058c	29	29	29	27	3	0	0	0	0	Cj1058c	29	29	8	0	0	0	0	0	0	*gua*B (inosine 5'-monophosphate dehydrogenase)	DNA metabolism and repair

Continued on next page

Table A.8 Continued from previous page

Gene	16-mer coverage									Gene	32-mer coverage									Gene name (putative function)	Functional categories
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9		
										Cj1096c	29	28	1	1	1	0	0	0	0	*met*K (S-adenosylmethionine synthetase)	Metabolism of cofactors and vitamins
										Cj1108	28	28	1	1	1	0	0	0	0	*clpA (Clp ATPase)*	stress response (general)
										Cj1109	25	25	12	12	12	11	11	11	11	*aat* (leucyl/phenylalanyl-tRNA–protein transferase)	Protein synthesis, Amino Acid metabolism
Cj1110c	29	29	29	29	29	29	29	29	0	Cj1110c	29	29	29	29	29	29	29	29	4	Putative MCP-type signal transduction protein	signal transduction
Cj1111c	29	29	29	29	29	29	29	29	0	Cj1111c	29	29	29	29	29	29	29	0	0	integral membrane protein	membrane transport
Cj1112c	29	29	29	29	29	29	29	29	0	Cj1112c	29	29	29	29	29	29	29	28	0	*mrs*B; peptide-methionine (R)-S-oxide reductase	stress response (oxidative)
Cj1113	29	29	29	29	29	29	0	0	0	Cj1113	29	29	29	29	29	0	0	0	0	hypothetical protein	unknown
Cj1114c	29	29	0	0	0	0	0	0	0	Cj1114c	29	29	0	0	0	0	0	0	0	*pss*A (CDP-diacylglycerol–serine O-phosphatidyltransferase)	Lipid metabolism
Cj1128c	28	27	20	15	3	3	0	0	0	Cj1128c	28	28	20	20	14	3	0	0	0	*pg*II (glycosylation)	Cell wall/membrane/-capsule
Cj1129c	29	29	29	29	28	5	0	0	0	Cj1129c	29	29	29	29	5	0	0	0	0	*pgl*H (GalNAc-alpha-(1->4)-GalNAc-alpha-(1->3)-diNAcBac-PP-undecaprenol alpha-1,4-N-acetyl-D-galactosaminyltransferase)	Cell wall/membrane/-capsule
Cj1130c	28	28	26	25	8	0	0	0	0	Cj1130c	28	28	25	13	1	0	0	0	0	*pgl*K (protein glycosylation K)	Cell wall/membrane/-capsule
Cj1131c	25	25	25	25	12	0	0	0	0	Cj1131c	25	25	12	9	0	0	0	0	0	*gne* (UDP-GlcNAc/Glc 4-epimerase)	Cell wall/membrane/-capsule
Cj1134	25	25	25	25	11	0	0	0	0	Cj1134	25	25	0	0	0	0	0	0	0	*htr*B (lipid A biosynthesis lauroyl acyltransferase); survival harsh environment	stress response (general)
Cj1182c	29	28	10	10	10	0	0	0	0	Cj1182c	29	28	28	10	10	10	0	0	0	*rps*B (30S ribosomal protein S2)	Protein synthesis, Amino Acid metabolism

Continued on next page

Table A.8 Continued from previous page

Gene	16-mer coverage									Gene	32-mer coverage									Gene name (putative function)	Functional categories
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9		
Cj1188c	27	27	27	18	0	0	0	0	0	Cj1188c	27	27	27	0	0	0	0	0	0	*gid*A (tRNA uridine 5-carboxymethylaminomethyl modification protein GidA; glucose-inhibited cell division protein A	Protein synthesis, Amino Acid metabolism
Cj1220	29	29	29	29	29	0	0	0	0	Cj1220	29	29	29	29	29	29	29	0	0	10 kD chaperonin (cpn10); *gro*ES	stress response (general)
Cj1221	29	29	27	9	0	0	0	0	0	Cj1221	29	29	27	1	0	0	0	0	0	*gro*EL (cpn60)	stress response (general)
Cj1227c	29	29	20	0	0	0	0	0	0	Cj1227c	29	29	0	0	0	0	0	0	0	two-component regulator (3' of *htr*A)	signal transduction
Cj1228c	29	29	29	27	27	0	0	0	0	Cj1228c	29	29	29	27	10	0	0	0	0	*htr*A (serine protease), virulence factor and HtrA may protect oxidatively damaged proteins; chaperone activity	stress response (general)
Cj1257c	28	28	28	28	28	9	0	0	0	Cj1257c	28	28	28	28	9	0	0	0	0	efflux pump protein	membrane transport
Cj1258	28	28	0	0	0	0	0	0	0	Cj1258	28	28	0	0	0	0	0	0	0	phosphotyrosine protein phosphatase	signal transduction
Cj1267c	28	27	0	0	0	0	0	0	0											*hyd*A (Ni/Fe-hydrogenase small subunit)	stress response (oxidative)
Cj1283	29	29	29	29	29	29	29	0	0	Cj1283	29	29	29	29	29	29	21	0	0	*ktrB (putative K+ uptake protein)*	membrane transport
Cj1284	29	29	29	29	29	29	29	29	0	Cj1284	29	29	29	29	29	29	0	0	0	*ktrA (putative K+ uptake protein)*	membrane transport
Cj1364c	29	29	8	4	3	3	1	0	0	Cj1364c	29	29	7	4	3	3	0	0	0	*fumC (fumarate hydratase, class II family (does not require metal); functions in the TCA cycle)*	Carbohydrate metabolism
Cj1385	29	29	22	20	3	0	0	0	0	Cj1385	29	29	22	21	3	3	0	0	0	*katA (katalase)*	stress response (oxidative)
Cj1386	29	29	29	29	29	12	2	0	0	Cj1386	29	29	29	29	29	29	29	12	2	atypical hemin-binding protein, mediates hemin trafficking to KatA	stress response (oxidative)
Cj1387c	29	29	29	29	29	29	29	17	0	Cj1387c	29	29	29	29	29	29	27	16	0	YheO-like PAS6 domain linked to a helix-turn-helix domain modulates post-translational modification of the flagella	cell motility
Cj1388	29	25	0	0	0	0	0	0	0	Cj1388	29	29	0	0	0	0	0	0	0	endoribonuclease L-PSP (Endoribonuclease active on single-stranded mRNA. Inhibits protein synthesis by cleavage of mRNA)	Protein synthesis, Amino Acid metabolism
Cj1399c	29	28	27	27	0	0	0	0	0	Cj1399c	29	28	27	0	0	0	0	0	0	*hyd*A2 (Ni/Fe-hydrogenase small subunit)	stress response (oxidative)

Continued on next page

Table A.8 Continued from previous page

Gene	16-mer coverage									Gene	32-mer coverage									Gene name (putative function)	Functional categories
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9		
Cj1400c	28	28	22	0	0	0	0	0	0	Cj1400c	28	28	25	0	0	0	0	0	0	*fab*I (enoyl-ACP reductase; fatty acid biosynthesis)	Lipid metabolism
Cj1401c	29	29	28	28	0	0	0	0	0	Cj1401c	28	28	0	0	0	0	0	0	0	*tpi*A (triosephosphate isomerase)	Carbohydrate metabolism
Cj1402c	29	29	0	0	0	0	0	0	0	Cj1402c	29	29	4	0	0	0	0	0	0	*pgk* (phosphoglycerate kinase)	Carbohydrate metabolism
Cj1404	29	29	26	0	0	0	0	0	0											*nad*D (nicotinate-nucleotide adenylyltransferase; central role in the synthesis of the redox cofactor NAD+)	stress response (oxidative)
Cj1413c	29	29	29	29	24	0	0	0	0	Cj1413c	29	29	29	29	24	9	0	0	0	*kps*S (capsule polysaccharide modification protein)	Cell wall/membrane/-capsule
Cj1416c	29	29	0	0	0	0	0	0	0	Cj1416c	29	29	0	0	0	0	0	0	0	sugar nucleotidyltransferase	Cell wall/membrane/-capsule
Cj1443c	29	29	0	0	0	0	0	0	0											*kps*F (D-arabinose 5-phosphate isomerase)	Cell wall/membrane/-capsule
Cj1444c	29	29	29	29	29	28	0	0	0	Cj1444c	29	29	29	29	0	0	0	0	0	*kps*D (capsule polysaccharide ABC transporter substrate-binding protein)	Cell wall/membrane/-capsule
Cj1445c	28	28	28	22	0	0	0	0	0	Cj1445c	28	28	22	0	0	0	0	0	0	*kps*E (capsule polysaccharide ABC transporter permease)	Cell wall/membrane/-capsule
Cj1447c	28	28	0	0	0	0	0	0	0											*kps*T (capsule polysaccharide ABC transporter ATP-binding protein)	Cell wall/membrane/-capsule
Cj1449c	28	26	0	0	0	0	0	0	0											hypothetical protein	unknown
Cj1450	28	28	25	0	0	0	0	0	0	Cj1450	28	28	28	28	17	0	0	0	0	ATP/GTP-binding protein	signal transduction
Cj1451	28	27	20	0	0	0	0	0	0	Cj1451	28	25	16	0	0	0	0	0	0	*dut* (dUTPase, dNTP biosynthesis)	DNA metabolism and repair
Cj1459	26	26	1	0	0	0	0	0	0	Cj1459	29	26	26	1	0	0	0	0	0	hypothetical protein	unknown
Cj1534c	26	26	25	24	22	0	0	0	0	Cj1534c	26	26	25	25	22	22	0	0	0	bacterioferritin (iron binding)	stress response (oxidative)
Cj1536c	29	29	29	29	29	29	0	0	0	Cj1536c	29	29	29	29	29	28	0	0	0	*gal*U (UTP–glucose-1-phosphate uridylyltransferase)	Cell wall/membrane/-capsule
Cj1574c	26	26	26	25	25	0	0	0	0	Cj1574c	26	26	25	0	0	0	0	0	0	hypothetical protein	unknown
										Cj1575c	29	29	0	0	0	0	0	0	0	hypothetical protein	unknown

Continued on next page

Table A.8 Continued from previous page

Gene	16-mer coverage									Gene	32-mer coverage									Gene name (putative function)	Functional categories
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9		
Cj1577c	29	29	28	28	0	0	0	0	0	Cj1577c	29	29	28	28	0	0	0	0	0	*nuo*C (NADH-quinone oxidoreductase subunit C, catalyzes the transfer of electrons from NADH to ubiquinone)	stress response (oxidative)
Cj1578c	28	28	26	26	26	0	0	0	0	Cj1578c	28	28	26	26	0	0	0	0	0	*nuo*B (NADH-quinone oxidoreductase subunit B; The point of entry for the majority of electrons that traverse the respiratory chain eventually resulting in the reduction of oxygen)	stress response (oxidative)
Cj1579c	28	28	28	28	28	28	0	0	0	Cj1579c	28	28	28	28	28	28	0	0	0	*nuo*A (NADH dehydrogenase I chain A)	stress response (oxidative)
Cj1586	29	29	1	1	0	0	0	0	0											*cgb* (single domain hemoglobin)	stress response (oxidative)
Cj1588c	26	26	21	4	4	1	0	0	0	Cj1588c	26	26	26	5	4	4	3	0	0	Major facilitator transport protein for small solutes	membrane transport
Cj1638	29	29	29	29	1	0	0	0	0	Cj1638	29	29	29	12	0	0	0	0	0	*dna*G (DNA primase)	DNA metabolism and repair
										Cj1640	29	27	6	0	0	0	0	0	0	hypothetical protein	unknown
Cj1641	29	29	29	25	25	20	10	0	0	Cj1641	29	29	29	25	25	16	1	0	0	*mur*E (peptidoglycane synthesis)	Cell wall/membrane/-capsule
Cj1642	25	25	25	3	0	0	0	0	0	Cj1642	25	25	14	3	0	0	0	0	0	nucleoid-associated protein	unknown
Cj1643	29	29	29	29	29	29	29	25	0	Cj1643	29	29	29	29	29	29	25	0	0	putative periplasmic protein	unknown
Cj1644	29	29	29	28	28	28	0	0	0	Cj1644	29	29	28	28	1	0	0	0	0	*isp*A (geranyltranstransferase)	Metabolism of cofactors and vitamins
Cj1645	29	29	25	1	1	1	1	0	0	Cj1645	29	29	18	1	1	1	0	0	0	*tkt* (transketolase)	Carbohydrate metabolism
Cj1650	28	28	28	28	28	28	28	0	0	Cj1650	28	28	28	28	28	28	1	0	0	hypothetical protein	unknown
Cj1651c	29	29	29	29	29	29	29	28	0	Cj1651c	29	29	29	29	29	29	28	2	0	methionine aminopeptidase	Protein synthesis, Amino Acid metabolism
Cj1652c	29	29	29	29	29	28	28	0	0	Cj1652c	29	29	29	29	29	29	28	0	0	glutamate racemase	Cell wall/membrane/-capsule

Continued on next page

Table A.8 Continued from previous page

Gene	16-mer coverage									Gene	32-mer coverage									Gene name (putative function)	Functional categories
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9		
										Cj1681c	29	29	0	0	0	0	0	0	0	*cys*Q (3'(2'),5'-bisphosphate nucleotidase CysQ)	other
Cj1684c	29	29	29	29	0	0	0	0	0	Cj1684c	29	29	29	29	0	0	0	0	0	transmembrane transport protein	membrane transport
Cj1687	29	29	29	29	29	29	29	0	0	Cj1687	29	29	29	29	29	29	0	0	0	putative efflux protein	membrane transport
										Cj1704c	29	29	0	0	0	0	0	0	0	*rpl*B (50S ribosomal protein L2)	Protein synthesis, Amino Acid metabolism

End of table

Bibliography

[1] P. M. Burnham and D. R. Hendrixson, "Campylobacter jejuni: collective components promoting a successful enteric lifestyle," *Nat. Rev. Microbiol.*, vol. 16, no. 9, pp. 551–565, Sep. 2018.

[2] T. Humphrey, S. O'Brien, and M. Madsen, "Campylobacters as zoonotic pathogens: a food production perspective," *Int. J. Food Microbiol.*, vol. 117, no. 3, pp. 237–257, Jul. 2007.

[3] C. R. Hale, E. Scallan, A. B. Cronquist, J. Dunn, K. Smith, T. Robinson, S. Lathrop, M. Tobin-D'Angelo, and P. Clogher, "Estimates of enteric illness attributable to contact with animals and their environments in the united states," *Clin. Infect. Dis.*, vol. 54 Suppl 5, pp. S472–9, Jun. 2012.

[4] A. N. Jensen, A. Dalsgaard, D. L. Baggesen, and E. M. Nielsen, "The occurrence and characterization of campylobacter jejuni and c. coli in organic pigs and their outdoor environment," *Veterinary Microbiology*, vol. 116, no. 1-3, pp. 96–105, 2006.

[5] N. O. Kaakoush, N. Castaño-Rodríguez, H. M. Mitchell, and S. M. Man, "Global epidemiology of campylobacter infection," *Clin. Microbiol. Rev.*, vol. 28, no. 3, pp. 687–720, Jul. 2015.

[6] C. C. Tam, S. J. O'Brien, D. S. Tompkins, F. J. Bolton, L. Berry, J. Dodds, D. Choudhury, F. Halstead, M. Iturriza-Gómara, K. Mather, G. Rait, A. Ridge, L. C. Rodrigues, J. Wain, B. Wood, J. J. Gray, and IID2 Study Executive Committee, "Changes in causes of acute gastroenteritis in the united kingdom over 15 years: microbiologic findings from 2 prospective, population-based studies of infectious intestinal disease," *Clin. Infect. Dis.*, vol. 54, no. 9, pp. 1275–1286, May 2012.

[7] H. Fernández and G. Pérez-Pérez, "Campylobacter: fluoroquinolone resistance in latin-american countries," *Archivos de Medicina Veterinaria*, vol. 48, no. 3, pp. 255–259, 2016.

[8] L. M. Hodges, C. D. Carrillo, J. P. Upham, A. Borza, M. Eisebraun, R. Kenwell, S. K. Mutschall, D. Haldane, E. Schleihauf, and E. N. Taboada, "A strain comparison of campylobacter isolated from retail poultry and human clinical cases in atlantic canada," *PLoS One*, vol. 14, no. 5, p. e0215928, May 2019.

[9] S. V. R. Epps, R. B. Harvey, M. E. Hume, T. D. Phillips, R. C. Anderson, and D. J. Nisbet, "Foodborne campylobacter: infections, metabolism, pathogenesis and reservoirs," *Int. J. Environ. Res. Public Health*, vol. 10, no. 12, pp. 6292–6304, Nov. 2013.

[10] R. Louwen, P. van Baarlen, A. H. M. van Vliet, A. van Belkum, J. P. Hays, and H. P. Endtz, "Campylobacter bacteremia: a rare and under-reported event?" *Eur. J. Microbiol. Immunol.*, vol. 2, no. 1, pp. 76–87, Mar. 2012.

[11] J. H. Rees, N. A. Gregson, P. L. Griffiths, and R. A. C. Hughes, "Campylobacter jejuni and Guillain-Barre syndrome," *QJM*, vol. 86, no. 10, pp. 623–634, 1993.

[12] E. P. Marder, P. R. Cieslak, A. B. Cronquist, J. Dunn, S. Lathrop, T. Rabatsky-Ehr, P. Ryan, K. Smith, M. Tobin-D'Angelo, D. J. Vugia, S. Zansky, K. G. Holt, B. J. Wolpert, M. Lynch, R. Tauxe, and A. L. Geissler, "Incidence and trends of infections with pathogens transmitted commonly through food and the effect of increasing use of culture-independent diagnostic tests on surveillance — foodborne diseases active surveillance network, 10 U.S. sites, 2013–2016," *MMWR. Morbidity and Mortality Weekly Report*, vol. 66, no. 15, pp. 397–403, 2017.

[13] D. M. Tack, E. P. Marder, P. M. Griffin, P. R. Cieslak, J. Dunn, S. Hurd, E. Scallan, S. Lathrop, A. Muse, P. Ryan, K. Smith, M. Tobin-D'Angelo, D. J. Vugia, K. G. Holt, B. J. Wolpert, R. Tauxe, and A. L. Geissler, "Preliminary incidence and trends of infections with pathogens transmitted commonly through food — foodborne diseases active surveillance network, 10 U.S. sites, 2015–2018," *American Journal of Transplantation*, vol. 19, no. 6, pp. 1859–1863, 2019.

[14] S. Hoffmann, B. Maculloch, and M. Batz, *Economic Burden of Major Foodborne Illnesses Acquired in the United States*. CreateSpace, May 2015.

[15] E. P. o. B. H. (biohaz) and EFSA Panel on Biological Hazards (BIOHAZ), "Scientific opinion on campylobacterin broiler meat production: control options and performance objectives and/or targets at different stages of the food chain," *EFSA Journal*, vol. 9, no. 4, p. 2105, 2011.

[16] E. F. S. Authority, E. F. S. Authority, E. C. for Disease Prevention, and Control, "The european union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2015," *EFSA Journal*, vol. 14, no. 12, 2016.

[17] M.-J. J. Mangen, D. Plass, A. H. Havelaar, C. L. Gibbons, A. Cassini, N. Mühlberger, A. van Lier, J. A. Haagsma, R. John Brooke, T. Lai, C. de Waure, P. Kramarz, M. E. E. Kretzschmar, and on behalf of the BCoDE consortium, "Correction: The pathogen- and Incidence-Based DALY approach: An appropriated methodology for estimating the burden of infectious diseases," *PLoS ONE*, vol. 8, no. 12, 2013.

[18] S. K. Sheppard, X. Didelot, G. Meric, A. Torralbo, K. A. Jolley, D. J. Kelly, S. D. Bentley, M. C. J. Maiden, J. Parkhill, and D. Falush, "Genome-wide association study identifies vitamin b5 biosynthesis as a host specificity factor in campylobacter," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 29, pp. 11 923–11 927, Jul. 2013.

[19] C. J. Buchanan, A. L. Webb, S. K. Mutschall, P. Kruczkiewicz, D. O. R. Barker, B. M. Hetman, V. P. J. Gannon, D. Wade Abbott, J. E. Thomas, G. Douglas Inglis, and E. N. Taboada, "A genome-wide association study to identify diagnostic markers for human pathogenic campylobacter jejuni strains," *Frontiers in Microbiology*, vol. 8, 2017.

[20] S. P. de Vries, S. Gupta, A. Baig, E. Wright, A. Wedley, A. N. Jensen, L. L. Lora, S. Humphrey, H. Skovgård, K. Macleod, E. Pont, D. P. Wolanska, J. L'Heureux, F. M. Mobegi, D. G. E. Smith, P. Everest, A. Zomer, N. Williams, P. Wigley, T. Humphrey, D. J. Maskell, and A. J. Grant, "Genome-wide fitness analyses of the foodborne pathogen campylobacter jejuni in in vitro and in vivo models," *Sci. Rep.*, vol. 7, no. 1, p. 1251, Apr. 2017.

[21] E. Mourkas, A. J. Taylor, G. Méric, S. C. Bayliss, B. Pascoe, L. Mageiros, J. K. Calland, M. D. Hitchings, A. Ridley, A. Vidal, K. J. Forbes, N. J. C. Strachan, C. T. Parker, J. Parkhill, K. A. Jolley, A. J. Cody, M. C. J. Maiden, D. J. Kelly, and S. K. Sheppard, "Agricultural intensification and the evolution of host specialism in the enteric pathogen," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 117, no. 20, pp. 11 018–11 028, May 2020.

[22] C. Murphy, C. Carroll, and K. N. Jordan, "Environmental survival mechanisms of the foodborne pathogen campylobacter jejuni," *J. Appl. Microbiol.*, vol. 100, no. 4, pp. 623–632, Apr. 2006.

[23] K. E. Dingle, F. M. Colles, D. Falush, and M. C. J. Maiden, "Sequence typing and comparison of population

biology of campylobacter coli and campylobacter jejuni," *J. Clin. Microbiol.*, vol. 43, no. 1, pp. 340–347, Jan. 2005.

[24] X. Didelot and D. Falush, "Inference of bacterial microevolution using multilocus sequence data," *Genetics*, vol. 175, no. 3, pp. 1251–1266, Mar. 2007.

[25] S. K. Sheppard, N. D. McCarthy, D. Falush, and M. C. J. Maiden, "Convergence of campylobacter species: implications for bacterial evolution," *Science*, vol. 320, no. 5873, pp. 237–239, Apr. 2008.

[26] J. Parkhill, B. Wren, K. Mungall, J. Ketley, C. Churcher, D. Basham, T. Chillingworth, R. Davies, T. Feltwell, S. Holroyd *et al.*, "The genome sequence of the food-borne pathogen campylobacter jejuni reveals hypervariable sequences," *Nature*, vol. 403, no. 6770, pp. 665–668, 2000.

[27] A. Mira, A. B. Martín-Cuadrado, G. D'Auria, and F. Rodríguez-Valera, "The bacterial pan-genome:a new paradigm in microbiology," *Int. Microbiol.*, vol. 13, no. 2, pp. 45–57, Jun. 2010.

[28] H. Tettelin, V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. Deboy, T. M. Davidsen, M. Mora, M. Scarselli, I. Margarit y Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. B. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli, and C. M. Fraser, "Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: implications for the microbial 'pan-genome'," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 39, pp. 13 950–13 955, Sep. 2005.

[29] S. K. Sheppard, D. S. Guttman, and J. Ross Fitzgerald, "Population genomics of bacterial host adaptation," *Nature Reviews Genetics*, vol. 19, no. 9, pp. 549–565, 2018.

[30] E. Zuckerkandl and L. Pauling, "Evolutionary divergence and convergence in proteins," in *Evolving genes and proteins*. Elsevier, 1965, pp. 97–166.

[31] H. Ochman and A. C. Wilson, "Evolution in bacteria: evidence for a universal substitution rate in cellular genomes," *J. Mol. Evol.*, vol. 26, no. 1-2, pp. 74–86, 1987.

[32] M. C. J. M. Samuel K. Sheppard, "The evolution of campylobacter jejuni and campylobacter coli," *Cold Spring Harb. Perspect. Biol.*, vol. 7, no. 8, Aug. 2015.

[33] A. Eyre-Walker and P. D. Keightley, "The distribution of fitness effects of new mutations," *Nat. Rev. Genet.*, vol. 8, no. 8, pp. 610–618, Aug. 2007.

[34] D. J. Wilson, E. Gabriel, A. J. H. Leatherbarrow, J. Cheesbrough, S. Gee, E. Bolton, A. Fox, C. A. Hart, P. J. Diggle, and P. Fearnhead, "Rapid evolution and the importance of recombination to the gastroenteric pathogen campylobacter jejuni," *Mol. Biol. Evol.*, vol. 26, no. 2, pp. 385–397, Feb. 2009.

[35] A. Mira, R. Pushker, and F. Rodríguez-Valera, "The neolithic revolution of bacterial genomes," *Trends Microbiol.*, vol. 14, no. 5, pp. 200–206, May 2006.

[36] S. K. Sheppard, X. Didelot, K. A. Jolley, A. E. Darling, B. Pascoe, G. Meric, D. J. Kelly, A. Cody, F. M. Colles, N. J. C. Strachan, I. D. Ogden, K. Forbes, N. P. French, P. Carter, W. G. Miller, N. D.

McCarthy, R. Owen, E. Litrup, M. Egholm, J. P. Affourtit, S. D. Bentley, J. Parkhill, M. C. J. Maiden, and D. Falush, "Progressive genome-wide introgression in agricultural campylobacter coli," *Molecular Ecology*, vol. 22, no. 4, pp. 1051–1064, 2013.

[37] C. d. Mazancourt, C. de Mazancourt, E. Johnson, and T. G. Barraclough, "Biodiversity inhibits species' evolutionary responses to changing environments," *Ecology Letters*, vol. 11, no. 4, pp. 380–388, 2008.

[38] A. G. Mathew, R. Cissell, and S. Liamthong, "Antibiotic resistance in bacteria associated with food animals: A united states perspective of livestock production," *Foodborne Pathogens and Disease*, vol. 4, no. 2, pp. 115–133, 2007.

[39] J. C. Golz, L. Epping, M.-T. Knüver, M. Borowiak, F. Hartkopf, C. Deneke, B. Malorny, T. Semmler, and K. Stingl, "Whole genome sequencing reveals extended natural transformation in campylobacter impacting diagnostics and the pathogens adaptive potential," *Scientific Reports*, vol. 10, no. 1, 2020.

[40] L. Epping, E.-M. Antão, and T. Semmler, "Population biology and comparative genomics of campylobacter species." *Current Topics in Microbiology and Immunology*, vol. 431, pp. 59–78, 2021.

[41] L. H. Taylor, S. M. Latham, and M. E. Woolhouse, "Risk factors for human disease emergence," *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, vol. 356, no. 1411, pp. 983–989, Jul. 2001.

[42] E. J. Richardson, R. Bacigalupe, E. M. Harrison, L. A. Weinert, S. Lycett, M. Vrieling, K. Robb, P. A. Hoskisson, M. T. Holden, E. J. Feil *et al.*, "Gene exchange drives the ecological success of a multi-host bacterial pathogen," *Nature ecology & evolution*, vol. 2, no. 9, pp. 1468–1478, 2018.

[43] S. K. Sheppard, L. Cheng, G. Méric, C. P. A. de Haan, A.-K. Llarena, P. Marttinen, A. Vidal, A. Ridley, F. Clifton-Hadley, T. R. Connor, N. J. C. Strachan, K. Forbes, F. M. Colles, K. A. Jolley, S. D. Bentley, M. C. J. Maiden, M.-L. Hänninen, J. Parkhill, W. P. Hanage, and J. Corander, "Cryptic ecology among host generalist campylobacter jejuni in domestic animals," *Mol. Ecol.*, vol. 23, no. 10, pp. 2442–2451, May 2014.

[44] N. Lupolova, T. J. Dallman, N. J. Holden, and D. L. Gally, "Patchy promiscuity: machine learning applied to predict the host specificity of salmonella enterica and escherichia coli," *Microbial genomics*, vol. 3, no. 10, 2017.

[45] H. Song, J. Hwang, H. Yi, R. L. Ulrich, Y. Yu, W. C. Nierman, and H. S. Kim, "The early stage of bacterial genome-reductive evolution in the host," *PLoS Pathog*, vol. 6, no. 5, p. e1000922, 2010.

[46] N. A. Moran and G. R. Plague, "Genomic changes following host restriction in bacteria," *Current opinion in genetics & development*, vol. 14, no. 6, pp. 627–633, 2004.

[47] S. Watford and S. J. Warrington, *Bacterial DNA Mutations*. StatPearls Publishing, Treasure Island (FL), 2020. [Online]. Available: http://europepmc.org/books/NBK459274

[48] A. K. Hottes, P. L. Freddolino, A. Khare, Z. N. Donnell, J. C. Liu, and S. Tavazoie, "Bacterial adaptation through loss of function," *PLoS Genetics*, vol. 9, no. 7, p. e1003617, 2013.

[49] H. Goodarzi, B. D. Bennett, S. Amini, M. L. Reaves, A. K. Hottes, J. D. Rabinowitz, and S. Tavazoie, "Regulatory and metabolic rewiring during laboratory evolution of ethanol tolerance in e. coli," *Molecular systems biology*, vol. 6, no. 1, p. 378, 2010.

[50] A. J. Westermann, K. U. Förstner, F. Amman, L. Barquist, Y. Chao, L. N. Schulte, L. Müller, R. Reinhardt, P. F. Stadler, and J. Vogel, "Dual rna-seq unveils noncoding rna functions in host–pathogen interactions," *Nature*, vol. 529, no. 7587, pp. 496–501, 2016.

[51] D. Moradigaravand and J. Engelstädter, "The effect of bacterial recombination on adaptation on fitness landscapes with limited peak accessibility," *PLoS Comput Biol*, vol. 8, no. 10, p. e1002735, 2012.

[52] T. F. Cooper, "Recombination speeds adaptation by reducing competition between beneficial mutations in populations of escherichia coli," *PLoS Biol*, vol. 5, no. 9, p. e225, 2007.

[53] L. Boto, "Horizontal gene transfer in evolution: facts and challenges," *Proceedings of the Royal Society B: Biological Sciences*, vol. 277, no. 1683, pp. 819–827, 2010.

[54] M. Vos and X. Didelot, "A comparison of homologous recombination rates in bacteria and archaea," *The ISME journal*, vol. 3, no. 2, pp. 199–208, 2009.

[55] N. Basic-Hammer, V. Vogel, P. Basset, and D. S. Blanc, "Impact of recombination on genetic variability within staphylococcus aureus clonal complexes," *Infection, Genetics and Evolution*, vol. 10, no. 7, pp. 1117–1123, 2010.

[56] S. Suerbaum, J. M. Smith, K. Bapumia, G. Morelli, N. H. Smith, E. Kunstmann, I. Dyrek, and M. Achtman, "Free recombination within helicobacter pylori," *Proceedings of the National Academy of Sciences*, vol. 95, no. 21, pp. 12 619–12 624, 1998.

[57] K. E. Dingle, F. M. Colles, D. R. Wareing, R. Ure, A. J. Fox, F. E. Bolton, H. J. Bootsma, R. J. Willems, R. Urwin, and M. C. Maiden, "Multilocus sequence typing system for campylobacter jejuni," *J. Clin. Microbiol.*, vol. 39, no. 1, pp. 14–23, Jan. 2001.

[58] U. Nübel, J. Dordel, K. Kurt, B. Strommenger, H. Westh, S. K. Shukla, H. Žemličková, R. Leblois, T. Wirth, T. Jombart *et al.*, "A timescale for evolution, population expansion, and spatial spread of an emerging clone of methicillin-resistant staphylococcus aureus," *PLoS Pathog*, vol. 6, no. 4, p. e1000855, 2010.

[59] P. R. Reeves, B. Liu, Z. Zhou, D. Li, D. Guo, Y. Ren, C. Clabots, R. Lan, J. R. Johnson, and L. Wang, "Rates of mutation and host transmission for an escherichia coli clone over 3 years," *PloS one*, vol. 6, no. 10, p. e26907, 2011.

[60] G. W. Blakely, "Mechanisms of horizontal gene transfer and dna recombination," in *Molecular medical microbiology*. Elsevier, 2015, pp. 291–302.

[61] C. W. Birky and J. B. Walsh, "Effects of linkage on rates of molecular evolution," *Proceedings of the National Academy of Sciences*, vol. 85, no. 17, pp. 6414–6418, 1988.

[62] A.-K. Llarena, E. Taboada, and M. Rossi, "Whole-genome sequencing in epidemiology of campylobacter jejuni infections," *Journal of clinical microbiology*, vol. 55, no. 5, pp. 1269–1275, 2017.

[63] M. C. Maiden, J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt, "Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 95, no. 6, pp. 3140–3145, Mar. 1998.

[64] W. G. Miller, S. L. W. On, G. Wang, S. Fontanoz, A. J. Lastovica, and R. E. Mandrell, "Extended multilocus sequence typing system for campylobacter coli, c. lari, c. upsaliensis, and c. helveticus," *Journal of Clinical Microbiology*, vol. 43, no. 5, pp. 2315–2329, 2005.

[65] K. E. Dingle, F. M. Colles, R. Ure, J. A. Wagenaar, B. Duim, F. J. Bolton, A. J. Fox, D. R. A. Wareing, and M. C. J. Maiden, "Molecular characterization of campylobacter jejuni clones: a basis for epidemiologic investigation," *Emerg. Infect. Dis.*, vol. 8, no. 9, pp. 949–955, Sep. 2002.

[66] X. Zhang, F. Li, S. Cui, L. Mao, X. Li, F. Awan, W. Lv, and Z. Zeng, "Prevalence and distribution characteristics of blakpc-2 and blandm-1 genes in klebsiella pneumoniae," *Infection and Drug Resistance*, vol. 13, p. 2901, 2020.

[67] L. Riley, "Pandemic lineages of extraintestinal pathogenic escherichia coli," *Clinical Microbiology and Infection*, vol. 20, no. 5, pp. 380–390, 2014.

[68] G. Manning, C. G. Dowson, M. C. Bagnall, I. H. Ahmed, M. West, and D. G. Newell, "Multilocus sequence typing for comparison of veterinary and human isolates of campylobacter jejuni," *Applied and Environmental Microbiology*, vol. 69, no. 11, pp. 6370–6379, 2003.

[69] B. L. Dearlove, A. J. Cody, B. Pascoe, G. Méric, D. J. Wilson, and S. K. Sheppard, "Rapid host switching in generalist campylobacter strains erodes the signal for tracing human infections," *ISME J.*, vol. 10, no. 3, pp. 721–729, Mar. 2016.

[70] P. J. Biggs, P. Fearnhead, G. Hotter, V. Mohan, J. Collins-Emerson, E. Kwan, T. E. Besser, A. Cookson, P. E. Carter, and N. P. French, "Whole-genome comparison of two campylobacter jejuni isolates of the same sequence type reveals multiple loci of different ancestral lineage," *PLoS One*, vol. 6, no. 11, p. e27121, Nov. 2011.

[71] A.-K. Llarena, J. Zhang, V. Minna, N. Välimäki, M. Hakkinen, M.-L. Hänninen, M. Roasto, M. Mäesaar, E. Taboada, D. Barker, G. Garfolo, C. Cammà, E. Di Giannatale, J. Corander, and M. Rossi, "Monomorphic genotypes within a generalist lineage of campylobacter jejuni show signs of global dispersion."

[72] T. Azarian, I.-T. Huang, and W. P. Hanage, "Structure and dynamics of bacterial populations: Pangenome ecology," in *The Pangenome*. Springer, Cham, 2020, pp. 115–128.

[73] N. Nguyen, G. Hickey, D. R. Zerbino, B. Raney, D. Earl, J. Armstrong, D. Haussler, and B. Paten, "Building a pangenome reference for a population," in *International Conference on Research in Computational Molecular Biology*. Springer, 2014, pp. 207–221.

[74] A. J. Cody, J. E. Bray, K. A. Jolley, N. D. McCarthy, and M. C. J. Maiden, "Core genome multilocus sequence typing scheme for stable, comparative analyses of campylobacter jejuni and c. coli human disease isolates," *Journal of Clinical Microbiology*, vol. 55, no. 7, pp. 2086–2097, 2017.

[75] S. F. Altekruse, N. J. Stern, P. I. Fields, and D. L. Swerdlow, "Campylobacter jejuni—an emerging foodborne pathogen," *Emerging Infectious Diseases*, vol. 5, no. 1, pp. 28–35, 1999.

[76] S. Suerbaum, M. Lohrengel, A. Sonnevend, F. Ruberg, and M. Kist, "Allelic diversity and recombination incampylobacter jejuni," *Journal of Bacteriology*, vol. 183, no. 8, pp. 2553–2559, 2001.

[77] B. Pascoe, G. Méric, K. Yahara, H. Wimalarathna, S. Murray, M. D. Hitchings, E. L. Sproston, C. D. Carrillo, E. N. Taboada, K. K. Cooper, S. Huynh, A. J. Cody, K. A. Jolley, M. C. J. Maiden, N. D. McCarthy, X. Didelot, C. T. Parker, and S. K. Sheppard, "Local genes for local bacteria: Evidence of allopatry in the genomes of transatlantic campylobacter populations," *Mol. Ecol.*, vol. 26, no. 17, pp. 4497–4508, Sep. 2017.

[78] J. Revez, M. Rossi, P. Ellström, C. de Haan, H. Rautelin, and M.-L. Hänninen, "Finnish campylobacter jejuni strains of multilocus sequence type ST-22 complex have two lineages with different characteristics," *PLoS One*, vol. 6, no. 10, p. e26880, Oct. 2011.

[79] H. Asakura, H. Brüggemann, S. K. Sheppard, T. Ekawa, T. F. Meyer, S. Yamamoto, and S. Igimi, "Molecular evidence for the thriving of campylobacter jejuni ST-4526 in japan," *PLoS One*, vol. 7, no. 11, p. e48394, Nov. 2012.

[80] S. M. McTavish, C. E. Pope, C. Nicol, K. Sexton, N. French, and P. E. Carter, "Wide geographical distribution of internationally rare campylobacter clones within new zealand," *Epidemiol. Infect.*, vol. 136, no. 9, pp. 1244–1252, Sep. 2008.

[81] V. Mohan, M. Stevenson, J. Marshall, P. Fearnhead, B. R. Holland, G. Hotter, and N. P. French, "Campylobacter jejuni colonization and population structure in urban populations of ducks and starlings in new zealand," *Microbiologyopen*, vol. 2, no. 4, pp. 659–673, Aug. 2013.

[82] F. M. Colles, K. Jones, R. M. Harding, and M. C. J. Maiden, "Genetic diversity of campylobacter jejuni isolates from farm animals and the farm environment," *Applied and Environmental Microbiology*, vol. 69, no. 12, pp. 7409–7413, 2003.

[83] S. K. Sheppard, F. M. Colles, N. D. McCarthy, N. J. C. Strachan, I. D. Ogden, K. J. Forbes, J. F. Dallas, and M. C. J. Maiden, "Niche segregation and genetic structure of campylobacter jejuni populations from wild and agricultural host species," *Mol. Ecol.*, vol. 20, no. 16, pp. 3484–3490, Aug. 2011.

[84] F. M. Colles, N. D. McCarthy, J. C. Howe, C. L. Devereux, A. G. Gosler, and M. C. J. Maiden, "Dynamics of campylobacter colonization of a natural host, sturnus vulgaris (european starling)," *Environ. Microbiol.*, vol. 11, no. 1, pp. 258–267, Jan. 2009.

[85] M. N. Price, P. S. Dehal, and A. P. Arkin, "Fasttree 2–approximately maximum-likelihood trees for large alignments," *PloS one*, vol. 5, no. 3, p. e9490, 2010.

[86] A. B. Vidal, F. M. Colles, J. D. Rodgers, N. D. McCarthy, R. H. Davies, M. C. J. Maiden, and F. A. Clifton-Hadley, "Genetic diversity of campylobacter jejuni and campylobacter coli isolates from conventional broiler flocks and the impacts of sampling strategy and laboratory method," *Appl. Environ. Microbiol.*, vol. 82, no. 8, pp. 2347–2355, Apr. 2016.

[87] F. M. Colles, T. A. Jones, N. D. McCarthy, S. K. Sheppard, A. J. Cody, K. E. Dingle, M. S. Dawkins, and M. C. J. Maiden, "Campylobacter infection of broiler chickens in a free-range environment," *Environ. Microbiol.*, vol. 10, no. 8, pp. 2042–2050, Aug. 2008.

[88] S. M. Kovanen, R. I. Kivistö, M. Rossi, T. Schott, U.-M. Kärkkäinen, T. Tuuminen, J. Uksila, H. Rautelin, and M.-L. Hänninen, "Multilocus sequence typing (MLST) and whole-genome MLST of campylobacter

jejuni isolates from human infections in three districts during a seasonal peak in finland," *J. Clin. Microbiol.*, vol. 52, no. 12, pp. 4147–4154, Dec. 2014.

[89] C. P. A. Skarp-de Haan, A. Culebro, T. Schott, J. Revez, E. K. H. Schweda, M.-L. Hänninen, and M. Rossi, "Comparative genomics of unintrogressed campylobacter coli clades 2 and 3," *BMC Genomics*, vol. 15, p. 129, Feb. 2014.

[90] W. G. Miller, M. D. Englen, S. Kathariou, I. V. Wesley, G. Wang, L. Pittenger-Alley, R. M. Siletz, W. Muraoka, P. J. Fedorka-Cray, and R. E. Mandrell, "Identification of host-associated alleles by multilocus sequence typing of campylobacter coli strains from food animals," *Microbiology*, vol. 152, no. 1, pp. 245–255, 2006.

[91] S. Thakur, W. E. M. Morrow, J. A. Funk, P. B. Bahnson, and W. A. Gebreyes, "Molecular epidemiologic investigation of campylobacter coli in swine production systems, using multilocus sequence typing," *Appl. Environ. Microbiol.*, vol. 72, no. 8, pp. 5666–5669, Aug. 2006.

[92] A. J. Cody, N. M. McCarthy, H. L. Wimalarathna, F. M. Colles, L. Clark, I. C. J. W. Bowler, M. C. J. Maiden, and K. E. Dingle, "A longitudinal 6-year study of the molecular epidemiology of clinical campylobacter isolates in oxfordshire, united kingdom," *J. Clin. Microbiol.*, vol. 50, no. 10, pp. 3193–3201, Oct. 2012.

[93] A. Nohra, A. Grinberg, A. C. Midwinter, J. C. Marshall, J. M. Collins-Emerson, and N. P. French, "Molecular epidemiology of campylobacter coli strains isolated from different sources in new zealand between 2005 and 2014," *Appl. Environ. Microbiol.*, vol. 82, no. 14, pp. 4363–4370, Jul. 2016.

[94] B. Duim, T. M. Wassenaar, A. Rigter, and J. Wagenaar, "High-resolution genotyping of campylobacter strains isolated from poultry and humans with amplified fragment length polymorphism fingerprinting," *Appl. Environ. Microbiol.*, vol. 65, no. 6, pp. 2369–2375, Jun. 1999.

[95] S. K. Sheppard, J. F. Dallas, D. J. Wilson, N. J. C. Strachan, N. D. McCarthy, K. A. Jolley, F. M. Colles, O. Rotariu, I. D. Ogden, K. J. Forbes, and M. C. J. Maiden, "Evolution of an agriculture-associated disease causing campylobacter coli clade: Evidence from national surveillance data in scotland," *PLoS ONE*, vol. 5, no. 12, p. e15708, 2010.

[96] E. W. Myers, "A Whole-Genome assembly of drosophila," *Science*, vol. 287, no. 5461, pp. 2196–2204, 2000.

[97] S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy, "Canu: scalable and accurate long-read assembly via adaptive -mer weighting and repeat separation," *Genome Res.*, vol. 27, no. 5, pp. 722–736, May 2017.

[98] D. R. Zerbino and E. Birney, "Velvet: algorithms for de novo short read assembly using de bruijn graphs," *Genome research*, vol. 18, no. 5, pp. 821–829, 2008.

[99] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzner, "Spades: a new genome assembly algorithm and its applications to single-cell sequencing," *J. Comput. Biol.*, vol. 19, no. 5, pp. 455–477, May 2012.

[100] D. E. Wood and S. L. Salzberg, "Kraken: ultrafast metagenomic sequence classification using exact alignments," *Genome Biol.*, vol. 15, no. 3, p. R46, Mar. 2014.

[101] D. E. Wood, J. Lu, and B. Langmead, "Improved metagenomic analysis with kraken 2," *Genome Biol.*, vol. 20, no. 1, p. 257, Nov. 2019.

[102] H. Hasman, D. Saputra, T. Sicheritz-Ponten, O. Lund, C. A. Svendsen, N. Frimodt-Møller, and F. M. Aarestrup, "Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples," *J. Clin. Microbiol.*, vol. 52, no. 1, pp. 139–146, Jan. 2014.

[103] M. V. Larsen, S. Cosentino, O. Lukjancenko, D. Saputra, S. Rasmussen, H. Hasman, T. Sicheritz-Ponten, F. M. Aarestrup, D. W. Ussery, and O. Lund, "Benchmarking of methods for genomic taxonomy," *Journal of Clinical Microbiology*, vol. 52, no. 5, pp. 1529–1539, 2014.

[104] P. T. L. C. Clausen, F. M. Aarestrup, and O. Lund, "Rapid and precise alignment of raw reads against redundant databases with KMA," *BMC Bioinformatics*, vol. 19, no. 1, p. 307, Aug. 2018.

[105] B. D. Ondov, T. J. Treangen, P. Melsted, A. B. Mallonee, N. H. Bergman, S. Koren, and A. M. Phillippy, "Mash: fast genome and metagenome distance estimation using MinHash," *Genome Biol.*, vol. 17, no. 1, p. 132, Jun. 2016.

[106] R. Chikhi and P. Medvedev, "Informed and automated k-mer size selection for genome assembly," *Bioinformatics*, vol. 30, no. 1, pp. 31–37, Jan. 2014.

[107] D. R. Zerbino, "Using the velvet de novo assembler for short-read sequencing technologies," *Current protocols in bioinformatics*, vol. 31, no. 1, pp. 11–5, 2010.

[108] C. X. Chan and M. A. Ragan, "Next-generation phylogenomics," *Biology Direct*, vol. 8, no. 1, 2013.

[109] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins, "Clustalw and clustalx version 2.0," *Bioinformatics*, vol. 23, no. 21, pp. 2947–2948, Nov. 2007.

[110] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, no. 3, Oct. 1990.

[111] O. Bonham-Carter, J. Steele, and D. Bastola, "Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis," *Brief. Bioinform.*, vol. 15, no. 6, pp. 890–905, Jul. 2013.

[112] K. M. Wong, M. A. Suchard, and J. P. Huelsenbeck, "Alignment uncertainty and genomic analysis," *Science*, vol. 319, no. 5862, pp. 473–476, 2008.

[113] S. C. Manekar and S. R. Sathe, "A benchmark study of k-mer counting methods for high-throughput sequencing," *GigaScience*, vol. 7, no. 12, p. giy125, 2018.

[114] G. Marçais and C. Kingsford, "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers," *Bioinformatics*, vol. 27, no. 6, pp. 764–770, Mar. 2011.

[115] P. Melsted and J. K. Pritchard, "Efficient counting of k-mers in DNA sequences using a bloom filter," *BMC Bioinformatics*, vol. 12, no. 1, 2011.

[116] M. Roberts, W. Hayes, B. R. Hunt, S. M. Mount, and J. A. Yorke, "Reducing storage requirements for biological sequence comparison," *Bioinformatics*, vol. 20, no. 18, pp. 3363–3369, 2004.

[117] S. Deorowicz, M. Kokot, S. Grabowski, and A. Debudaj-Grabysz, "Kmc 2: fast and resource-frugal k-mer counting," *Bioinformatics*, vol. 31, no. 10, pp. 1569–1576, 2015.

[118] R. C. Eagle, "Complement factor H polymorphism in Age-Related macular degeneration," *Yearbook of Ophthalmology*, vol. 2006, pp. 245–248, 2006.

[119] J. MacArthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales *et al.*, "The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog)," *Nucleic acids research*, vol. 45, no. D1, pp. D896–D901, 2017.

[120] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang, "10 years of gwas discovery: biology, function, and translation," *The American Journal of Human Genetics*, vol. 101, no. 1, pp. 5–22, 2017.

[121] D. Ellinghaus, L. Jostins, S. L. Spain, A. Cortes, J. Bethune, B. Han, Y. R. Park, S. Raychaudhuri, J. G. Pouget, M. Hübenthal *et al.*, "Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci," *Nature genetics*, vol. 48, no. 5, pp. 510–518, 2016.

[122] T. A. Manolio, "Bringing genome-wide association findings into clinical use," *Nature Reviews Genetics*, vol. 14, no. 8, pp. 549–558, 2013.

[123] S. F. Kingsmore, I. E. Lindquist, J. Mudge, D. D. Gessler, and W. D. Beavis, "Genome-wide association studies: progress and potential for drug discovery and development," *Nature Reviews Drug Discovery*, vol. 7, no. 3, pp. 221–230, 2008.

[124] D. E. Reich and E. S. Lander, "On the allelic spectrum of human disease," *Trends in Genetics*, vol. 17, no. 9, pp. 502–510, 2001.

[125] W. Y. S. Wang, B. J. Barratt, D. G. Clayton, and J. A. Todd, "Genome-wide association studies: theoretical and practical concerns," *Nature Reviews Genetics*, vol. 6, no. 2, pp. 109–118, 2005.

[126] S. A. McCarroll, "Extending genome-wide association studies to copy-number variation," *Human molecular genetics*, vol. 17, no. R2, pp. R135–R142, 2008.

[127] C. C. Spencer, Z. Su, P. Donnelly, and J. Marchini, "Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip," *PLoS Genet*, vol. 5, no. 5, p. e1000477, 2009.

[128] J. A. Lees, M. Vehkala, N. Välimäki, S. R. Harris, C. Chewapreecha, N. J. Croucher, P. Marttinen, M. R. Davies, A. C. Steer, S. Y. Tong *et al.*, "Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes," *Nature communications*, vol. 7, no. 1, pp. 1–8, 2016.

[129] Y. Zhang, "On the use of p-values in genome wide disease association mapping," *J Biom Biostat*, vol. 7, p. 297, 2016.

[130] D. J. Schaid, W. Chen, and N. B. Larson, "From genome-wide associations to candidate causal variants by statistical fine-mapping," *Nature Reviews Genetics*, vol. 19, no. 8, pp. 491–504, 2018.

[131] R. A. Power, J. Parkhill, and T. de Oliveira, "Microbial genome-wide association studies: lessons from human gwas," *Nat. Rev. Genet.*, vol. 18, no. 1, pp. 41–50, Jan. 2017.

[132] B. Aslam, W. Wang, M. I. Arshad, M. Khurshid, S. Muzammil, M. H. Rasool, M. A. Nisar, R. F. Alvi, M. A. Aslam, M. U. Qamar, M. K. F. Salamat, and Z. Baloch, "Antibiotic resistance: a rundown of a global crisis," *Infect. Drug Resist.*, vol. 11, pp. 1645–1658, Oct. 2018.

[133] D. Falush and R. Bowden, "Genome-wide association mapping in bacteria?" *Trends in Microbiology*, vol. 14, no. 8, pp. 353–355, 2006.

[134] S. E. James, T. de Oliveira, S. Baichoo, V. Fonseca, A. M. Kanzi, Y. Moosa, R. Lessells, and R. Power, "Current affairs of microbial genome-wide association studies: approaches, bottlenecks and analytical pitfalls," *Frontiers in Microbiology*, vol. 10, p. 3119, 2019.

[135] B. C. Young, S. G. Earle, S. Soeng, P. Sar, V. Kumar, S. Hor, V. Sar, R. Bousfield, N. D. Sanderson, L. Barker *et al.*, "Panton–valentine leucocidin is the key determinant of staphylococcus aureus pyomyositis in a bacterial gwas," *Elife*, vol. 8, p. e42486, 2019.

[136] Y. Li, B. J. Metcalf, S. Chochua, Z. Li, H. Walker, T. Tran, P. A. Hawkins, R. Gierke, T. Pilishvili, L. McGee *et al.*, "Genome-wide association analyses of invasive pneumococcal isolates identify a missense bacterial mutation associated with meningitis," *Nature communications*, vol. 10, no. 1, pp. 1–11, 2019.

[137] M. Galardini, O. Clermont, A. Baron, B. Busby, S. Dion, S. Schubert, P. Beltrao, and E. Denamur, "Major role of the high-pathogenicity island (hpi) in the intrinsic extra-intestinal virulence of escherichia coli revealed by a genome-wide association study," *BioRxiv*, p. 712034, 2019.

[138] M. R. Farhat, L. Freschi, R. Calderon, T. Ioerger, M. Snyder, C. J. Meehan, B. de Jong, L. Rigouts, A. Sloutsky, D. Kaur *et al.*, "Gwas for quantitative resistance phenotypes in mycobacterium tuberculosis reveals resistance genes and regulatory regions," *Nature communications*, vol. 10, no. 1, pp. 1–11, 2019.

[139] Z. Wang, M. Cabrera, J. Yang, L. Yuan, B. Gupta, X. Liang, K. Kemirembe, S. Shrestha, A. Brashear, X. Li *et al.*, "Genome-wide association analysis identifies genetic loci associated with resistance to multiple antimalarials in plasmodium falciparum from china-myanmar border," *Scientific reports*, vol. 6, p. 33891, 2016.

[140] D. Falush, "Bacterial genomics: Microbial gwas coming of age," *Nat Microbiol*, vol. 1, p. 16059, Apr. 2016.

[141] S. J. Bush, D. Foster, D. W. Eyre, E. L. Clark, N. De Maio, L. P. Shaw, N. Stoesser, T. E. Peto, D. W. Crook, and A. S. Walker, "Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism–calling pipelines," *GigaScience*, vol. 9, no. 2, p. giaa007, 2020.

[142] P. E. Chen and B. J. Shapiro, "The advent of genome-wide association studies for bacteria," *Current opinion in microbiology*, vol. 25, pp. 17–24, 2015.

[143] S. G. Earle, C.-H. Wu, J. Charlesworth, N. Stoesser, N. C. Gordon, T. M. Walker, C. C. Spencer, Z. Iqbal, D. A. Clifton, K. L. Hopkins *et al.*, "Identifying lineage effects when controlling for population structure improves power in bacterial association studies," *Nature microbiology*, vol. 1, no. 5, pp. 1–8, 2016.

[144] A. T. Marees, H. de Kluiver, S. Stringer, F. Vorspan, E. Curis, C. Marie-Claire, and E. M. Derks, "A tutorial

on conducting genome-wide association studies: Quality control and statistical analysis," *International journal of methods in psychiatric research*, vol. 27, no. 2, p. e1608, 2018.

[145] M. Slatkin, "Linkage disequilibrium—understanding the evolutionary past and mapping the medical future," *Nature Reviews Genetics*, vol. 9, no. 6, pp. 477–485, 2008.

[146] M. M. Saber and B. J. Shapiro, "Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes," *Microbial genomics*, vol. 6, no. 3, 2020.

[147] B. A. Wilson, N. R. Garud, A. F. Feder, Z. J. Assaf, and P. S. Pennings, "The population genetics of drug resistance evolution in natural populations of viral, bacterial and eukaryotic pathogens," *Molecular ecology*, vol. 25, no. 1, pp. 42–66, 2016.

[148] T. Wein and T. Dagan, "The effect of population bottleneck size and selective regime on genetic diversity and evolvability in bacteria," *Genome biology and evolution*, vol. 11, no. 11, pp. 3283–3290, 2019.

[149] M. R. Farhat, B. J. Shapiro, S. K. Sheppard, C. Colijn, and M. Murray, "A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens," *Genome medicine*, vol. 6, no. 11, pp. 1–14, 2014.

[150] S. Chatterjee and A. S. Hadi, *Regression analysis by example*. John Wiley & Sons, 2015.

[151] G. McVean, "A genealogical interpretation of principal components analysis," *PLoS Genet*, vol. 5, no. 10, p. e1000686, 2009.

[152] C. Widmer, C. Lippert, O. Weissbrod, N. Fusi, C. Kadie, R. Davidson, J. Listgarten, and D. Heckerman, "Further improvements to linear mixed models for genome-wide association studies," *Scientific reports*, vol. 4, no. 1, pp. 1–13, 2014.

[153] A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson, "New approaches to population stratification in genome-wide association studies," *Nature Reviews Genetics*, vol. 11, no. 7, pp. 459–463, 2010.

[154] L. Epping, B. Walther, R. M. Piro, M.-T. Knüver, C. Huber, A. Thürmer, A. Flieger, A. Fruth, N. Janecko, L. H. Wieler *et al.*, "Genome-wide insights into population structure and host specificity of campylobacter jejuni," *Scientific reports*, vol. 11, no. 1, pp. 1–15, 2021.

[155] L. Epping, J. C. Golz, M.-T. Knüver, C. Huber, A. Thürmer, L. H. Wieler, K. Stingl, and T. Semmler, "Comparison of different technologies for the decipherment of the whole genome sequence of campylobacter jejuni bfr-ca-14430," *Gut pathogens*, vol. 11, no. 1, pp. 1–8, 2019.

[156] International Organization for Standardization, *Microbiology of the Food Chain - Horizontal Method for Detection and Enumeration of Campylobacter Spp*, 2017.

[157] A. M. Mayr, S. Lick, J. Bauer, D. Thärigen, U. Busch, and I. Huber, "Rapid detection and differentiation of campylobacter jejuni, campylobacter coli, and campylobacter lari in food, using multiplex real-time PCR," *J. Food Prot.*, vol. 73, no. 2, pp. 241–250, Feb. 2010.

[158] E. L. Best, E. J. Powell, C. Swift, K. A. Grant, and J. A. Frost, "Applicability of a rapid duplex real-time PCR assay for speciation of campylobacter jejuni and campylobacter coli directly from culture plates," *FEMS Microbiol. Lett.*, vol. 229, no. 2, pp. 237–241, Dec. 2003.

[159] C. E. Miller, P. H. Williams, and J. M. Ketley, "Pumping iron: mechanisms for iron uptake by campylobacter," *Microbiology*, vol. 155, no. Pt 10, pp. 3157–3165, Oct. 2009.

[160] A. Thrash, M. Arick, and D. G. Peterson, "Quack: A quality assurance tool for high throughput sequence data," *Analytical Biochemistry*, vol. 548, pp. 38–43, 2018.

[161] M. Dodt, J. Roehr, R. Ahmed, and C. Dieterich, "FLEXBAR—Flexible barcode and adapter processing for Next-Generation sequencing platforms," *Biology*, vol. 1, no. 3, pp. 895–905, 2012.

[162] S. I. Nikolenko, A. I. Korobeynikov, and M. A. Alekseyev, "BayesHammer: Bayesian clustering for error correction in single-cell sequencing," *BMC Genomics*, vol. 14 Suppl 1, p. S7, Jan. 2013.

[163] P. A. Pevzner, H. Tang, and M. S. Waterman, "An eulerian path approach to DNA fragment assembly," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 17, pp. 9748–9753, Aug. 2001.

[164] R. R. Wick, L. M. Judd, C. L. Gorrie, and K. E. Holt, "Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads," *PLoS Comput. Biol.*, vol. 13, no. 6, p. e1005595, Jun. 2017.

[165] H. Li, "Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences," *Bioinformatics*, vol. 32, no. 14, pp. 2103–2110, Jul. 2016.

[166] B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, and A. M. Earl, "Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement," *PLoS One*, vol. 9, no. 11, p. e112963, Nov. 2014.

[167] Z. Zhou, N.-F. Alikhan, M. J. Sergeant, N. Luhmann, C. Vaz, A. P. Francisco, J. A. Carriço, and M. Achtman, "GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens," *Genome Res.*, vol. 28, no. 9, pp. 1395–1404, Sep. 2018.

[168] T. Seemann, "Prokka: rapid prokaryotic genome annotation," *Bioinformatics*, vol. 30, no. 14, pp. 2068–2069, Jul. 2014.

[169] D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser, "Prodigal: prokaryotic gene recognition and translation initiation site identification," *BMC Bioinformatics*, vol. 11, no. 1, 2010.

[170] A. J. Page, C. A. Cummins, M. Hunt, V. K. Wong, S. Reuter, M. T. G. Holden, M. Fookes, D. Falush, J. A. Keane, and J. Parkhill, "Roary: rapid large-scale prokaryote pan genome analysis," *Bioinformatics*, vol. 31, no. 22, pp. 3691–3693, Nov. 2015.

[171] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, Dec. 2012.

[172] K. Katoh, K. Misawa, K.-I. Kuma, and T. Miyata, "MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform," *Nucleic Acids Res.*, vol. 30, no. 14, pp. 3059–3066, Jul. 2002.

[173] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[174] S. Argimón, K. Abudahab, R. J. E. Goater, A. Fedosejev, J. Bhai, C. Glasner, E. J. Feil, M. T. G. Holden,

C. A. Yeats, H. Grundmann, B. G. Spratt, and D. M. Aanensen, "Microreact: visualizing and sharing data for genomic epidemiology and phylogeography," *Microb Genom*, vol. 2, no. 11, p. e000093, Nov. 2016.

[175] L. Uelze, J. Grützke, M. Borowiak, J. A. Hammerl, K. Juraschek, C. Deneke, S. H. Tausch, and B. Malorny, "Typing methods based on whole genome sequencing data," *One Health Outlook*, vol. 2, no. 1, pp. 1–19, 2020.

[176] J. Goris, K. T. Konstantinidis, J. A. Klappenbach, T. Coenye, P. Vandamme, and J. M. Tiedje, "DNA-DNA hybridization values and their relationship to whole-genome sequence similarities," *Int. J. Syst. Evol. Microbiol.*, vol. 57, no. Pt 1, pp. 81–91, Jan. 2007.

[177] L. G. Wayne, W. E. C. Moore, E. Stackebrandt, O. Kandler, R. R. Colwell, M. I. Krichevsky, H. G. Truper, R. G. E. Murray, P. A. D. Grimont, D. J. Brenner, M. P. Starr, and L. H. Moore, "Report of the ad hoc committee on reconciliation of approaches to bacterial systematics," *International Journal of Systematic and Evolutionary Microbiology*, vol. 37, no. 4, pp. 463–464, 1987.

[178] S. Ciufo, S. Kannan, S. Sharma, A. Badretdin, K. Clark, S. Turner, S. Brover, C. L. Schoch, A. Kimchi, and M. DiCuccio, "Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the ncbi," *International journal of systematic and evolutionary microbiology*, vol. 68, no. 7, p. 2386, 2018.

[179] C. Jain, L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis, and S. Aluru, "High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries," *Nat. Commun.*, vol. 9, no. 1, p. 5114, Nov. 2018.

[180] C. Jain, A. Dilthey, S. Koren, S. Aluru, and A. M. Phillippy, "A fast approximate algorithm for mapping long reads to large reference databases," *Journal of Computational Biology*, vol. 25, no. 7, pp. 766–779, 2018.

[181] R. Kolde and M. R. Kolde, "Package 'pheatmap'," *R Package*, vol. 1, no. 7, p. 790, 2015.

[182] A. P. Stamatakis, T. Ludwig, and H. Meier, "A fast program for maximum likelihood-based inference of large phylogenetic trees," in *Proceedings of the 2004 ACM symposium on Applied computing - SAC '04*, 2004.

[183] A. Stamatakis, "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies," *Bioinformatics*, vol. 30, no. 9, pp. 1312–1313, May 2014.

[184] B. Chor and T. Tuller, "Maximum likelihood of evolutionary trees: hardness and approximation," *Bioinformatics*, vol. 21 Suppl 1, pp. i97–106, Jun. 2005.

[185] F. J. Rohlf and F. James Rohlf, "J. felsenstein, inferring phylogenies, sinauer assoc., 2004, pp. xx 664," *J. Classification*, vol. 22, no. 1, pp. 139–142, 2005.

[186] J. Felsenstein, "Evolutionary trees from DNA sequences: A maximum likelihood approach," *J. Mol. Evol.*, vol. 17, no. 6, pp. 368–376, 1981.

[187] R. M. Miura, *Some Mathematical Questions in Biology: DNA Sequence Analysis*. American Mathematical Soc., Dec. 1986.

[188] Z. Yang, "Among-site rate variation and its impact on phylogenetic analyses," *Trends Ecol. Evol.*, vol. 11, no. 9, pp. 367–372, Sep. 1996.

[189] X. Didelot and D. J. Wilson, "ClonalFrameML: Efficient inference of recombination in whole bacterial genomes," *PLOS Computational Biology*, vol. 11, no. 2, p. e1004041, 2015.

[190] J. Corander and J. Tang, "Bayesian analysis of population structure based on linked molecular information," *Mathematical Biosciences*, vol. 205, no. 1, pp. 19–31, 2007.

[191] G. Tonkin-Hill, J. A. Lees, S. D. Bentley, S. D. W. Frost, and J. Corander, "Rhierbaps: An r implementation of the population clustering algorithm hierbaps," *Wellcome Open Research*, vol. 3, p. 93, 2018.

[192] J. A. Lees, M. Galardini, S. D. Bentley, J. N. Weiser, and J. Corander, "pyseer: a comprehensive tool for microbial pangenome-wide association studies," *Bioinformatics*, vol. 34, no. 24, pp. 4310–4312, Dec. 2018.

[193] M. Kokot, M. Długosz, and S. Deorowicz, "Kmc 3: counting and manipulating k-mer statistics," *Bioinformatics*, vol. 33, no. 17, pp. 2759–2761, 2017.

[194] M. Kokot, S. Deorowicz, and A. Debudaj-Grabysz, "Sorting data on Ultra-Large scale with RADULS," in *Communications in Computer and Information Science*, 2017, pp. 235–245.

[195] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, Jul. 2009.

[196] J. Huerta-Cepas, D. Szklarczyk, D. Heller, A. Hernández-Plaza, S. K. Forslund, H. Cook, D. R. Mende, I. Letunic, T. Rattei, L. J. Jensen *et al.*, "eggnog 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses," *Nucleic acids research*, vol. 47, no. D1, pp. D309–D314, 2019.

[197] P. Marttinen, W. P. Hanage, N. J. Croucher, T. R. Connor, S. R. Harris, S. D. Bentley, and J. Corander, "Detection of recombination events in bacterial genomes from large population samples," *Nucleic Acids Res.*, vol. 40, no. 1, p. e6, Jan. 2012.

[198] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2," *Nat. Methods*, vol. 9, no. 4, pp. 357–359, Mar. 2012.

[199] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup, "The sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009.

[200] A. R. Quinlan and I. M. Hall, "Bedtools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, pp. 841–842, Mar. 2010.

[201] D. Hermans, K. Van Deun, A. Martel, F. Van Immerseel, W. Messens, M. Heyndrickx, F. Haesebrouck, and F. Pasmans, "Colonization factors of campylobacter jejuni in the chicken gut," *Vet. Res.*, vol. 42, p. 82, Jun. 2011.

[202] K. Yahara, G. Méric, A. J. Taylor, S. P. W. de Vries, S. Murray, B. Pascoe, L. Mageiros, A. Torralbo, A. Vidal, A. Ridley, S. Komukai, H. Wimalarathna, A. J. Cody, F. M. Colles, N. McCarthy, D. Harris, J. E. Bray, K. A. Jolley, M. C. J. Maiden, S. D. Bentley, J. Parkhill, C. D. Bayliss, A. Grant, D. Maskell,

X. Didelot, D. J. Kelly, and S. K. Sheppard, "Genome-wide association of functional traits linked with campylobacter jejuni survival from farm to fork," *Environ. Microbiol.*, vol. 19, no. 1, pp. 361–380, Jan. 2017.

[203] A. Thépault, G. Méric, K. Rivoal, B. Pascoe, L. Mageiros, F. Touzain, V. Rose, V. Béven, M. Chemaly, and S. K. Sheppard, "Genome-Wide identification of Host-Segregating epidemiological markers for source attribution in campylobacter jejuni," *Appl. Environ. Microbiol.*, vol. 83, no. 7, Apr. 2017.

[204] F. J. Gormley, M. Macrae, K. J. Forbes, I. D. Ogden, J. F. Dallas, and N. J. C. Strachan, "Has retail chicken played a role in the decline of human campylobacteriosis?" *Appl. Environ. Microbiol.*, vol. 74, no. 2, pp. 383–390, Jan. 2008.

[205] S. Lévesque, E. Frost, R. D. Arbeit, and S. Michaud, "Multilocus sequence typing of campylobacter jejuni isolates from humans, chickens, raw milk, and environmental water in quebec, canada," *J. Clin. Microbiol.*, vol. 46, no. 10, pp. 3404–3411, Oct. 2008.

[206] B. M. Korczak, M. Zurfluh, S. Emler, J. Kuhn-Oertli, and P. Kuhnert, "Multiplex strategy for multilocus sequence typing, fla typing, and genetic determination of antimicrobial resistance of campylobacter jejuni and campylobacter coli isolates collected in switzerland," *J. Clin. Microbiol.*, vol. 47, no. 7, pp. 1996–2007, Jul. 2009.

[207] I. Habib, M. Uyttendaele, and L. De Zutter, "Survival of poultry-derived campylobacter jejuni of multilocus sequence type clonal complexes 21 and 45 under freeze, chill, oxidative, acid and heat stresses," *Food Microbiol.*, vol. 27, no. 6, pp. 829–834, Sep. 2010.

[208] T. Alter and K. Scherer, "Stress response of campylobacter spp. and its role in food processing," *Journal of Veterinary Medicine Series B*, vol. 53, no. 8, pp. 351–357, 2006.

[209] A. Noormohamed and M. K. Fakhr, "Molecular typing of campylobacter jejuni and campylobacter coli isolated from various retail meats by mlst and pfge," *Foods*, vol. 3, no. 1, pp. 82–93, 2014.

[210] E. Gripp, D. Hlahla, X. Didelot, F. Kops, S. Maurischat, K. Tedin, T. Alter, L. Ellerbroek, K. Schreiber, D. Schomburg *et al.*, "Closely related campylobacter jejuni strains from different sources reveal a generalist rather than a specialist lifestyle," *Bmc Genomics*, vol. 12, no. 1, pp. 1–21, 2011.

[211] J. Waldenström, T. Broman, I. Carlsson, D. Hasselquist, R. P. Achterberg, J. A. Wagenaar, and B. Olsen, "Prevalence of campylobacter jejuni, campylobacter lari, and campylobacter coli in different ecological guilds and taxa of migrating birds," *Applied and Environmental Microbiology*, vol. 68, no. 12, pp. 5911–5917, 2002.

[212] L. Morley, A. McNally, K. Paszkiewicz, J. Corander, G. Méric, S. K. Sheppard, J. Blom, and G. Manning, "Gene loss and lineage-specific restriction-modification systems associated with niche differentiation in the campylobacter jejuni sequence type 403 clonal complex," *Applied and Environmental Microbiology*, vol. 81, no. 11, pp. 3641–3647, 2015.

[213] G. Schröder and E. Lanka, "TraG-Like proteins of type IV secretion systems: Functional dissection of the multiple activities of TraG (RP4) and TrwB (r388)," *Journal of Bacteriology*, vol. 185, no. 15, pp. 4371–4381, 2003.

[214] F. Poly, D. Threadgill, and A. Stintzi, "Genomic diversity in campylobacter jejuni: Identification of c. jejuni 81-176-specific genes," *Journal of Clinical Microbiology*, vol. 43, no. 5, pp. 2330–2338, 2005.

[215] K.-Y. Lee, K.-Y. Lee, J.-H. Kim, I.-G. Lee, S.-H. Lee, D.-W. Sim, H.-S. Won, and B.-J. Lee, "Structure-based functional identification of helicobacter pylori hp0268 as a nuclease with both dna nicking and rnase activities," *Nucleic acids research*, vol. 43, no. 10, pp. 5194–5207, 2015.

[216] R. Hershberg, "Mutation–The engine of evolution: Studying mutation and its role in the evolution of bacteria," *Cold Spring Harb. Perspect. Biol.*, vol. 7, no. 9, p. a018077, Sep. 2015.

[217] E. Berthenet, A. Thépault, M. Chemaly, K. Rivoal, A. Ducournau, A. Buissonnière, L. Bénéjat, E. Bessède, F. Mégraud, S. K. Sheppard, and P. Lehours, "Source attribution of campylobacter jejuni shows variable importance of chicken and ruminants reservoirs in non-invasive and invasive french clinical isolates," *Sci. Rep.*, vol. 9, no. 1, p. 8098, May 2019.

[218] M. C. Brandley, D. L. Warren, A. D. Leaché, and J. A. McGuire, "Homoplasy and clade support," *Systematic Biology*, vol. 58, no. 2, pp. 184–198, 2009.

[219] A. Hassanin, G. Lecointre, and S. Tillier, "The 'evolutionary signal' of homoplasy in proteincoding gene sequences and its consequences for a priori weighting in phylogeny," *Comptes Rendus de l'Académie des Sciences - Series III - Sciences de la Vie*, vol. 321, no. 7, pp. 611–620, 1998.

[220] Z. Han, T. Willer, L. Li, C. Pielsticker, I. Rychlik, P. Velge, B. Kaspers, and S. Rautenschlein, "Influence of the gut microbiota composition on campylobacter jejuni colonization in chickens," *Infect. Immun.*, vol. 85, no. 11, Nov. 2017.

[221] I. Indikova, T. J. Humphrey, and F. Hilbert, "Survival with a helping hand: Campylobacter and microbiota," *Front. Microbiol.*, vol. 6, p. 1266, Nov. 2015.

[222] R. Motiejūnaitė, J. Armalytė, A. Markuckas, and E. Sužiedėlienė, "Escherichia coli dinJ-yafQgenes act as a toxin-antitoxin module," *FEMS Microbiology Letters*, vol. 268, no. 1, pp. 112–119, 2007.

[223] L. Buts, J. Lah, M.-H. Dao-Thi, L. Wyns, and R. Loris, "Toxin-antitoxin modules as bacterial metabolic stress managers," *Trends Biochem. Sci.*, vol. 30, no. 12, pp. 672–679, Dec. 2005.

[224] K. Gerdes, S. K. Christensen, and A. Løbner-Olesen, "Prokaryotic toxin–antitoxin stress response loci," *Nature Reviews Microbiology*, vol. 3, no. 5, pp. 371–382, 2005.

[225] I. J. Fijalkowska, R. M. Schaaper, and P. Jonczyk, "DNA replication fidelity inescherichia coli: a multi-DNA polymerase affair," *FEMS Microbiology Reviews*, vol. 36, no. 6, pp. 1105–1121, 2012.

[226] D. Vandewiele, A. R. F. de Henestrosa, A. R. Timms, B. A. Bridges, and R. Woodgate, "Sequence analysis and phenotypes of five temperature sensitive mutator alleles of dnae, encoding modified α-catalytic subunits of escherichia coli DNA polymerase III holoenzyme," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 499, no. 1, pp. 85–95, 2002.

[227] S.-O. Shan, R. M. Stroud, and P. Walter, "Mechanism of association and reciprocal activation of two GTPases," *PLoS Biology*, vol. 2, no. 10, p. e320, 2004.

[228] I. Yosef, E. S. Bochkareva, and E. Bibi, "Escherichia coli SRP, its protein subunit ffh, and the ffh M domain are able to selectively limit membrane protein expression when overexpressed," *mBio*, vol. 1, no. 2, 2010.

[229] M. Balaban, S. N. Joslin, and D. R. Hendrixson, "FlhF and its GTPase activity are required for distinct processes in flagellar gene regulation and biosynthesis in campylobacter jejuni," *J. Bacteriol.*, vol. 191, no. 21, pp. 6602–6611, Nov. 2009.

[230] S. Budroni, E. Siena, J. C. Dunning Hotopp, K. L. Seib, D. Serruto, C. Nofroni, M. Comanducci, D. R. Riley, S. C. Daugherty, S. V. Angiuoli, A. Covacci, M. Pizza, R. Rappuoli, E. R. Moxon, H. Tettelin, and D. Medini, "Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 11, pp. 4494–4499, Mar. 2011.

[231] N. D. McCarthy, F. M. Colles, K. E. Dingle, M. C. Bagnall, G. Manning, M. C. J. Maiden, and D. Falush, "Host-associated genetic import in campylobacter jejuni," *Emerg. Infect. Dis.*, vol. 13, no. 2, pp. 267–272, Feb. 2007.

[232] National Research Council, *Nutrient Requirements of Swine*, Jul. 2012.

[233] G. Schröder, S. Krause, E. L. Zechner, B. Traxler, H.-J. Yeo, R. Lurz, G. Waksman, and E. Lanka, "TraG-Like proteins of DNA transfer systems and of the helicobacter pylori type IV secretion system: Inner membrane gate for exported substrates?" *Journal of Bacteriology*, vol. 184, no. 10, pp. 2767–2779, 2002.

[234] S. Kienesberger, C. S. Trummler, A. Fauster, S. Lang, H. Sprenger, G. Gorkiewicz, and E. L. Zechner, "Interbacterial macromolecular transfer by the campylobacter fetus subsp. venerealis type IV secretion system," *Journal of Bacteriology*, vol. 193, no. 3, pp. 744–758, 2011.

[235] J. Velayudhan and D. J. Kelly, "Analysis of gluconeogenic and anaplerotic enzymes in campylobacter jejuni: an essential role for phosphoenolpyruvate carboxykinase," *Microbiology*, vol. 148, no. 3, pp. 685–694, 2002.

[236] B. M. Korczak, R. Stieber, S. Emler, A. P. Burnens, J. Frey, and P. Kuhnert, "Genetic relatedness within the genus campylobacter inferred from rpob sequences," *Int. J. Syst. Evol. Microbiol.*, vol. 56, no. Pt 5, pp. 937–945, May 2006.

[237] A. González-González, S. M. Hug, A. Rodríguez-Verdugo, J. S. Patel, and B. S. Gaut, "Adaptive mutations in RNA polymerase and the transcriptional terminator rho have similar effects on escherichia coli gene expression," *Mol. Biol. Evol.*, vol. 34, no. 11, pp. 2839–2855, Nov. 2017.

[238] S. A. Richards, "The significance of changes in the temperature of the skin and body core of the chicken in the regulation of heat loss," *J. Physiol.*, vol. 216, no. 1, pp. 1–10, Jul. 1971.

[239] J. Iranzo, Y. I. Wolf, E. V. Koonin, and I. Sela, "Gene gain and loss push prokaryotes beyond the homologous recombination barrier and accelerate genome sequence divergence," *Nat. Commun.*, vol. 10, no. 1, p. 5376, Nov. 2019.

[240] C. Riedel, K. U. Förstner, C. Püning, T. Alter, C. M. Sharma, and G. Gölz, "Differences in the transcriptomic response of campylobacter coli and campylobacter lari to heat stress," *Frontiers in Microbiology*, vol. 11, 2020.

[241] E. P. on Biological Hazards (BIOHAZ), K. Koutsoumanis, A. Allende, A. Alvarez-Ordóñez, D. Bolton, S. Bover-Cid, R. Davies, A. De Cesare, L. Herman, F. Hilbert *et al.*, "Update and review of control options for campylobacter in broilers at primary production," *EFSA Journal*, vol. 18, no. 4, p. e06090, 2020.

[242] G. Wang, C. G. Clark, T. M. Taylor, C. Pucknell, C. Barton, L. Price, D. L. Woodward, and F. G. Rodgers, "Colony multiplex PCR assay for identification and differentiation of campylobacter jejuni, c. coli, c. lari, c. upsaliensis, and c. fetus subsp. fetus," *Journal of Clinical Microbiology*, vol. 40, no. 12, pp. 4744–4747, 2002.

[243] M. J. Figueras, R. Beaz-Hidalgo, M. J. Hossain, and M. R. Liles, "Taxonomic affiliation of new genomes should be verified using average nucleotide identity and multilocus phylogenetic analysis," *Genome Announcements*, vol. 2, no. 6, 2014.

[244] M. J. Jansen van Rensburg, C. Swift, A. J. Cody, C. Jenkins, and M. C. J. Maiden, "Exploiting bacterial Whole-Genome sequencing data for evaluation of diagnostic assays: Campylobacter species identification as a case study," *J. Clin. Microbiol.*, vol. 54, no. 12, pp. 2882–2890, Dec. 2016.

[245] M. J. J. van Rensburg, C. Swift, A. J. Cody, C. Jenkins, and M. C. Maiden, "Exploiting bacterial whole-genome sequencing data for evaluation of diagnostic assays: Campylobacter species identification as a case study," *Journal of clinical microbiology*, vol. 54, no. 12, pp. 2882–2890, 2016.

[246] M. F. M. Ahmed, J. Schulz, and J. Hartung, "Survival of campylobacter jejuni in naturally and artificially contaminated laying hen feces," *Poult. Sci.*, vol. 92, no. 2, pp. 364–369, Feb. 2013.

[247] X. T. Bui, A. Wolff, M. Madsen, and D. D. Bang, "Reverse transcriptase real-time PCR for detection and quantification of viable campylobacter jejuni directly from poultry faecal samples," *Res. Microbiol.*, vol. 163, no. 1, pp. 64–72, Jan. 2012.

[248] A. Flint and A. Stintzi, "Cj1386, an atypical hemin-binding protein, mediates hemin trafficking to KatA in campylobacter jejuni," *J. Bacteriol.*, vol. 197, no. 5, pp. 1002–1011, Mar. 2015.

[249] J. M. Atack and D. J. Kelly, "Contribution of the stereospecific methionine sulphoxide reductases MsrA and MsrB to oxidative and nitrosative stress resistance in the food-borne pathogen campylobacter jejuni," *Microbiology*, vol. 154, no. Pt 8, pp. 2219–2230, Aug. 2008.

[250] H. Al-Haideri, M. A. White, and D. J. Kelly, "Major contribution of the type II beta carbonic anhydrase CanB (cj0237) to the capnophilic growth phenotype of campylobacter jejuni," *Environ. Microbiol.*, vol. 18, no. 2, pp. 721–735, Feb. 2016.

[251] R. J. Jackson, K. T. Elvers, L. J. Lee, M. D. Gidley, L. M. Wainwright, J. Lightfoot, S. F. Park, and R. K. Poole, "Oxygen reactivity of both respiratory oxidases in campylobacter jejuni: the cydAB genes encode a Cyanide-Resistant, Low-Affinity oxidase that is not of the cytochrome bd type," *Journal of Bacteriology*, vol. 189, no. 5, pp. 1604–1615, 2007.

[252] K. T. Baek, C. S. Vegge, J. Skórko-Glonek, and L. Brøndsted, "Different contributions of HtrA protease and chaperone activities to campylobacter jejuni stress tolerance and physiology," *Appl. Environ. Microbiol.*, vol. 77, no. 1, pp. 57–66, Jan. 2011.

[253] M. Boehm, B. Hoy, M. Rohde, N. Tegtmeyer, K. T. Bæk, O. A. Oyarzabal, L. Brøndsted, S. Wessler,

and S. Backert, "Rapid paracellular transmigration of campylobacter jejuni across polarized epithelial cells without affecting TER: role of proteolytic-active HtrA cleaving e-cadherin but not fibronectin," *Gut Pathogens*, vol. 4, no. 1, p. 3, 2012.

[254] M. Boehm, J. Lind, S. Backert, and N. Tegtmeyer, "Campylobacter jejuni serine protease HtrA plays an important role in heat tolerance, oxygen resistance, host cell adhesion, invasion, and transmigration," *Eur. J. Microbiol. Immunol.*, vol. 5, no. 1, pp. 68–80, Mar. 2015.

[255] V. Phongsisay, V. N. Perera, and B. N. Fry, "Expression of the htrb gene is essential for responsiveness of salmonella typhimurium and campylobacter jejuni to harsh environments," *Microbiology*, vol. 153, no. Pt 1, pp. 254–262, Jan. 2007.

[256] K. T. Elvers, G. Wu, N. J. Gilberthorpe, R. K. Poole, and S. F. Park, "Role of an inducible single-domain hemoglobin in mediating resistance to nitric oxide and nitrosative stress in campylobacter jejuni and campylobacter coli," *J. Bacteriol.*, vol. 186, no. 16, pp. 5332–5341, Aug. 2004.

[257] B. Chaban, K. M. Musil, C. G. Himsworth, and J. E. Hill, "Development of cpn60-based real-time quantitative PCR assays for the detection of 14 campylobacter species and application to screening of canine fecal samples," *Appl. Environ. Microbiol.*, vol. 75, no. 10, pp. 3055–3061, May 2009.

[258] P. Kralik and M. Ricchi, "A basic guide to real time pcr in microbial diagnostics: definitions, parameters, and everything," *Frontiers in microbiology*, vol. 8, p. 108, 2017.

[259] M. Pérez-Losada, M. Arenas, and E. Castro-Nallar, "Microbial sequence typing in the genomic era," *Infection, Genetics and Evolution*, vol. 63, pp. 346–359, 2018.

[260] S. Shams, B. Bakhshi, and T. Tohidi Moghadam, "In silico analysis of the cadf gene and development of a duplex polymerase chain reaction for Species-Specific identification of campylobacter jejuni and campylobacter coli," *Jundishapur J Microbiol*, vol. 9, no. 2, p. e29645, Feb. 2016.

[261] N. Toplak, M. Kovač, S. Piskernik, S. S. Možina, and B. Jeršek, "Detection and quantification of campylobacter jejuni and campylobacter coli using real-time multiplex PCR," *J. Appl. Microbiol.*, vol. 112, no. 4, pp. 752–764, Apr. 2012.

[262] K. Chan, D. Elhanafi, and S. Kathariou, "Genomic evidence for interspecies acquisition of chromosomal DNA from campylobacter jejuni by campylobacter coli strains of a turkey-associated clonal group (cluster II)," *Foodborne Pathog. Dis.*, vol. 5, no. 4, pp. 387–398, Aug. 2008.

[263] A. Igwaran and A. I. Okoh, "Human campylobacteriosis: A public health concern of global importance," *Heliyon*, vol. 5, no. 11, p. e02814, 2019.

[264] J. A. Lees, B. Ferwerda, P. H. Kremer, N. E. Wheeler, M. V. Serón, N. J. Croucher, R. A. Gladstone, H. J. Bootsma, N. Y. Rots, A. J. Wijmega-Monsuur *et al.*, "Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis," *Nature communications*, vol. 10, no. 1, pp. 1–14, 2019.

[265] N. Arning, S. K. Sheppard, D. A. Clifton, and D. J. Wilson, "Machine learning to predict the source of campylobacteriosis using whole genome data," *bioRxiv*, 2021.

[266] M. Baym, T. D. Lieberman, E. D. Kelsic, R. Chait, R. Gross, I. Yelin, and R. Kishony, "Spatiotemporal microbial evolution on antibiotic landscapes," *Science*, vol. 353, no. 6304, pp. 1147–1151, 2016.

[267] T. Semmler, E. M. Harrison, A. Lübke-Becker, R. G. Ulrich, L. H. Wieler, S. Guenther, I. Stamm, A.-M. Hanssen, M. A. Holmes, S. Vincze *et al.*, "A look into the melting pot: The mec c-harboring region is a recombination hot spot in staphylococcus stepanovicii," *PLoS One*, vol. 11, no. 1, p. e0147150, 2016.

[268] E. Bessède, O. Solecki, E. Sifré, L. Labadi, and F. Mégraud, "Identification of campylobacter species and related organisms by matrix assisted laser desorption ionization–time of flight (MALDI-TOF) mass spectrometry," *Clinical Microbiology and Infection*, vol. 17, no. 11, pp. 1735–1739, 2011.

[269] M. F. Emele, S. S. Možina, R. Lugert, W. Bohne, W. O. Masanta, T. Riedel, U. Groß, O. Bader, and A. E. Zautner, "Proteotyping as alternate typing method to differentiate campylobacter coli clades," *Sci. Rep.*, vol. 9, no. 1, p. 4244, Mar. 2019.

[270] Y.-H. Hsieh, Y. F. Wang, H. Moura, N. Miranda, S. Simpson, R. Gowrishankar, J. Barr, K. Kerdahi, and I. M. Sulaiman, "Application of MALDI-TOF MS systems in the rapid identification of campylobacter spp. of public health importance," *Journal of AOAC INTERNATIONAL*, vol. 101, no. 3, pp. 761–768, 2018.

Zusammenfassung

Verschiedene Spezies der Gattung *Campylobacter* (*C.*) sind zoonotische Krankheitserreger, die zu den Hauptverursachern von durch Lebensmittel übertragbare Infektionskrankheiten weltweit gehören. Obwohl *C. jejuni* und *C. coli* unterschiedliche Wirte wie Geflügel, Vieh und Wildtiere kolonisieren, sind die Mechanismen, die es diesen Bakterien ermöglichen, sich an neue ökologische Nischen anzupassen, nicht vollständig geklärt. In dieser Arbeit wurden neue *k-mer*-basierte Methoden für Hochdurchsatzanalysen von Ganzgenomsequenzierungen von *C. jejuni* und *C. coli* entwickelt, erweitert und angewendet, um das Anpassungspotenzial der Spezies an unterschiedliche Wirte und Umgebungen zu untersuchen.

In der ersten Studie wurde eine auf *k-meren* basierende mikrobielle genomweite Assoziationsstudie (GWAS) durchgeführt, um wirtsspezifische genomische *C. jejuni*-Signaturen von Isolaten aus Hühnern, Rindern, Schweinen und humanen klinischen Proben zu identifizieren. Die GWAS zeigte eine starke Assoziation sowohl des Kern- wie auch des akzessorischen *C. jejuni* Genoms mit verschiedenen Wirtstieren. Durch die *in silico* Prädiktion von Veränderungen in Peptiden bzw. Proteinen ist es gelungen, mehrere adaptive metabolische Pfade zu identifiziert, welche potentiell die Evolution der Wirtspräferenz von phylogenetisch unterschiedlichen *C. jejuni* an verschiedene Lebensräume ermöglichen.

In einem zweiten Ansatz wurden Ganzgenomsequenzen von *Campylobacter* Isolaten analysiert, die in der Routinediagnostik mittels Polymerase-Kettenreaktionen (PCR) nicht eindeutig einer genauen Spezies zuzuordnen waren. Die *Campylobacter* Genome aus diesen Proben wurden hinsichtlich ihres genomischen Aufbaus untersucht. Zu diesem Zweck wurde eine *k-mer*-basierte Methode entwickelt, um Rekombinationsereignisse zwischen *C. jejuni* und *C. coli* zu identifizieren, welche maßgeblich die Ergebnisse der PCR beeinflussten. Die auf diese Weise identifizierten Gene kodieren häufig Proteine mit wichtiger Funktion in der Chromosomenerhaltung bzw. DNA Reparatur, im Membrantransport und Stressabwehr.

Die in dieser Arbeit vorgestellten Ergebnisse leisten einen Beitrag zur routinemäßigen Überwachung und schnellen Diagnostik von *Campylobacter* Ausbrüchen im Sinne einer integrierten molekularen Surveillance. Wirtsspezifische Allele, die in *Campylobacter* mit unterschiedlichen phylogenetischen Hintergründen identifiziert wurden, können dabei als wichtige Markergene dienen, um die ursprüngliche Quelle des Ausbruchs schnell und präzise retrograd entlang der Lebensmittelkette zu identifizieren.

Eigenständigkeitserklärung

Ich versichere, dass ich die hier vorgelegte Dissertation selbstständig angefertigt habe und die benutzten Quellen und Hilfsmittel vollständig angegeben sind. Ein Promotionsverfahren wurde zu keinem früheren Zeitpunkt an einer anderen in- oder ausländischen Hochschule oder bei einem anderen Fachbereich beantragt. Die Bestimmungen der Promotionsordnung sind mir bekannt.

Lennard Epping, Berlin, 08. Oktober 2021