The evolution of the pan-genome of Shiga-toxin (Stx) producing *Escherichia coli* and the Stx_2 bacteriophage

CHAD LAING
University of Lethbridge

A Thesis
Submitted to the School of Graduate Studies
of the University of Lethbridge
in Partial Fulfillment of the
Requirements for the Degree

DOCTOR OF PHILOSOPHY

Department of Biological Sciences
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

Sooner or later you'll have to see
The cause and effect
So many things still left to do
But we haven't made it yet
–Neil Young, 'Transformer Man'

Abstract

Human infections with Shiga-toxin (Stx)-producing *E. coli* (STEC) vary in severity of illness. The pan-genome of a bacterial species contains a shared, essential core genome, and a variably distributed accessory genome. While single nucleotide changes likely influence virulence in STEC, horizontal gene transfer (HGT) on elements such as bacteriophage are thought to be most important. My thesis objectives were to: 1) develop tools for the pan-genomic analyses of bacterial genomes; 2) describe the phylogeny of STEC and; 3) determine if the evolution of the Stx_2-bacteriophage parallels that of its bacterial host. For this thesis, the software program Panseq was created and used to identify pan-genomic differences among STEC. Whole-genome phylogenies showed all serotypes as discrete clusters, with O157:H7 having three distinct lineages and grouping separately from all other STEC. Finally, the phylogenies of Stx2-bacteriophage and their bacterial hosts were largely concordant, with occasional instances of HGT having led to novel pathogen emergence.

Acknowledgments

Thank you to everyone involved.

Contents

List of Tables

List of Figures

List of Abbreviations

AAF	Aggregative Adherence Fimbriae
A/E	Attaching and Effacing
AIEC	Adherent-invasive *E. coli*
AMR	Anti-microbial Resistance
bp	base pair
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-like Alignment Tool
BLV	Bovine Leukemia Virus
CAGF	Core / Accessory Genome Finder
CGF	Comparative Genomic Fingerprinting
CFU	Colony Forming Unit
mCGH	microarray-Comparative Genomic Hybridization
DAEC	Diffusely Adherent *E. coli*
EAEC	Enteroaggregative *E. coli*
EHEC	Enterohemorrhagic *E. coli*
EIEC	Enteroinvasive *E. coli*
ELISA	Enzyme-linked Immunosorbent Assay
EPEC	Enteropathogenic *E. coli*
ER	Endoplasmic Reticulum
ERIC	Enterobacterial Repetitive Intergenic Consensus
GAS	Group A *Streptococcus*
GISSH	Genomic *In Silico* Subtractive Hybridization
HGT	Horizontal Gene Transfer
HUS	Hemolytic Uremic Syndrome
IHF	Integration Host Factor
kbp	kilo base pair
LEE	Locus for Enterocyte Attachment and Effacement
LS	Loci Selector
MLEE	Multi-locus Enzyme Electrophoresis
MLST	Multi-locus Sequence Typing
MLVA	Multi-locus Variable Number Tandem Repeat Analysis
mbp	Mega base pair
m	Minute
ML	Maximum Likelihood

MP	Maximum Parsimony
MRSA	Methicillin-resistant *Staphylococcus aureus*
NLE	non-LEE encoded effector
NRF	Novel Region Finder
OBGS	Octamer-based Genome Scanning
PCR	Polymerase Chain Reaction
PFGE	Pulsed-field Gel Electrophoresis
PGM	Personal Genome Machine
POD	Points of Discrimination
PT	Phage Type
QS	Quorum Sensing
RAPD	Random Amplification of Polymorphic DNA
RT	Room Temperature
s	Second
SNP	Single Nucleotide Polymorphism
ST	Sequence Type
STEC	Shiga-toxin producing *E. coli*
SVG	Scalable Vector Graphic
VNTR	Variable Number Tandem Repeats
VTEC	Verocytotoxin-producing *E. coli*

Chapter 1

Introduction

My thesis was broadly constructed around the following objectives: 1) Develop tools for the analyses of bacterial genomes in a pan-genomic context; 2) Describe the overall phylogeny of *Eschericha coli* O157:H7 and other Shiga toxin-producing *E. coli* (STEC); 3) Determine if the evolution of the Shiga-toxin 2 producing bacteriophage parallels that of its bacterial STEC host. The following is an overview of my thesis, which briefly highlights how each chapter addresses these objectives.

In Chapter 2, 'Everything at once: Comparative analyses of the genomes of bacterial pathogens', I review the comparative genomics of bacterial pathogens that affect humans. I discuss how whole-genome sequencing (WGS) is influencing the definition of clusters of genetically and phenotypically related organisms and how it is used in molecular epidemiology to 'fingerprint' disease agents and determine sources. The use of WGS in identifying genes or genomic regions that may be responsible for phenotypic attributes such as virulence or niche specialization is discussed, as is elucidating the mechanisms of emergence and evolution of pathogenic microorganisms. Lastly, the evolution of both epidemiology and novel gene discovery, current methods of whole-genome analyses and the influence of low-cost, high-throughput sequencing on microbiology are discussed.

In Chapter 3, 'A review of Shiga-toxin producing *E. coli* virulence and genomic evolution', I discuss the major virulence factors of STEC, including Shiga-toxins, the locus of enterocyte attachment and effacement and other chromosome- and plasmid-encoded virulence factors. The composition of the STEC genome is discussed in terms of parallel evolution and the dynamics of STEC evolution are briefly discussed. This chapter highlights the heterogeneous makeup of STEC and the need for whole-genome comparisons to fully understand the population structure and evolution of the pathogen.

In Chapter 4, ‘*In silico* genomic analyses reveal three distinct lineages of *Escherichia coli* O157:H7, one of which is associated with hyper-virulence’, I discuss the *E. coli* O157:H7 lineages and their stepwise emergence from *E. coli* O55:H7, as well as the molecular typing methods that have traditionally been used to assess population structure and diversity of the serotype. In this study an *in silico* comparison of six different genotyping approaches was performed on 19 *E. coli* genome sequences from 17 O157:H7 strains and single O145:NM and K12 MG1655 strains to provide an overall picture of diversity of the *E. coli* O157:H7 population, and to compare genotyping methods for O157:H7 strains. I determined that by combining six individual typing methods *in silico* into a supernetwork representation that three distinct clusters of O157:H7 strains were observed, and that each of these O157:H7 strain clusters was lineage-specific. Additionally, a clade of O157:H7 associated with hyper-virulence was found to be part of O157:H7 lineage I/II. These lineage I/II strains clustered closest on the supernetwork to *E. coli* K12 and *E. coli* O55:H7, O145:NM and sorbitol-fermenting O157 strains. The results of my study highlighted the similarities in relationships derived from multi-locus genome sampling methods and suggested that a ‘common genotyping language’ should be used for population genetics and epidemiological studies. It is now clear that WGS has become the *de facto* standard genotyping language for bacteria, and the subsequent chapters of my thesis adopt this paradigm for analyses.

In Chapter 5, ‘Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions’, I discuss the concept of the pan-genome and how WGS allows an unprecedented examination of bacterial pan-genomes. The pan-genome of a bacterial species contains a shared, essential core genome, and a variably distributed accessory genome. The remainder of the chapter describes Panseq, the pan-genomic software that I created, and its use in analyzing the pan-genome of bacterial groups, identifying regions unique to a genome or group of genomes, identifying single-nucleotide polymorphisms (SNPs) in the core genome, and constructing phylogenies based on the presence / absence of accessory regions as well as SNPs within shared core genomic regions. I also describe the use of Panseq as a loci selector that calculates the most variable and discriminatory loci among sets of accessory loci or core genome SNPs. I demonstrate the utility of Panseq for a number of bacterial species and show that branch support successively increased between trees created from multi-locus sequence typing, core-genome, and pan-genome loci among *Salmonella enterica* strains.

In Chapter 6, ‘A comparison of Shiga-toxin 2 bacteriophage from classical enterohemorrhagic *Escherichia coli* serotypes and the German *E. coli* O104:H4 outbreak strain’, I discuss how Stx_2-bacteriophage are mobile and often contain ‘passenger genes’ that can account for large amounts of horizontal-gene transfer among STEC. Because of this I hypothesized that the Stx_2-bacteriophage evolutionary history would not necessarily mimic that of their bacterial hosts. To test this, I used Panseq to analyze the phylogeny of 52 Stx_2-bacteriophage from 42 STEC, and compared the phylogeny to that of the whole-genome STEC phylogeny. I found that all serotypes in the whole-genome tree formed discrete clusters, and that the tree was divided into two separate branches, with O157:H7 strains comprising one branch, and all other STEC strains the other. Within the large O157:H7 branch, three sub-groups that corresponded to genetic lineages I, I/II and II were observed. Despite the expectation that Stx_2-phage would be highly heterogeneous, most were highly concordant with the core-genome phylogenetic trees of their bacterial hosts, suggesting that the host and phage had stably co-evolved for significant periods of time. I found that of the 42 whole-genomes examined, only *E. coli* serotype O111 strains, and an O145:H2 strain were placed differently on the whole-genome tree compared to the phylogeny of the Stx_2-phage, likely indicating horizontal acquisition of the phage from bacteria with different evolutionary histories. Lastly I examined the WGS from the German *E. coli* O104:H4 strain and its Stx2-bacteriophage from the outbreak in 2011 and found that among the *E.*

coli Stx_2-phage sequences studied, the Stx_2-bacteriophage from O111:H- strain JB1-95 was most closely related phylogenetically to the Stx_2-phage from the O104:H4 outbreak strain. This observation indicated that Stx_2-phage are capable of integrating into diverse *E. coli* genomic backgrounds and of qualitatively transforming health risks associated with *E. coli* strains from other pathogroups such as enteropathogenic *E. coli* and enteroaggregative *E. coli*.

In Chapter 7, the 'Thesis summary' chapter, I discuss the major contributions of my thesis. I place the thesis into the context of current whole-genome sequencing availability and pricing, and software that will be needed in the future to perform routine biological studies. I also place the results of my STEC investigations into the wider picture of the *E. coli* species as a whole and suggest avenues for continued work in both bioinformatics and comparative genomics of STEC.

Chapter 2

Everything at once: Comparative analyses of bacterial pathogen genomes

2.1 Preface

This chapter was previously published as: Laing CR, Zhang Y, Thomas JE, Gannon VP. 'Everything at once: comparative analysis of the genomes of bacterial pathogens'. Vet Microbiol. 2011 Nov 21;153(1-2):13-26. doi: 10.1016/j.vetmic.2011.06.014. [5]

2.2 Introduction

This review focuses on the comparative genomics of both human and animal bacterial pathogens. Comparative genomics has found a growing use in: (1) the definition of clusters of genetically and phenotypically related organisms that previously were not captured by taxonomic schemes based on a limited number of attributes such as biotype, serotype, phage type, etc., (2) molecular epidemiology for use in the 'fingerprinting' of disease agents and in determining their likely source, (3) in identifying particular genes or genomic regions that may be responsible for phenotypic attributes such as virulence or niche specialization, and (4) for elucidating the mechanisms of emergence and evolution of pathogenic microorganisms. While simple phenotypic and genetic assays that rely on a limited number of informative loci will continue to play an important role in medical and veterinary microbiology, the availability of whole-genomic sequences at reduced costs has revolutionized our ability to discriminate among and characterize bacterial strains and bacterial strain clusters. In addition, comparative genomics enhances our ability to understand the complex genetic systems that are responsible for differences in phenotypes among bacteria beyond what has been possible with simple gene deletion and complementation studies. While high-throughput genomic sequencing technology has presented us with many opportunities, software tools capable of handling these data are just beginning to be developed. In this review we will discuss the evolution of both epidemiology and novel gene discovery, discuss current methods of whole-genome analysis and briefly examine the profound influence low-cost, high-throughput sequencing will have in the field of veterinary microbiology.

2.3 The Pan-genome

The sum of unique genes in all genomes of a species is referred to as the pan-genome [6]. Variation within the pan-genome forms the basis of genotyping techniques, such as pulsed-field gel electrophoresis (PFGE) [7] and repetitive-sequence-based (rep)-PCR [8, 9]. However, the large amount of variation that has been observed through whole-genome sequencing of pathogenic bacteria was not anticipated in the pre-genomics era. This variation among the genomes of strains from the same or related species is largely attributable to horizontal gene transfer (HGT) [10]. In addition to single nucleotide polymorphisms (SNPs) and small insertion / deletion (indel) regions, large-scale indel events were observed when the first two completely sequenced bacterial genomes of the same species, *Helicobacter pylori*, were compared [11]. Subsequent comparisons of strains from the same bacterial species found HGT to be commonplace among taxa such as *Escherichia coli* [12, 13], *Streptococcus* [14, 15] and *Salmonella* [16]. Such variation in the ‘accessory’ genome, or those genes outside the ‘core’ set shared by all members of the species, are presumed not to be necessary for basic cellular function; however, they may contribute to phenotypes that allow survival in specific niches or environmental conditions [17]. These large insertions of DNA containing clusters of genes are known as ‘genomic islands’, and those specifically associated with virulence as ‘pathogenicity islands’ to differentiate them from ‘fitness islands’, which are thought to play more of a role in niche adaptation [18]. Occasionally, fitness enhancing genes in one host may increase the virulence of some strains in other incidental hosts such as humans [19]. Genomic islands conferring ecological or metabolic advantages can therefore be thought of generally as adaptive islands. Not all bacterial species studied have exhibited such genomic mosaicism, for example strains of the obligate intracellular pathogen *Chlamydia* [20] and human / animal pathogen *Bacillus anthracis* [21] differ mainly in core gene mutations rather than the gain or loss of large accessory genomic regions. Such genomic similarity is thought to be due to genetic isolation in specific niches, which dampens the ability of bacteria to acquire new genes from related organisms [10]. The mechanisms by which HGT takes place among many bacteria are thought to be largely bacteriophage related, although examples of plasmid- and transposon-mediated exchange are also observed [22, 23]. Temperate bacteriophages are distinguished by their propensity to transfer virulence factors. For example, the Shiga-toxin(s) in Shiga-toxin producing *E. coli* [24], the cholera toxin in *Vibrio cholera* [25] and the diphtheria toxin in *Corynebacterium diphtheria* [26] all reside on bacteriophage-related genomic islands. These accessory genome elements are often the reason why certain members of a taxonomic group are pathogenic, while other closely related members are not; therefore the accessory genome is of great importance to microbiologists. During the evolutionary process, accessory genes that provide a selective advantage in a particular niche soon become established in clusters of related strains and thus become part of the core genome for these specific groups. Despite the frequently advantageous role of HGT, not all foreign DNA is beneficial to a microorganism. Many bacteria with high rates of HGT, such as *Pseudomonas* and *E. coli*, contain genetic mechanisms to silence genes acquired through HGT, as the acquired genes may negatively affect the cell through disturbance of essential gene expression networks, or otherwise create genome instability [27, 28].

While the importance of ‘novel-gene’ acquisition cannot be overstated, SNPs in the core genome can also influence phenotype, such as fluoroquinolone resistance in *Salmonella enterica* serovar Typhi [29] and the decreased ability to cause necrotizing fasciitis in group A *Streptococcus* [30]. Additionally, the loss of accessory genes can sometimes lead to the expression of a pathogenic phenotype through the activation of a pathway that was inhibited by a particular gene. In this way, not only gene acquisition, but also gene loss may give rise to a more virulent phenotype, which may have occurred with species of *Rickettsia* [31]. Other more virulent subgroups of a species often show genome reduction, but gene loss cannot be absolutely said to explain increased virulence; for example *Mycobacterium ulcerans*, the causative agent of Buruli ulcer has a reduced genome in comparison to the fish-pathogen and likely-ancestor, *Mycobacterium marinum*, but has also acquired a large plasmid that produces the toxin present in ulcer-producing strains [32]. In this case, genome loss and plasmid acquisition may both be associated with adaptation to a specific niche. Identifying differences within the pan-genome of virulent and avirulent members of the same species can help elucidate potential genetic mechanisms that explain phenotypic differences among strains. This type of comparison has been and will continue to be greatly facilitated by high-throughput genomic sequencing.

2.4 Molecular Genotyping via Phenotypic Differences

Differences in biochemical activities among bacteria, or biotyping, have been used since the beginnings of bacteriology to separate strains into phenotypically distinct groups. Relatively few biochemical tests are needed to identify groups of bacteria at the species level and occasionally they can be used to identify sub-types within a species [33]. In the case of animal or human pathogens, pathogenic and non-pathogenic strains can occasionally be identified by biotyping, but often pathogens and non-pathogens belong to the same biotype.

While a small number of biochemical tests may not be suitable for pathogen discrimination, interest in biotyping has been revived with the use of phenotypic arrays, such as the OmniLog system (Biolog, Hayward, CA), which can test isolates for nearly 2000 phenotypic traits. Despite this impressive coverage, in many cases single-gene molecular fingerprinting methods such as 16S rRNA gene sequence analysis have been shown to be superior to phenotypic analysis in the discrimination among groups of bacterial strains [34]. Other strain typing methods that provide additional discrimination include serotyping, phage typing, plasmid profiling, colicin typing and antimicrobial resistance typing. Plasmid based methods of bacterial typing, such as plasmid profiling [35] and colicin typing [36] of Enterobacteriaceae strains have proven useful, though they have been largely supplanted by other more discriminatory, and largely chromosomal, based methods.

Serotyping is based upon the reactions of antibodies with cell-surface antigens. It is useful in delineating subgroups within species; however, it is time-consuming, expensive and requires designated reference laboratories to reliably carry out the procedure. Despite these limitations, serotyping is widely used in the typing of human and animal pathogens. Phage typing relies on a standard set of bacteriophage and is based on the ability of each specific phage in the panel to lyse the bacterial strain under investigation [37]. The particular phage type (PT) of a strain depends on a number of factors including the ability to replicate once inside the bacterium. Phage typing does not require specialized training or equipment, but the upkeep of a standardized set of bacteriophage is required. Phage typing is often useful in grouping phenotypically related strains, but less useful in determining relationships in space and time, as is required in outbreak investigations. Additionally, changes may occur in PT during laboratory storage [38], due to loss of a plasmid [39] or a change in super-infection immunity through loss of an integrated bacteriophage [40].

Antimicrobial resistance (AMR) typing is used to test the resistance of bacterial strains to multiple antibiotics through growth on agar plates supplemented with antibiotic disks

[41]. Similar to phage typing, AMR typing is useful for broad categorization of phenotypically related bacteria, but also suffers from the possibility that the AMR type might change over time, as AMR is often encoded on plasmids or mobile elements that can easily be lost or exchanged among bacteria in a population [42].

2.5 Molecular genotyping using the accessory genome

Methods of genotyping that rely on the accessory genome include pulsed-field gel electrophoresis (PFGE), amplified-fragment length polymorphism (AFLP) analysis, multi-locus variable number tandem repeat analysis (MLVA), microarray comparative genomic hybridization (mCGH) and comparative genomic fingerprinting (CGF). The basis for all of these methods is variation in the genome at the specific sites that are used in the analysis. The most important drawback for these analyses is that phenotypically important differences that exist among genomes may not be captured by the analysis of variation in only a limited number of loci. PFGE is still considered the 'gold standard' of DNA-based genotyping methods and involves digesting whole-genomes with rare-cutting enzymes and visualizing the resulting banding patterns through gel electrophoresis; the choice of enzyme(s) varies depending on the organism and the amount of resolution required [43, 44]. The banding pattern for a strain is then used as a fingerprint to identify and compare to other similar strains. PFGE has been an invaluable tool in outbreak detection and intervention. Government-sponsored organizations, such as PulseNet, provide an international database that stores PFGE patterns, and to which contributors can rapidly compare samples [7]. Despite its proven usefulness, PFGE requires specially trained staff, is highly labor intensive, is dependent upon complex graphical analysis and requires extensive standardization to ensure that results can be accurately compared among laboratories [45, 46]. Additionally, studies have found that temporally and spatially unrelated strains of *E. coli*

O157:H7 occasionally have identical fingerprints [47]. In addition, for species such as *Staphylococcus aureus*, closely related subtypes are often impossible to distinguish [48]. Another, less labor intensive method, AFLP also relies on restriction digests, where both a rare- and frequent-cutting enzyme are used to digest genomic DNA, after which DNA-adapters of known sequence are ligated to the ends of the resulting fragments [49]. Based on the adapter sequences, these fragments created from both a rare- and frequent-cutting enzyme are amplified by polymerase-chain reaction (PCR) and the amplified DNA visualized on an agarose gel. Just as in PFGE, the pattern is the unique fingerprint for a strain. AFLP typing schemes exist for a number of pathogens, with PFGE generally giving better discrimination within a phenotypically related group of strains [50]; however, a study on *S. enterica* serovar Enterica found AFLP and PFGE to be comparable [51]. The usefulness of a particular analytical method therefore also depends on the species under investigation, with bacteria harboring fewer evolutionary changes requiring a deeper genomic analysis to identify differences. A number of PCR-based discrimination procedures have been used to exploit the variation in repeated DNA sequences in the bacterial genome. These techniques include random amplification of polymorphic DNA (RAPD) [52], enterobacterial repetitive intergenic consensus sequence (ERIC) [53], repetitive extragenic palindromic (REP) and BOX element PCR assays [54], analysis of the 16S-5S rDNA intergenic spacer regions [55] and the analysis of variable number tandem repeats (VNTR) [56]. Among these, multilocus VNTR analysis, or MLVA, has become one of the most popular because it provides a high level of discrimination among strains [57]. It has resolution equal to or greater than that of PFGE while being significantly easier to perform, and since only a limited number of alleles need to be analyzed per locus, it does not require software capable of complex image analysis. For these reasons, it was suggested that MLVA will be the successor to PFGE as the new ‘gold standard’ in PulseNet [58]. MLVA, when used in tracking the source of *E. coli* O157:H7 during human-associated outbreaks related to leafy greens [59], and for determining the source of *Listeria monocytogenes* [60], was found to be more discriminatory than PFGE. MLVA was also able to discriminate among *B. anthracis* strains, a pathogen known to be extremely homogeneous genetically [61]. Although it provides very good discrimination among strains, it only captures a small proportion of the genetic variability among them and is unlikely to be useful for localizing phenotypically relevant differences among the genomes. Methods for examining the presence / absence of the entire accessory genome of strains, or a subset of genes therein include mCGH and CGF respectively [62]. In mCGH, a solid support (frequently glass) contains an ordered matrix of oligonucleotide probes representative of every gene for a known strain or a number of strains. By fluorescently labeling DNA from the reference strains used to create the slide and comparing them through hybridization to a test strain labeled with a different fluorescent dye, the presence and absence / divergence of every known gene can be obtained for the test strain. The mCGH approach has been used to examine populations of *Campylobacter jejuni* [63], enterohemorrhagic *E. coli* (EHEC) [64], and the important but little-studied animal and human pathogen *Lactococcus garvieae* [65], among others. Although mCGH offers excellent resolution, it is limited to knowledge about the presence or absence of the genes present with respect to the reference strain(s) used for probe design; it is also expensive and labor intensive to perform, all of which limit its applications to basic science and it has of yet not been used as an epidemiological tool.

CGF utilizes the most variable regions of the accessory genome based on whole-genome comparisons using mCGH or whole-genome sequencing [66]. A small set (usually 2040) of loci are chosen from those found to be the most variable, that also provide the best discrimination among strains and are best able to differentiate among clades of the organism identified through phylogenetic analysis of whole-genome comparisons. CGF has been shown to discriminate between molecular sub-types and evolutionary lineages of *E. coli* O157:H7 [66] and to detect clusters of epidemiological significance among populations of

C. jejuni strains [67, 68, 69]. It has also been used to identify lineages and clades of *E. coli* O157:H7 that appear to differ in virulence, such as those that are bovine-associated and SNP clade 8 strains that are thought to be 'hyper-virulent' [70, 71]. Due to its ease of use, relative low cost, and binary data generation, CGF can be performed in most modern microbiology laboratories and data can easily be shared among laboratories without the need for image analysis or extensive standardization. Drawbacks associated with CGF include the need for whole-genome population data from which to identify the most variable loci. Comprehensive population genomic studies using Sanger sequencing technology and microarrays are time consuming and expensive. However, as the cost of genome sequencing continues to decrease and consequently, as the rate of whole-genome sequence deposition in public databases continues to increase in an exponential manner, the initial comparative genomics step has become trivial.

2.6 Molecular genotyping using the core genome

Much phenotypic variation can be explained by examination of the accessory genome; however, many researchers feel that selectively neutral changes in the core genome such as single nucleotide alterations in codons that do not result in a change in the amino acid sequence of polypeptides (synonymous mutations) represent a molecular clock that provides a more accurate record of strain evolution. The most common genotyping methods utilizing the core genome include multi-locus sequence typing (MLST), SNP genotyping and whole-genome sequence analysis. The choice of typing tool depends upon the sequence being analyzed and the questions being asked; genetically monomorphic bacteria with highly conserved genomes such as *B. anthracis* may require an analysis of the entire genome to find loci which allow differentiation among strains [21], while highly recombination-prone bacteria such as *Streptococcus pneumoniae* can be effectively differentiated using the presence or absence of a small number of genomic loci [72]. Much of bacterial taxonomy has been based solely upon the 16S ribosomal RNA gene (rRNA). This single-gene rRNA approach has been useful in constructing the complete tree of life, as the rRNA genes are required by all organisms for protein synthesis [73]. The presence of multiple copies of the rRNA operon in bacterial genomes and the high level of identity among their sequences suggests that they are the result of duplication events and not lateral gene transfer. Therefore, they are assumed to be conserved and relatively stable among members of a given species. However, owing to this stability they are not generally useful for finding differences among closely related strains [74]. MLST attempts to overcome single-gene limitations by analyzing a number of genes with essential function or housekeeping genes. They are believed to be conserved because of their essential nature to the survival of the organism, but unlike rRNA genes, they usually occur in single copies [57]. MLST has been used to differentiate *S. pneumoniae* isolates and estimate rates of recombination within them [75]; to offer support for the parallel evolution of virulence traits in enterohemorrhagic *E. coli* [76]; to uncover the phylogeny and population structure of *Vibrio parahaemolyticus*, a pathogen frequently associated with raw seafood and the leading cause of human food-poisoning [77]; and to demonstrate clonal population structure of *Clostridium perfringens* isolates from both healthy chickens and broiler chicken populations associated with outbreaks of necrotic enteritis [78]. Recent work has shown that a recapitulation of the historical migration of populations in the Pacific region is provided in the phylogenies of *H. pylori* strains isolated from people in the region today. These phylogenies, based on sequence variations in seven conserved genes of *H. pylori*, show that two waves of migration occurred; one to Australia and New Guinea and a much later migration from Taiwan into Melanesia and Polynesia [79]. However, MLST is not without its limitations. At least five to seven loci are typically used in the analysis because phylogenies for individual loci are not always congruent and may not only differ from each other but also from phylogenies inferred by rRNA sequences. To get around this, an average phylogeny can be determined by simply splicing house keeping gene sequences together in a continuous string, known as a concatenome. Intriguingly, it has been shown that there is a significant difference among individual MLST loci in their ability to predict the phylogeny based on analysis of the concatenome of all core loci. Certain ‘best performing genes’ have been identified based on the results of whole-genome comparative studies. Konstantinidis et al. (2006) showed that as few as three loci were capable of reproducing a phylogeny consistent with the whole-genome analysis [80]. Loci that do not perform well with respect to reproducing the phylogeny may well have been derived from lateral gene transfer from unrelated members of the same species. As a result of this, *C. jejuni* strains that have different MLST types have been shown to be highly similar in other regions of their genomes and strains with the same MLST type have been shown to have significant differences in their genomes when analyzed by mCGH [81]. On the other hand, differentiation among strains within a taxon using MLST may not be possible for bacteria that have relatively stable core genomes but have high rates of accessory HGT such as with *E. coli* O157:H7 [82]. SNP genotyping is the core genome analogue of CGF, where informative SNPs from either whole-genome comparisons or multiple SNPs in core genes of interest are analyzed to determine if they are ‘informative’ with respect to strain differentiation. A number of software programs have been designed to assist with choosing the SNPs with the appropriate discriminatory power [83, 84]. The SNP approach is much more discriminatory than MLST because it relies on sequence differences in many more core genes. SNP typing has been used to identify a number of clades within *E. coli* O157, including a ‘hyper-virulent’ clade 8 associated with high levels of hospitalization and the hemolytic uremic syndrome [70]; has led to the development of a rapid and highly discriminatory genotyping method for *C. jejuni* and *C. coli* [85]; and provided a way to easily differentiate *Mycobacterium tuberculosis* strains (which like *B. anthracis* strains are highly similar genetically), into one of the eight known clonal complexes and to identify those associated with greater virulence [86]. The ease and low cost with which whole-genome sequence data can now be obtained from multiple isolates in bacterial populations has led to a blossoming of both short- and long-term epidemiological studies that have harnessed the power of this technology to allow levels of differentiation / discrimination among genomes that were previously not possible. Fifteen recent studies in which this high-throughput sequencing has been performed are provided in Table 2.2. It is apparent from these studies that the sequencing of hundreds of strains per study is now possible, as with the study of *S. pneumoniae* PMEN clone 1, where 240 isolates were sequenced [72]. Also apparent is that the most informative studies combine whole-genome SNP analysis and the presence / absence of specific factors in the accessory genome with additional metrics, such as the study that used social network analysis in the elucidation of *M. tuberculosis* transmission [87], and geographical distribution in the intercontinental transmission of hospital-associated methicillin-resistant *S. aureus* strains among patients [88]. Lastly, even the investigation of outbreaks is changing due to whole-genome analysis, with single-molecule sequencing and SNP typing used to identify the likely geographical source of the Haitian cholera outbreak [89] and the real-time outbreak investigation in Canada of a processed-meat associated outbreak of listeriosis [90]. SNP genotyping relies on mutations in conserved genes, and is useful for differentiating among members of highly clonal taxa that rarely undergo HGT. However, in other species, HGT is responsible for ten fold or more genetic changes than that attributable to single base mutations and makes SNP genotyping less suited for strain fingerprinting and for use in outbreak investigations than methods that sample more frequently occurring types of genomic variation. Whole-genome sequencing allows the study of the entire pan-genome *in silico*, and is based on sequence alignments of the core genes, the binary presence / absence of accessory genes, or a combination of the two. It is the most data-rich of all genotyping methods, and until recently was also the most expensive to perform. Additionally, software tools for the large-scale comparison of many genomes are just beginning to be created, as many programs designed for the analysis of one or a few genomes do not scale well to pan-genomic studies. The use of variation in the hundreds and in some cases thousands of genes in the core genome should provide a better estimate of phylogeny than methods that only sample a few loci such as MLST. This type of broad-based genomic analysis suggests a step-wise evolution for *E. coli* O157:H7 [91, 92]; that the genus *Listeria* has evolved with limited gene acquisition and loss [93]; and indicates how serovar-associated differences in the genome of the porcine respiratory pathogen *Actinobacillus pleuropneumoniae* contribute to virulence [94].

As previously discussed, the accessory genome can have a significant influence on phenotype, and whole-genome sequencing allows a comparison of accessory genes among groups of strains. *Pseudomonas aeruginosa* strains can metabolize a large number of substrates; however, only some of these strains are associated with human disease. An analysis of the *P. aeruginosa* accessory genome identified specific genes responsible for adaptation to particular niches [28]. *S. aureus* genome phylogenies constructed using core sequences suggest that strains from sequence type (ST) ST5, which were responsible for a pandemic infection of broiler chickens, emerged from a human clade of the organism within the last 60 years [95]. These ST5 *S. aureus* strains are thought to have acquired avian-specific accessory genome elements responsible for the resistance to chicken heterophils, and to have lost human-specific genes performing a similar function. Both aligned core data and binary presence / absence data were used to construct a phylogeny of *L. monocytogenes* and identify defective / absent genes in *L. monocytogenes* lineage III, strains of which are rarely isolated from human clinical cases [96].

While comparative genomics is now firmly based on whole-genome DNA sequence analysis, comparisons also extend to gene expression and the emerging field of comparative transcriptomics, which has been facilitated by recently developed approaches such as

RNA-Seq analysis [97]. The choice of comparative genomics tool for a particular experiment depends on the needs of the investigator, but should be balanced against the cost and technical requirements to generate data, and need for data dissemination and assay reproducibility, particularly where image-based methods are used. In addition, many methods require the use of sets of reference pathogenic microorganisms as controls and standards. The maintenance and acquisition of these reference strains has become increasingly difficult as a result of international import and export restrictions and the need to meet enhanced laboratory biosecurity requirements.

While single-locus methods for strain genotyping, such as serotype, are less informative than multi-locus ones, these methods often provide a reasonable estimate of strain relationship and virulence. Their wide-spread use is likely to continue for some time for historical as well as legislative reasons. A recent study comparing many of the common molecular typing methods for *E. coli* O157:H7 found that most suggested very similar phylogenies and relationships among strains; therefore a method that is easy to perform and from which data can easily be shared could theoretically be selected as a 'common typing language', so that comparative genomics data from multiple genotyping methods could be easily compared and built upon [71]. However, in the very near future the *de facto* common molecular typing language for bacteria will most likely be whole-genome sequencing, and so efforts to best utilize these data should be undertaken, rather than determinations of which lower-resolution method is best. We have recently demonstrated that results which would be expected from low-resolution genotyping methods can readily be derived from whole-genome sequence data *in silico* [71].

2.7 Software for Whole-genome Analyses

Whole-genome sequencing was the inevitable outcome of an increasingly large number of gene sequencing projects. The ability to determine nucleotide sequences was first developed by Sanger using dye-termination fluorescent sequencing [98], and by Maxam and Gilbert using chemical modification of nucleotides [99]. Due to its use of toxic chemicals and inefficiency in scaling, the Maxam and Gilbert method was largely abandoned and the Sanger method became the standard sequencing technique until high-throughput random shotgun sequencing methods became available. Sanger sequencing generates reads of approximately 800 bp and has the benefit of a known starting position for the sequencing read; thus, if subsequent sequencing reads continue from the end of previous reads, a complete sequence with no gaps is produced. In contrast, massively parallel sequencing platforms such as the Roche (GS-FLX Titanium) [100], Illumina Solexa (HiSeq 2000) [101] and Applied Biosystems (5500xl) [102] are based on amplification of short regions and generate anywhere from 500 Mbp of 400 – 600 bp reads for the GS-FLX Titanium machine, to over 200 Gbp of up to 150 bp reads for the HiSeq 2000 machine.

IonTorrent offers a semi-conductor based sequencing platform called the Personal Genome Machine (PGM) that produces a projected 1 Gbp of 200 bp reads per run. This low-cost platform is upgradeable, can generate genome sequence data in two days and has lower memory and processing requirements than light-based systems. This and a number of other companies hope to entice users away from the use of high-throughput core facilities to laboratory-based sequencing in a manner analogous to the move away from mainframe to personal computing. The PGM played a key role in the near real-time sequencing of the *E. coli* O104:H4 strain responsible for the large and deadly outbreak of hemorrhagic colitis and hemolytic uremic syndrome in Germany in 2011. Single molecule sequencing, which requires no amplification, is currently available from Pacific Biosystems, with read lengths on average of ¿1000 bp, with the possibility of increasing this to 10 kbp in the near future.

In addition, at least fifteen other companies are in the process of bringing single molecule sequencing technology to market [103].

Moores law states that the number of transistors that can be fit onto a circuit board doubles every two years, leading to a concordant increase in computing power. Until the advent of multi-parallel sequencing technology, the rate of DNA sequence acquisition also doubled approximately every two years, which allowed information technology to keep pace with the processing requirements for data analysis. Now, sequence information is being gathered at rates orders of magnitudes faster than corresponding growth in computer processor speed and the availability of data storage, creating an information bottleneck at the level of raw data analysis. New more efficient software algorithms are now needed to compensate for the lack of computer processing brute force. For example, the cost of massively parallel sequencing has fallen to the point that on the Illumina platform, which currently generates the most sequence information for the least amount of money, the cost to sequence 1 Mbp is $0.10 [104].

Although the draft genome assemblies of the relatively short reads contain gaps that cannot be fully resolved, the data are of sufficient quality to allow meaningful whole-genome comparisons. Typically, multiple large contigs are obtained that cannot be unequivocally joined because of sequence ambiguities associated with short reads of repetitive sequence. While a closed genome would be much preferred to a collection of contigs, the cost and labor to join these ambiguous regions using Sanger sequencing is often not felt to be worth the effort.

Irrespective of the platform used, the main areas that still require improvement are in the correct identification of bases (reducing sequencing error) and in the assembly of small reads into large contiguous regions [105]. Even in systems with low sequencing error rates, it is advisable that regions with rare-variants be re-sequenced for confirmation [106]. With the trend to longer and longer reads for all platforms, reads of 10- or 20-kbp may become routine and the assembly of a completed genome free of gaps will become trivial. It is also thought that base-calling will be significantly improved and that there will no longer be a need to resequence. Some of the most interesting and powerful scientific studies now require analyses of whole genomes, and often comparisons among and between groups of genomic sequences. To fully understand a taxonomic group or a species, sometimes hundreds of whole-genome sequences are required to ensure that the coverage of genetic variation in the population is complete. Multiple-sequence alignment programs such as MUSCLE [107], T-coffee [108], CLUSTAL W [109] and others perform very well when aligning one gene with a large number of variants of that gene; however, due to the size of whole-genomes and insertions / deletions within them, these programs do not scale well. MUSCLE is particularly efficient in searching due to its progressive alignment algorithm, which requires less time and computer memory to generate results equivalent to those using other programs. Searching a database for a particular sequence is a common task in genomics-based research and was pioneered by the FASTA and FASTP programs [110], and is now nearly synonymous with the basic local alignment search tool (BLAST) suite of programs [111]. These are hosted online by the National Center for Biotechnology Information in the United States and also available for download for standalone use [112]. Implementations of the BLAST algorithm utilizing graphical processing units (GPUs) or alternative search strategies have been released and include GPU-BLAST [113], which utilizes the GPU of a computer for faster traditional BLAST searches. The BLAST-like alignment tool (BLAT) is optimized for fast protein alignments [114] and WU-BLAST, from Washington University, provides additional statistical information about searches and offers a unique gapped-alignment algorithm [115]. New algorithms continue to be developed for local database searches; promising new ones include UBLAST and UCLUST, which are part of a standalone package reportedly hundreds of times faster than BLAST, with near-identical specificity for closely related searches, but not for more distant ones

[116]. Allowing for longer sequence alignments than BLAST has also been an area of study, with the recent Burrows–Wheeler transform algorithm allowing up to 1 MB of contiguous sequence to be aligned, while being more accurate and several times faster than BLAT [117].

Although BLAST is one of the most cited resources for searching and alignment of similar sequences, it does not perform *in silico* subtractive hybridization. That is, rather than searching a sequence against a database for regions of similarity, searching a sequence against a database for regions that do not match. Programs with this exact functionality using the BLAST algorithm have been created; FindTraget [118] and nWayComp [119] are standalone downloads with this ability, but they have largely been superseded by the very fast and massively parallel web-server, mGenomeSubtractor [120]. Programs designed specifically for whole-genome comparisons built upon previous sequence alignment knowledge include Mauve [121] and progressiveMauve [4], which allow for the identification of genome duplications and rearrangements as well as core genome SNPs; MUMmer [122] and MUMmer GPU [123], which perform very efficient whole-genome alignments based on suffix-trees and can find SNPs in aligned sequences; MISHIMA [124] is a new method for whole-genome comparison that uses a divide-and-conquer heuristic approach that is faster than the traditional pairwise-comparison method, but is only useful if the genomes are closely related. Another segmentation-based aligner to recently emerge is Mugsy [125], which relies on the suffix-tree algorithm of MUMmer and an implementation of the T-coffee tool for segmentation alignment [126]; Mugsy is very fast and useful in the alignment of many genomes. The software Panseq also makes use of the MUMmer suffix-tree algorithm for whole-genome alignments, but presents results in the context of the pan-genome [127]. Panseq allows unique genomic regions to be found in genome data from groups of strains, determines the non-redundant pan-genome of a group of genomes based on a percent sequence identity cutoff and finds the SNPs in the derived core genome and the binary presence / absence of each locus in the accessory genome.

2.8 Examples of WGS Comparisons Using Current Methods

We will consider pan-genomic analyses performed with the programs Mauve, MUMmer and Panseq. Mauve provides multiple genome alignments allowing for rearrangements using local co-linear blocks. These blocks are nicely visualized and provide a graphical overview of the similarities between strains. This approach works well for genomes from a small number of strains that can be grouped easily into a single-page graphic, as has been done with five *Francisella tularensis* strains, where sequence inversions between high and low virulence strains are readily apparent [128]. In a separate study, where two strains of *Mycoplasma hyopneumoniae*, one pathogenic and the cause of enzootic pneumonia in pigs and the other non-pathogenic, as well as one strain of the poultry pathogen *Mycoplasma synoviae* were compared, Mauve was used for global genome alignments [129]. This helped identify regions of horizontal gene transfer, and genomic regions that differed between the pathogenic and non-pathogenic isolates. Specific regions identified in this way were probed using protein-specific BLAST searches.

MUMmer is command-line based only, and can produce graphics such as a dot-plot of the sequence similarity between strains. It has been used to align the genomes of the sheep pathogen *Brucella ovis*, to the zoonotic members of the genus *Brucella* [130]. The results from this MUMmer-based comparison showed that *B. ovis* has undergone genome reduction, which could account for its reduced host range and tissue specificity. In a recent study of the plague-associated *Yersinia pestis* strain Angola, use of MUMmer allowed comparisons to other *Yersinia* genomes from *Y. pestis* and *Yersinia pseudotuberculosis* [131]. Using SNPs and differences in the accessory genome determined through MUMmer, strain

Angola was shown to have a genome composition and an inferred evolutionary position intermediate between most other *Y. pestis* and *Y. pseudotuberculosis* strains. The chimeric virulence plasmid of strain Angola appears to comprise portions of two other *Y. pestis* specific plasmids, and several novel genomic sequences that significantly increased the size of the *Y. pestis* pan-genome.

Panseq allows for single-step analyses of a group of genomic sequences. It determines either the composition and distribution of the pan-genome, or genomic regions specific to a strain or group of strains. In the first instance, Panseq was used to find the pan-genome of six *L. monocytogenes* strains, where the phylogeny built from the concatenated core genome was shown to be very similar to the concatenated binary values for the presence / absence of the accessory genome, suggesting that both the core and accessory genome reflect the evolutionary history of these organisms [127]. In the second instance, the genomic sequence of an adherent-invasive *E. coli* (AIEC) associated with Crohns disease was compared with other AIEC strains and to other pathotypes of *E. coli* using the Novel Region Finder of Panseq [132]. It was shown that this AIEC strain possessed 21 genomic regions not found in other members of the same phylogenetic cluster, and that it contained two specific regions not found in any other *E. coli* genomes in Genbank. While these programs can perform similar types of analyses, the ultimate choice among them comes down to the unique features required by the investigator to address specific scientific questions. For identifying SNPs for evolutionary reconstructions, and for identifying genes responsible for unique phenotypic traits, any of the three above programs can be used. Mauve provides both agraphical user interface and graphical output of whole-genome alignments, but its alignment algorithm is slower than MUMmer. Panseq automates genomic subtraction, which can be done manually from the output of the other programs. Although MUMmer provides one of the fastest alignment algorithms, it is completely command-line based and runs exclusively on Linux operating systems. Where the programs differ substantially is in defining clusters of strains that are genomically similar and in creating fingerprints useful in molecular epidemiology. Mauve can be used for a broad graphical overview of similar clusters, but only Panseq can specifically produce pan-genomic output. Additionally, Panseq can also be used for determining the most discriminatory set of loci from SNP or accessory genome presence / absence data, something which the other programs do not currently do.

2.9 Conclusion

The goals of genotyping and sequence comparison are to identify unique signatures or fingerprints for individual strains while at the same time identifying closely related groups of strains. These goals are extremely important and need to be addressed with as much information as possible. In an epidemiological context, source-attribution and outbreak investigations require as much certainty as possible, not rough estimates based on limited information, as the outcomes of these analyses have wide-ranging economic and social consequences. In a population genomics context, the identification of core and accessory regions is imperative if accurate phylogenies are to be constructed and if accessory genomic regions implicated in novel phenotypes are to be discovered. Future population-based studies of microbial pathogen populations will in some cases be based on hundreds of fully sequenced genomes. The problem has now shifted from sequencing the full genome of strains to acquiring the appropriate isolates to sequence, as the difficulty in international strain transfer, and acquiring isolates from specific individuals or environments often interferes with this. Further, the downstream task of effective analyses of the myriad genomic data generated by modern sequencing platforms will require increasingly innovative software tools.

2.10 Conflict of interest statement

The authors are also the creators of the Panseq software tool.

2.11 Acknowledgements

We would like to thank the Canadian Food Inspection Agency for allowing this work to be conducted in their building. This work was funded by the Public Health Agency of Canada, the Natural Sciences and Engineering Research Council of Canada and Alberta Innovates Technology Futures.

Table 2.1: A summary of 15 recent whole-genome based studies.

Table 2.2: A summary of 15 recent whole-genome based studies.

Organism/disease	Strains sequenced	Platform	Summary	References
Salmonella enterica Typhi Outbreaks / Septicaemia, typhoid fever	19 strains	Roche 454 GS-20 (8 strains), Roche 454 GS-FLX (3 strains) 250 bp reads, Illumina Genome Analyzer (12 strains) 25 bp reads	1964 SNPs identified. Monomorphic genome characterized by loss of gene function. Small effective population size. Strong selection for antimicrobial resistance	[133]
Streptococcus pneumoniae PMEN clone 1 / Endemic Pneumonia, pharyngitis, otitis	240 strains of ST81 and serotype 23F, multi-antimicrobial resistant strains	Illumina Genome Analyzer II multiplexed libraries	23% of the 39,107 polymorphic sites were homoplasies and 88% of the SNPS were introduced by 702 recombination events. Core base substitution rate was estimated to be 1.57x10^{-6} per year. Antibiotics and vaccine usage likely have been important forces driving selection	[72]
Methicillin-resistant *Staphylococcus aureus* (MRSA) sequence type 239 / Endemic hospital-acquired wound infections	43 strains from a global collection (1982–2003), 20 strains from patients in a Thai hospital over 7 months	Illumina Genome Analyzer II multiplexed libraries	4310 SNPs were identified. While homoplastic SNPs were infrequent (0.88%), 10 of these SNPs were in genes associated with antimicrobial resistance. Strain clusters tended to be geographically restricted; however, several cases of intercontinental transmission were demonstrated. A transmission event from one ward to another in a Thai hospital was also demonstrated. The estimate for the core base substitution rate was 3.3x10^{-6} per year	[88]

Table 2.2: A summary of 15 recent whole-genome based studies.

Organism/disease	Strains sequenced	Platform	Summary	References
Multi drug-resistant *Acinetobacter baumannii* (MDR-Aci) / Endemic Hospital-acquired wound infections	Pooled DNA from 7 and 4 strains, and 3 individual strains	Roche 454 Titanium FLX	Three SNP loci were identified which were useful in differentiating among outbreak strains. Workers were able to distinguish between strains that were identical using PFGE and VNTR genotyping and to determine the likely source of transmission from one patient to another within a hospital	[134]
Group A *Streptococcus* (GAS) / Endemic septicaemia, scarlet fever, pharyngitis	95 strains	Illumina Genome Analyzer II, 36 bp reads	280 of 801 bi-allelic core SNPs were analyzed by mass spectroscopy in 344 GAS strains. SNPs and indels were used in the construction of Neighbour-Joining phylogenies. Spatially defined / geographically limited clusters were identified. Micro and macro bursts of multiple emergent strains were noted. There were an overabundance of SNPs in virulence related genes such as *ropB* suggesting strong selection at these loci. Small differences in genomes among strains appeared to result in large changes in their transcriptomes	[135]

Table 2.2: A summary of 15 recent whole-genome based studies.

Organism/disease	Strains sequenced	Platform	Summary	References
Listeria monocytogenes / outbreaks , septicemia, abortion, encephalitis	Two strains	Roche Genome Sequencer FLX System	27 SNPs differentiated two outbreak strains. Differential occurrence of a prophage accounted for differences in PFGE patterns among outbreak isolates. Three days were taken for sequencing, five weeks for closure and three weeks for resequencing to confirm SNPs and indels.	[90]
Mycobacterium tuberculosis / endemic, outbreaks / Lymphadentitis, pneumonia, osteomyelitis	Three strains	Roche Genome Sequencer FLX System, average of 250 bp reads	Strains were sequenced from the start and end of the transmission chain from five patients over 12 years. Six SNPs, a tandem repeat and an IS6110 site differed among the strains	[136]
	32 case and four reference strains, 50 bp paired-end sequence reads	Illumina Genome Analyzer II	204 SNPs were identified. Social network analysis was used together with whole genome based SNP genotyping to track community transmission and relate a rise in cases to an increase in crack cocaine usage	[87]
Bacillus anthracis /Endemic, outbreaks, septicaemia, pneumonia, sudden death	Two Japanese strains	Illumina Genome Analyzer II 50 bp reads	The two genome sequences were compared to 17 reference genomes. Among the 2965 SNPs identified, 80 ‘tag’ SNPs were better able to discriminate among *B. anthracis* strains using a PCR format than a 25 loci MLVA assay	[137]

Table 2.2: A summary of 15 recent whole-genome based studies.

Organism/disease	Strains sequenced	Platform	Summary	References
Francisella tularensis / endemic, outbreaks, lymphadentitis, septicaemia	13 *F. tularensis* subsp. holoarctica, 26 *F. tularensis* subsp. tularensis and one strain of *F. tularensis* subsp. novicida	Affymetrix Genechip resequencing array	Compared genomes to the reference sequences, LVS and SCHU S4. The number of SNPs ranged from 15 to 12,407 per strain. B strains had from 15 to 2915 SNPs, A1 strains had an average of 6362 and A2 strains 5096 SNPs. The single *F. tularensis* subsp. novicida had 12,407 SNPs. Strains clustered according to their subspecies	[138]
	One strain of *F. tularensis*	Illumina GAII 36 bp single-end reads	A comparison was made of the genomes of closely related laboratory and naturally propagated strains	[139]
Yersinia pestis / epidemic, endemic, lymphadenitis, septicaemia, pneumonia, sudden death	Three new strains and 14 genomes	ABI3730xl	33 SNPs were identified and evaluated in 284 strains using Mass Array technology. Based on the analysis, *Y. pestis* likely originated in China and then spread to other parts of Asia, Europe and the Americas through multiple radiations. Each region now has its own lineage of the organism	[140]

Table 2.2: A summary of 15 recent whole-genome based studies.

Organism/disease	Strains sequenced	Platform	Summary	References
	Two strains	Roche 454 Titanium FLX, Illumina Genome Analyzer II	The genome of a virulent *Yerinia pestis* strain KIM 10+ was differentiated from an attenuated strain D27 that had lost a 102 kb chromosomal locus following multiple passages. Combined use of these two sequencing platforms and their associated software (Newbler and CLC Genomics Workbench) was required to resolve all sequence ambiguities	[106]
Helicobacter pylori / endemic , stomach ulcers	12 strains from Colombia	Roche 454 Titanium FLX	Strains differed by 27-232 SNPSs and 16-441 polymorphisms were associated with recombination. Recombination was most marked in genes from the *hop* gene family which are associated with bacterial adhesion. Changes in the genome were noted over the course of multi-year infections and after early colonization in a volunteer	[141]
Vibrio cholera / epidemic diarrhea, dehydration	Five Haitian strains	Third-generation single molecule real-time SMRT 1100 bp reads, 3045 min run	30 SNPs were identified. Haitian outbreak strains were most closely related to Asian rather than American isolates	[89]

Chapter 3

A review of Shiga-toxin producing *E. coli* virulence and genomic evolution

3.1 Introduction

Escherichia coli are gram-negative, facultative anaerobic bacteria normally found as commensal organisms within the intestines of warm-blooded animals [142]. In humans, the composition of normal microbiota, including *E. coli* and other bacterial strains, is determined in the first week of life [143], with continuous changes in composition occurring due to the environment and consumed food, with meat being a particularly large contributor [144]. The *E. coli* species has been classified into four large phylogenetic groups based on multi-locus enzyme electrophoresis profiles, designated A, B1, B2 and D. Commensal strains comprise most of group A, and pathogenic isolates are distributed among the other phylogenetic groups [145]. While most *E. coli* are a normal constituent of the human microbiota, certain serotypes have been shown to be pathogens responsible for gastrointestinal and extra-intestinal infections in humans [146].

E. coli that cause diarrhea and intestinal illness have been classified into six categories: enteropathogenic *E. coli* (EPEC), enterotoxigenic *E. coli* (ETEC), enteroinvasive *E. coli* (EIEC), enteroaggregative *E. coli* (EAEC), diffusely adherent *E. coli* (DAEC), and Shiga-toxin producing *E. coli* (STEC), also known as verocytotoxin-producing *E. coli* (VTEC) [147]. A subset of STEC that cause hemorrhagic colitis (HC) are known as enterohemorrhagic *E. coli* (EHEC). Many of these pathogenic *E. coli* are of human origin while others such as STEC are zoonotic pathogens that are found as part of the microflora of healthy ruminants. Cattle in particular have been identified as reservoirs of pathogenic *E. coli* associated with human disease [146].

STEC are of particular concern for humans due to the frequency and severity of disease that they cause. Human infection has been documented from the consumption of undercooked meat [147], raw milk and cheeses [148], vegetables [149], fruit juices [150] and recreational and drinking water [151, 152]. STEC are characterized by the production of Shiga-toxins (Stx), which are compound toxins characterized by a catalytic A subunit that inhibits host ribosome function and a pentameric B subunit that binds to host GB3 receptors in intestinal cells [153]. In humans, Stx damages endothelial cells in the gut, causes cytokine release, and inflammation, which can lead to the formation of microthrombi that cause blockage of small blood vessels in the brain, pancreas, kidney and intestine [154]. STEC infection in humans can be asymptomatic or cause disease ranging from mild diarrhea to HC and the hemolytic uremic syndrome (HUS).

STEC by definition are capable of producing one or more Stx, but not all STEC cause disease in humans [155]. The subset that cause bloody diarrhea, possess a 60 kDa large virulence plasmid and produce attaching and effacing (A/E) lesions on epithelial cells are referred to as enterohemorrhagic *E. coli* (EHEC) [156]. Karmali et al. (2003) defined five groupings of STEC strains based on their frequency and severity of human disease, deemed ‘seropathotypes’, which are defined as follows: Seropathotype A strains consist only of serotype O157:H7 and O157:NM strains, which are frequently associated with outbreaks and the most serious forms of disease, such as HUS; seropathotype B include strains of serotypes O26:H11, O103:H2, O111:NM, O121:H19 and O145:NM, and are associated with outbreaks and severe illness, but less frequently than group A; seropathotype C includes strains from ten serotypes, among them O91:H21 and O113:H21, which have been implicated in sporadic cases of HUS but rarely with outbreaks; seropathotype D includes strains from serotypes associated with diarrhea only; and seropathotype E strains are from serotypes that have not been implicated in human disease [157].

Many STEC appear to be part of the normal microflora of adult cattle, but certain

STEC serotypes also cause disease in young food-production animals, creating substantial economic loss. In the dairy industry, neonatal calf diarrhea, and in the swine industry, post-weaning diarrhea and edema are frequently associated with specific STEC serotypes [158, 159]. Despite the presence of Stx in strains associated with neonatal diarrhea in animals, the role of Stx itself is unclear, as strains lacking Stx from serogroups such as O5, O26, and O111 are also frequently associated with diarrhea in calves [160].

The most studied STEC serotype to date has been O157:H7, which is responsible for approximately half of all STEC related illness in humans, and the majority of HUS cases worldwide [161]. Serotype O157:H7 has been the focus of numerous studies, and its genomic composition as well as its evolution from a toxin-negative EPEC O55:H7 ancestor is supported by numerous studies [162, 163, 164, 165]. While the importance of O157:H7 strains is clear, the importance of non-O157 STEC has most likely been underestimated. It is currently unclear whether this has been due to the inadequacy of the methods historically used for the isolation and detection of non-O157 STEC in human infection, or whether illness associated with non-O157 STEC is actually increasing [166].

Human infections caused by non-O157 STEC are generally less frequent and severe than those caused by O157:H7 strains, but the infections are generally indistinguishable from *E. coli* O157:H7 infections [156]. Although historically under-reported, improvements and use of cultural methods now allow a better determination of the frequency of illness associated with particular non-O157 serotypes [167, 142, 168].

Studies in Europe have long recognized the importance of many serovars of STEC in human disease [169, 170, 171, 172]. In North America, the predominance of O157:H7 and relative paucity of other serotypes associated with human illness led to a focus on serotype O157:H7. Recently the Centers for Disease Control in the United States reported that the following serotypes were responsible for cases of gastrointestinal illness among 43 states at the following prevalence: O26 (22%), O111 (16%), O103 (12%), O121 (8%), O45

(7%), and O145 (5%) [167]. This, and other reports, prompted the United States Department of Agriculture to implement routine verification testing for six non-O157 serotypes: O26, O45, O103, O111, O121, and O145 in raw beef trimmings [173]. To facilitate this, rapid detection schemes based on both biochemical and molecular markers have been devised, though there is currently no single, standardized method to easily detect all possible serotypes associated with human infection [174, 175, 176, 177].

Although humans may become seriously ill when infected with STEC, the virulence factors that contribute to this illness likely initially arose to enhance the ability of the bacteria to survive in their natural environment, the bovine intestinal tract. Humans are thought to be only 'accidental' hosts and a 'sink' habitat that receives bacteria from the bovine 'source'. The evolutionary pressures in the bovine reservoir are thought to be the governing forces shaping the acquisition and retention of novel virulence attributes [178].

In this review I will discuss the common virulence attributes, examine their apparent evolutionary acquisition and discuss the factors acting to shape the genomes of STEC, a diverse group of human pathogens.

3.2 Main Virulence Factors

3.2.1 Shiga-toxins

STEC are characterized by the production of Stx, which are so called due to their similarity to the toxin produced by *Shigella dysenteriae* type 1. Stxs are alternatively referred to as Vero-toxins due to their damaging effects on Vero cells (derived from the kidney epithelial cells of the African Green Monkey) [179]. Stx are thought to be the greatest contributors to severe human illness such as HC and HUS. In both *Shigella* and *E. coli*, the Stx genes are thought to have been introduced through lysogenic conversion by temperate lambdoid bacteriophage [161, 180]. In certain STEC lineages these prophage have become defective and can no longer be induced into the lytic cycle. There is also evidence that while many STEC lineages have maintained stable prophage over multiple generations, other lineages and species have acquired Stx through horizontal gene transfer [181, 161].

In addition to *E. coli* and *Shigella*, Stx genes have also been found in species of *Citrobacter*, *Enterobacter* and *Acinetobacter* [181, 182]. There are two main antigenic forms of the toxin: Stx1 and Stx_2 [183]. Stx1 differs by a single amino acid from the Stx produced by *Shigella dysenteriae*, but shares only 55-60% identity with Stx_2 at the nucleotide and amino acid level respectively [184, 183]. Toxicity studies using Vero cells have shown Stx1 to be more cytotoxic than Stx_2 on an equivalent weight basis; however, when toxicity was tested on human renal endothelial cells, Stx_2 was found to be a thousand times more toxic than Stx1 [185]. Although the toxins are thought to share the same receptors, the differential toxicity has been shown to be based on the affinity for the membrane-bound GB3 receptor, and discrete responses to the activated components of the endoplasmic reticulum stress response [186, 187].

It is thought that because of this difference in toxicity, STEC producing Stx_2 are more frequently associated with human disease, and with more severe manifestations of human disease such as HUS. In the bovine environment stx_1 is most often found in *E. coli* strains isolated from calves with diarrhea [188], whereas stx_2 is most often found in isolates from adult cattle [189].

Multiple sequence variants exist within each toxin type. Stx1 gene subtypes include stx_1, $stx_{1\mathrm{c}}$ and $stx_{1\mathrm{d}}$, with the stx_1 variant associated with human disease, while the latter two variants have only been found in STEC from cattle [161]. Stx_2 gene subtypes include $stx_{2\mathrm{a}}$, $stx_{2\mathrm{b}}$, $stx_{2\mathrm{c}}$, $stx_{2\mathrm{d}}$, $stx_{2\mathrm{e}}$, $stx_{2\mathrm{f}}$ and $stx_{2\mathrm{g}}$, with $stx_{2\mathrm{a}}$ being the most frequently associated with human disease [190].

STEC O157:H7 strains encode only two variants, $stx_{2\mathrm{a}}$ and $stx_{2\mathrm{c}}$, with the others being found among non-O157 STEC strains. Variant *stx*$_{2c}$ is usually associated with strains isolated from the bovine host (O157:H7 lineage II), but has also been found in combination with *stx*$_2$a in certain O157:H7 lineage I/II clade 8 strains. Those infected with the latter type have an increased risk of developing HUS [70]. Also implicated in an increased risk of HUS is *stx*$_{2d}$, an activatable Stx that is cleaved by elastase in the intestinal mucus, increasing its activity a thousand fold [191]. STEC producing *stx*$_{2e}$ are associated with edema disease in pigs, and are only occasionally isolated from humans with gastrointestinal illness [183]. Stx$_{2f}$ producing STEC have traditionally been most frequently isolated from pigeons, and only sometimes isolated from human disease [192, 193]; however, recent work has shown this Stx-subtype to be prevalent among many *E. coli* strains originally classified as atypical EPEC, indicating it could be a more important virulence factor in human disease than previously thought [194]. Stx$_{2g}$ is a recently described variant carried by STEC of serotypes O2, O15, O136 and O175 strains that also express a heat-stable enterotoxin and may constitute an emerging pathogenic group [195].

Stxs are bipartite molecules consisting of two distinct subunits. The catalytic A subunit of 32 kDa enzymatically cleaves host rRNA, inhibiting host ribosome function and leading to cell death. The five B subunits in each toxin molecule form a pentameric hollow ring, to accommodate a single A subunit. Each of these 7.7 kDa B-subunits bind to globotriaosylceramide (GB3) receptors, which are embedded in lipid rafts present in the host cytoplasmic membrane [153]. In contrast to humans, tissues in the gastrointestinal tract of cattle have been found to lack the GB3 receptor, and it is thought that for this reason they can act as asymptomatic hosts for STEC [196].

Serotypes O26 and O111 have been associated with gastrointestinal disease in calves, and O157:H7 causes epithelial cell damage in the bovine intestine, inducing an innate immune response and production of antibodies [197, 198]. However, it is likely that this response is caused by LEE-associated factors rather than Stx, as Stx-negative, LEE-positive

O26 and O111 strains also cause diarrhea in calves and Stx-negative O157:H7 mutants have been found to colonize the recto-anal junction of cattle [199, 200].

Human intestinal epithelial cells contain globotetraosylceramide (GB4), as well as less abundant quantities of GB3 [201]. GB4 is an alternate receptor of lower affinity for Stx and to which subtype stx_{2e}, the porcine edema disease-associated toxin, preferentially binds.

Stxs belong to a family of ribosome-inactivating proteins, which are found nearly universally in plants, and include well known toxins such as ricin, produced by the castor oil plant *Ricinus communis* [202]. Upon being released in the intestinal tract, the Stx B-subunit binds to the GB3 gangliosides found in lipid rafts of the host epithelial plasma membrane. After binding, Stx enters the cell through receptor-mediated endocytosis [203]; either through a clathrin-coated pit-dependent mechanism, or by inducing plasma membrane invaginations without the aid of host-cell machinery [204, 205].

Once taken into the host cell, Stx is passed through the *trans*-Golgi network via retrograde transport from the endosome, where Stx localizes, through the Golgi apparatus to the endoplasmic reticulum (ER)[206, 207]. The toxin is activated by cleavage of the C-terminus of StxA by the endoprotease furin, which is membrane associated and thus likely takes place in the early endosome [208]. The activated StxA subunit remains attached via a disulphide bond to the remaining StxA/B complex, which is eventually reduced in the ER lumen, where the cleaved and reduced StxA subunit is transported into the cytosol via the ER-associated protein degradation pathway [209, 210]. Interestingly, only about 4% of the transported toxin is reduced and transported trans-membrane into the cytosol [211]. Once in the cytosol, 28S rRNA is cleaved at the 4324 base adenine residue by the glycosidase activity of StxA, inhibiting amino acid chain elongation [212, 213]. The mechanism of action of Stx is thought to be similar to that of ricin, where cleavage of 28S rRNA results in the ‘ribotoxic stress response’ [214]. This response activates the stress-activated protein kinases, which in turn upregulate a number of proinflammatory cytokines. For example, internalized Stx has been shown to induce the release of interleukin-8 by host enterocytes [215], and results in inflammation and leukocyte migration into the sub-mucosa [216]. It is thought that the toxins are bound to the leukocytes, which then enter the blood stream and are free to bind other sites in the body rich in the GB3 receptor, such as the small vessels in the kidneys and brain [217].

This results in death of endothelial cells lining these vessels and activation of the inflammatory response, which is characterized by up-regulation of cellular kinase pathways and expression of P-selectin on the membranes of epithelial cells. P-selectin binds to and activates complement factor 3 via an alternate inflammatory pathway[218]. Initiation of the complement-mediated coagulation cascade culminates in attachment of fibrin to vessel walls, clumping of platelets and the formation of microthrombi. These thrombi obstruct micro-circulation blood flow and lead to hypoxic damage of tissues and organs such as the large-intestine, kidney, pancreas and brain [219]. Clinical manifestations of this include HC, kidney failure and neurological disturbances.

Regulation of toxin production in STEC is not fully understood, and significnat differences in toxin production between two strains of the same toxin-variant can be observed [220].

Although Stx has adverse effects on incidental human hosts, in the natural ruminant reservoir, Stx-production has been linked to enhanced bacterial fitness in terms of increased colonization [221], and in surviving predation by protozoan bacteriovores inside the bovine host [222, 223]. A possible benefit to both sheep and cattle harboring Stx-producing bacteria may be the anti-viral activity of the toxin against the oncogenic bovine leukemia virus (BLV), which, although widely distributed among ruminants, rarely exhibit clinical signs of BLV disease [224].

3.2.2 Stx-containing Bacteriophage

Stx production is mediated by lytic conversion of temperate bacteriophage integrated in the chromosome of STEC strains [225]. These bacteriophage are of the lambda phage family, comprised of dsDNA, have an icosahedral head, and a tail of heterogeneous size. The lambda phage itself is approximately 48.5 kb in size; however, most of the characterized Stx-phage are up to 50% larger and contain additional DNA that is thought to be of benefit to the host bacterium.

STEC can harbor one or more Stx-containing inducible prophage [226]. The lambda bacteriophage has been described as an ordered set of interchangeable modules, which increases the propensity of ‘mosaic’ bacteriophages to occur among related bacterial strains [227]. Many of the Stx-phages encode the Red recombinase system, which is an order of magnitude more efficient than the native *E. coli* recombinase system, increasing the potential for multiple genetic mosaics to be created [228]. Due to the highly efficient recombination machinery and the possibility of multiple Stx-phages within a single genome, recombination between different phage within a single cell is possible.

Stx genes are found within the late-gene region of the bacteriophage, downstream of the Q anti-terminator gene [228]; thus, the toxin genes are expressed whenever the phage enter the lytic cycle, without need for a separate secretion system to export the toxin, as the host bacterial cell is no longer intact following lysis. Two variants of the Q anti-terminator have been identified in *E. coli* O157:H7 strains, with the largely cattle-associated strains that possess Q21 found to produce lower amounts of Stx_2 than strains with the more frequently human-associated Q933 allele [229].

Activation of the bacterial SOS response implies the impending death of the bacterium, and as such there is an advantage to the temperate bacteriophage in entering the lytic cycle [230]. In STEC, the bacterial SOS stress response is activated by factors including environmental stress, DNA damage, exposure to fluoroquinolone drugs, and reactive oxygen species, which are released by predatory protozoa and cells that are components of the human innate immune system, such as neutrophils and leukocytes [223, 231, 232, 233]. Antimicrobials are not recommended in the treatment of STEC infection for this reason, as their usage is often accompanied by a surge in toxin production. Additionally, Stx-production has been shown to increase survivability of STEC O157:H7 within human macrophages [234].

The integrated bacteriophage of STEC are also unstable, and are completely lost from the bacterial host genome approximately 10% of the time during the course of human infection, in strains associated with HUS [235]. Thus, after host-passage, a former STEC strain may no longer harbor the defining feature of its virulence. In addition to escaping their STEC hosts, Stx-bacteriophage are frequently transferred to other susceptible bacteria in the intestinal environment, even converting former commensal organisms into toxin-producers [236]. During an infection, the amount of Stx produced can be increased by an order of magnitude or more simply by the *stx*-containing bacteriophage causing lytic conversion of additional bacterial hosts [237]. Bacteriophage are more resistant to chlorination and pasteurization than their bacterial hosts, and recovered phage from the environment are capable of infecting new cells [228]. Thus, the potential for Stx-bacteriophage transfer among bacteria means that phage can outlive their original bacterial hosts.

The mechanism by which Stx-phages recognize compatible cells differs between phage with short and long tails. Phage with long tails have previously been shown to recognize bacterial surface receptors such as the OmpC protein [238], while short-tailed Stx-phages recognize the YaeT membrane protein, which is involved in inserting other proteins into the cell membrane, and is highly conserved across the gram-negative bacteria due to its essential function [239]. Around 70% of all short tailed Stx-phages possess tail-spike proteins with identical or highly conserved YaeT sequences. The ubiquity of this surface protein is thought to have facilitated the spread of Stx-phages among gram-negative bacteria.

3.2.3 Role of Shiga-toxin

The production of Stx confers an advantage to a population of bacteria as a whole while being lethal to the individual cells that undergo lytic conversion to facilitate Stx production. Non-Stx containing bacteria in the same environment tend to undergo lytic, rather than lysogenic conversion; the sacrifice of a few STEC can cause an orders-of-magnitude increase in the amount of environmental Stx, while at the same time decimating Stx-naive bacterial populations [240, 241]. This allows the population to harbor the bacteriophage in a lysogenic state until a response to a threat is required, be it competing bacteria, protozoan predators or human immune cells.

The resistance to predation conferred to a bacterial population by expression of Stx may also facilitate increased environmental persistence and re-uptake of the bacteria by their bovine hosts, allowing a single-clone to dominate in a herd [242]. The increased environmental persistence of these strains also makes them more prone to infecting humans.

3.2.4 The Locus of Enterocyte Attachment and Effacement (LEE)

STEC are defined by their major virulence factor, but it is not solely responsible for the variation in virulence observed among different seropathotypes. It is thought that mobile genetic elements acquired through horizontal gene transfer (HGT) combine in serotype and strain-specific combinations that lead to different pathogenic outcomes [161]. Perhaps the most critical ability related to causing human disease is that of the bacteria to intimately attach to enterocytes (simple columnar epithelial cells) in the intestine. In many pathogenic STEC, most notably in EHEC, this characteristic is encoded by a chromosomal gene cluster known as the locus for enterocyte attachment and effacement (LEE) [243]. The LEE is comprised of five operons: LEE1-3 encode a type III secretion system (T3SS), seven effector proteins and the LEE-encoded regulator (Ler), which is encoded by the first gene of LEE1 and regulates expression of the entire locus; LEE4 encodes the EPEC-secreted-proteins EspA, EspB, EspD, and EspF (EspC is encoded outside of the LEE and is cytotoxic, but requires a type V secretion system working in concert with the T3SS to be translocated into epthelial cells [244]); LEE5 encodes intimin, the translocated intimin receptor (Tir) and the Tir-chaperone protein CesT [245, 246].

Attaching and effacing lesions are characteristic of colonization by LEE-encoding bacteria, and require the T3SS, which acts as a molecular syringe [247]. These lesions form following the creation of a pore in the host-cell membrane, where EscF molecules form the needle and EspA polymerizes to form a syringe linking EscF to the host cell membrane [248, 249]. EspB and EspD are then injected through the EspA syringe and integrate into the host membrane, which completes the pore formation between the host and bacterium, Tir and a number of other LEE and non-LEE encoded effectors (Nles) are injected into the cell using this secretion system. Once in the host-cell cytoplasm, Tir localizes to the plasma membrane, which it spans and subsequently exposes its central domain to the cell-surface; this domain interacts with the LEE-encoded intimin protein present on the surface of the bacteria to tightly attach the bacterium to the host cell [250, 243]. Microtubules supporting the epithelial cell-surface structure are altered to form a cup-like structure and the microvilli are destroyed (effaced) in sites of bacterial colonization, and an actin-based pedestal is formed, which is known as an attaching and effacing (A/E) lesion. Symptoms of watery diarrhea in patients are thought to be caused by inflammation resulting from the effacement of microvilli and the secretion of chloride into the large intestine. There are 19 known variants of the *eae* gene, which encode intimin within STEC; this variation may in part be responsible for the different host-tissue and host-organism specificities that are observed [251, 252]. The most commonly isolated serotypes from humans have intimin types with a high specificity for both human and bovine tissue [253].

The expression of LEE genes is under global repression by a histone-like nucleoid-structuring (H-NS) protein, which can be relieved through interaction with Ler, itself an H-NS-like protein [254]. The expression of *ler* is regulated by a number of global and LEE-specific regulators, with integration host factor (IHF) being essential to the activation of *ler* by binding upstream of the promoter [255].

Other global positive regulators of LEE expression include the ribosome binding protein BipA [256], the factor for inversion stimulation, IHF [257], and the stress-response alternate sigma factor RpoS [258]. Hha appears to be a global negative regulator (so called due to the strain it was discovered in) [259]. GrlA and GrlR are an activator and repressor of LEE expression, respectively, encoded within the LEE, which also regulate expression of flagellar genes and enterohemolysin (*ehx*) [260]. GrlA and Ler are also activators for each other's expression, mutually activating the transcription of the other and creating a positive-feedback expression loop [261] .

Induction of LEE expression can be in response to contact of fimbrial adhesins (Hcp, Ecp, Efa) with host enterocytes, high osmolarity, bicarbonate ions, epinephrine, norepinephine, the bacterial SOS response and quorum sensing [185]. Quorum sensing (QS) is a form of communication within and between members of bacterial species [262]. Hormone-like chemicals are produced as signals to other bacteria regarding environmental conditions such as bacterial population density and nutrient availability. The QS molecules act as regulators of gene expression. In STEC, QS in carried out using the *luxS* system, where QseA (quorum-sensing *E. coli* regulator A), activates transcription of a number of genes including *ler* in the LEE locus [263]. Other quorum sensing factors include the histidine kinases, QseC and QseE, which respond to epinephrine and norepinephrine and regulate LEE expression positively and negatively, respectively [264]. It has been hypothesized that QS suppresses expression of LEE until the organism reaches a suitable site for colonization in the bovine intestine. Regulatory factors include RgdR, which requires the expression of Ler, but enhances LEE expression [265]. Classical EPEC strains that also possess LEE also typically have a plasmid-encoded PerC locus, which is needed for *ler* to be expressed. EHEC strains do not have a PerC-encoding plasmid, but five chromosomally encoded *perC* homologues (*pch*) have been identified [266, 267]. Two of these genes, *pchA* and *pchB* directly control LEE expression and are under positive regulation from the LysR-homologue A gene product LrhA, which is also involved in regulation of chemotaxis, the production of flagella, and motility [268, 269]. In addition to the Tir protein injected by T3SS of the bacteria, host-cell encoded nucleolin is thought to act as a receptor for LEE-encoded intimin [270]. Stx has been shown to up-regulate LEE gene expression and promote intestinal colonization in mice, in part by causing enhanced expression of nucleolin by host-cells. However, it has also been shown that $\text{Stx}_{2\text{-}}$bacteriophage integrated into the genome of *E. coli* O157:H7, and specifically the CII gene encoded by this phage, repress LEE expression. These apparently contradictory findings may at least in part be explained by the need for a coordinated, staged approach to allow colonization. It is theorized that initially Stx up-regulates nucleolin expression by host cells and suppresses LEE expression. Subsequently, in the absence of Stx expression, LEE-encoded intimin is produced and binds to nucleolin. This regulation and coordination of LEE operons by Stx-bacteriophage elements may be important in limiting the host immune response and in determining the anatomical site and extent of colonization by these pathogens [271].

Lastly, post-transcriptional and post-translational control of LEE gene expression may allow fine-level adjustment of transcript levels [272]. Regulatory elements responsible for this may include RNA-binding proteins such as carbon-storage regulator A (CsrA), which in EPEC acts in a concentration dependent manner to activate (low CsrA levels) and repress (high CsrA levels) LEE expression [273]. In EHEC, Hfq represses LEE expression through two mechanisms: in the exponential growth phase Hfq interacts with the transcribed *grlRA* operon, causing lower levels of GrlA and thus reduced expression of Ler; in the stationary phase, Hfq represses translation of the *ler* transcript directly [274]. RNase E also regulates LEE expression, by degrading mRNA transcripts of the LEE4 operon [275]. The non-coding RNA DsrA up-regulates expression by pairing with H-NS and RpoS mRNAs in the presence of Hfq [276, 277]. The most well characterized post-translational regulator of LEE expression is ClpXP, which is an ATP-dependent protease, which is thought to induce expression of *ler* by regulating the levels of GrlR and RpoS expression [258].

In addition to LEE-encoded proteins, non-LEE encoded effectors (NLE) operate by modifying signal transduction pathways and can influence the amount of bacterial attachment, cytotoxicity and protection from phagocytosis [278]. Over 100 such proteins have been identified or proposed to exist [279, 280]. Nles are encoded on the genome outside of the LEE pathogenicity island, typically in prophage or genomic islands with lambda phage remnants. Nles exploit the LEE T3SS apparatus and are injected into the host cell via the LEE T3SS [157, 281]. The function of the majority of these effectors is unknown. The largest family of NLEs are the NleG homologues, of which 14 have been identified [282]. These effectors have been identified as E3 ubiquitin ligases, but their role in human infection has not yet been discovered. Other NLEs appear to suppress the host innate immune response, such as NleC and NleH1 / NleH2, which suppresses the pro-inflammatory transcription factor NF-kB interleukin-8 release [283, 284]. NleA localizes to the Golgi apparatus and has been shown to play a role in the destruction of the tight junctions of epithelial cells; transcriptional activation of this effector is enhanced by bacterial starvation conditions [285, 286].

3.3 The STEC Genome

The first sequenced genome of *Escherichia coli* was that of K-12 strain MG1655 in 1997 [287]. This strain was originally isolated from the feces of a healthy human and has become a common laboratory strain [288]. This 4,639,221 bp genome was found to contain many areas associated with horizontal gene transfer (HGT), including insertion sequence elements and bacteriophage elements. The extent to which HGT was responsible for altering the *E. coli* genome and contributing to pathogenicity was not fully apparent until the first two pathogenic strains of *E. coli* were sequenced and compared to K-12 MG1655.

Both of these strains were of serotype O157:H7; one was isolated from a foodborne outbreak related to radish sprouts in 1996 from the city of Sakai, Japan [12]; the second was isolated from a hamburger-related outbreak in 1982 from the state of Michigan in the United States [13]. The O157:H7 genomes were found to be 5,498,450 bp and 5,528,445 bp in size respectively. The O157:H7 and K-12 genomes shared approximately 4.1 Mbp, and there were nearly 1400 genes unique to the O157:H7 strains, and roughly 530 genes unique to K-12, with an estimated most recent common ancestor 4.5 million years ago [287]. As more genomes of *E. coli* have been sequenced, nearly 16 000 gene families have been identified, with only 6% being shared among all genome sequences of *E. coli*, and genomic regions specific to individual strains nearly always identified [289].

The unique regions interrupting the syntenic backbone were called ‘genomic islands’ and were found to largely contain genes known or thought to be related to virulence in humans. These islands also suggested that large-scale changes in phenotype could be acquired via a single gene transfer event among phylogenetically unrelated organisms. Such an occurrence was most recently observed in the 2011 *E. coli* O104:H4 outbreak in Germany, where the bacteriophage encoding Stx$_2$ was transfered from an EHEC strain into the human-adapted enteroaggregative *E. coli* O104:H4 genomic background, resulting in previously unseen levels of STEC-associated human illness [290].

Many STEC serotypes are thought to be distantly related to *E. coli* O157:H7 based on multi-locus enzyme electrophoresis (MLEE); however, they may have adapted to existence in the ruminant gastrointestinal tract through the acquisition of highly related genetic elements such as Stx-encoding phage, the LEE and adherence-conferring pathogenicity islands and plasmids [291]. This intra-STEC transfer may be facilitated by the close spatial proximity of numerous STEC strains within the bovine intestinal tract.

Even though these groups diverged long ago, they have independently acquired the requisite virulence attributes that seem to enhance survival in the ruminant intestine and / or the environment that ruminants live in. The LEE pathogenicity island, which is inserted in the *selC* locus in both *E. coli* O55:H7 EPEC and O157:H7 EHEC strains, encodes the gamma-intimin variant. This locus is instead inserted in the *pheU* locus in O111:H8 and O26:H11 strains and encodes the beta-intimin variant. The apparent parallel evolution suggests that similar selective pressures have acted on bacterial populations of different evolutionary lineages to form common features in STEC [76].

Parallel evolution is evident not only in LEE-containing, but also in LEE-negative STEC [292]. LEE-negative STEC are the most common STEC isolated from adults in Germany [293], and consist of serotypes such as O113:H21 that are associated with severe disease indistinguishable from that caused by STEC O157. A comparison of 17 diverse LEE-negative STEC using MLST showed multiple evolutionary lineages, suggesting LEE-negative pathogenic STEC have emerged multiple times [292]. In a separate study of LEE-negative STEC O174 strains, similar virulence gene content in unrelated clonal groups of the serovar was demonstrated [294]. Genes from the urease genomic island were rarely found in these strains, suggesting that the selective pressures in the ruminant environment act to select a particular phenotype, rather than the same genomic makeup in every STEC group. When examining individual genes on the large plasmid of LEE-negative STEC, different evolutionary histories were found for each of those studied, highlighting the role of HGT in the acquisition of common virulence determinants. Additionally, putative adhesins and toxins found on the large plasmid of LEE-negative strains, that are not found on the large plasmid of O157:H7 strains, were shown to have homologs on the O157:H7 chromosome. Such multiple occurrences of pathogenically similar, yet evolutionarily distinct groups, suggest that STEC virulent to humans may arise when a given set of virulence genes arrive in an *E. coli* host, and that these virulence genes likely play a common role in the ecology of the organism, and subsequently the pathogenesis of human infections.

STEC are etymologically confined to the *E. coli* group possessing at least one Shiga-toxin; however, just as there are STEC that are LEE-negative, there are EPEC that possess the LEE but do not produce Stx. EPEC also typically possess a plasmid encoded bundle forming pili gene (*bfp*) that helps stabilize microcolony formation [295]. Similarly, when the evolution of the *E. coli* enterohemolysin gene, *ehx* was examined, it appeared to have been acquired on at least three separate occasions by HGT: the first by *eae*-positive strains (including both EPEC and STEC); the second by most *eae*-negative STEC strains; and the third by a specific subgroup of *eae*-negative STEC. Although six subtypes were identified, selective pressures have maintained a highly conserved locus [296]. The *ehxA* gene on the pO113 and pO157 plasmids were also found to have been acquired separately, and the plasmids themselves were shown to have separate evolutionary histories [292]. When EPEC and STEC isolates from New Zealand were characterized by intimin typing, strain groups comprised of isolates with the same intimin type (beta, zeta or theta), were identified from both EPEC and STEC groups, and from humans, cattle and sheep, with the exception of theta strains, which were only isolated from cattle and sheep [297]. This suggests that HGT has occurred between strains harboring highly related LEE loci to create strains that we currently classify as both STEC and EPEC.

Examinations of other virulence genes have revealed the same extensive HGT among distinct clonal groups. Phylogenetic analysis of both *stx*$_1$ and *stx*$_2$ from STEC associated with disease outbreaks showed an uneven distribution of these *stx* genes, suggesting that a specific background was required for the maintenance and expression of *stx*. When the phylogeny of EspP was examined, four distinct alleles that were not serotype specific were discovered among 56 separate STEC serotypes, with only two alleles encoding proteins capable of proteolysis [298].

Combined, these independent observations are highly suggestive that multiple events, rather than a single one. were necessary for the evolution of STEC virulence, and that STEC have evolved in parallel through HGT of multiple virulence genes. It also suggests that a phylogeny created from only shared genes among all STEC may not be able to identify strain groups that have the same virulence profiles or disease potential to humans, including STEC that have already been identified as clinically important.

3.3.1 Acid Resistance

STEC have evolved a number of acid resistance mechanisms that allow them to survive transit from the stomach into the small intestine. Acid resistance is thought to be an important determinant of infectious dose. Epidemiological studies suggest that the infectious dose for certain STEC may be fewer than 100 CFU [299]. STEC have acquired at least three acid resistance systems that allow this: ie., arginine and glutamate decarboxylase-dependent systems and an oxidative glucose catabolite system [300]. The latter two are mediated by the expression of an alternate RNA polymerase sigma factor RpoS, which is a master regulator of over 30 genes that are expressed in response to environmental stress [301]. Despite this master role of RpoS, clinical STEC isolates with rpoS mutants have been identified and a range of functional heterogeneity of *rpoS* expression has been documented, suggesting the existence of an independent acid response system [302]. These RpoS mediated systems are of particular concern for food-production and food preservation. Acid resistance also plays an important role in survival of bacteria within animal reservoirs [302].

Additional acid resistance can be conferred through the urease gene cluster (*ureA* - *ureG*), which can help neutralize gastric acid through the enzymatic breakdown of urea; however, only a small subset of STEC produce functional urease [303]. Despite this functional paucity, the urease gene cluster is significantly associated with increased intestinal colonization and its presence is highly linked with the presence of the LEE [304, 303].

3.3.2 Adherence and Additional Virulence Factors

Adherence to epithelial cell surfaces such as those lining the gastrointestinal tract, is essential for bacterial colonization of the host. As previously discussed, most EHEC possess the LEE T3SS and adhere to host cells through this A/E mechanism. Non-O157 STEC, LEE-negative strains appear to have evolved alternative adherence factors to make up for this deficiency [305]. These other adhesins include the chromosomally encoded *iha*, the iron-regulated gene A homologue adhesin, which is present in a wide range of STEC serotypes [306], and *efa1* and *cah*, which have been associated with increased attachment and with STEC serotypes most commonly associated with severe disease [157, 307]. A novel immunoglobulin-binding protein, EibG, associated with chain-like adherence to human-epithelial cells has also been identified in non-O157 STEC strains [308].

PagC appears to be important in establishing infections. Wild type STEC strains that harbor this gene have higher population densities or competitive indices in mice when co-administered with isogenic pagC-mutants [234]. The presence of this gene has also been shown to aid in the survival of bacteria inside human macrophages in strains of STEC as well as *Salmonella* [309]. ChuA and FepC are involved in heme-transport [310], which along with genes *irp2* and *fyuA*, found on the non-O157 STEC high-pathogenicity island, may enhance the fitness of strains under iron limitation [311]. This probable toxin gene has a high level of similarity to *set*, encoded by *Shigella* spp and is found on a genomic island along with the Nle factors, *nleB*, and *nleE* [312]. A factor associated with A/E lesions in pigs, the porcine A/E associated factors, *paa*, has also been identified in STEC [313].

3.4 Plasmid-encoded Virulence Factors

STEC typically carry one or two large plasmids, ranging from 90 to 205 kbp in size, which are typically named after the serogroup that they were originally found in, e.g. pO157, pO113, pO26 etc. [161, 314]. These large virulence plasmids are highly heterogeneous and mosaic in form, and while they are variable, encode a subset of virulence factors that are common to many large STEC plasmids. A recent study found that 95% of all STEC strains carried at least one plasmid, with approximately half carrying a single plasmid 90 kbp in size and the other half containing additional plasmids, some with as many as six plasmids [315]. Small plasmids that code for little more than plasmid replication and maintenance genes, as well as those containing antibiotic resistance determinants have also been characterized in STEC [316].

Large plasmid associated virulence factors include *ehxA* (also known as *hlyA*), which codes for an enterohemolysin that frees hemoglobin and iron from red blood cells for bacterial use [317]. It may be required for human infections and is often found in bovine STEC strains, and is more commonly isolated from STEC strains bearing the intimin (*eae*) gene rather than LEE-negative STEC [296, 318]. The *ehxA* gene is also commonly found in environmental *E. coli* that lack well recognized virulence factors such as Stx and intimin, suggesting a role in survival and / or persistence outside of the intestinal tract, where trace elements such as iron are required but rare [296].

The plasmid borne EHEC enterohemolysin operon has $>$ 60% identity to the alpha-hemolysin gene of *E. coli*. Like alpha-hemolysin, enterohemolysin has the gene order ehxCABD [319]. EhxA is synthesized in an inactive form that must be acylated by EhxC, and its secretion from the cell requires EhxB and EhxD. The toxin forms multimeric, ring-shaped pores in the host-cell membrane. Enterohemolysin has been shown to lyse sheep red blood cells and bovine lymphoma cell line leukocytes, as well as provoke the release of a pro-inflammatory cytokine interleukin-1B response in human leukocytes. The cytotoxic extracellular serine protease and autotransporter, EspP, is another EHEC plasmid-encoded gene. The activity of EspP is thought to be synergistic with that of ExhA during infection [320]. EspA cleaves pepsin A and human coagulation factor V and has at least four distinct alleles; however, only two of these alleles encode functional proteins. These functional alleles of the protein have been found in the non-O157 serovars O26, O111 and O145, which are classified as seropathotype B and frequently associated with human infections.

A novel AB(5) STEC toxin, encoded by *subAB* is similar in sequence and mechanism of action to subtilase, found in *Bacillus subtilis*. The gene encoding the toxin is found on the large virulence plasmid of certain non-O157 STEC [298]. This toxin is normally associated with STEC that lack the LEE, but has recently been found in the chromosome of *E. coli* strains lacking *stx* genes and is associated with a phenotype typical of enteroinvasive *E. coli* [321]. Rather than disrupting ribosome function like Stx, this toxin cleaves the endoplasmic reticulum chaperone protein BiP/GRP78 [298].

Another virulence determinant of STEC is a catalase peroxidase coded for by *katP*, which is in addition to two other peroxidases, KatG and KatE, is found in most *E. coli*, including non-pathogenic strains [322]. KatP is active only when bacteria are exposed to high levels of peroxide, and helps protect the bacterium from peroxide-mediated oxidative damage. It has been shown that KatP is statistically more efficient at scavenging peroxide than the KatG and KatE, even though the latter two were sufficient for complete protection against peroxide damage [323]. KatP protection is independent of the *rpoS* regulator, while

KatG and KatE are *rpoS*-dependent.

In addition to peroxidases, *E. coli* O157:H7 has been found to contain a plasmidic second copy of lipid A myristoyl transferase on the *ecf* (*E. coli* attaching and effacing gene-positive conserved fragments) operon. This operon is responsible for altering the lipid-membrane structure of the bacterium, and has been shown to be critical for passage through the bovine gut, survivability in water troughs, susceptibility to antibiotics and detergents, and motility [324]. A zinc metalloprotease, StcE, under control of the LEE master regulator, Ler, also contributes to the alteration of the bacterial membrane by recruiting an inflammatory regulator, C1 esterase inhibitor, and increasing intimate adherence of STEC through the destruction of host glycoproteins [325, 326].

In the absence of LEE, other factors have been acquired that allow STEC to attach to host cells. These include the additional adherence factor ToxB, which is required for full adherence to host cells [327]; and the STEC autoagglutinating adhesin (*saa*), long polar fimbriae (*lpf*), and an auto-transporter involved in biofilm formation (*sab*) [12, 57]. The presence of *eae* and *saa* appear to be mutually exclusive in all strains studied to date [92]. These factors allow for bacterial attachment in the absence of the LEE and therefore strains expressing them do not cause the characteristic A/E lesions of LEE-positive strains.

The rapid accumulation of anti-microbial resistance (AMR) by certain STEC is increasing and is of concern from a public-health perspective. STEC worldwide have shown resistance to multiple, structurally unrelated antibiotics including aminoglycosides, fluoroquinolones, and cephalosporins [328, 329, 330, 331]. The prevalence of non-O157 STEC human infections, as well as AMR, has been increasing in Japan, where AMR in non-O157 and O157 STEC was found to be at 40% and 10%, respectively [332]. A study of dairy calves in India found that all STEC strains isolated were resistant to at least three of 11 antibiotics tested, most of which were cephalosporins and fluoroquinolones [159]. Extensive resistance to enrofloxacin, a fluroquinolone prescribed for human use was also observed.

In a study of cow-calf herds in Canada, 89% of all *E. coli* isolated contained resistance to at least one antibiotic, with 48% of the strains containing *stx* genes and / or the *eae* gene [333]. A separate study of AMR in STEC isolated from human and animal waste-water in Spain found 92% of strains from 32 STEC serotypes resistant to at least one antibiotic, and 50% resistant to 5 or more [334]. Interestingly, strains that were susceptible to the most antibiotics carried more than one *stx* gene.

Interestingly, both induction of temperate phage into the lytic cycle and HGT can be enhanced in the presence of antibiotics such as fluoroquinolones. Phage induction is mediated through the bacterial SOS response and would be expected to increase the amount of lytic bacteriophage in the gut following antibiotic exposure, including those encoding Stx [334]. This in turn increases the chances of disseminating toxin genes to other bacteria. Use of antibiotics in animal production as a growth promoting factor, and in veterinary medicine provides selective pressure to acquire AMR, which may be inadvertently enhancing the dissemination of bacteriophage-encoded toxins and other virulence genes, contributing to the emergence of new pathogens through the spread of bacteriophage-encoded virulence genes. This is especially important for STEC, as human-pathogenic strains appear to have emerged independently multiple times by acquiring virulence factors through HGT.

3.5 Evolutionary Dynamics

The natural STEC reservoir consists of ruminants, of which cattle are the major contributors of STEC to the environment and contaminated consumables based on their abundance [335, 149, 336]. Therefore, the evolutionary forces acting to select for particular traits are those that confer an advantage in the bovine host, or the ‘source’ habitat. Humans are thought to be largely incidental hosts. While cases of human to human transmission occur during outbreaks, STEC typically persists in human stools for only a few weeks following infection. Since the organism cannot sustain itself in humans and eventually succumbs to the harsh environment, traits that favor survival in humans are unlikely to be maintained in the population and therefore humans represent a ‘sink’ habitat [178]. In the source / sink evolutionary model, the source habitat selects for long-term evolutionary change and the sink habitat allows proliferation over a limited time-scale, with clearance from the sink habitat being the eventual result. Thus, even though small-scale evolutionary changes may occur in a sink environment, they will be lost along with the sink population.

The bovine intestinal environment allows ample opportunities for horizontal gene transfer to facilitate adaptation of the organism to this niche in the reservoir host. A multitude of prokaryotic and eukaryotic organisms, along with bacteriophage inhabit the intestinal tract. The intestinal lining is covered with a protective mucus layer, which consists of extracellular polymeric substances (polysaccharides, DNA, proteins, lipids) and provides a physical barrier for the bacteria that reside within and represents an impediment to pathogens attempting to reach the host cells [337]. This protective layer is important in signaling and ensuring nutrient availability for both pathogens and beneficial, host-associated bacteria. In the host, it is mucosal signaling that initializes the inflammatory response to a perceived pathogen.

The ability of human pathogens to form their own biofilms has been linked to increased virulence. In *E. coli* O157:H7, it was found that strains that were deficient in forming biofilms adhered poorly to intestinal epithelial cells [338]. Other work has shown that clades of *E. coli* that are good biofilm producers are able to survive in the environment longer than clades that do not produce biofilms [339]. In the context of a cattle herd, biofilm formation may also allow environmentally-adapted strains to become established within the cattle population by persisting in niches outside of the bovine host. Additionally, the virulence of the 2011 O104:H4 outbreak strain was thought to in part be due to its ability to form aggregative biofilms within the host gastrointestinal tract [340], and it was recently shown that the virulence gene expression by this strain was highest within an established biofilm [341].

The rate of HGT is up to 10^4 times more frequent in a biofilm than in the planktonic form of bacteria, owing to closer proximity of strains and the fact that one of the major components of the biofilm extra-cellular matrix is extra-cellular DNA [342]. The bovine intestinal environment has been home to the evolution of similar virulence characteristics among STEC strains multiple times. It is not currently known whether different serotypes of STEC compete with each other for sites within the ruminant host, or if infection by one type allows a greater chance of infection by another type. Future research is needed to understand the relationship among STEC, other members of gastrointestinal microflora and the host.

In addition to facilitating HGT, lytic bacteriophage also prey upon STEC, a relationship that has been shown to exhibit classic predator / prey relationships [343, 344]. Thus, selective forces are acting not only for optimal interaction with the bovine host, but also for avoidance of lytic bacteriophage. To this end, *E. coli*, and bacteria in general, have evolved methods of predator evasion that include: preventing adsorption of the bactriophage to the bacterial cell surface; restriction-modification systems that degrade viral DNA upon entry into the cell; abortive infection systems that lead to cell death and prevent viral replication; superinfection immunity conferred to the bacterial cell by a previously integrated bacteriophage, which prevents entry of other similar bacteriophage ; and the clustered regularly interspaced short palindromic repeats (CRISPRs) and the CRISPR-associated (*cas*) genes that target foreign nucleic acids in a manner similar to RNA-interference [345].

These microorganisms appear to be engaged in an arms-race situation, where bacteria evolve defenses against bacteriophage, which in turn evolve means around these defenses. Experimentally, co-propagation of *E. coli* and predatory bacteriophage have resulted in increased mutation rates in both bacteria and bacteriophage and observable resistance to infection and increased growth rates by the bacteria [346]; multiple outcomes have been observed from these experiments ranging from complete bacterial resistance to the phage to the persistence of both phage and bacteria [347]. Examples of this arms-race include the CRISPR / *cas* evolution in bacteria, and the retaliative anti-CRISPR genes recently discovered in bacteriophage [348]; and the toxin / anti-toxin systems that confer resistance to bacteriophage and the production of novel anti-toxins by bacteriophage that overcome these defenses [349].

3.6 Conclusions

Humans are the unfortunate incidental hosts of STEC that appear to be continually evolving to best suit their source environment. STEC are capable of causing devastating disease via virulence attributes that have been acquired multiple times through parallel evolutionary history, and through the use of different virulence factor combinations that yield similar pathogenic outcomes in humans. This suggests selective pressures in the bovine host for the phenotypes conferred by these genetic elements. Due to the wide diversity of genomic elements and virulence factors within STEC, they pose a difficult public health problem, with many cases of infection likely being overlooked simply because important characteristics of the etiological agent are not being targeted by laboratory tests. Many of the current laboratory tests and detection methods are targeted to detect specific serotypes such as O157:H7, and are not ubiquitous among STEC. In the future, as surveillance and knowledge of virulence mechanisms become more complete, it may be possible to reduce the human-health cost associated with all STEC.

Chapter 4

In silico genomic analyses reveal three distinct lineages of *Escherichia coli* O157:H7, one of which is associated with hyper-virulence.

4.1 Preface

This chapter was previously published as Laing CR, Buchanan C, Taboada EN, Zhang Y, Karmali MA, Thomas JE, Gannon VP. '*In silico* genomic analyses reveal three distinct lineages of *Escherichia coli* O157:H7, one of which is associated with hyper-virulence.' BMC Genomics. 2009 Jun 29;10:287. doi: 10.1186/1471-2164-10-287 [71].

4.2 Introduction

Escherichia coli O157:H7 is the most commonly implicated serotype of Shiga-toxin producing *E. coli* (STEC) or enterohemorrhagic *E. coli* (EHEC) associated with hemorrhagic colitis and the hemolytic uremic syndrome (HUS) [350, 351]. It is recognized world-wide as an important cause of both sporadic cases and outbreaks of food- and waterborne disease. Genomic diversity within populations of this pathogen is extensive and genome comparisons have revealed many DNA insertion / deletion and recombination events which are thought to be driven primarily through bacteriophage-mediated and other mechanisms of lateral gene transfer [352, 228, 22].

Three genetic lineages of *E. coli* O157:H7 have been described using octamer-based genome scanning (OBGS) and microarray-based comparative genomic hybridization (mCGH). Lineage I strains are commonly isolated from both cattle and clinically ill humans, lineage II strains are isolated primarily from cattle and intermediate lineage I/II strains are less well characterized with respect to phenotype and host distribution [353, 354, 355]. Additional studies on the evolution, population structure and genetic diversity of *Escherichia coli* O157:H7 have been carried out using a number of different genotyping approaches, each of which is based on targeting polymorphisms in a particular locus or set of loci. For example, the presence of Shiga-toxin (Stx) containing bacteriophage integration sites has been used to describe sixteen *E. coli* O157:H7 genotypes (with the majority associated with strains of bovine origin) [356]. Multi-locus variable number tandem repeat analysis (MLVA), which targets the number of repeats at nine genetic loci in the method proposed for PulseNet, has recently been used for epidemiological typing of O157:H7 strains [58]. mCGH, which examines the presence or absence of every gene in the genome of a reference strain or strains, has been used to elucidate the stepwise emergence of O157:H7 strains from an O55:H7-like ancestor [163] and by our laboratory to first characterize lineage I/II strains [354]. Single-nucleotide polymorphism (SNP) typing examines the single nucleotide changes throughout the *E. coli* O157:H7 genome; SNPs in 96 loci have been used to delineate 39 genotypes in over 500 strains, which have been partitioned into nine evolutionary clades [70]. One of these clades (clade 8) was found to contain putative hypervirulent strains, including those implicated in the 2006 spinach- and lettuce-associated outbreaks in the United States. Comparative genomic fingerprinting (CGF), which examines the presence or absence of the most common variable loci within the genome, has recently been used with 23 loci to analyze 79 O157:H7 strains, and to group them into lineages and epidemiologically and phenotypically related clusters [66].

The various methods have proved useful on their own but in most cases the strain groupings from one method are not easily relatable to another, and none of these methods has yet achieved the ‘gold standard’ status to be used as a common genotyping method. Because each of these genotyping approaches can provide important information on isolates with different pathogenic and virulence characteristics, the need to replicate each of these methods using a common group of reference strains has arisen. However, present restrictions on international strain exchange make the acquisition of certain reference strains nearly impossible for researchers outside the country of origin of the strains. While the stepwise emergence of *E. coli* O157:H7 from an O55:H7-like serotype has been well supported [163, 357], the relationship among *E. coli* O157:H7 lineages and clades incorporating data from standard and more advanced genotyping methods has not been evaluated. Thus, discoveries such as the recently described hyper-virulent clade of *E. coli* O157:H7 strains [70] are framed only in the context of the individual study.

Recently, three complete *E. coli* O157:H7 strain sequences and 11 whole genome shotgun sequences have become available in Genbank. With the additional sequencing of one O157:H7 and one O145:NM strain by our group, the availability of multiple whole-genome sequences has allowed us to apply several molecular typing schemes *in silico* to a group of *E. coli* O157:H7 and related strains. We have applied six molecular typing methods to differentiate among *E. coli* O157:H7 strains to a panel of nineteen genome-sequenced strains: Stx-bacteriophage insertion site typing [356], CGF [66], mCGH [354, 163], SNP genotyping [70], genomic *in silico* subtractive hybridization (GISSH) (Laing et al. in preparation) and MLVA [58].

In this study we provide a view of the relationships among *E. coli* O157:H7 strains based on a supernetwork representation of these six separate molecular typing methods. We also show how hyper-virulent clade 8 *E. coli* O157:H7 strains fit into this scheme and suggest the use of genotyping approaches that are relatable as part of a ‘common genotyping language’ until whole genome sequencing becomes routine in bacterial strain genotyping.

4.3 Materials and Methods

4.3.1 Genome-sequenced Strains Used in the Analyses

Our analyses utilized the complete and fully annotated genomes of K12 MG1655 and *E. coli* O157:H7 strains EDL933 and Sakai. Additionally, 14 *E. coli* O157:H7 whole-genome shotgun sequences from GenBank were included in the study, along with two whole-genome shotgun sequences that were sequenced in-house: EC71074, an *E. coli* O157:H7 strain and EC33264, an *E. coli* O145:NM strain (Table 4.1).

4.3.2 In silico Lineage Typing

The LSPA6 primer sequences [355] were utilized for *in silico* polymerase chain reaction (is-PCR) experiments to determine the lineage types of the strains in Table 4.1. The is-PCR method involved BLASTN searches [111] of the primer sequences combined with data processing in Microsoft Excel to determine the expected band size; any primer sequence with less than 80% sequence identity in a BLASTN search was considered absent.

4.3.3 Shiga-toxin Encoding Bacteriophage Insertion Site Typing

As described by Besser et al. [356] the primer sequences targeting regions of bacteriophage insertion were subjected to is-PCR, using a sequence identity of 80% as a positive threshold. Data were given discrete character scores, with 0 representing absence and 1 representing an amplified band. In cases where bands of more than one size were found for the same primer set, the digits 2 through 5 were used to denote bands of additional size.

4.3.4 Comparative Genomic Fingerprinting (CGF)

Twenty-three of the most variable loci among *E. coli* O157:H7 strains were targeted by the comparative genomic fingerprinting method of Laing et al. [66]. The 80 strain binary dataset from the Laing et al. study was used, along with is-PCR results from the strains in Table 4.1, using a sequence identity of 80% as a positive threshold. The is-PCR data were converted to binary characters, with 0 for absence and 1 for presence.

4.3.5 Genomic in silico Subtractive Hybridization (GISSH)

The method for determining the existence of 417 genomic regions larger than 500 bp that are absent from the genomes of EDL933 and Sakai but present in other *E. coli* O157:H7 whole genome shotgun sequences will be described in detail elsewhere (Laing et al. in preparation). Each of the novel regions was split into contiguous segments of 500 bp, creating 1456 segments representing the original set of novel regions.

BLASTN was used to determine the distribution and these 1456 novel regions among the strains in Table 4.1. A segment containing less than 80% total sequence identity was considered absent and denoted by a 0, while a region containing greater than or equal to 80% total sequence identity was considered present and denoted by a 1.

4.3.6 Single Nucleotide Polymorphism (SNP) Analysis

Manning et al. [70] recently described the existence of 96 SNPs among *E. coli* O157:H7 strains. The sequence of each SNP-containing locus in *E. coli* O157:H7 strain Sakai was used to conduct BLASTN comparisons of all 96 SNPs in each genomic sequence. Data were scored as single-letter nucleotides for either version of the SNP as put forth by Manning et al. with the following exception: in the Manning et al. study, *E. coli* O157:H7 strain

Sakai locus ECs0606 is listed as having a SNP of A117G, whereas our data analysis found the SNP to be A117C.

4.3.7 Microarray Comparative Genomic Hybridization (mCGH)

Zhang et al. [354] conducted a comprehensive microarray analysis of 31 *E. coli* O157:H7 strains using probe hybridization data for 6057 open reading frames. Analysis of those data found that the 50 mer probes used in the microarray experiments shared sequence identity of at least 80% to one or more of the three control strains (K12 MG1655, EDL933 and Sakai) in 6021 of the 6057 probes. The positive threshold used for *in silico* microarray hybridization was 80% and therefore the 36 ORFs with probes having less than 80% homology in all three control strains were not used. The *in silico* microarray hybridization used BLASTN to test the presence of each ORF in every strain in Table 4.1. Probes with identity greater than or equal to 80% were scored as present and everything else as divergent. Results were converted to a binary format for subsequent analyses with 1 representing a present ORF and 0 a divergent one.

Zhang et al. used the used the same MWG probe set as Wick et al. [163], which was used to determine the stepwise emergence of O157:H7 strains from an O55:H7-like ancestor. In their study, Wick et al. included two sorbitol-fermenting strains (CB2755 and 493/89) and three O55:H7 strains (Dec 5d, TB182A and 5905) that were absent from the study by Zhang et al. To expand the scope of our study, binary data for these five strains were included, with any loci not included in the Wick et al. study scored as missing and denoted by '?'.

4.3.8 Multi-locus Variable Number Tandem Repeat Analysis (MLVA)

The MLVA PCR scheme described by Hyytia-Trees et al. [58] was used *in silico* to identify the number of repeats at each locus, which is based on the number of repeats at nine variable loci within the O157:H7 genome. In accordance with the reference, partial repeats were rounded to the nearest whole number to generate the final set of data.

4.3.9 Construction of Dendrograms

The character data for each of the six methods were converted to Nexus format [358] and imported to PAUP* v4.0 [359], where the data were analyzed using maximum parsimony. The tree space was thoroughly sampled using 1000 random sequence additions, with 10 trees kept per search using the following PAUP* command: ‘hsearch enforce = no start = stepwise addseq = random nreps = 1000 nchuck = 5 chuckscore = 1’. The maximum parsimony tree for each set of data was saved and visualized in SplitsTree v4.0, where K12 was used as the outgroup.

4.3.10 Construction of Supernetworks

The six maximum parsimony trees representing the six datasets comprised of the 19 strains in Table 4.1 were imported to Splitstree v4.0 and a supernetwork was created using Z-closure with 1000 iterations and the unweighted mean distance of each tree to generate an equal angle representation of the data. A second supernetwork was created in an analogous fashion; however the CGF and microarray trees included data from the previously published studies of Laing et al. [66], Wick et al. [163] and Zhang et al. [354].

4.4 Results

4.4.1 Lineage Typing

The is-PCR LSPA6 typing of the *E. coli* O157:H7 strains in Table 4.1 designated strains EC4501 and TW14588 as lineage I, eleven of the 14 GenBank strains and the in-house strain EC71074 as lineage I/II, and strain EC869 as lineage II.

4.4.2 Stx-phage Insertion Site Typing

Besser et al. [356] found greater diversity within O157:H7 strains isolated from cattle than those from human clinical illness, but the relationship between their typing scheme and lineage groupings was not explored. In Table 4.1 it is evident from *in silico* analysis that all genotypes based on the Stx-phage integration site typing can be related to mCGH lineage.

All lineage I strains were of genotype 3 in that they possessed both *stx*$_1$ and *stx*$_2$ and both bacteriophage integration sites *yehV* and *wrbA* were occupied. Additionally, they all possessed the 'K' form of the N135K polymorphism in the FimH mannose binding pocket, lacked *stx*$_{2c}$, had the 'T' form of the T238A polymorphism in *tir* and were positive for the Q-*stx*$_2$ junction of phage 933W.

All lineage I/II strains were of genotype 1 in that they contained *stx*$_2$ but not *stx*$_1$ and contained an occupied *yehV* and an intact *wrbA* bacteriophage integration site. Additionally, they all possessed the 'N' form of the N135K polymorphism in the FimH mannose binding pocket and the 'T' form of the T238A polymorphism in *tir*. However, the Q-*stx*$_2$ junction of phage 933W and *stx*$_{2c}$ were variably present among the lineage I/II strains.

Lineage II strain EC869 was of genotype 6. It contained both *stx*$_1$ and *stx*$_2$ had the variant-R form of the *yehV* bacteriophage right junction and an intact *wrbA* integration site. EC869 possessed *stx*$_{2c}$, the 'A' form of the T238A polymorphism in *tir*, the 'N' form of the N135K polymorphism in the FimH mannose binding pocket and was negative for the Q-*stx*$_2$ junction of phage 933W. The raw data for this and all other *in silico* typing methods are provided (supplementary material at: `http://www.biomedcentral.com/content/supplementary/1471-2164-10-287-s1.xls`).

4.4.3 SNP Genotyping

The clade of each strain based on SNP typing was determined as described by Riordan et al. [360] where SNPs in the four genes ECs2357, ECs2521, ECs3881 and ECs4130 allow the delineation of the four most common clades of *E. coli* O157:H7. Our analysis typed EDL933 as clade 3 (SNP profile in the gene order given above as CGCT), Sakai as clade 1 (SNP profile CTTT), TW14588 and EC4501 as clade 2 (SNP profile CGTT), EC869 as one of clades 47 (SNP profile CGCC) and the remaining O157:H7 strains as members of the hyper-virulent clade 8 (SNP profile AGCC) (Table 4.1).

4.4.4 GISSH-based Novel Region Distribution

Zhang et al. used a microarray based on lineage I strains EDL933 and Sakai and the K12 strain MG1655, so it is not surprising that the lineage I strains were found to contain more of the lineage I markers than strains from the other lineages [354] . To account for novel genomic regions present in other O157:H7 strains, but not found in EDL933 or Sakai, an *in silico* subtractive hybridization was performed on every O157:H7 sequence in Table 4.1 against EDL933 and Sakai [127]. The study found 417 separate regions comprising 1456 segments of approximately 500 bp in length (0.8 Mbp of novel DNA sequence). The distribution of these segments *in silico* is shown in Figure 4.1 and highlights the fact that lineage I strains possess genomic regions not represented in the original microarray probe set, as well as the fact that there are other lineage I/II and lineage II specific genomic regions.

4.4.5 Combined Analysis

The maximum parsimony trees from each method shared a number of commonalities (supplementary material at: `http://www.biomedcentral.com/content/supplementary/1471-2164-10-287-s2.pdf`). All three lineages grouped distinctly with the following methods: CGF, SNP genotyping, mCGH and GISSH-based novel region typing. Stx-phage insertion site typing distinguished lineage I strains as a separate cluster but lineage II strain EC869 did not form a distinct cluster from the lineage I/II strains. Similarly, MLVA put lineage II strain EC869 on its own branch, whereas the lineage I/II and lineage I strains did not group separately with this method. All of the genotyping methods except mCGH identified lineage I/II strains EC508 and EC71074 as a separate cluster, which was placed into a group close to the lineage I strains in the CGF and Stx-phage insertion site based maximum parsimony trees and was placed between lineage I strain Sakai and the other three lineage I strains in the maximum parsimony tree based on MLVA data.

In order to obtain a more complete understanding of the relationships among *E. coli* O157:H7 strains, the maximum parsimony trees derived using *in silico* data from the six separate molecular typing methods included in this study were combined to form a supernetwork. We accomplished this by combing the six maximum parsimony trees generated using PAUP* instead of concatenating the data into a single matrix and generating a single most parsimonious tree. This ensured a method with many data points such as mCGH did not outweigh a method like MLVA that contains few data points. The supernetwork based on the six trees given equal weighting (Figure 4.2) shows competing signals, rather than an inferred absolute tree [361]. As can be seen, the 19 strains were distributed into four main groups. One contained strains K12 MG1655 and O145:NM EC33264 and another the single lineage II strain EC869. The remaining two groups of the supernetwork were lineage-specific, comprising one cluster of lineage I strains and another of lineage I/II strains. The distribution of the four lineage I strains was consistent with the clade breakdown of Manning et al. [70], as the clade 2 strains TW14588 and EC4501 were more closely related to each other than clade 3 strain EDL933 or clade 1 strain Sakai.

Because only a relatively small number of strains have been sequenced, we repeated the analysis including experimental microarray [354, 163] and CGF data [66] for which lineage information was available to provide a larger context for our *in silico* analysis. This added 83 strains to the 19 strain *in silico* dataset and included data for three *E. coli* O55:H7 strains and two sorbitol-fermenting *E. coli* O157:H- strains (Table 4.2), allowing us to frame the resulting supernetwork within the context of the proposed stepwise emergence of *E. coli* O157:H7 [163]. The supernetwork in Figure 4.3 again showed four main groups. All lineage I and II strains formed discrete branches in the supernetwork, while the lineage I/II strains formed a cluster that was closest to *E. coli* K12 and strains of O55:H7, O145:NM and sorbitol-fermenting O157:H-.

4.5 Discussion

The acquisition and loss of genetic elements in *E. coli* O157:H7 is thought to affect the virulence of this pathogen in humans [362, 363, 64]. While attributes like the Stxs [364, 365, 366] and the locus of enterocyte attachment and effacement (LEE) [367] are well known, other less well characterized elements such as non-LEE effectors (NLE)s [279] may contribute to the spectrum of virulence that has been captured for STEC within the seropathotype classification [157].

This genomic diversity can largely be attributed to bacteriophages [228, 368] and other mobile elements [369] that cause DNA segment insertion and deletion events, rather than single-nucleotide changes [368]. The best characterization of diversity within *E. coli* O157:H7 must therefore take both single-nucleotide changes and large region turnover events into consideration. Whole genome sequencing is the best method for such analysis, but until the number of completely sequenced strains increases and whole genome sequencing becomes both routine and cost-effective, an estimation of whole genomic change based on sampling of polymorphisms or variability at a number of specific loci will have to suffice. A number of molecular genotyping methods have recently emerged that target different regions and types of variation, so the combination of data from these methods can offer a picture that is greater than the sum of their individual views. It can be argued that lateral gene transfer obscures the relationship among bacterial strains; however, once mobile elements are acquired and if they are stably maintained, they can be especially valuable in assessing strain relationships. We therefore advocate the use of genotyping methods that rely on multiple spatially distinct loci to provide a robust view of the O157:H7 population structure. Each individual multi-locus method in our study pointed to a similar tree structure and the combination of methods in the supernetwork (Figure 4.3) showed three distinct lineages, two of which were originally proposed by Kim et al. [353] and the third by Zhang et al. [354], adding confidence to the conclusion that these lineages exist.

The results of SNP genotyping over 500 O157:H7 isolates found hyper-virulent clade 8 strains, the causative agents in the spinach- and lettuce-related outbreaks of 2006 in the United States, to be most closely related to sorbitol fermenting *E. coli* O157:NM strains [70]. Based on our *in silico* analyses, these clade 8 strains, which appear to be increasing in prevalence and have a greater association with HUS than other strains [70], are members of lineage I/II (Table 4.1). Interestingly, *E. coli* O157:H7 strain TW14588 was designated as clade 8 in the initial publication on variation in virulence among the clades [70], but all *in silico* analyses conducted during this study on the whole-genome shotgun sequence available from GenBank (NZ_ABKY) suggests it belongs to clade 2 (lineage I).

It must be recognized that the publicly available genome sequences are not error-free and that this could have affected the architecture of trees that are highly dependent on single nucleotide changes in sequences of target genes, such as SNP-genotyping. However, the *in silico* tree architecture was very similar to that described by Manning et al. [70] based on experimental data; therefore we suspect that this was not a significant source of error in this study.

Differences between lineage I and lineage II strains have been described [370, 371, 372], but much less is known about lineage I/II strains with respect to host / disease association or expression of virulence attributes. We have recently demonstrated in a study of *E. coli* O157:H7 strains in Canada that certain phage types (PT)s are specific to O157:H7 lineages [370]. In the latter study, PT2 strains were shown to belong exclusively to lineage I/II, however, others strains in lineage I/II belonged to PTs that were not lineage-restricted e.g. PT23, PT8 and PT1, suggesting that this lineage is widespread and also diverse. Such diversity was also apparent in the examination of the novel regions of O157:H7 DNA, presented in Figure 4.1.

The Stxs, which are bacteriophage-encoded and the primary virulence factors of *E. coli* O157:H7 [364] show differential distribution among the lineages. The stx_2 gene was found in nearly all O157:H7 lineage I and lineage I/II strains [70, 370], while the stx_1 gene was absent in lineage I/II strains but present in nearly all other strains studied. Additionally, Ziebell et al. [370] found the stx_{2c} gene in 96.7% of lineage II strains, 50.0% of lineage I/II strains and 1.8% of lineage I strains, while Manning et al. [70] found stx_{2c} to be present in 57.6% of clade 8 strains. This SNP genotyping study also found a significant relationship between the presence of stx_2 in conjunction with stx_{2c} among strains of clade 8, in that no other clade associated with human illness displayed this combination. This is not surprising as lineage I strains only rarely contain stx_{2c} despite their high association with human disease and lineage II strains nearly always contain *stx*$_{2c}$ despite a rare association with human disease [353]. It is unlikely that the combination of *stx*$_2$ and *stx*$_{2c}$ alone is the reason for the hyper-virulence of lineage I/II strains, as the presence of *stx*$_{2c}$ is nearly ubiquitous among bovine-associated lineage II strains. The SNP genotyping study cited above only considered isolates associated with human disease; therefore it is likely that few lineage II strains were included in the study. A study by Friedrich et al. [373] examining *stx*$_2$ subtypes and their association with clinical symptoms found *stx*$_{2c}$ to be the only subtype besides *stx*$_2$ present in strains isolated from cases of HUS, but found no correlation between the presence of *stx*$_{2c}$ and the development of HUS. It has recently been shown that the level of Stx2 production is greater in lineage I strains than lineage II strains [33] so it may be that lineage I/II strains implicated in cases of HUS simply produce more toxin than other O157:H7 strains. However, it is possible that other factors possessed by the hyper-virulent lineage I/II group strains are responsible for their greater virulence in humans, and remain to be discovered.

The findings in this study highlight the need for a common genotyping approach, as it is evident that the same groups of genetically related strains have been given multiple designations based on the use of different comparative genotyping methods. This need exists for epidemiological studies of outbreak strains, where strain discrimination is the primary focus, as well as for population genetic studies where genomic information is of central importance. The value of being able to compare a pattern produced in a particular laboratory to one found in a national central database has made PFGE and PulseNet very useful in tracking outbreaks that are widely disseminated [59], despite the fact that PFGE is labour intensive and difficult to standardize [46]. As this common approach in identifying outbreak strains has been useful for a pattern-based method such as PFGE, approaches based on a multi-locus sampling of the genome could take this centralized database concept and extend it to contain presence / absence data for specific loci, SNPs, and other measures of heterogeneity. In this way, whether the goal is identifying an outbreak source or using information for a population-based study, the data would be available in a central repository. This type of system would be important in monitoring and identifying the emergence of new clones of O157:H7, such as the hyper-virulent lineage I/II / clade 8 strains and recognizing other changes in the population when genotyping *E. coli* O157:H7 strains associated with disease outbreaks.

The availability of whole genome sequence information has led to the development of new genotyping methods that are easier to perform than traditional methods, more discriminatory and more informative with respect to genotype and phenotype. It is interesting that the approaches targeting genetic polymorphisms within conserved genes and those targeting genetic changes based on gene insertion / deletion events converge to give a similar picture of *E. coli* O157:H7 strain relationships. Such concordance of methods has been previously demonstrated with mCGH and MLST for *Campylobacter jejuni* [81] and *Streptococcus pneumoniae* [374]. However, given that methods differ greatly in terms of the time required for the analysis, labour and equipment required, need for expertise, freedom from subjectivity in interpretation of the data and portability of the genotyping results from one laboratory to another, there is considerable advantage in selecting a method that is simple, extensible and easily portable. While some typing methods are better than others in these aspects, most of the multi-locus typing methods examined produced a similar tree architecture. This suggests a ‘common typing language’ is possible at least in the context that genotypes derived using different methods can be integrated and communicated in a broader framework. While it was shown that multiple methods converge to provide a similar picture of the O157:H7 population structure, we are not advocating the routine use of multiple genotyping methods but rather the use of methods based on comparative genomics.

With the genomic sequencing revolution well under way, the ability to harvest novel sequence information in a timely fashion from new genomic sequences will become increasingly important and the ability to include and compare *in silico* results to those from traditional laboratory experiments will become necessary.

The results of this study suggest that genotyping approaches based on common comparative genomic data are likely to form the basis for the next-generation of analytical tools used for both population-based comparative genotyping and epidemiological studies.

Table 4.1: The 19 *E. coli* strains analyzed *in silico* in this study.

Strain	Serotype	Sequence source	LSPA6 lineage	SNP clade	Stx-phage insertion site genotype
EC4024	O157:H7	NZ_ABJT00000000	I/II	8	1
EC4042	O157:H7	NZ_ABHM00000000	I/II	8	1
EC4045	O157:H7	NZ_ABHL00000000	I/II	8	1
EC4076	O157:H7	NZ_ABHQ00000000	I/II	8	1
EC4113	O157:H7	NZ_ABHP00000000	I/II	8	1
EC4115	O157:H7	NZ_ABHN00000000	I/II	8	1
EC4196	O157:H7	NZ_ABHO00000000	I/II	8	1
EC4206	O157:H7	NZ_ABHK00000000	I/II	8	1
EC4401	O157:H7	NZ_ABHR00000000	I/II	8	1
EC4486	O157:H7	NZ_ABHS00000000	I/II	8	1
EC508	O157:H7	NZ_ABHW00000000	I/II	8	1
EC71074	O157:H7	Public Health Agency of Canada	I/II	8	1
EC869	O157:H7	NZ_ABHU00000000	II	uncertain	6
EC4501	O157:H7	NZ_ABHT00000000	I	2	3
TW14588	O157:H7	NZ_ABKY00000000	I	2	3
EDL933	O157:H7	NC_002655.2	I	3	3
Sakai	O157:H7	NC_002695.1	I	1	3
EC33264	O145:NM	Public Health Agency of Canada	NA	NA	NA
MG1655	K12	NC_000913.2	NA	NA	NA

Table 4.2: The 102 *E. coli* O157:H7 strains used in the construction of the supernetwork in Figure 4.3

Strain	Lineage	Strain	Lineage	Strain	Lineage
AA10002	I	H2727	I	EC4113	I/II
AA10021	I	H2731	I	EC4115	I/II
APF593	I	H432	I	EC4196	I/II
2328	I	H435	I	EC4206	I/II
23339	I	H4420	I	EC4401	I/II
58212	I	H451	I	EC4486	I/II
63154	I	H453	I	EC508	I/II
70490	I	H454	I	EC71074	I/II
813601	I	H568	I	R1388	I/II
93111	I	H571	I	Zap0046	I/II
97701	I	H572	I	AA6192	II
EC4501	I	H573	I	AA9952	II
EC980120	I	H574	I	E12491	II
EC980121	I	LN6374	I	EC19920026	II
EC980122	I	LRH6	I	EC869	II
EC980125	I	LRH73	I	EC970520	II
EDL933	I	LS110	I	F1081	II
F1082	I	LS236	I	F12	II
F1095	I	M01MD3265	I	F1305	II
F1103	I	OK1	I	FRIK1985	II
F1299	I	R1195	I	FRIK1990	II
F2	I	S23021	I	FRIK1999	II
F30	I	S2628	I	FRIK2001	II
F5	I	S3722	I	FRIK920	II
F732	I	Sakai	I	LRH13	II
F744	I	TS97	I	LS68	II
H2160	I	TW14588	I	R1797	II
H2161	I	09601Fe046.1	I/II	493/89	O157:H-
H2163	I	32511	I/II	CB2755	O157:H-
H2164	I	59243	I/II	Dec5d	O55:H7
H2176	I	EC4024	I/II	TB182A	O55:H7
H2704	I	EC4042	I/II	5905	O55:H7
H2718	I	EC4045	I/II	EC33264	O145:NM
H2723	I	EC4076	I/II	MG1655	K12

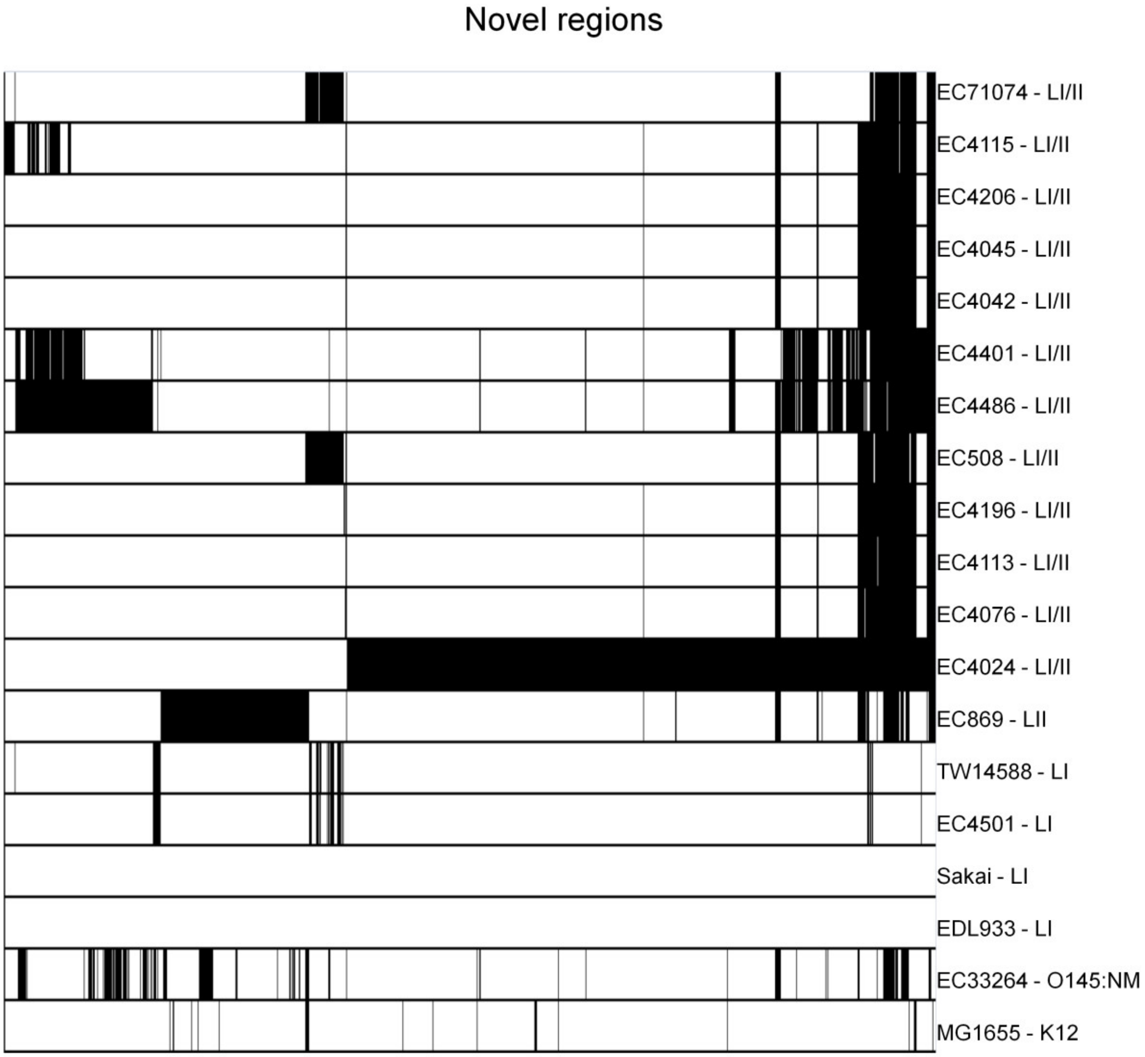


Figure 4.1: The distribution of 1456 regions approximately 500 bp in size among 17 *E. coli* O157:H7 strains and the O145:NM strain EC33264 and K12 strain MG1655. Regions are not necessarily contiguous and are defined as novel based on less than 80% sequence identity to the genome of either EDL933 or Sakai. Black indicates the presence of a region and white indicates the absence of a region.

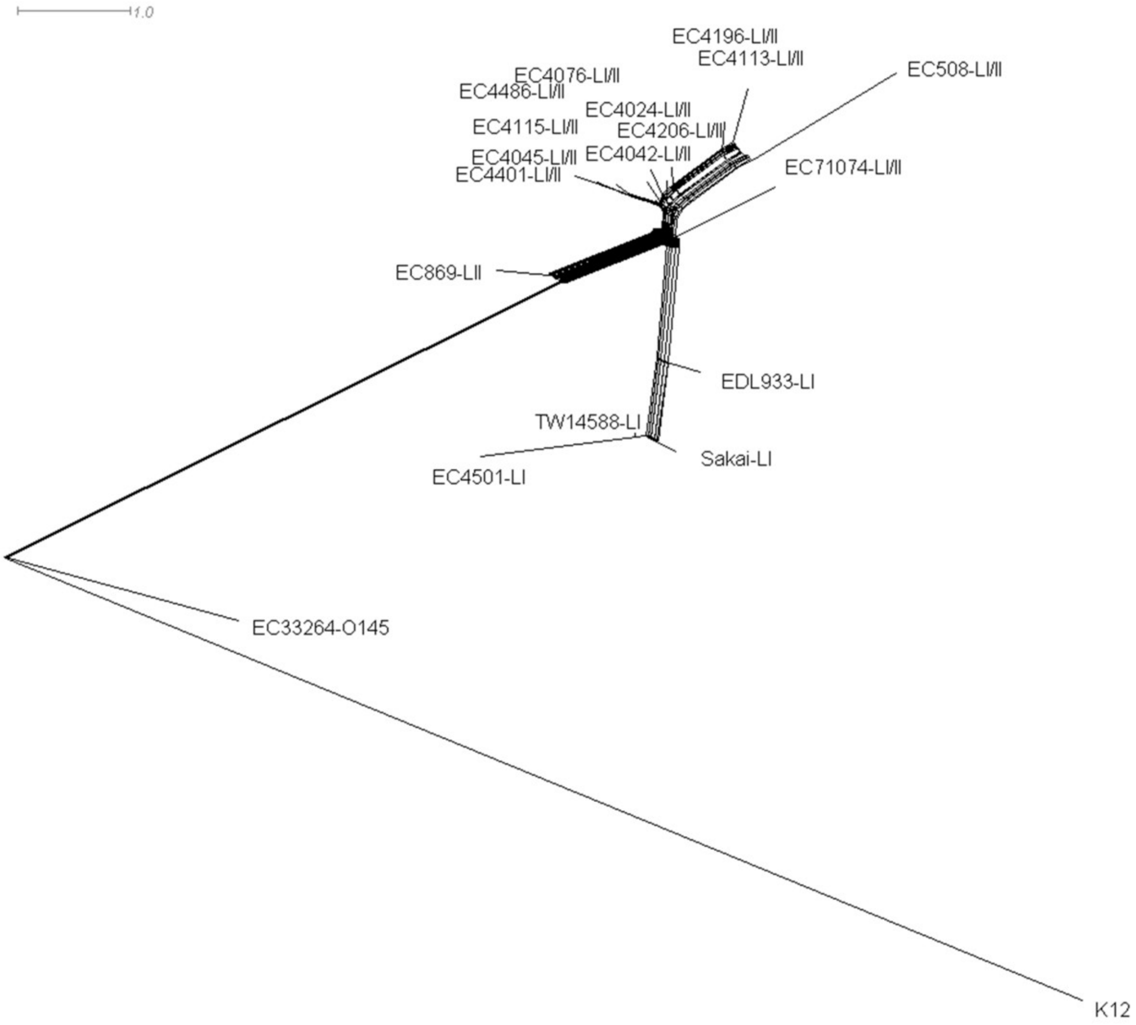


Figure 4.2: The supernetwork created from the combination of each maximum parsimony tree from the following typing methods: Stx-phage insertion site typing, MLVA, CGF, SNP genotyping, mCGH, and GISSH-based novel region distribution typing. Maximum parsimony trees were combined using the unweighted mean distance and Z-closure with 1000 iterations; the resulting supernetwork was displayed using the equal angle method.

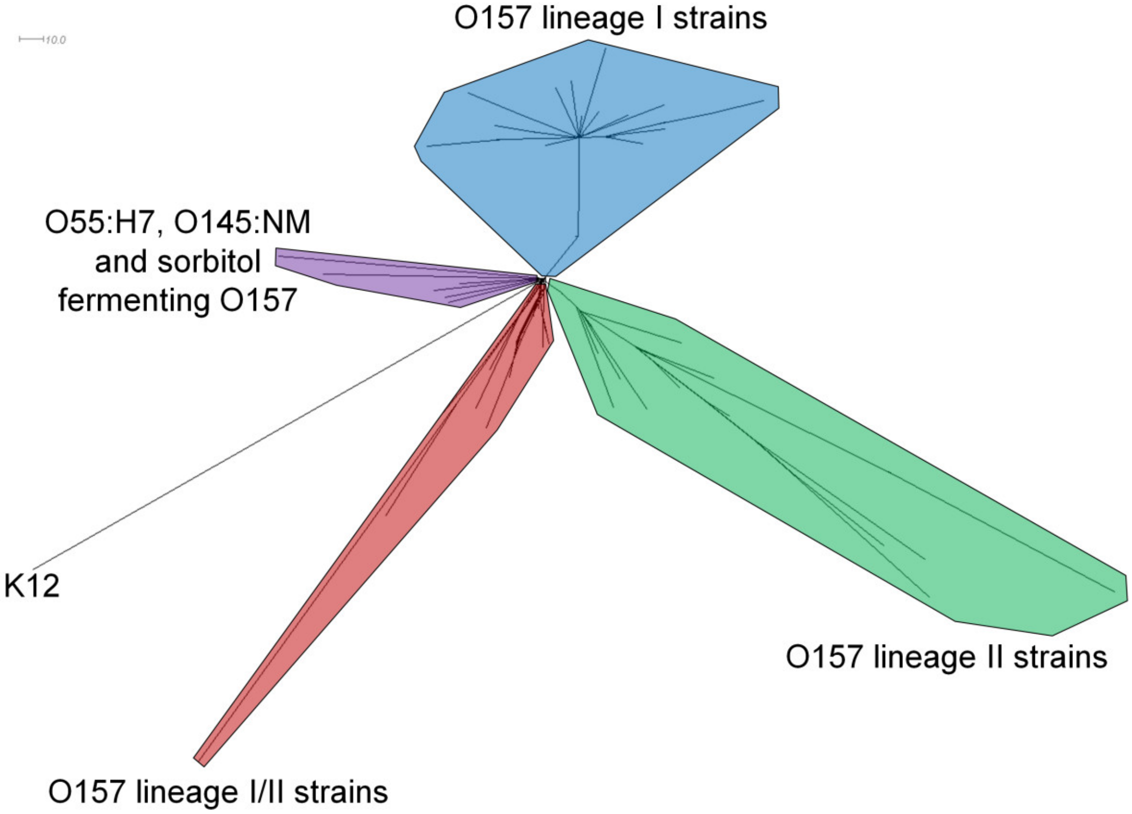


Figure 4.3: The supernetwork created from the combination of each maximum parsimony tree from the following typing methods: Stx-phage insertion site typing, MLVA, CGF, SNP genotyping, mCGH, and GISSH-based novel region distribution typing. Both mCGH and CGF datasets included the *in silico* data from the strains in Table 4.1 in addition to experimental data from the remaining strains in Table 4.2. Maximum parsimony trees were combined using the unweighted mean distance and Z-closure with 1000 iterations; the resulting supernetwork was displayed using the equal angle method.

Chapter 5

Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions

5.1 Preface

This chapter contains the previously published paper Laing C, Buchanan C, Taboada EN, Zhang Y, Kropinski A, Villegas A, Thomas JE, Gannon VP. ‘Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions.’ BMC Bioinformatics. 2010 Sep 15;11:461. doi: 10.1186/1471-2105-11-461. [127] and Laing C, Villegas A, Taboada EN, Kropinski A, Thomas JE, Gannon VP. ‘Identification of *Salmonella enterica* species- and subgroup-specific genomic regions using Panseq 2.0’ Infect Genet Evol. 2011 Dec;11(8):2151-61. doi: 10.1016/j.meegid.2011.09.021. Epub 2011 Oct 1. [375].

5.2 Introduction

The field of genomics has blossomed as a result of the fast rate of whole-genome sequence data acquisition. The pace of genome data growth continues to increase as the cost to acquire the data continues to decrease. This has been led in large part by massively parallel sequencing platforms such as the 454 Genome Sequencer FLX (Roche Applied Science), the Illumina (Solexa) Genome Analyzer and the ABI SOLiD System (Applied Biosystems), which generate tens of millions of base pairs of information in short reads 30 to several hundred base pairs in length [376, 377]. These reads must be combined into large contiguous DNA sequences by dedicated software such as Newbler (Roche) and MAQ [377].

Although these ‘contigs’ can stretch into the megabase-pair (mb) range, the sequencing of an entire organism by any one of these techniques invariably leaves gaps in the reassembled sequence [378]. The finishing of a sequence requires gap-closure by sequencing of PCR products and the resolution of sequencing errors. Sequencing efforts are primarily driven by the discovery of novel genes and, as gap closure is time-consuming and expensive, many researchers now use unfinished draft sequences of genomes in their analyses [379].

Tettelin et al. [15] used the term ‘pan-genome’ to refer to the full complement of genes within a bacterial species. Comprising the pan-genome are the core complement of genes common to all members of a species and a dispensable or accessory genome that is present in at least one but not all members of a species. As more whole-genome sequences of a species or group within a species become available, the size of the pan-genome of that species or group will usually increase, due to an increase in the number of accessory genes. Based on mathematical models, it is predicted that new genes will be discovered within the pan-genome of many free-living bacterial species even after hundreds or possibly thousands of complete genome sequences have been characterized [6]. While originally applied to an entire species, any group of related strains can be said to contain a ‘core’ and ‘accessory’ set of genes. As such, tools that extract the new pieces of information from an extremely large pool of data and that can be used to determine the pan-genome and its distribution among strains will be invaluable in the study of genotypic and phenotypic traits in bacterial populations.

Regardless of whether an investigator uses a draft or finished sequence, software tools able to efficiently extract relevant information are critical. A number of programs have been designed to assist with the analysis of DNA sequence data. These include programs designed for multiple sequence alignments such as CLUSTAL W [109], T-COFFEE [108] and MUSCLE [107]; programs for local sequence comparison including FASTA [380] and BLAST [111]; and programs designed for whole-genome comparisons such as

MUMmer [122], MAUVE [121] and MISHIMA [124]. Pre-computed alignments of completed genomes are available online from Web-ACT [381] (which also allows up to five user-submitted comparisons), the map-viewer and gMap database features from the National Center for Biotechnology Information [382], the MOSAIC web-server [383] and the prokaryotic gene-order database PSAT [384].

Programs designed specifically to find mobile genomic islands as opposed to regions of differing sequence between or among genomes using comparative genomic approaches include MobilomeFINDER [385] and IslandPick [386]. Three programs offering *in silico* subtractive hybridization among genomes using BLAST are available: FindTarget [118], mGenomeSubtractor [120] and nWayComp [119]. See (supplementary material at: `http://www.biomedcentral.com/content/supplementary/1471-2105-11-461-s1.xls`) for a comparison of features among many of these programs.

Despite the myriad program options available, there is no comprehensive package for pan-genome analysis. Prior to the advent of the current generation of sequencing platforms, the number of genome sequences available for intra-species comparative genomic analysis and for the determination of 'accessory' genes was limiting; consequently, tools specifically designed for the analysis of the pan-genome have not been available.

Most studies have used one or two reference genomes, which include the core elements but are missing much of the accessory genome of the species in study. In the laboratory, selective subtractive hybridization [371, 387] and population surveys using microarray comparative genomics [354, 388] have been used to examine / define accessory genome content. With the current explosion in availability of whole-genome draft sequences, it would be highly desirable to exploit this information through *in silico* analysis to separate the novel accessory component of the genome from previously identified core sequence without the requirement for a finished assembly. Identifying novel accessory sequences in this way has application in characterizing novel metabolic pathways, virulence attributes [389], and molecular fingerprinting targets useful in epidemiological and population genetic studies [66]. Finally, both the core and accessory genomes can be helpful in elucidating the evolutionary history of organisms [390].

Until recently, studying the pan-genome of a taxonomic-group was limited by genome sequence availability; however, thanks to the low costs associated with high-throughput genomic sequencing, this technological limitation no longer exists. This has opened the door to studies on the evolutionary history of many taxonomic groups. Whole-genome evolutionary analysis of the genus *Listeria* has demonstrated that it evolved from a pathogenic ancestor into multiple non-pathogenic clades, and that overall the evolution of the genus was characterized by a loss of virulence traits [93]. This is in contrast to groups like *Escherichia coli* O157:H7, which as demonstrated by Leopold et al. seem to have evolved via the acquisition of virulence traits [91, 92].

New epidemiological insights have been made possible due to the availability of complete genome sequence information. For example, the world-wide dissemination of *Yersinia pestis* from China to the rest of the world occurred in a series of distinct radiations, and each geographical location was found to contain a population of strains distinguished by radiation-specific SNPs [140]. In a study combining both whole-genome comparisons and social network analysis, the transmission routes of a tuberculosis outbreak in a Canadian community were identified through SNPs in the genomes of strains from infected individuals and specific social interactions and events such as increased crack cocaine usage in the community [87]. These long- and short-term epidemiological studies would not have been possible without low-cost, high-throughput sequencing technology.

In this study we describe a pan-genome sequence analysis program, Panseq, that extracts novel regions with respect to a sequence or group of sequences, determines the core and accessory regions of sequences based on sequence identity and segmentation length parameters, creates files based on the core and accessory genome for use in phylogeny programs and determines the most discriminatory and variable set of loci from a dataset. The software allows for the core genome to be user-defined, where an investigator might wish to look for group-dominant (<100%), rather than group-specific (100%) genetic loci in collections of strains (present in many, rather than all of the strains). This feature helps deal with strains that may have lost 'core' elements and are not phenotypically representative of the group, strains that may have lost genetic elements in the laboratory that are maintained by selective pressure in nature, and draft genome assemblies with incomplete sequence coverage. More importantly, it allows new sub-groups to be identified which share a subgroup-specific core that is distinct from the core possessed by the entire group of strains. These subgroups are not only likely to share a common ancestry, but to also be phenotypically divergent from other strains in the group.

5.3 Implementation and Methods

The pan-genome sequence analysis program (Panseq) was written in Perl with BioPerl [391] modules and is available at `http://lfz.corefacility.ca/panseq/`. As a web-server it is platform independent and makes use of the NCBI GenBank database for pre-existing nucleotide FASTA files. The core genome threshold option has the ability to output the accessory genome table as percent sequence identity or nucleotide sequence. The software incorporates the use of multiple-cores, allowing parallel processing and the reduction of computation time. A standalone version configurable by the simple editing of a single file has been made available, allowing the full functionality of the web-server version on a standalone machine. Extension of the scripts are possible by those with knowledge of Perl, provided proper attribution is cited. The Perl scripts and all future updates are available from: `https://github.com/chadlaing/Panseq`

5.3.1 Novel Region Finder (NRF)

The NRF module compares an input sequence(s) to a database of sequence(s), and contiguous regions not present in the database but present in a user-defined combination of input sequences are extracted. This is accomplished using the MUMmer alignment program [122] with the 'novel' sequences extracted into a file with sequence location information added to the FASTA header. This process is iterative, adding the initial sequence examined to the database before examining the second sequence, and so on until all sequences have been examined. Because the algorithm examines all matches given user-specified criteria, regions of high sequence identity will also be matched independent of their order (non-syntenic). A summary file indicating the size distribution of fragments and total number of novel nucleotides is created, as well as a graphical representation of the novel regions distributed among all input sequences in scalable vector graphic (SVG) form. Default Nucmer parameters used by the NRF module are: b:200, c:50, d:0.12, g:100 and l:20; where according to the MUMmer manual (`http://mummer.sourceforge.net/manual/\#nucmer`), b: the distance an alignment will extend poor scoring regions; c: the minimum cluster length; d: the maximum diagonal difference (diagonal difference/match separation); g: the maximum gap between two adjacent matches in a cluster; l: the minimum length of an exact match.

5.3.2 Core and Accessory Genome Finder (CAGF)

The CAGF module considers for the purposes of analyses the 'pan-genome' to be comprised of sequences selected as input to the program. Panseq initiates using a single sequence file as a seed to which all other sequences are compared using MUMmer. If a segment greater than the 'Minimum Sequence Size' is found in a sequence other than the seed, that segment is added to the 'pan-genome'. This newly-added-to 'pan-genome' is used as the reference for subsequent comparisons and the process continues iteratively until all sequences have been examined. Panseq next fragments the entire pan-genome into segments of user-defined length and determines the presence or absence of each of these fragments in each of the original sequences based on the percent sequence identity cutoff using the BLASTn algorithm [111], with the following default parameters: blastall -p blastn -W 11, -b (2*number of input sequences) -v (2*number of input sequences) -e 0.001, -F F. Fragments above the cutoff found in every original sequence are considered part of the 'core' genome, while fragments below the cutoff in at least one strain are considered part of the 'accessory' genome.

The core genome for each input sequence is concatenated into a single sequence and a multiple sequence alignment is produced. The accessory genome is reported in a tab-delimited table, where binary (0 for absence, 1 for presence) data indicate the state of each fragment in the original sequences. A NEXUS formatted file [358] for both the accessory and core genomes are output for use in downstream phylogenetic applications. Panseq also produces a SNP file containing core segments with sequence variability; a tabular file listing each SNP, its position, and value among each original sequence; 'core' and 'accessory' genomes output to separate FASTA files; and a scalable vector graphic depicting the pan-genome and its presence/absence among all of the original input sequences.

5.3.3 Loci Selector (LS)

The LS module constructs loci sets that are maximized with respect to the unique number of fingerprints produced among the input sequences as well as the discriminatory power of the loci among the input sequences. The LS module iteratively builds the final loci set, in the following steps, given a tab-delimited table with loci names in the first column, sequence names in the first row, and single character data filling the matrix:

(1) Each potential available locus is evaluated for the number of unique fingerprints that would result from its addition to the final loci set. All loci that would generate the maximum number of unique fingerprints in this respect are evaluated in step (2).

(2) All loci from step (1) are evaluated for their discriminatory power among the sequences, which is given as points of discrimination (POD). The POD for a locus is calculated as follows.

A listing of all possible pair-wise comparisons is constructed; for example, if the input table consisted of three sequences, A, B and C, the list would consist of A-B, A-C and B-C. Next, it is determined whether or not the sequences in each pair-wise comparison contain the same single character denoting the locus state. If they do, a value of 0 is assigned; if they differ a value of 1 is assigned. The POD is then the summation of all pair-wise comparisons that differ for that locus. With our previous example, if A-B = 1, A-C = 1 and B-C = 0, the POD for that locus would be 2.

(3) The locus with the highest value from step (2) is selected for addition to the final loci set and removed from the pool of candidate loci. If two or more loci tie in value, one is randomly selected. If all possible unique fingerprints have been found, the algorithm continues with (4); if additional unique fingerprints are possible, the algorithm continues with (5).

(4) Sequence pairs for which the allele of the locus chosen in (3) differ are excluded from the analysis. This ensures loci that differ between other pairs of strains are preferentially considered. Consider our A, B and C example with pair-wise comparisons of A-B = 1, A-C = 1 and B-C = 0. In the case of this locus being chosen, the sequence pairs A-B and A-C would be temporarily removed from the analysis (‘masked’), leaving only loci that differed between B-C as viable options.

(5) Once a locus has been chosen:

a) the specified number of loci has been reached (all unique fingerprints in the case of ‘best’) and the algorithm terminates; or

b) the specified number of loci has not been reached and there are remaining fingerprints possible, or sequence pairs for which differences exist. The algorithm returns to (1); or

c) there are no remaining fingerprints possible and no sequence pairs for which differences exist. At such time, all sequence pairs are again considered part of the analysis (‘unmasked’). If no differences among any sequence pairs exist at this point, the algorithm terminates; if differences remain, the algorithm returns to (1).

5.3.4 Construction of Salmonella enterica Phylogenies

Aligned core genomes were generated using the Panseq Core / Accessory Genome Analysis module. All aligned sequences were analyzed under maximum likelihood in PhyML with the approximate likelihood ratio test for branch support using the following command: PhyML_3.0_linux64 –p –m HKY85 [392]; under maximum parsimony in TNT with 100 bootstrap replicates using the following command: ‘hold 10000, taxname=, resample boot replications 100 savetrees’ [393]; under the neighbour-joining algorithm in Phylip v 3.69 [2] with 100 bootstrap replications by using the seqboot program with default settings to generate 100 bootstrap replicates, then by using the dnadist program with default settings to generate a distance matrix for each bootstrap replicate, and finally by using the neighbour program to generate a neighbour-joining tree for each replicate. The consense program of Phylip was used to generate a majority-rule tree for each dataset. Phylogenies from each method were visualized using Dendroscope [394].

5.3.5 In Silico MLST of Salmonella enterica genomes

The *in silico* multi-locus sequence typing was carried out utilizing the seven genes described by Kidgell et al. [3]. The reference genes were obtained from the serovar *S.* Typhimurium strain LT2 (NC_003197) and used to BLASTn query all 39 genomes, with an e-value of 0.1 and a word size of 7 [111]. Alignments based on the BLAST results were used to create a seven-gene concatenome which was analyzed and visualized in an identical manner to the core genomes.

5.4 Results and Discussion

We have used a number of examples to highlight the functionality of Panseq, many of which could be carried forward as complete studies of their own; however, our intention is to demonstrate that Panseq is capable of finding and extracting useful data from sequences, which can be used as the basis for hypothesis generation and future investigations.

5.4.1 Novel Region Finder (NRF) Module

Alignment programs are capable of finding regions of similarity between sequences, and regions of uniqueness can be inferred from the gaps between areas with high sequence similarity. However, a number of steps are required using alignment programs to identify genomic regions that are unique with respect to other sequence(s). These steps include the location of the sequence coordinates for each sequence of interest and the subsequent location of the corresponding sequence in a sequence editor. Panseq automates this process, creating FASTA files of all unique regions of a given sequence or sequences as well as presenting a graphical overview of the locations of the novel regions, based on the results of sequence comparisons made using the MUMmer algorithm. MUMmer was chosen as the sequence alignment engine because of its use of suffix-trees, which allow it to perform operations up to 100 times faster than similar alignment programs in whole-genome comparisons [122].

Determining putative functions for the regions identified by the NRF module can be accomplished by comparing the translated nucleotide sequences to known protein sequences using the Panseq links to NCBI BLASTx [111] or the UniProt database [395]. This linking allows the genomic information to be easily queried and is a logical first step in connecting genotype to phenotype in comparative genomics analyses.

5.4.1.1 The NRF Module in Genomic Island Identification

Novel regions have the potential to affect virulence and niche specificity of pathogenic microorganisms. In 2000, Perna et al. [13] published the complete genomic sequence of the pathogenic *E. coli* O157:H7 strain EDL933 and compared it to the previously sequenced, non-pathogenic laboratory *E. coli* K12 strain MG1655. Genomic regions found in *E. coli* O157:H7 EDL933 but not in K12 were called ‘O-islands’ and genomic regions present in K12 but not EDL933 were termed ‘K-islands’. Perna et al. found 177 O-islands greater than 50 bp in *E. coli* O157:H7 EDL933, constituting 1.34 Mb and 234 K-islands greater than 50 bp in the K12 strain MG1655, representing 0.53 Mb. They determined the presence of these islands using a custom modification of the MUMmer program [122] and found that many of the genomic regions in strain EDL933 were bacteriophage-related and suggested that they may play a role in the virulence of the organism.

Using Panseq, we re-analyzed the genome sequence data used in these experiments with the NRF module, checking ‘Unique among the sequences selected’ to generate both sets of genomic islands in a single step. Perna et al. [13] defined islands as regions with less than 90% sequence identity over 90% of the sequence length. We found that genomic islands identified in this previous analysis can extend into conserved regions, combining multiple islands into a single contiguous sequence that is interspersed with regions of high sequence identity. Similarly, genomic islands may be misclassified as core regions simply because of heterogeneous composition with interspersed core and accessory sequences in regions of high heterogeneity. Panseq uses parameters optimized for high-resolution comparison of genomic sequences and we found that it more stringently defined islands unique to a genome sequence.

With default program settings (Nucmer values of b:200, c:50, d:0.12, g:100, l:20) and a minimum novel region size of 50 bp, Panseq identified 214 K- and 304 O-islands, compared with the 234 K- and 177 O-islands of Perna et al. [13]; a detailed comparison of the findings can be found (supplementary material at: `http://www.biomedcentral.com/content/supplementary/1471-2105-11-461-s2.xls`). As a result of differences in the stringency between the two methods, present but heterogeneous islands such as K-Island #4 were not identified as islands by Panseq; K-island #4 matches at 90% sequence identity for 84 bp at positions 111429 - 11511 in K-12 and 116035 - 116118 in EDL933. This represents an example of modest differences in core sequence rather than of 'novel' regions that comprise the accessory genome (see also below the discussion of identity thresholds). Additionally, islands identified by Perna et al. with interspersed regions of high sequence identity were split and refined into separate islands by Panseq; for example O-island #7 was split into four unique islands, eliminating the regions of high sequence identity. These results demonstrate that Panseq can correctly identify and rapidly extract novel genomic sequences.

Two-way comparisons similar to those used for *E. coli* strains K-12 and EDL933 have been used to identify genetic attributes which confer distinct phenotypes on many other taxonomically related strain pairs. This type of analysis has allowed researchers to identify elements unique to each strain and suggest putative functions for these elements in the life cycles of the respective organisms, their virulence and their ability to survive in different niches. Examples of such studies include the identification of genetic differences between the human-restricted *Salmonella enterica* subspecies enterica serovar Typhi, and *Salmonella enterica* subspecies enterica serovar Typhimurium, which is a murine pathogen but is not host-restricted [396]; and the differences between the host-restricted causative agent of plague, *Yersinia pestis* and the enteric pathogen *Yersinia pseudotuberculosis* [397]. While these types of two strain comparison studies have been extremely insightful, studies using programs such as Panseq will greatly facilitate these comparisons and also allow comparisons between multi-strain groups.

5.4.1.2 The NRF Module in Multiple Sequence Comparisons

To demonstrate a comparison of recently sequenced genomes to previously completed 'reference' genomic sequences, we used the two draft genome sequences of *Listeria monocytogenes* F6900 and 10403S and compared them to the four complete *L. monocytogenes* genomes in GenBank: Clip81459, EGD-e, HCC23 and 4bF2365 (Table 5.1). The NRF module found 45 novel regions 500 bp, constituting 126858 bp of genomic DNA not present in the four reference *L. monocytogenes* genomes. The size distribution of the novel regions from the output file is presented in Figure 5.1. In addition to the summary file, the novel regions are output in FASTA format with sequence location and size information found in the header. These files are suitable for additional bioinformatic or phylogenetic analysis. In addition to the FASTA file, an SVG graphics file showing the location of the novel regions in each strain is optionally provided. As can be seen in Figure 5.1, the greatest number of novel regions was found in the 1000 -2000 bp range, with 38 of the 45 regions less than 4000 bp in size.

5.4.1.3 Experimental Confirmation of Outputs from the NRF Module

We wished to experimentally confirm the uniqueness of the novel regions extracted by Panseq. To do this, we determined genomic regions present in at least one of 14 *E. coli* O157:H7 whole genome sequences (Table 5.2) but absent from the two O157:H7 reference genomes EDL933 and Sakai. We designed primers targeting 65 of these novel regions and examined the distribution of the regions among a population of 60 *E. coli* O157:H7 strains. All of the primer pairs generated amplicons in at least one but not all of the *E. coli* O157:H7 strains other than EDL933 and Sakai, and all failed to amplify DNA from *E. coli* O157 EDL933 and Sakai, as predicted by Panseq (data not shown). This experimentally demonstrated that the sequences identified by Panseq represent novel accessory regions not found in the two reference *E. coli* O157:H7 genomes.

5.4.2 Core/Accessory Genome Finder (CAGF) Module

The Panseq CAGF module uses a combination of fragmentation and sequence identity thresholding to define accessory and core genomic regions. To determine the effect of the sequence identity cutoff on core/accessory genome size, we examined groups of *L. monocytogenes*, *E. coli* O157:H7, *Clostridium difficile* and *C. jejuni* genomes with a fragmentation size of 500 bp (half the size of an average gene) over a range of sequence identity cutoffs (Figure 5.2). The size of the estimated accessory genome increased as the sequence identity threshold was raised, and the size of the estimated core genome decreased proportionally; however, this increase was observed to have two distinct phases: an initial linear growth in accessory genome size that was followed by an exponential increase in accessory genome size. The transition between the two stages occurred in the 80 - 90% sequence identity cutoff range for each species, suggesting that with values below 80% Panseq primarily identified accessory genome segments that were variably absent or present whereas above this threshold Panseq identified core genome segments with sequence heterogeneity.

5.4.2.1 The CAGF Module in SNP Analysis of the Core Genome

Accessing the core genome is important for phylogenetic studies, which have used 'core' gene concatenates ranging in size from a few genes in multi-locus sequence typing (MLST) schemes for *Campylobacter jejuni* [398] and *L. monocytogenes* [1] to all known 'core' genes in *E. coli* O157:H7 [91]. In addition to offering the best available data for assessing the phylogenetic reconstruction of the evolutionary history of an organism, a small number of single nucleotide polymorphisms (SNPs) can be useful in defining clusters of epidemiologically related strains [399, 400, 360].

We analyzed the six *L. monocytogenes* strains in Table 5.1 with the CAGF module of Panseq, using a fragmentation size of 500 bp and a sequence identity cutoff of 85%. The resulting concatenated core data, which include all conserved nucleotides and SNPs, were used to construct a maximum parsimony (MP) tree using Phylip v3.69 [2]. A MP tree was also generated *in silico* for the same six strains using the *L. monocytogenes* MLST protocol outlined by Nightingale et al. [1], for comparative purposes (Figure 5.3). The symmetric distance between the two trees shows that the overall topology of the trees differ between that created from MLST data and that based on the entire concatenated core. This is likely due to the fact that the MLST protocol only considers seven genes, where disproportionate variation among these few loci and the relative paucity of loci compared to that of the entire core genome can reduce the ability of this method to capture the overall relationships among strains. With the continued increase in sequencing throughput, settling for rough approximations of true tree topologies may no longer be necessary.

5.4.2.2 The CAGF Module in the Analysis of the Accessory Genome

While SNP analysis has proven to be an extremely useful tool, the ideal reconstruction of an evolutionary history would take into account not only the heterogeneity among all core genomic regions, but also the presence or absence of regions in the accessory gene pool, especially since the accessory genes can directly affect phenotype (e.g. niche specificity, antimicrobial resistance, virulence, etc.).

Traditional methods of estimating bacterial phylogenies rely on SNPs within the core genome, but it has been shown for *C. jejuni* [81], *E. coli* O157:H7 [289] and *Streptococcus pneumoniae* [374] that the distribution of accessory genes provides a very similar overall tree topology to methods based on variability in the core genome. Although variation in the core genome among a small number of loci may be sufficient for identifying clusters of related strains, discrimination among strains is often more difficult because there is less variation and fewer phylogenetically informative loci, leading to fewer genotypes. Accessory genome content, which can be highly variable among strains, appears in many cases to be consistent with phylogenetic analyses of core genes and as a result of greater variability provides a higher degree of discrimination among strains. Analysis of accessory gene content thus presents an opportunity to integrate valuable links to phenotype into a genetic classification scheme.

To examine the performance of accessory genome information for phylogenetic reconstruction, we used the binary presence/absence data of the accessory genome computed for the set of six *L. monocytogenes* strains above, and constructed a MP tree (Figure 5.3). When comparing the tree topology of this accessory-based tree to that of the SNP-based core tree and the MLST tree, we found that the core- and accessory-based MP trees had an identical topology, but differed from the MLST-based tree in the placement of the two strains 10403S and F6900. Further studies will be required to determine the extent of the phylogenetic concordance between the core genome and the accessory genome among other bacterial groups.

5.4.2.3 The CAGF Module in the Examination of Pan-Genomic Differences

As well as producing a concatenated core, the CAGF module also produces a table listing each SNP and its location and allele within each original sequence. This can be useful in examining individual differences or for hierarchical clustering of data. To demonstrate how the results of Panseq can be visualized, both tabular output files from the *Listeria* core / accessory analysis were used to create hierarchical clustering dendrograms: one based on the SNP character data from the core regions (Figure 5.4) and the other based on the binary presence / absence data of the accessory regions (Figure 5.5). Both dendrograms have the same tree topology, and the underlying data for both core and accessory regions shows only the differences among the genomic sequences, making comparisons between strains clearer than they might be from whole-genome comparisons where conserved as well as variable loci are considered in the analysis.

5.4.2.4 Selectable Core Stringency

Thirty-nine genomes of *Salmonella enterica*, 16 closed, 23 draft, were used to demonstrate the selectable core stringency of Panseq (Table 5.3).

Th phylogenetic analysis of all strains under study can reveal sub-groups that share a common evolutionary history. Such sub-groups often share similar genomic content that may be ‘core’ for the sub-group but not for all members of the species. It has previously been shown that sub-group specific loci often confer an advantageous phenotype to the sub-group, such as the ability to colonize and persist in a particular niche, the ability to metabolize novel molecules, or the propensity to contain specific virulence genes [401, 402]. Even if a phenotype is shared by a sub-group of strains, this does not mean that every strain in the group has the same genetic content; the accessory genome has been shown to vary even within an apparently phenotypically homogeneous bacterial group [403]. Additionally, the genes responsible for the predominant phenotypic characteristics within a sub-group of bacteria may not be present in all members of the group due to factors including inherent instability of the locus, redundancy of function, lack of ongoing selection and / or divergence in phenotype within the sub-group. Draft genomes may also not be completely closed, and loci that are shared by all members of a group may be missing.

Panseq therefore allows the user to adjust the number of genomes in which a locus must be present in for it to be designated ‘core’. This allows both group-dominant and group-specific genomic elements to be identified and is designated the ‘core genome threshold’. In contrast, if a locus is not present in the user-selected number of genomes, it will be considered part of the accessory genome. Using this feature of the CAGF module in Panseq, we determined how the size of the core genome of 39 *S. enterica* genomes was affected by varying the core genome threshold from 1 to 39 genomes. We also examined the effect on core genome size of varying the percent sequence identity cutoff from 40 100 % at each core genome threshold (Figure 5.6). As can be seen, the size of the pan-genome for these 39 strains was 9.03 mbp, with about 2.5 mbp belonging to strain-specific loci. Another approximately 1 mbp of core genome was specific to 6 or fewer of the strains examined. While the Typhi and Paratyphi genomes have been well described (Baker and Dougan, 2007; Holt et al., 2009), as have comparisons between *S. enterica* strains ([404, 405, 406, 407]), to our knowledge the complete pan-genome of this many diverse members of *S. enterica* has not been previously calculated.

Following the initial steep decline in the size of the core genome with an increase in the number of strains, the core genome size decreased approximately linearly with the increase in core genome threshold, such that from core genome thresholds of 6–34, the size of the core genome only dropped from $>$ 5 mbp to $>$ 4 mbp. At thresholds greater than 34 genomes, a marked decrease in core genome size was evident. This suggests that relatively few loci are core to all *S. enterica* strains; however, many loci are found in a large proportion of the strains. For example, the difference between core genome size using a threshold of all strains and a threshold of all but one strain at the same sequence identity cutoff is over 1 mbp in size. Such a large drop in core genome amount suggests that the core gene set present in all *S. eneterica* strains is relatively small, or alternatively some of these genomic areas were not present in the draft genome sequences. Based on our examination of the percentage of the total pan-genome present in each strain (Table 5.3), incomplete sequence coverage does not appear to be a problem, as no single strain stands out as having a disproportionately small genome size. The inclusion of *S. enterica* subsp. arizonae strain RSK2980 caused a large increase in the size of the pan-genome, with many genomic regions being specific to this subspecies or strain. This was the only *Salmonella* strain not of subspecies enterica to be analyzed and it also contained the largest percent of the entire *S. enterica* pan-genome [408]. These results indicate that significant amounts of genomic diversity exist between these subspecies of *S. enterica*. This genetic isolation has previously been described, where it was shown that recombination among *S. enterica* subsp. enterica serovars was common, but recombination between subspecies enterica and other subspecies was rare [409]. Additionally, a study using a resequencing array comprising approximately 10% of the *S. enterica* pan-genome showed five genetic lineages, and recombination was found predominantly among members of the same lineage, possibly indicating that these lineages are already incipient species that have diverged significantly [410].

We used Panseq to determine the effects on the inferred phylogeny that changing the core-genome threshold would have. We generated SNP-containing aligned concatenomes using the CAGF module of Panseq at a threshold of 39 genomes (the core genome) and at one genome (equivalent to the entire pan-genome). We also created aligned concatenomes based on the multi-locus sequence typing (MLST) scheme used to type *S. enterica* strains [3]. Each of these three datasets was used to construct phylogenies using maximum parsimony, maximum likelihood and the neighbour-joining algorithm with branch support values.

The dendrograms show that the phylogeny created based on the pan-genome (Figure 5.7-5.9), and the phylogeny created based only on the core genome of all 39 *S. enterica* strains (Figure 5.10-5.12) were largely concordant. In each dendrogram, all Typhi strains formed a monophyletic group, as did the Paratyphi A strains, which together formed a terminal branch separating the human host-restricted strains from all others. The bootstrap values for the pan-genome trees were significantly higher than the bootstrap values of the core genome trees. Strains of the same serovar tended to group together in both the core and pan-genome trees, and the overall tree architecture was very similar. Four strains found on branches bearing the lowest support values in the core-genome trees were found on topologically different branches with high branch support on the pan-genome trees: Hadar_RI_05P06601, Virchow_SL49102, Saintpaul_SARA2901 and Weltevreden_HI_N05_53701.

Although serovars tended to be clustered together, the overall tree architecture was much different in the MLST trees than the core- or pan-genome trees (Figure 5.13); most notably the human host-restricted strain groups Typhi and Paratyphi A no longer formed a distinct branch in all of the MLST trees. The large discrepancy in MLST trees compared with either the core or pan-genome trees is likely due to gene recombination within one or more of the seven loci. If recombination occurs with a phylogenetically unrelated strain, the resulting MLST tree can be greatly affected. In the core or pan-genome trees, many more loci are considered, and the erroneous phylogenetic signal of a few genes can be compensated for by hundreds or thousands of other loci that all share the same evolutionary history.

A phylogenetic tree created from the entire pan-genome would be expected to best describe strain relationships, as it takes into account the evolution of the group and subgroup specific core genomes, which are brought about by a range of events from single nucleotide changes to horizontal gene transfer. In our examination of the 39 *S. enterica* genomes, branch support successively increased between trees, from MLST to core-genome to pan-genome trees.

In GenBank there are significantly more draft bacterial genomes than closed bacterial genomes, and the gap will continue to increase as draft genomes provide enough information for most studies, which simply compare them to a closed reference genome, and that massively parallel sequencing makes obtaining draft sequences cheap and easy, while the process of closing gaps requires significantly more time and money. Panseq has been designed to effectively utilize both draft and closed genomes, and the results of this study demonstrate that building a phylogeny from all of the available information can lead to sub-groups with greater branch support than a phylogeny built only from core genomic regions present in every strain in a group. This has previously been demonstrated with the family *Pasteurellaceae*, where using $\geq$160 genes was shown to provide a robust phylogenetic tree with high branch support, even in the presence of missing data [411]. A concern commonly expressed is that horizontally acquired genes introduce erroneous phylogenetic signal; however, if the acquired elements provides a selective advantage they are likely to be maintained. These acquired elements often define a sub-group, and their phylogenetic signal would be lost if only core genes were considered. Another approach advocated by Philippe and Douady is the creation of a core phylogenetic tree based on nucleotide sequence differences, and subsequent addition of the binary presence / absence of the accessory genome to provide additional phylogenetic information [412]. This type of analysis is possible in Panseq by simply combining the aligned core sequence and accessory presence / absence output files from the CAGF module

5.4.3 The Loci Selector (LS) Module

Molecular fingerprinting methods such as SNP analysis, multi-locus variable number tandem repeat analysis (MLVA) and comparative genomic fingerprinting (CGF) [66] often rely on a small number of loci to differentiate among a large number of bacterial strains. Determining which loci to use in a scheme requires selecting from, in some cases, thousands of loci. While manual inspection of a dataset is required to determine biologically relevant loci, Panseq provides an automated way to empirically determine the most variable and discriminatory loci from an investigator-defined set of variable character data which can range from SNPs to sequence presence/absence data.

5.4.3.1 The LS Module in the Comparison of Randomly Generated Data

The approach of the LS module in Panseq is to iteratively build the final loci set, including only the loci that produce the most unique fingerprints, and offer the most variability among input sequences at each step. This allows it to efficiently examine datasets that would be computationally prohibitive if all possible combinations were considered.

To test the Panseq LS module, we created a random binary dataset of 100 loci among 10 sequences in Microsoft Excel (supplementary material at: `http://www.biomedcentral.com/content/supplementary/1471-2105-11-461-s3.txt`). We subsequently ran Panseq with the ‘best’ option and found that 4 loci generated a unique profile for all 10 sequences, and that the four loci: locus84, locus40, locus79 and locus29 provided the maximum possible discrimination of 100 POD for the dataset. While these loci cannot be guaranteed to be the only four loci to generate the same results, they will always be one of the most discriminatory sets.

We then ran the same dataset through a Perl script that generated all possible 4-loci combinations, outputting those that produced 10 unique sequence profiles (supplementary material at: `http://www.biomedcentral.com/content/supplementary/1471-2105-11-461-s4.txt`). We found that 99132 unique combinations of 4 loci from the dataset yielded 10 unique sequence profiles. Evaluating every possible combination required 449 s on a computer running Ubuntu 9.10 with 3.6 GB of available RAM and AMD Phenom 8450 triple-core processors. Panseq was able to sort through the data to generate a single group of loci that contained not only the most unique profiles, but that provided the most discrimination among the input sequences, completing the task in one s.

5.4.3.2 *The LS Module in the Analysis of SNP Genotyping Data*

Any variable character data can be used as inputs in the LS module. To illustrate the functionality of the LS module, we used a set of 96 SNPs identified from E. coli O157:H7 by Manning et al. [70], for which the nucleotide value of each SNP was determined among a set of 17 *E. coli* O157:H7 genomic sequences [71]. This dataset was first modified to represent any unknown character with the ‘?’ symbol, and to replace any locus described by two characters with a single character (supplementary material at: `http://www.biomedcentral.com/content/supplementary/1471-2105-11-461-s5.txt`).

This dataset was analyzed by the LS module to give the 20 best loci from the 96 available. The results are presented in Table 5.4, and show that the first locus selected by Panseq was ECs2696, a locus with three alleles and therefore more initial fingerprints than any other locus except ECs2006, which also had three alleles among the strains. However, ECs2696 provided more POD among the strains than ECs2006 and was thus selected by Panseq as the initial locus. The eighth locus added (ECs5359) was the last to provide unique strain sequence profiles. This locus differentiated the K-12 strain and *E. coli* O157:H7 strains EC508 and EC71074. Every subsequent locus (9-20) was chosen by the program for its ability to offer discrimination among the remaining strain pairs, while ensuring that highly variable loci that contain very similar allele patterns among strains (i.e are not informative) did not replace loci in the set that offered discrimination among fewer, but nevertheless diverse strains.

5.4.4 Advantages of Panseq over Other Related Programs

While many sequence analysis programs exists, some with overlapping capabilities, no two are identical with respect to the tasks they perform. With enough time and knowledge one can parse the output of a sequence alignment program such as BLAST or MAUVE manually, but there is a considerable time savings and ease of use with a program such as Panseq that automates the process. Panseq is unique in its single-step novel region finding options among groups and specific to individual sequences, which other *in silico* subtractive programs such as mGenomeSubtractor [120] and nWayComp [119] lack. Panseq also provides a comprehensive analysis of the pan-genome, automatically generating analyses that can only be partially accomplished by any other single program; e.g. With MAUVE [4] a list of SNPs can be generated and a display of the similarities/differences between the genomes is produced, but the underlying nucleotide sequence is not automatically extracted. With Panseq, the underlying sequence data is automatically extracted, the segments compared among all sequences and presented in tabular form and input files for phylogenetic programs are automatically created. Further, the SNP table or binary presence/absence table of the accessory genome can be used directly in the LS module, for the selection of the most discriminatory loci.

5.5 Conclusions

We have developed Panseq, a freely available online program to quickly find and extract strain- or group-specific novel accessory genomic information as well as the complete pan-genome for a group of genomic sequences. Panseq produces alignments of the core genome of each sequence and determines the distribution of accessory regions among all sequences analyzed. Panseq makes use of the MUMmer alignment algorithm for whole genome comparisons and the BLASTn algorithm for local sequence comparisons and can efficiently compute values for large numbers of sequences. Additionally, Panseq is able to rapidly identify the most variable and discriminatory locus set in an iterative manner from single character tabular data.

5.6 Availability and Requirements

Project name: Panseq

Project home page: `http://lfz.corefacility.ca/panseq`

Source code: `https://github.com/chadlaing/Panseq`

Operating system(s): Platform independent

Programming language: Perl

Other requirements: Firefox 2.0+, Internet Explorer 6.0+, Google Chrome or compatible web-browser

License: Freely available

5.7 Acknowledgements

We would like to thank the Canadian Food Inspection Agency at the Animal Diseases Research Institute in Lethbridge Alberta for allowing us the use of their facilities. This work was funded by the Public Health Agency of Canada and grants from the Natural Sciences and Engineering Research Council of Canada and Alberta Innovates Technology Futures.

Table 5.1: The six *Listeria monocytogenes* genomic sequences analyzed, with RefSeq accession numbers and genomic sequence status.

RefSeq Accession No.	Strain	Genome Status
NC_002973	4b F2365	Complete
NC_003210	EGD-e	Complete
NC_011660	HCC23	Complete
NC_012488	Clip81459	Complete
NZ_AARU02	F6900	Draft
NZ_AARZ02	10403S	Draft

Table 5.2: The 14 *E. coli* O157:H7 strains compared by Panseq to the *E. coli* O157:H7 reference strains EDL933 and Sakai for novel accessory genomic regions.

RefSeq Accession No.	Strain
NZ_ABJT	EC4024
NZ_ABHM	EC4042
NZ_ABHL	EC4045
NZ_ABHQ	EC4076
NZ_ABHP	EC4113
NZ_ABHO	EC4196
NZ_ABHK	EC4206
NZ_ABHR	EC4401
NZ_ABHS	EC4486
NZ_ABHT	EC4501
NZ_ABHW	EC508
NZ_ABHU	EC869
NZ_ABKY	TW14588
NC_011353.1	EC4115

Table 5.3: The 39 *Salmonella enterica* genome sequences analyzed, and the proportion of the 9.03 mbp pan-genome of the 39 strains present in each genome.

Serovar and strain	RefSeq ID	% of pan-genome
subsp. arizonae 62_z4_z23__RSK2980	NC_010067	59.23
Agona_SL483	NC_011149	46.29
Choleraesuis_SC_B67	NC_006905	46.62
Dublin_CT_02021853	NC_011205	45.10
Enteritidis_P125109	NC_011294	46.77
Gallinarum_287_91	NC_011274	47.02
Heidelberg_SL476	NC_011083	44.50
Newport_SL254	NC_011080	45.30
Paratyphi_A_AKU_12601	NC_011147	49.03
Paratyphi_A_ATCC_9150	NC_006511	49.04
Paratyphi_B_SPB7	NC_010102	45.01
Paratyphi_C_RKS4594	NC_012125	45.50
Schwarzengrund_CVM19633	NC_011094	46.97
Typhi_CT18	NC_003198	47.34
Typhi_Ty2	NC_004631	47.34
Typhimurium_LT2	NC_003197	43.82
4_5_12_i__CVM23701	NZ_ABAO	44.02
Hadar_RI_05P066	NZ_ABFG	46.07
Heidelberg_SL486	NZ_ABEL	46.56
Javiana_GA_MM04042433	NZ_ABEH	48.93
Kentucky_CDC_191	NZ_ABEI	47.28
Kentucky_CVM29188	NZ_ABAK	46.43
Newport_SL317	NZ_ABEW	44.74
Saintpaul_SARA23	NZ_ABAM	45.74
Saintpaul_SARA29	NZ_ABAN	44.88
Schwarzengrund_SL480	NZ_ABEJ	47.54
Tennessee_CDC07_0191	NZ_ACBF	45.85
Typhi_404ty	NZ_CAAQ	53.03
Typhi_AG3	NZ_CAAY	52.48
Typhi_E00_7866	NZ_CAAR	47.70
Typhi_E01_6750	NZ_CAAS	52.51
Typhi_E02_1180	NZ_CAAT	47.48
Typhi_E98_0664	NZ_CAAU	50.18
Typhi_E98_2068	NZ_CAAV	49.63
Typhi_E98_3139	NZ_CAAZ	48.92
Typhi_J185	NZ_CAAW	47.86
Typhi_M223	NZ_CAAX	47.02
Virchow_SL491	NZ_ABFH	45.31
Weltevreden_HI_N05_537	NZ_ABFF	44.58

Table 5.4: The 20 best loci as chosen by the LS module of Panseq from the original 96 loci. Loci are listed in order of choice and with points of discrimination (POD).

Locus	TW14588	Sakai	EDL933	EC4501	EC4486	EC4401	EC4206	EC4196	EC4115	EC4113	EC4076	EC4045	EC4042	EC4024	EC869	EC508	EC71074	EC33264	K12	POD
ECs2696	C	C	G	C	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	63
ECs2775	T	G	G	G	G	G	G	G	G	G	G	G	T	G	G	T	G	0	0	42
ECs2375	C	C	C	C	T	T	T	T	T	T	T	T	T	T	C	C	C	C	C	90
ECs4067	A	C	A	A	A	A	C	A	A	A	A	A	A	A	A	A	A	A	A	34
ECs1262	T	T	T	T	C	C	C	C	C	C	C	C	C	C	T	C	C	T	0	72
ECs1272	T	T	T	T	A	A	A	A	A	A	A	A	A	A	T	A	A	A	0	65
ECs1860	G	G	G	G	G	G	G	G	G	G	G	G	G	T	G	G	G	G	G	18
ECs5359	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	0	G	17
ECs3830	C	C	C	C	T	T	T	T	T	T	T	T	T	T	C	C	C	C	C	90
ECs2357	C	C	C	C	A	A	A	A	A	A	A	A	A	A	C	A	A	C	0	72
ECs4022	G	G	G	G	A	A	A	A	A	A	A	A	A	A	G	A	A	0	G	72
ECs0593	T	T	T	T	C	C	C	C	C	C	C	C	T	C	C	C	C	C	C	70
ECs4380	G	G	G	G	A	A	G	A	A	A	A	A	A	A	A	A	A	A	0	65
ECs0606	A	A	C	A	C	C	C	C	C	C	C	C	C	C	A	C	C	0	0	52
ECs2514	T	T	C	T	T	T	T	T	T	T	T	T	T	T	T	T	T	C	C	48
ECs2006	G	G	C	G	G	G	G	G	G	G	G	G	G	G	G	G	A	0	0	31
ECs2852	C	T	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	18
ECs4251	G	G	G	G	A	A	A	A	A	A	A	A	A	A	G	G	G	G	G	90
ECs4305	A	A	A	A	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	60
ECs4479	G	G	G	G	T	T	T	T	T	T	T	T	T	T	G	T	T	0	0	60

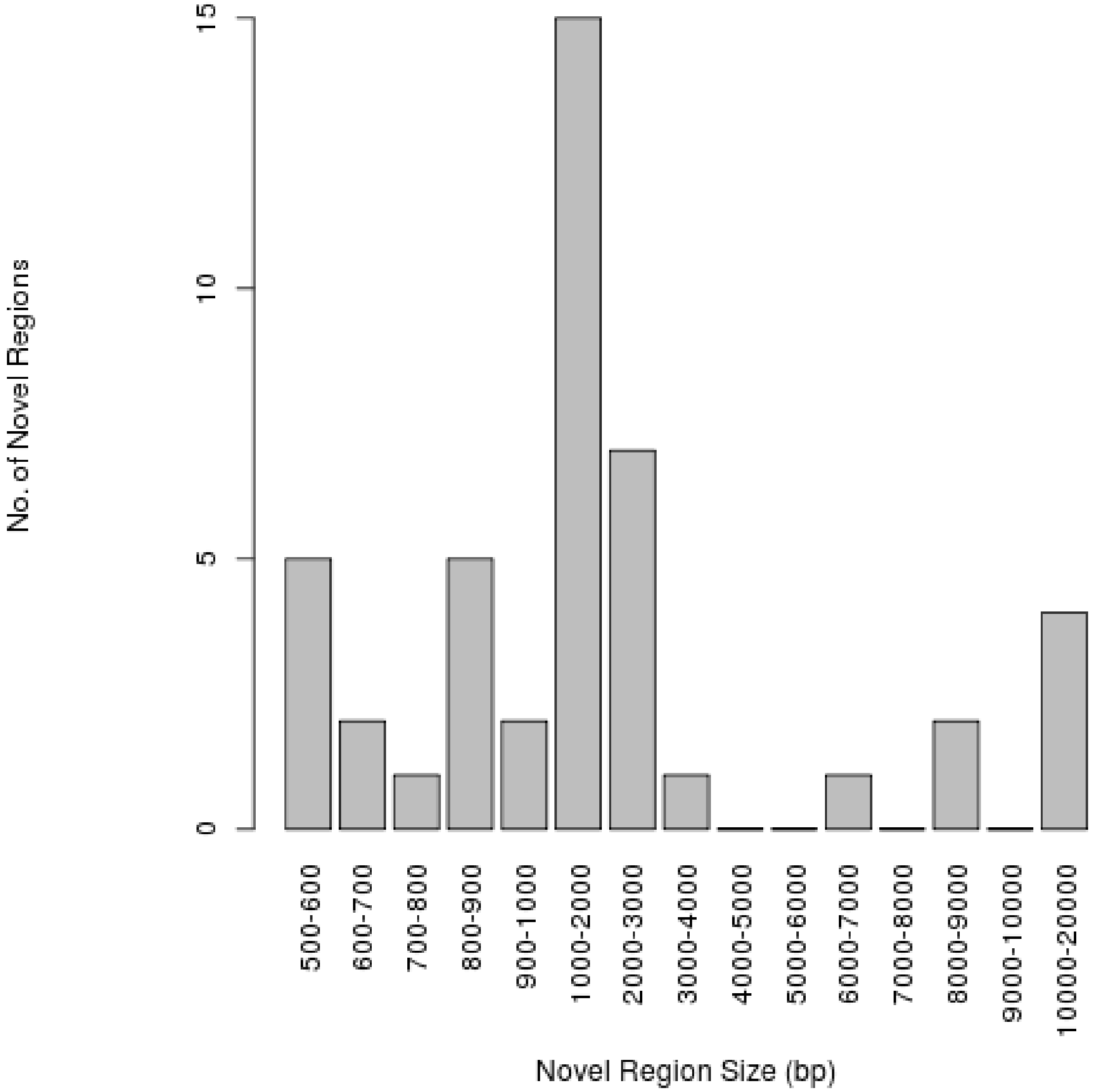


Figure 5.1: The size distribution of regions novel to one or both *Listeria monocytogenes* strains F6900 and 10403S with respect to the four *L. monocytogenes* strains Clip81459, EGD-e, HCC23 and 4bF2365. The minimum extracted novel region size was 500 bp.

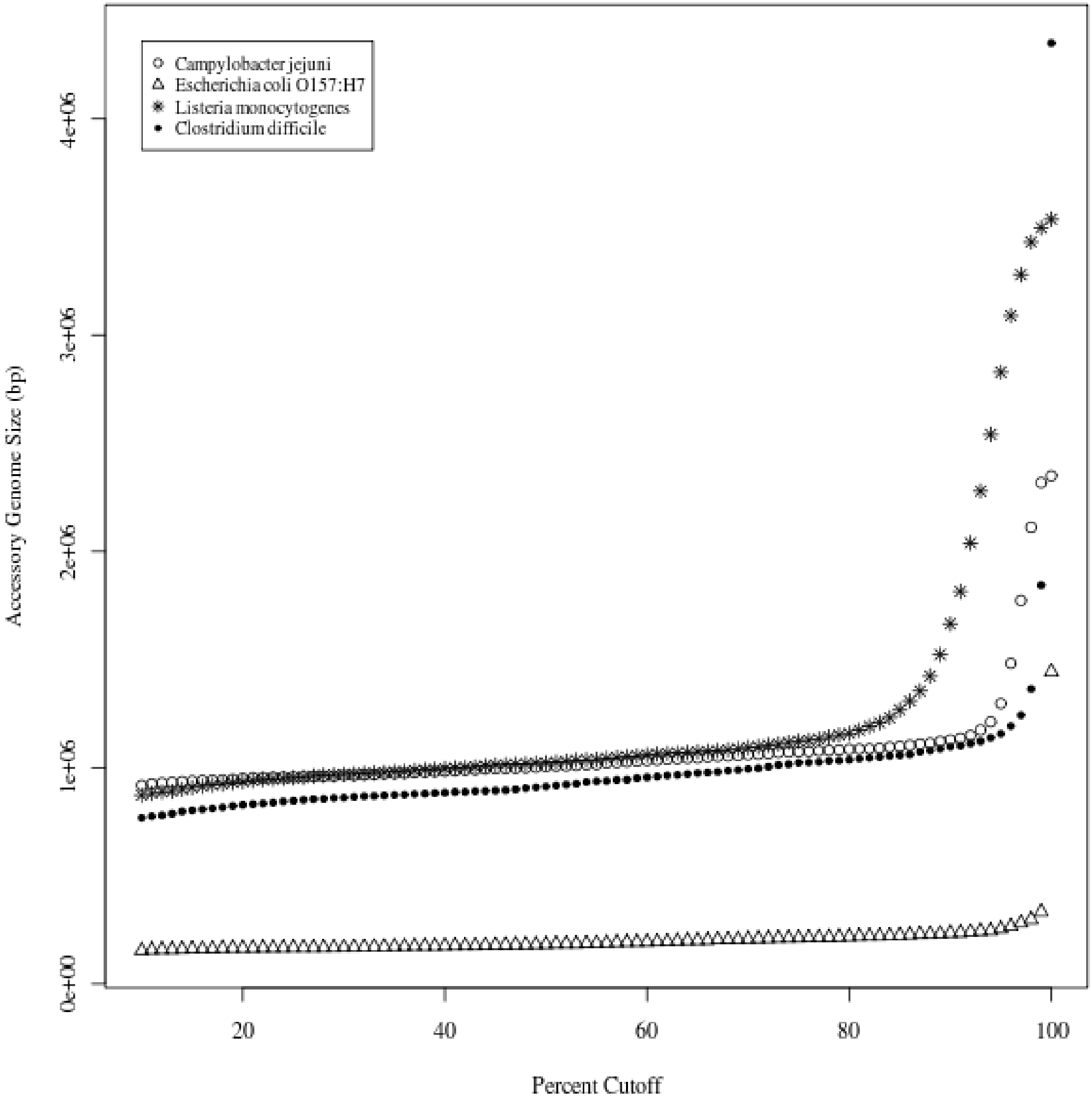
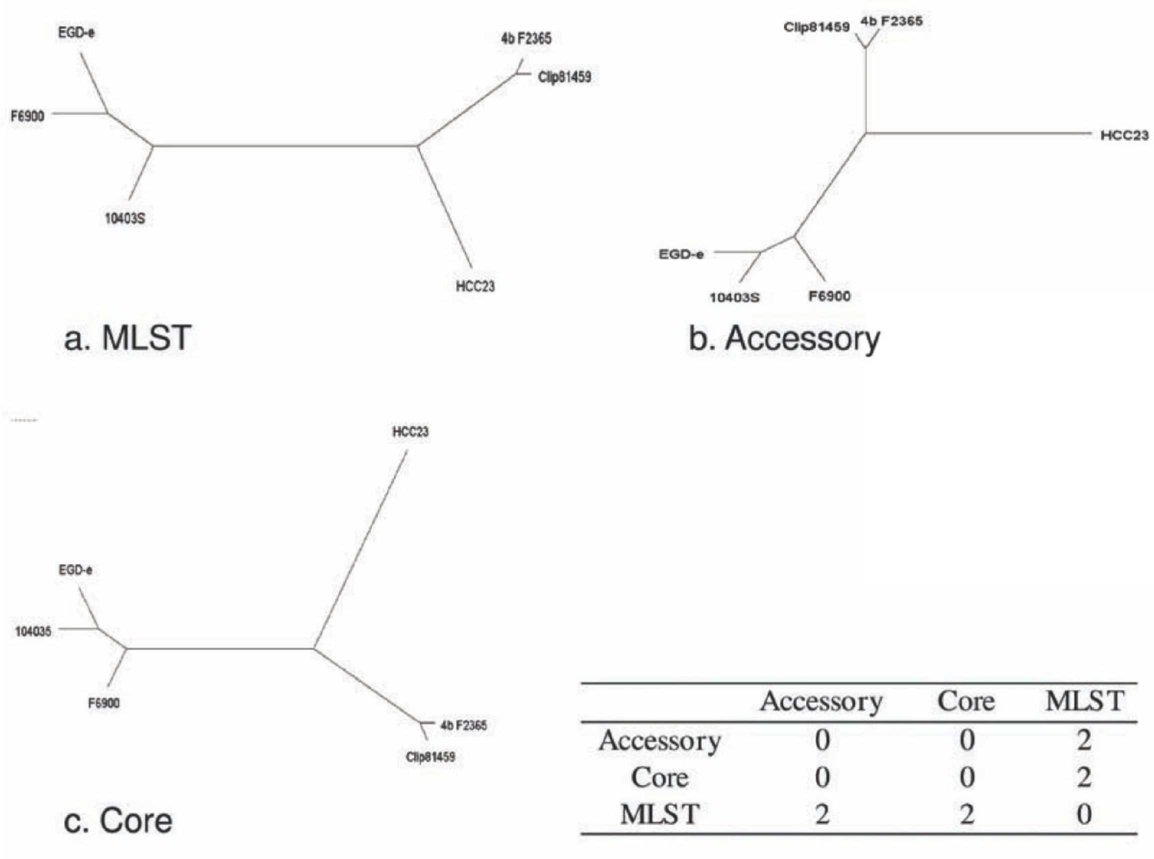


Figure 5.2: The size of the accessory genomes for groups of *Listeria monocytogenes* strains (F6900, 10403S, Clip81459, EGD-e, HCC23 and 4bF2365); *E. coli* O157:H7 strains (EDL933, Sakai, EC4115, TW14539); *Clostridium difficile* strains (630, CD196, R20291, BI9); *Campylobacter jejuni* strains (RM1221, 81-176, 81116, NCTC 11168, 269.97) over sequence identity cutoff values of 10 - 100%. Genomes were fragmented into 500 bp segments.



	Accessory	Core	MLST
Accessory	0	0	2
Core	0	0	2
MLST	2	2	0

Figure 5.3: The maximum parsimony (MP) trees generated for the six *Listeria monocytogenes* strains F6900, 10403S, Clip81459, EGD-e, HCC23 and 4bF2365 using a) multilocus sequence typing as described by Nightingale et al., 2005 [1]; b) the binary presence / absence data of the accessory genome found using the Panseq Core/Accessory Genome Analysis module with fragmentation size of 500 bp and sequence identity cutoff value of 85%; c) the aligned core genome found with the parameters of b). MP trees were created using Phylip v3.69 [2] with the PARS function. The inset table depicts the symmetrical tree distances between each pair of trees, calculated using the TREEDIST function of Phylip.

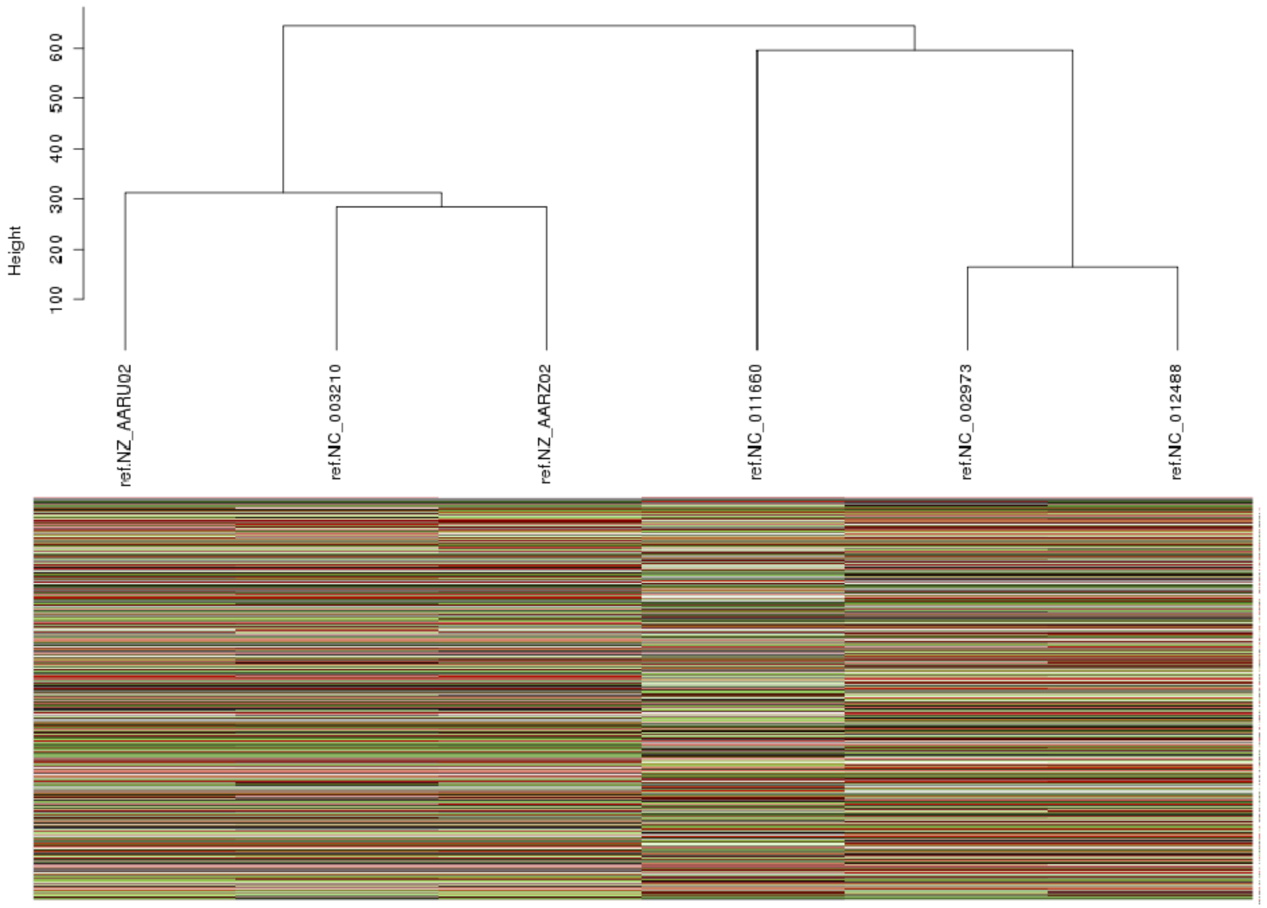


Figure 5.4: Hierarchical clustering of the SNPs within the core genome of the *Listeria monocytogenes* strains F6900, 10403S, Clip81459, EGD-e, HCC23 and 4bF2365. The core genome was generated using the Panseq Core/Accessory Genome Analysis module with fragmentation size of 500 bp and sequence identity cutoff value of 85%. The dendrogram was produced by the statistical package R, using the hclust function with Euclidean distance and average linkage after substituting the ACTG character values with 0,1,2,3 respectively; black = A, white = T, red = C and green = G.

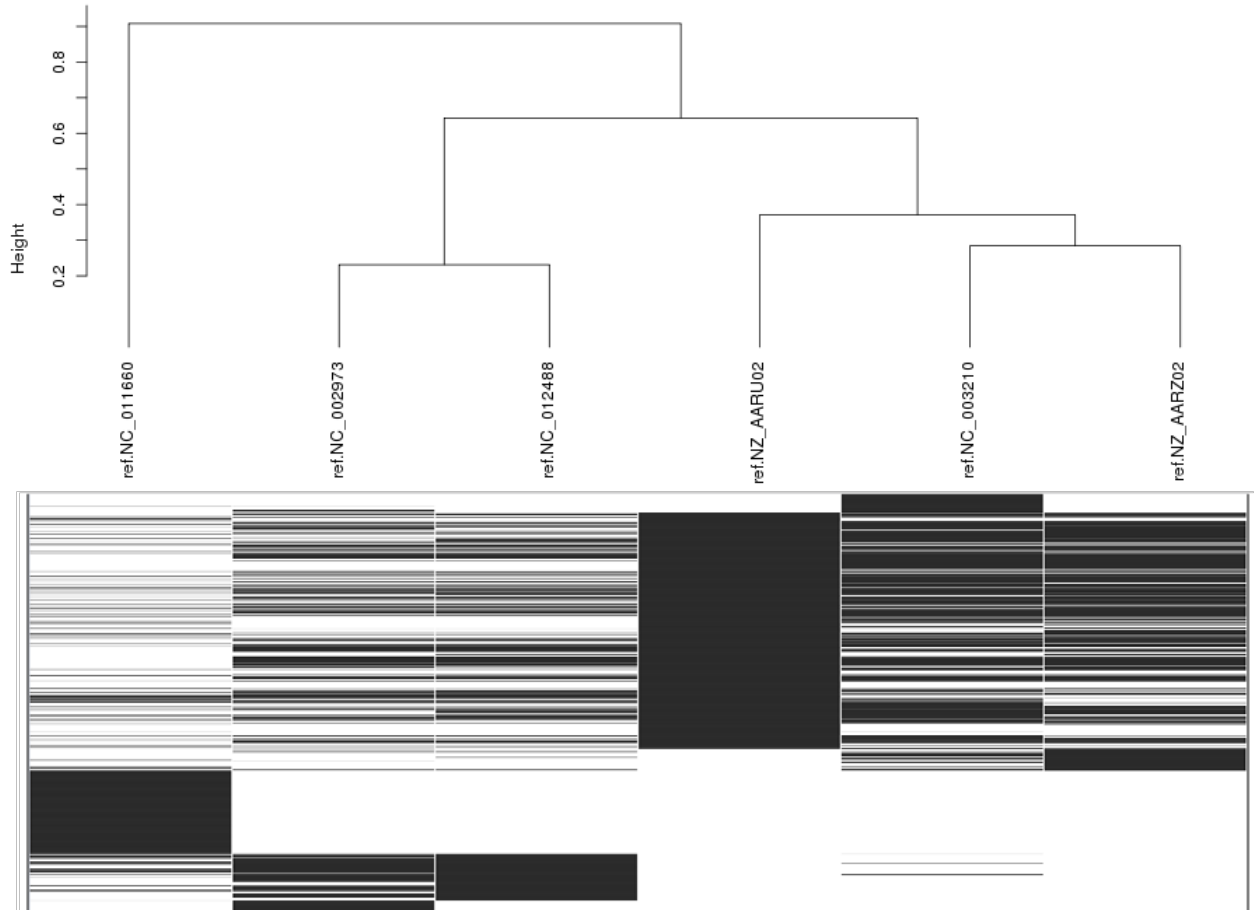


Figure 5.5: Hierarchical clustering of the binary presence/absence values of the accessory genome of the *Listeria monocytogenes* strains F6900, 10403S, Clip81459, EGD-e, HCC23 and 4bF2365. The accessory genome was generated using the Panseq Core/Accessory Genome Analysis module with fragmentation size of 500 bp and sequence identity cutoff value of 85%. The dendrogram was produced by the statistical package R, using the hclust function with binary distance and average linkage; black indicates presence of a locus and white the absence of a locus.

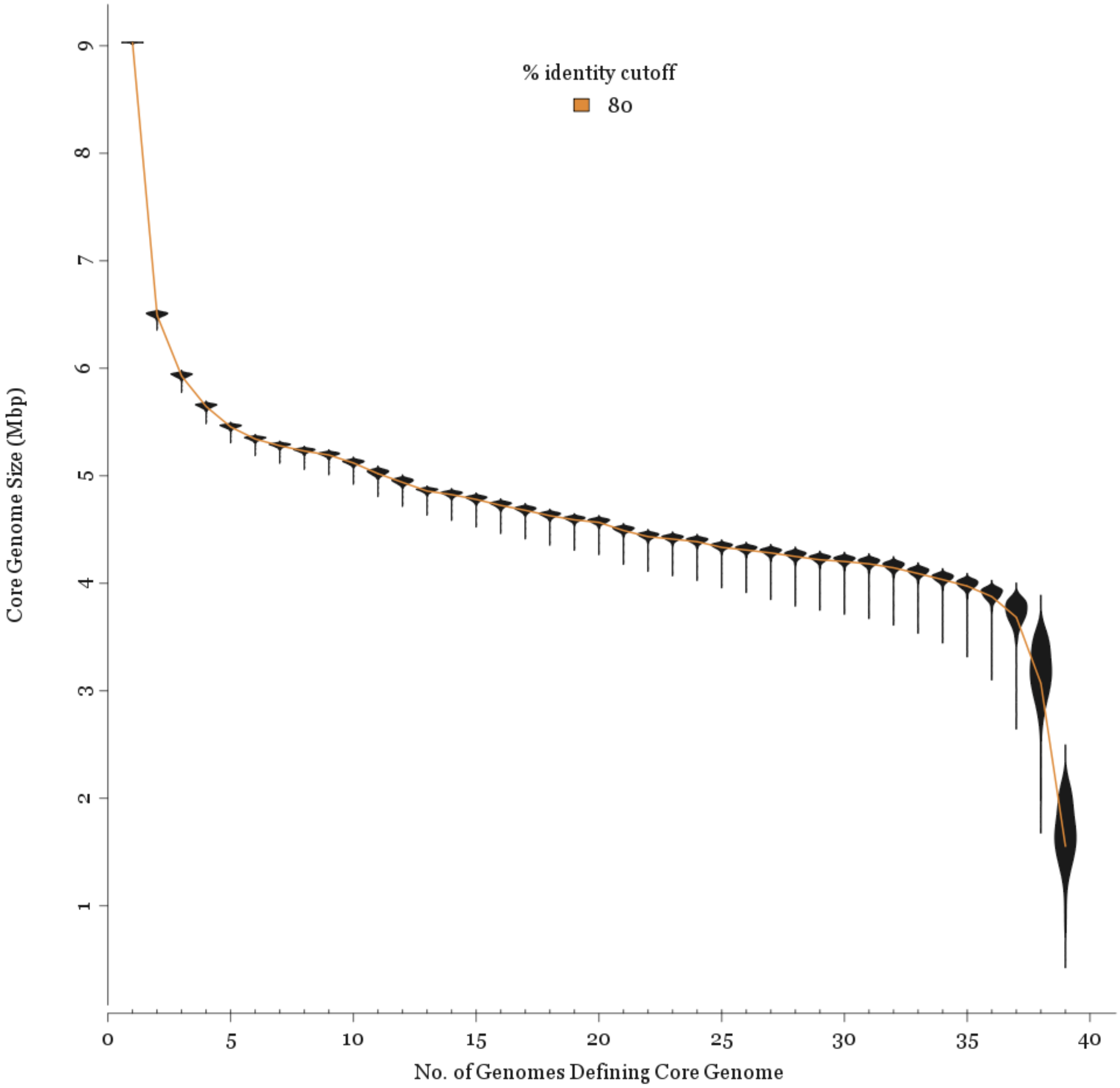


Figure 5.6: Core genome size of 39 *S. enterica* genomes over varying thresholds of core genome membership and percent sequence identity cutoff. Each plot represents the core genome size distribution at a particular core genome threshold, with the bottom of the plot representing a 100 % sequence identity cutoff and the top a 40 % sequence identity cutoff; the shape of the plot represents the size distribution among the range of intervening percent sequence identity cutoffs. For reference, the core genome size at an 80% sequence identity cutoff is shown for each core genome threshold.

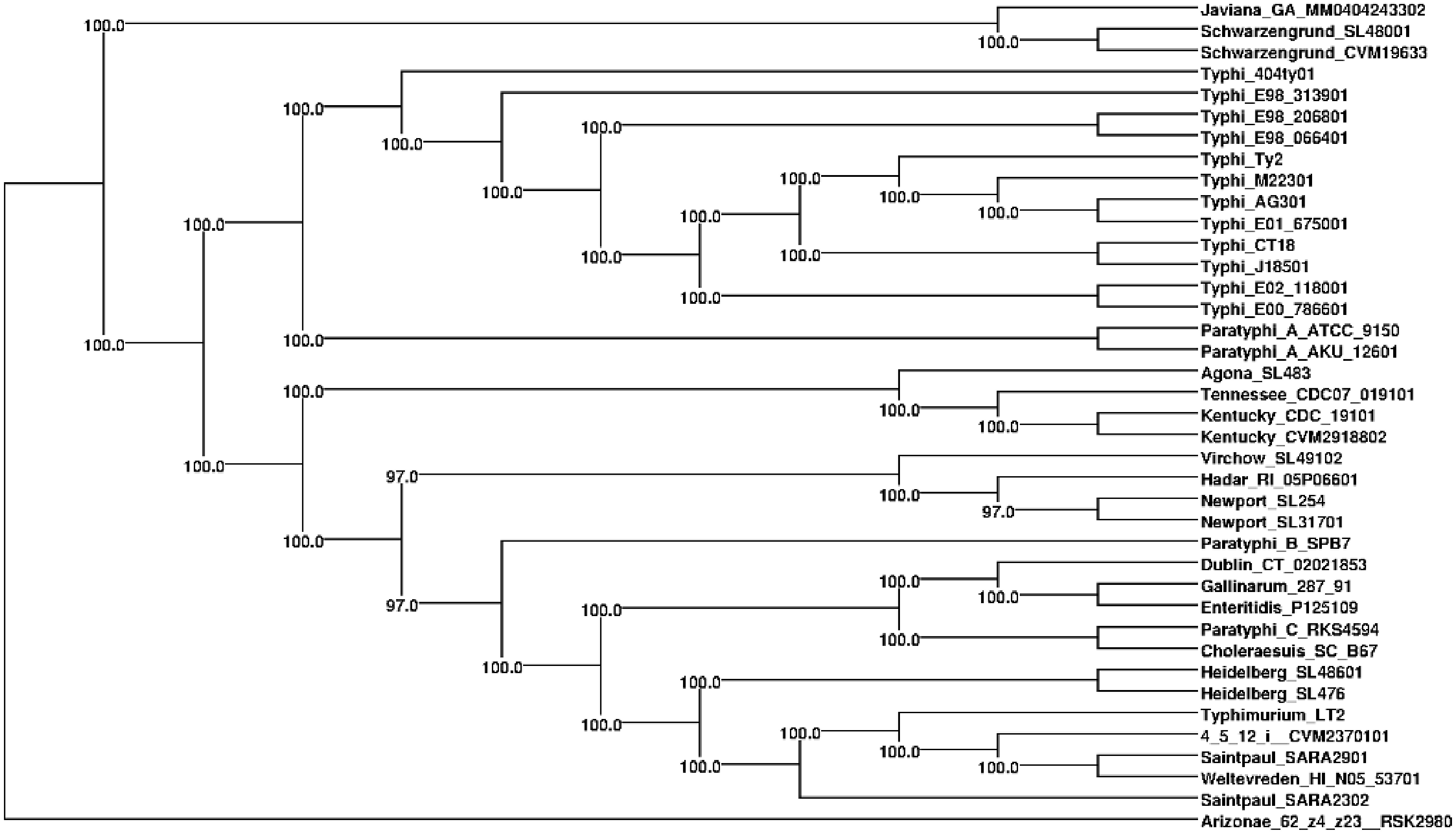


Figure 5.7: The phylogenetic tree of 39 *S. enterica* strains rooted with *S. enterica* subsp. arizonae strain RSK2980, using the concatenated core genome based on a core-genome threshold of 1 (the entire pan-genome); maximum parsimony tree with 100 replicate bootstrap support values

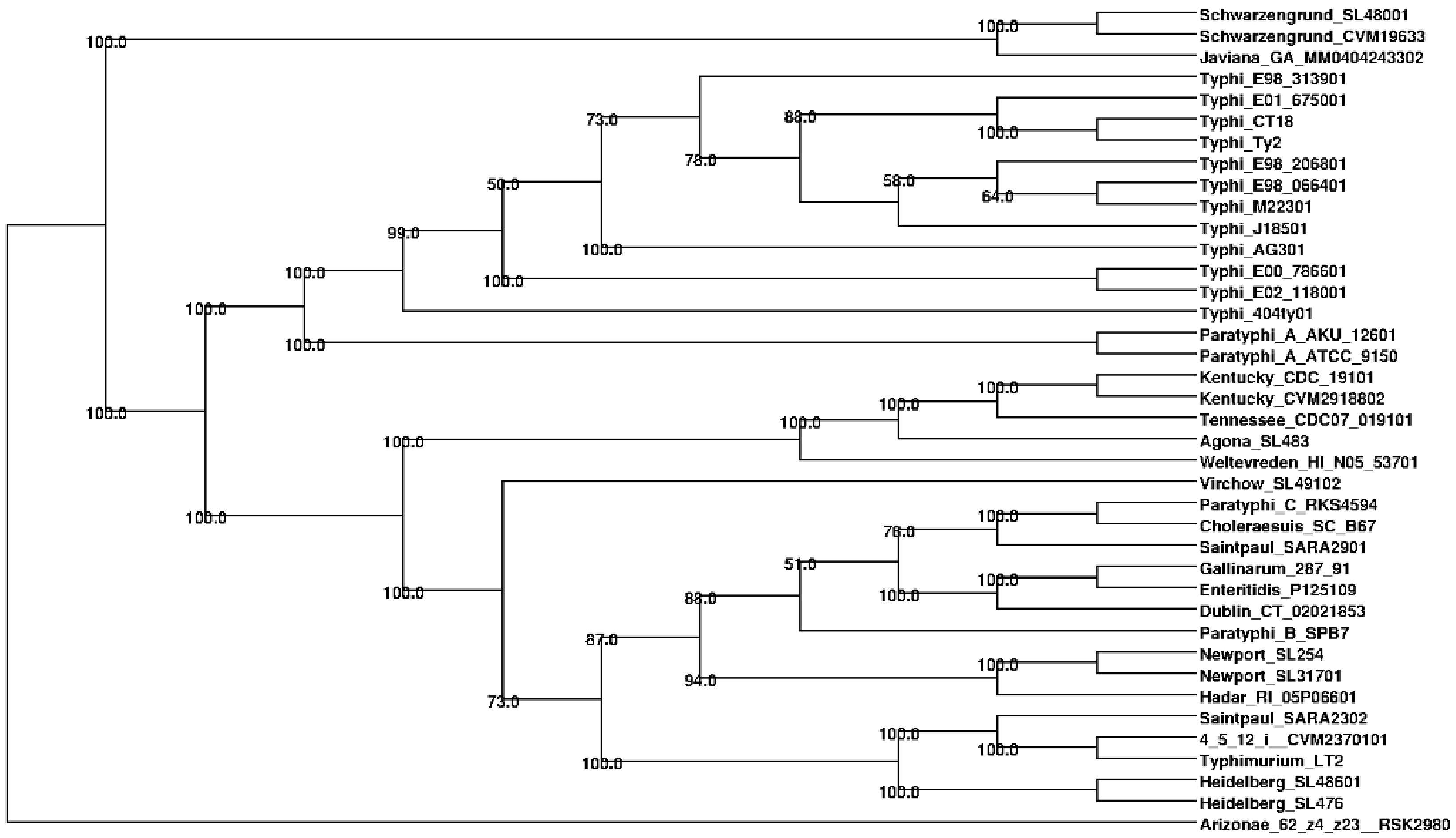


Figure 5.8: The phylogenetic tree of 39 *S. enterica* strains rooted with *S. enterica* subsp. arizonae strain RSK2980, using the concatenated core genome based on a core-genome threshold of 1 (the entire pan-genome); b) neighbour-joining tree with 100 replicate bootstrap support values;

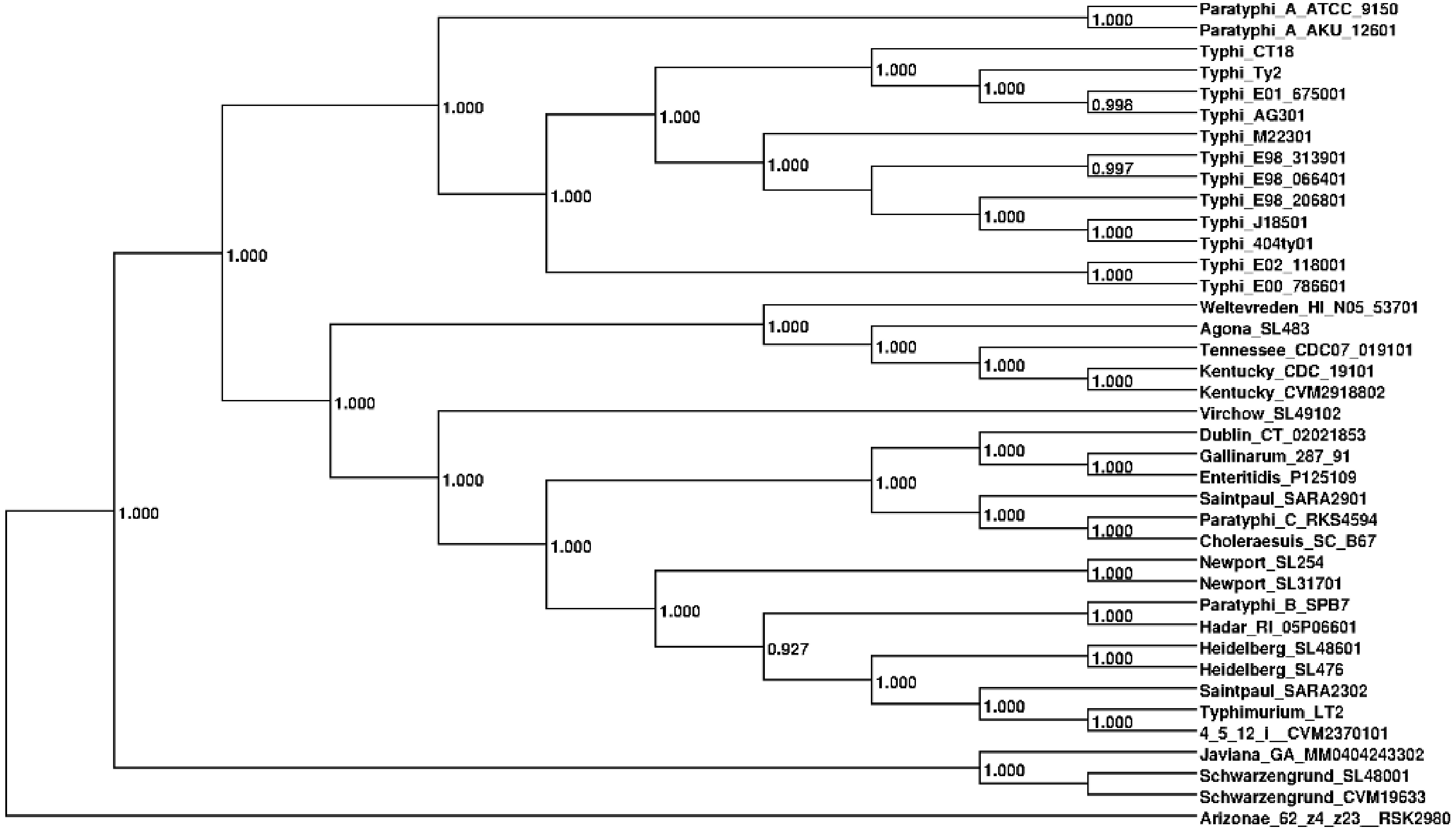


Figure 5.9: The phylogenetic tree of 39 *S. enterica* strains rooted with *S. enterica* subsp. arizonae strain RSK2980, using the concatenated core genome based on a core-genome threshold of 1 (the entire pan-genome); c) maximum likelihood tree with aLRT support values.

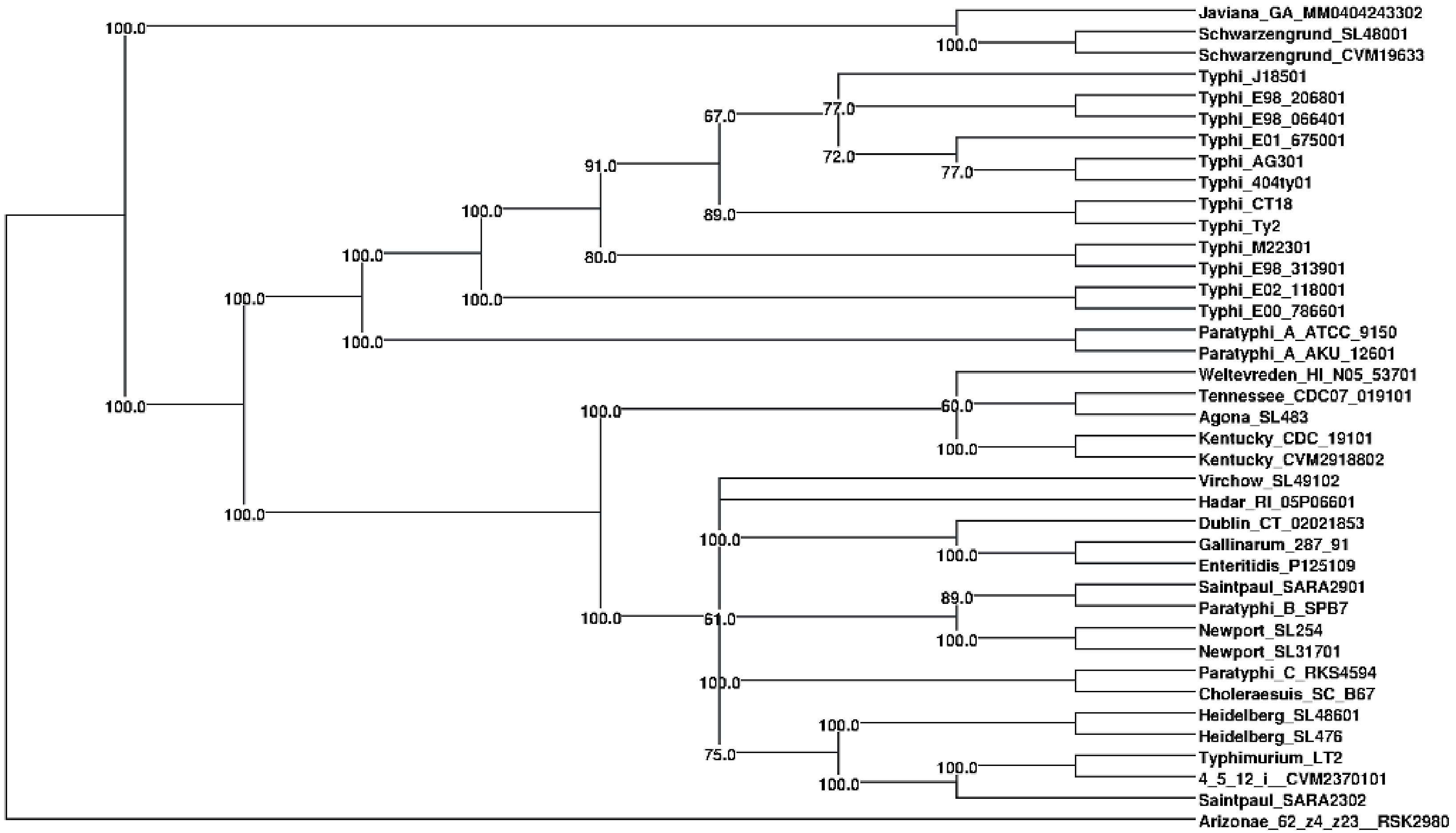


Figure 5.10: The phylogenetic tree of 39 *S. enterica* strains rooted with *S. enterica* subsp. arizonae strain RSK2980, using the concatenated core genome based on a core-genome threshold of 39 (the strict core genome); maximum parsimony tree with 100 replicate bootstrap support values.

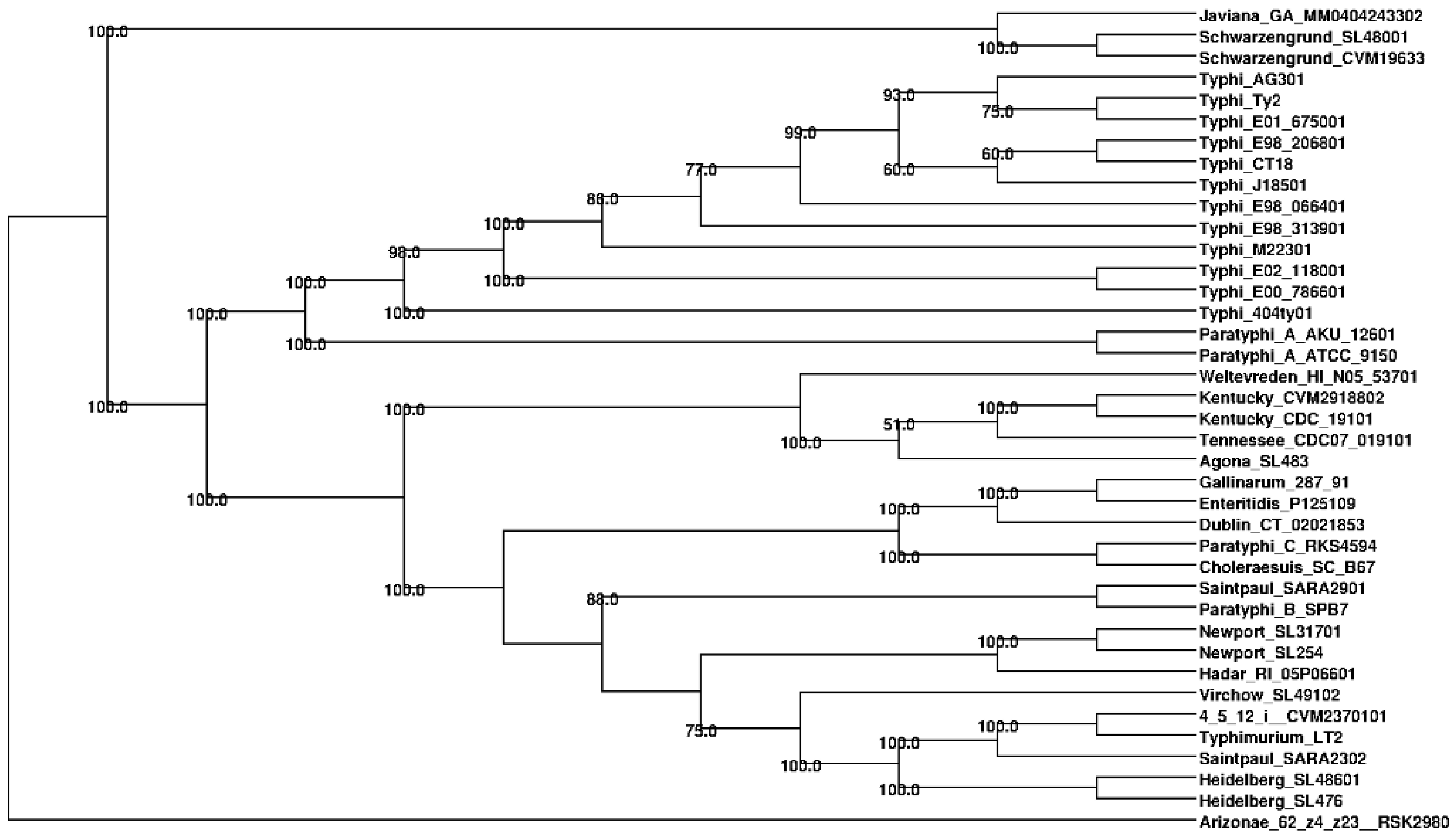


Figure 5.11: The phylogenetic tree of 39 *S. enterica* strains rooted with *S. enterica* subsp. arizonae strain RSK2980, using the concatenated core genome based on a core-genome threshold of 39 (the strict core genome); neighbour-joining tree with 100 replicate bootstrap support values.

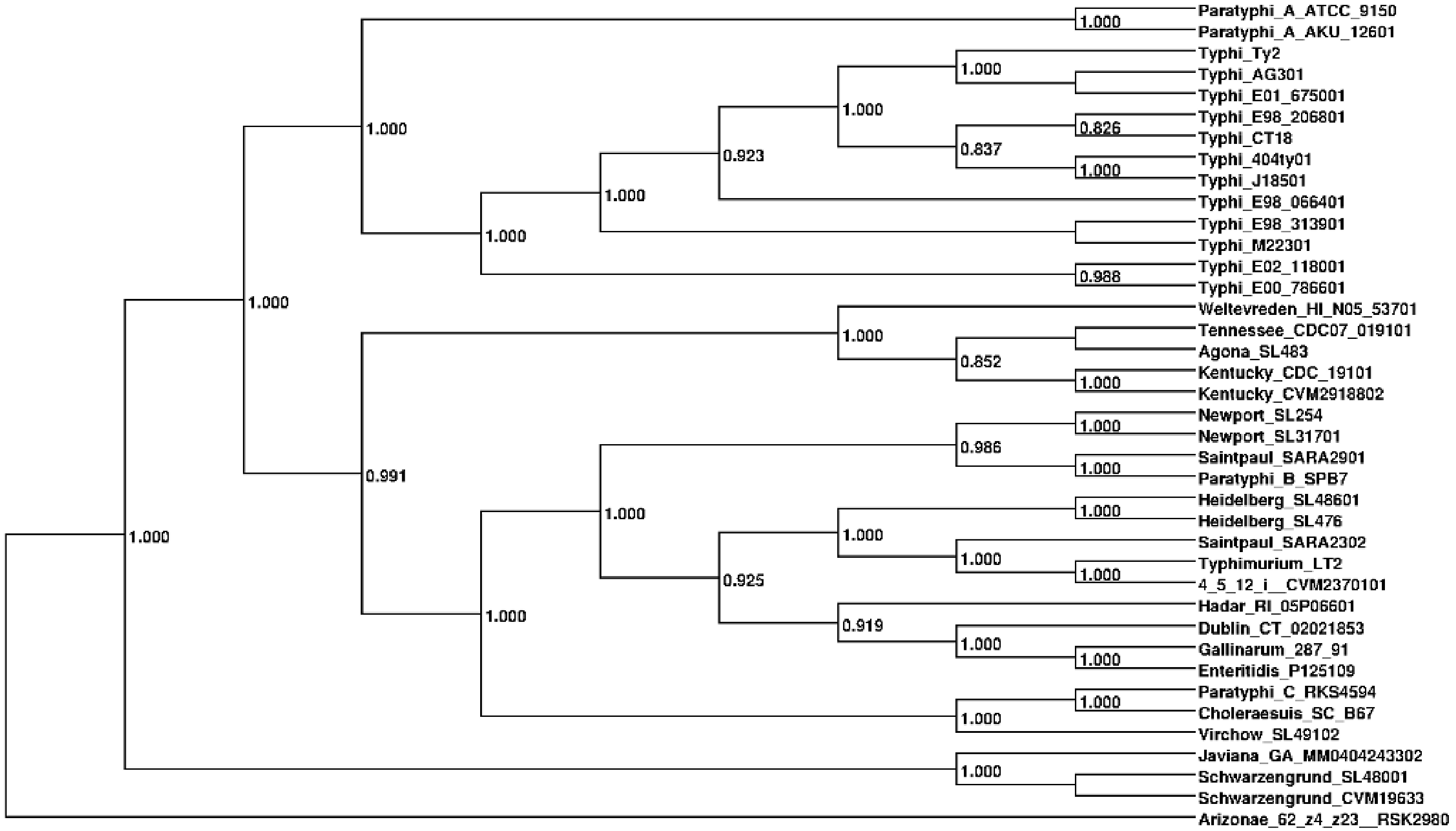


Figure 5.12: The phylogenetic tree of 39 *S. enterica* strains rooted with *S. enterica* subsp. arizonae strain RSK2980, using the concatenated core genome based on a core-genome threshold of 39 (the strict core genome); maximum likelihood tree with aLRT support values.

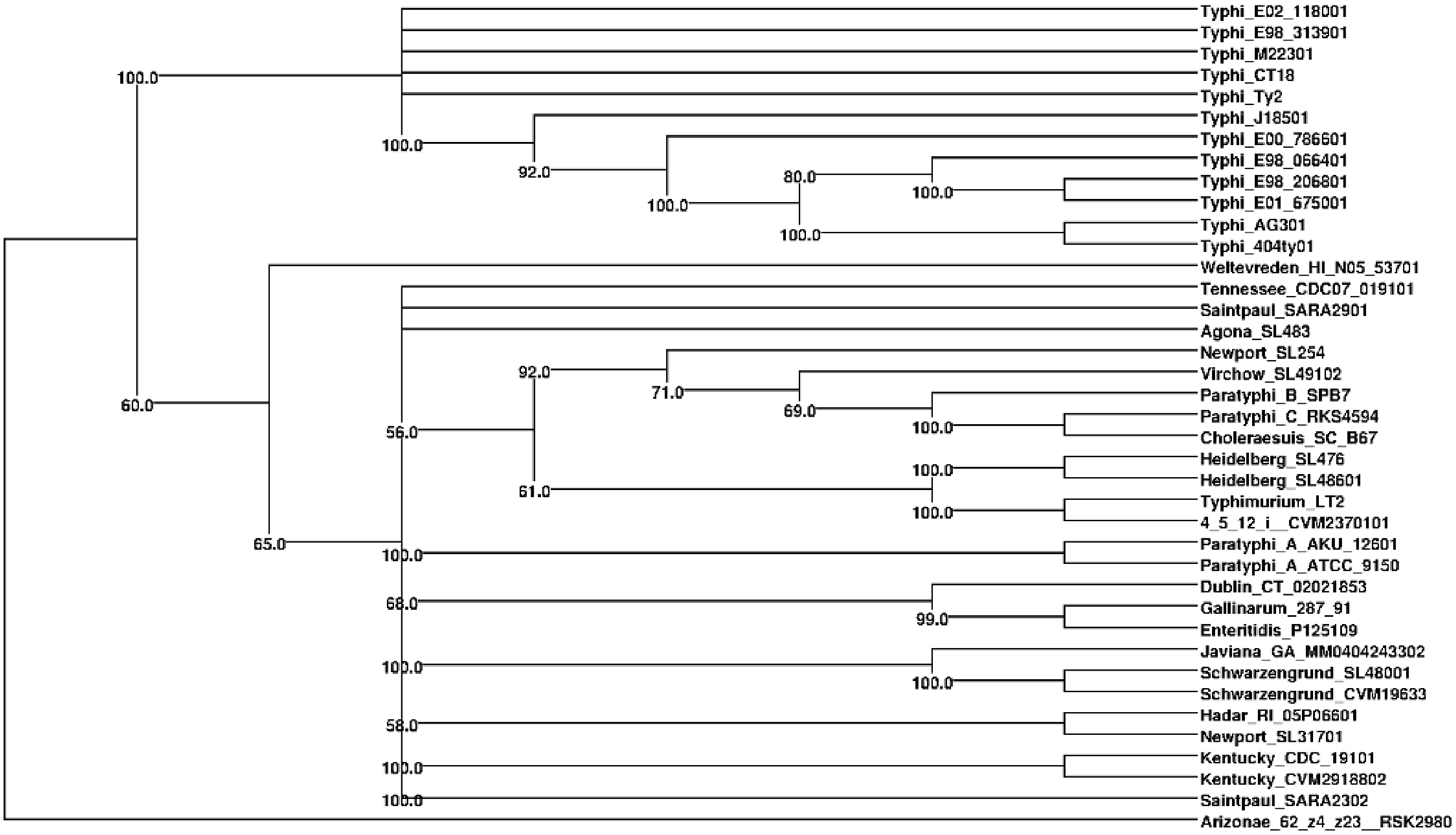


Figure 5.13: The phylogenetic tree of 39 *S. enterica* strains rooted with *S. enterica* subsp. arizonae strain RSK2980, using the concatenated MLST genes as defined in the *S. enterica* MLST scheme of Kidgell et al. [3]; maximum parsimony tree with 100 replicate bootstrap support values.

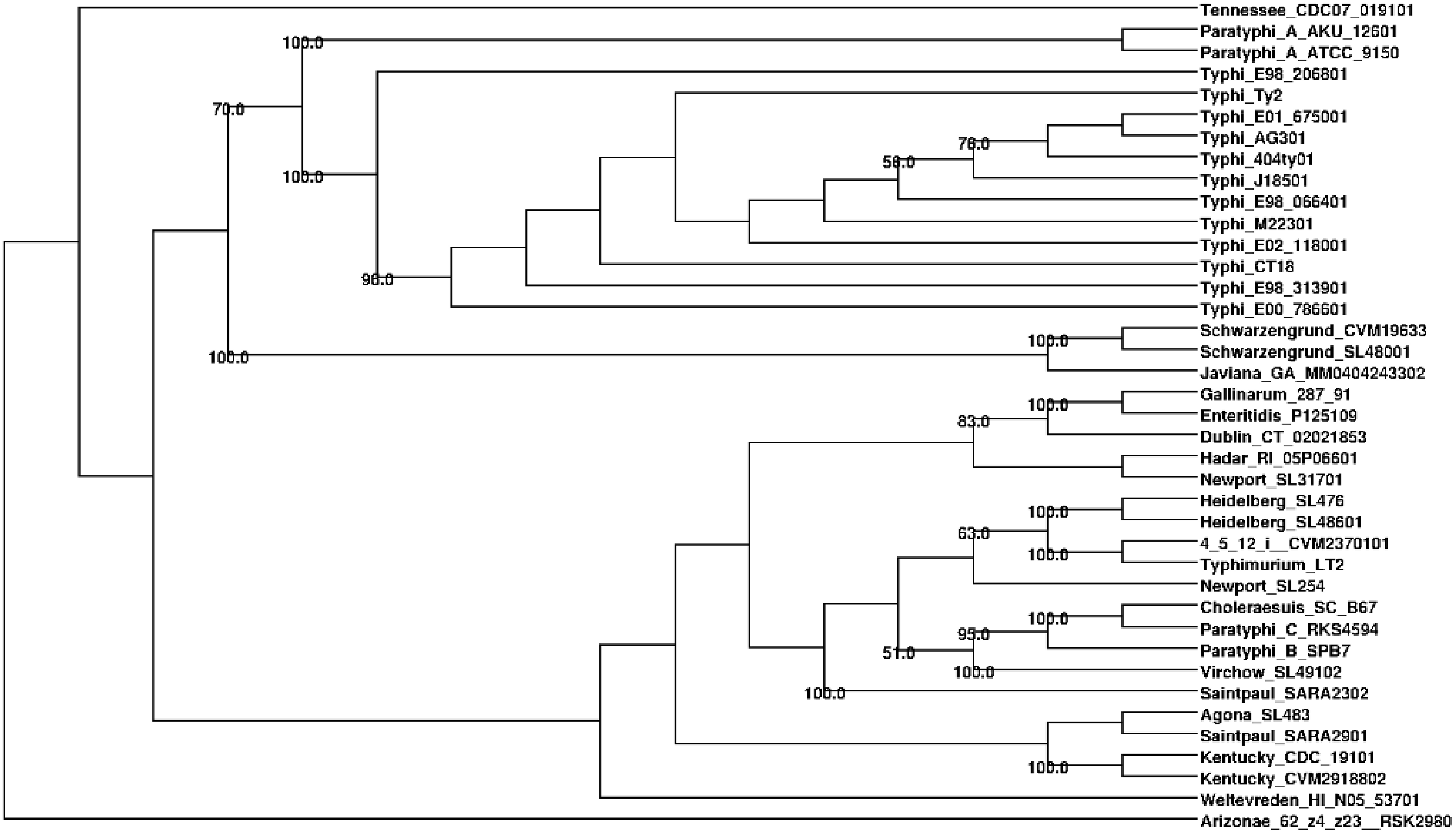


Figure 5.14: The phylogenetic tree of 39 *S. enterica* strains rooted with *S. enterica* subsp. arizonae strain RSK2980, using the concatenated MLST genes as defined in the *S. enterica* MLST scheme of Kidgell et al. [3]; neighbour-joining tree with 100 replicate bootstrap support values.

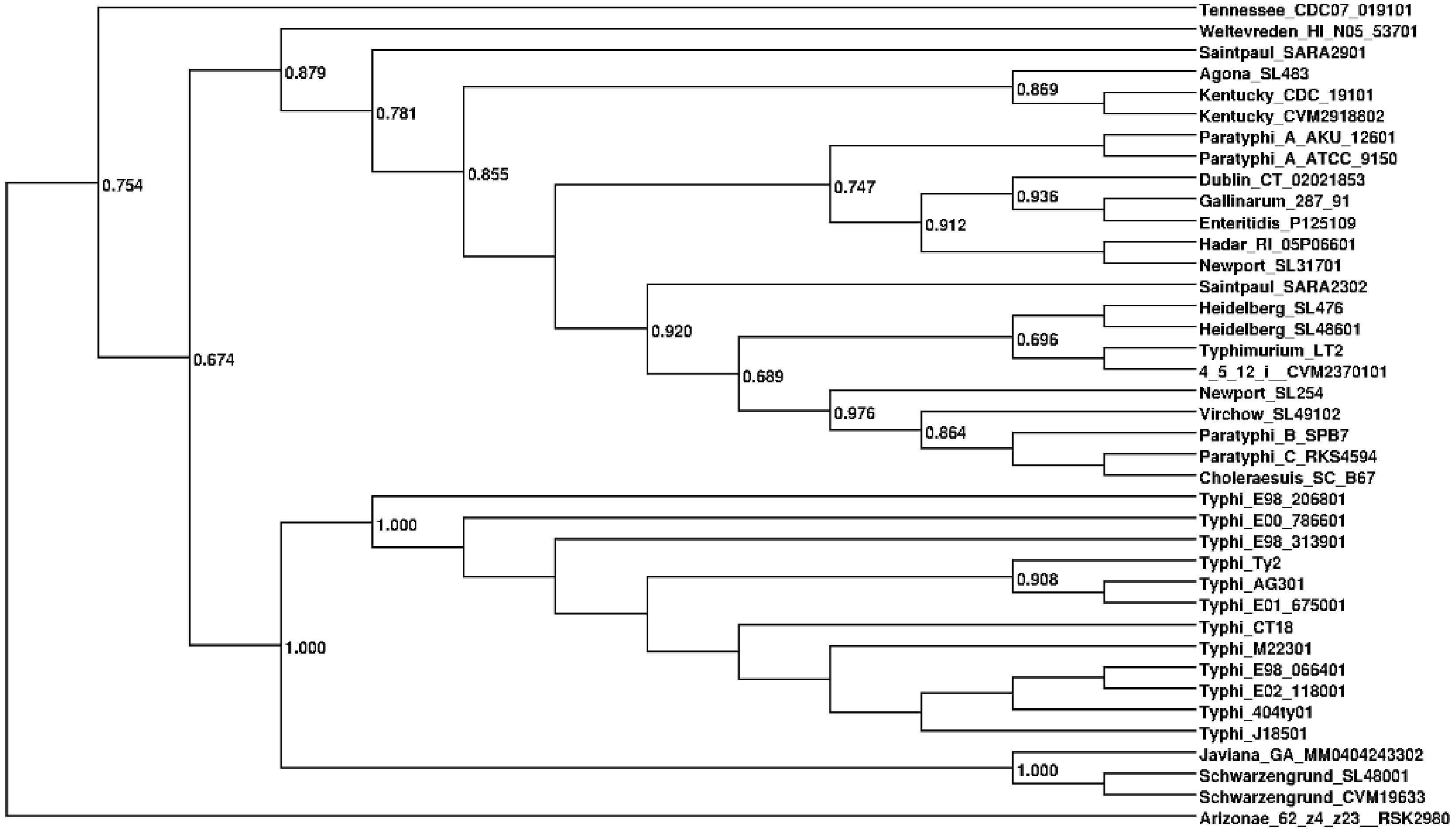


Figure 5.15: The phylogenetic tree of 39 *S. enterica* strains rooted with *S. enterica* subsp. arizonae strain RSK2980, using the concatenated MLST genes as defined in the *S. enterica* MLST scheme of Kidgell et al. [3]; maximum likelihood tree with aLRT support values.

Chapter 6

A comparison of Shiga-toxin 2 bacteriophage from classical enterohemorrhagic *Escherichia coli* serotypes and the German *E. coli* O104:H4 outbreak strain

6.1 Preface

This chapter was previously published as Laing CR, Zhang Y, Gilmour MW, Allen V, Johnson R, Thomas JE, Gannon VP. ‘A comparison of Shiga-toxin 2 bacteriophage from classical enterohemorrhagic *Escherichia coli* serotypes and the German *E. coli* O104:H4 outbreak strain.’ PLoS One. 2012;7(5):e37362. doi: 10.1371/journal.pone.0037362. [413]

6.2 Introduction

A novel *Escherichia coli* O104:H4 strain was associated with a widespread and severe foodborne disease outbreak in Germany between early May and July, 2011 [414]. Among the more than 4000 individuals from which the pathogen was isolated, approximately 800 people developed the life-threatening hemolytic uremic syndrome (HUS) and 50 people succumbed to their illness. Epidemiological studies eventually pointed to a single lot of imported Fenugreek seeds used to prepare sprouts for salads as the source of the organism [415].

The *E. coli* O104:H4 outbreak strain exhibited characteristics typical of other enteroaggregative *E. coli* (EAEC), including the presence of a paa-containing virulence plasmid, production of enteroaggregative adherence fimbriae (AAF) I, and formation of the stacked-brick adherence pattern on intestinal epithelial cells [416]. The genome of the outbreak strain was also very similar to the EAEC O104:H4 reference strain 55989, which was isolated in Africa in 2002 [417]. The 2011 outbreak isolates were shown to be clonal and part of multi-locus sequence type ST678, which is unique to *E. coli* O104:H4 strains, and part of the *E. coli* phylogenetic group B1, which contains a variety of other pathogenic serotypes [340]. The outbreak strains also acquired an antibiotic resistance plasmid, conferring resistance to all penicillins, cephalosporins and co-trimoxazole [416, 418]. Unusually, the outbreak-associated *E. coli* O104:H4 isolates also produced Shiga-toxin 2 (Stx$_{2)}$ and consequently were classified as Shiga-toxin producing *E. coli* (STEC). Stx $_{p}$roduction has only rarely been reported among EAEC strains. Prior to the outbreak, only seven sporadic cases of infections with *E. coli* O104:H4 strains that produced Stx$_2$ had been reported worldwide in the preceding ten years [419].

Prior to this outbreak, HUS in Germany and elsewhere was most frequently associated with STEC infections in children and the elderly, and only 1.5%10% of all reported STEC cases in Germany between 2006 and 2010 resulted in HUS in adults [420]. By contrast, in the *E. coli* O104:H4 outbreak, illness occurred predominantly in otherwise healthy adults and approximately 20% of the cases resulted in HUS, and severe neurological complications were also observed in a large number of cases [416, 421]. *E. coli* O104:H4 had previously been associated with only two cases of HUS; one in a woman from Tokyo and one from a child in Germany [416].

STEC can harbour one or more *stx*-genes which are encoded by inducible lambda phage integrated into their genomes, and the entire phage and specific regions within the phage can be gained or lost through horizontal gene transfer [225]. The *E. coli* O104:H4 outbreak isolates contained two integrated Stx$_{-l}$ike phages, one of which contained the Stx$_2$ A and B subunit genes and another that did not [340].

Whole-genome phylogenies place the *E. coli* O104:H4 outbreak strain as most closely related to other *E. coli* O104:H4 and EAEC strains, whether they have acquired the Stx$_{2-}$phage or not [340, 422]. Because Stx$_{-b}$acteriophage are mobile, their evolutionary history may differ from that of their bacterial host [236]. This raises the question as to the possible origin of the Stx$_{2-}$phage found in the *E. coli* O104:H4 outbreak strain.

In this study, we analyzed the Stx$_{2-}$phage sequences from the *E. coli* O104:H4 outbreak strain and compared it among 51 other Stx$_{2-}$phage sequences from GenBank and those obtained using *de novo* DNA sequence analysis. We demonstrate that the *E. coli* O104:H4 Stx$_{2-}$phage is most closely related to the Stx$_{2-}$phage from *E. coli* O111:H- strain JB1-95, and that a recent common ancestor likely gave rise to both it and the phage present in the outbreak strain. Additionally we show that

Stx$_2$ production by an *E. coli* O104:H4 outbreak-related isolate is significantly greater than that of two other outbreak-related *E. coli* O157:H7 strains following addition of mitomycin-C to cultures.

6.3 Materials and Methods

6.3.1 Identification and Isolation of Stx$_2$-bacteriophage Sequences

The Stx$_2$ subunit A nucleotide sequence from *E. coli* O157:H7 strain EDL933 (NC_002655.2 1352290..1353249) was used to identify Stx$_2$-bacteriophage containing bacteria within GenBank, both the closed and whole-genome shotgun sequence databases. These sequences were downloaded and the Stx$_2$-bacteriophage sequence was identified using PHAST [423]. Only continuous phage sequences of 20 kb from a single contig or closed-genome were used in subsequent analyses; Stx$_2$-positive draft genomes that did not meet these criteria were excluded. The Stx$_1$-like phage from *Shigella dysenteriae* was included as an outgroup for the analyses. The names, sources and accession numbers of all 52 bacteriophage sequences used in this study are presented in Table 6.2.

6.3.2 Phylogenetic Relationships Among Stx$_2$-bacteriophage

A non-redundant Stx$_2$-bacteriophage pan-genome of the 52 phage-genomes of this study was created, and fragments 500 bp present in at least three of the Stx$_2$-phage genomes were aligned using the standalone-version of the Panseq program (`http://lfz.corefacility.ca/panseq/`) with the following settings: ‘minimumNovelRegionSize’ = ‘500’, ‘fragmentationSize’ = ‘500’, ‘nucB’ = ‘200’, ‘nucC’ = ‘50’, ‘nucD’ = ‘0.12’, ‘nucG’ = ‘100’, ‘nucL’ = ‘11’, ‘percentIdentityCutoff’ = ‘85’, ‘coreGenomeThreshold’ = ‘3’ [127]. The aligned fragments were visualized in SplitsTree 4, using the uncorrected P distance and the neighbor-net algorithm [361]. We have previously shown these fragmentation and sequence identity threshold parameters to be appropriate choices for *E. coli*

[127]. Figure S1 depicts the neighbor-nets of the same data fragmented from 50 bp – 1000 bp at both an 85% and 95% sequence identity cutoff, all of which show similar relationships among strain groups.

A maximum-likelihood tree using the aligned Stx$_2$-bacteriophage pan-genome described above was created using PhyML 3.0 with the following settings: PhyML_3.0_linux64 -p -m HKY85 -s BEST –rand_start –n_rand_starts 5 –r_seed 5 [392]. The resulting tree file with approximate likelihood ratio test (aLRT) branch support values was visualized in Dendroscope [394].

The Stx$_2$-bacteriophage that were closest to the *E. coli* O104:H4 outbreak isolates in Figures 6.1 and 6.2 were aligned using MAFFT v6.857b with the command: mafft –maxiterate 1000 –clustalout –thread 22 [424]. Unlike the non-syntenic Panseq alignment, the MAFFT alignment was based on uninterrupted, whole bacteriophage genomes. The MAFFT-aligned genomes were subjected to maximum likelihood phylogenetic analysis as previously described and visualized in Dendroscope [394].

6.3.3 Whole-genome Phylogeny

A non-redundant pan-genome of 42 STEC and *Shigella dysenteriae* whole-genomes was created, and fragments ≥500 bp present in at least three of the genomes were aligned using the standalone-version of the Panseq program (`http://lfz.corefacility.ca/panseq/`) with the following settings: ‘minimumNovelRegionSize’ = ‘500’, ‘fragmentationSize’ = ‘500’, ‘nucB’ = ‘200’, ‘nucC’ = ‘50’, ‘nucD’ = ‘0.12’, ‘nucG’ = ‘100’, ‘nucL’ = ‘20’, ‘percentIdentityCutoff’ = ‘85’, ‘coreGenomeThreshold’ = ‘3’ [127]. A maximum-likelihood tree with aLRT branch support values was created using PhyML as described above [392].

6.3.4 Quantification of Stx$_2$ Production

Three Stx$_2$-producing strains were used for the quantification of Stx$_2$ production; two were *E. coli* O157:H7 strains associated with large outbreaks (strains Sakai [12] and EDL933 [13]) and were obtained from the American Type Culture Collection, and one a travel-associated isolate from the German 2011 *E. coli* O104:H4 outbreak (strain ON-2011 [425]). The amount of Stx_2 produced by each strain was quantified as previously described using an enzyme-linked immunosorbent assay (ELISA) [370], with the following modification: cells were lysed by incubation with 0.5 mg/ml polymyxin B (Sigma) at 37C for 60 min, rather than 1.5 mg/ml for 5 min. The values of three independent experiments, each with two replicates, were used to determine the average toxin production for each strain.

6.3.5 Antibiotic Induction

Overnight cultures of each strain from single colonies were inoculated in 1 ml of BHI medium and incubated at 37C with shaking at 150 rpm in 2 ml microfuge tubes for 17 hours. The overnight culture was diluted 1:100 in fresh BHI medium and incubated at 37C with shaking at 150 rpm for 3 hours after which optical density at 600 nm was measured. Cultures were diluted in two serial dilutions by factors of two to reach final optical densities 0.6. Antibiotic concentrations of 0.015625 g/ml ciprofloxacin and 0.125 g/ml norfloxacin were used for the induction of EDL933; 0.0625 g/ml ciprofloxacin and 0.125 g/ml norfloxacin were used for the induction of Sakai; 0.0625 g/ml ciprofloxacin and 0.25 g/ml norfloxacin were used for the induction of ON-2011. A concentration of 0.5 g/ml mitomycin was used for all three strains. Cultures were incubated at 37C with shaking at 150 rpm for 18 hours, which Vareille et al. (2007) found to be critical for maximum Stx_2 expression [426]. Optical density at 600 nm was measured with ten-fold dilutions of culture.

6.3.6 RNA Isolation

One ml of overnight culture was transferred to RNase-free 1.5 ml microfuge tubes and spun for 5 minutes at 13000 rpm in a Legend Micro 21 R (Sorvall). Following centrifugation, 990 l supernatant was removed, filtered using 0.20 um filters (Corning), and stored at -20C for use in ELISA. The bacterial pellet was dissolved in 1.0 ml of RNAprotect Bacteria Reagent (Qiagen) and incubated for

5 min at room temperature (RT) and subsequently centrifuged for 10 minutes at 5000 g to remove supernatant. The pellet was re-suspended in 100 l TE buffer (10 mM Tris:1 mM EDTA pH 8.0) with 1 mg/ml lysozyme, and incubated for 5 minutes at RT, with vortexing in 2 minute intervals. The RNeasy Mini Kit (Qiagen) was used to complete the RNA extraction and eluted RNA was stored at 80C.

6.3.7 cDNA synthesis

Reverse transcription of isolated RNA was performed using 1 g RNA. The genomic DNA elimination reaction consisting of 18 l template RNA in DNase, RNase-free water (Qiagen) and 3 l of 7 gDNA Wipeout Buffer from the QuantiTect Reverse Transcription Kit (Qiagen) was incubated for 2 minutes at 42C in a Mastercycler pro (Eppendorf). The reverse transcription reaction was conducted in a reaction mixture containing the 21 l gDNA elimination reaction, 6 l of 5X Quantiscript RT Buffer (Qiagen), 1.5 l RT Primer Mix (Qiagen), and 1.5 l Quantiscript Reverse Transcriptase. The mixture was incubated in a Mastercycler pro (Eppendorf) using a 2-stage program: 15 minutes of incubation at 42C followed by 3 minutes at 95C. The cDNA was stored at 20C until needed.

6.3.8 Detection of gnd and stx_2 Gene Expression by RT-PCR

RT-PCR reactions were performed using a Rotor Gene Q (Qiagen) in reaction volumes of 25 l consisting of: 9.25 l DNase, RNase-free water (Sigma), 12.5 l PerfeCTa FastMix II (Quanta Biosciences), 1 l of standard or cDNA, 7.5 pmol of each primer and probe. The probes were conjugated at the 5 end with fluorescent reporter dyes 6-carboxyfluorescein (FAM) and proprietary fluorophore VIC for *stx_2* and *gnd* probes respectively. Both probes were also conjugated at the 3end with the Black Hole Quencher BHQ-1 (Alpha DNA). The *gnd* housekeeping gene of *E. coli* was used to normalize values between RNA samples. Standard curves were generated using three concentrations

(0.1 ng/l, 0.01 ng/l, and 0.001 ng/l) of plasmids $pStx_{2R}$-1 and pAY01 containing *stx₂* and *gnd* genes respectively. The RT-PCR cycling conditions used were: a hold for 2 minutes at 55C preceding a hold for 2 minutes at 95C and 45 cycles of 15 seconds at 95C followed by 45 seconds at 60C. Gain optimization was set to the first tube for both Green (FAM) and Yellow (VIC) Channels and acquisition occurred during the 60C step in cycling. The values of three independent replicates were used to determine average mRNA copy numbers.

6.3.9 Statistical Analyses

Students unpaired two-tailed t-test in Microsoft Excel was used to test for significance the differences between Stx_{2-}toxin production and Stx_{2-}mRNA production among strains.

6.4 Results

The neighbor-net in Figure 6.1 depicts the Stx_{-p}hage distribution among STEC strains and *Shigella dysenteriae*. The phage in *S. dysenteriae* was most distantly related to those in the rest of the strains. The *E. coli* O157:H7 Stx_{2-}bacteriophage were clustered based on both genetic lineage and Stx_2 / Stx_{2c}-toxin sub-type. All Stx_{2c}-bacteriophage from *E. coli* O157:H7 lineage I/II strains were from hosts that also contained a second integrated Stx_{2-}bacteriophage. The *E. coli* O104:H4 outbreak Stx_{2-}phage were situated as a separate cluster, nearest to the Stx_{2-}phage from *E. coli* O111:H- strain JB1-95.

The maximum-likelihood tree of the same data used to construct Figure 6.1 depicts the relationship among the Stx_{2-}phage in tree form (Figure 6.2). Most Stx_{2-}phage grouped together with other Stx_{2-}phage isolated from bacteria of the same serotype, with the exception of *E. coli* serotype O111 strains. Again, the O157:H7 Stx_{2-}bacteriophage were clustered based on both genetic lineage and Stx_2 / Stx_{2c}-toxin sub-type. The *E. coli* O104:H4 outbreak isolate Stx_{2-}phage are most closely situated on the tree to the Stx_{2-}phage from *E. coli* O111:H- strain JB1-95, and more distantly to Stx_{2-}phage from other O111, O103 and O91 serotypes.

As Stx_{2}-phage are mobile, we compared the relationships among the host bacterial genomes to determine the concordance of the Stx_{2}-phage with that of their host-genome (Figure 6.3). The groupings of Stx_{2}-phage and host Stx_{2}-genomes were surprisingly similar, with the notable exception of Stx_{2}-phage and bacterial genomes from serogroup O111, which formed a discrete cluster in the whole-genome tree, but were more widely distributed in the Stx_{2}-phage tree. All serotypes in the whole-genome tree of Figure 6.3 formed discrete clusters. The tree was divided into two separate branches, with O157:H7 strains comprising one branch, and all other STEC strains the other. Within the large O157:H7 branch of Figure 6.3, three sub-groups that corresponded to genetic lineages I, I/II and II were observed. O145:H2 strain 4.0967 and O103:H2 strain 12009 formed a distinct group in contrast to the Stx_{2}-bacteriophage tree (Figure 6.2), where the O145:H2 strain 4.0967 Stx_{2}-phage was part of the O157:H7 lineage I Stx_{2} group and O103:H2 strain 12009 Stx_{2}-phage was most closely grouped to O111:H- strain 11128.

We wished to more thoroughly examine the relationships of the entire, intact Stx_{2}-phage sequences from those that most closely resembled the Stx_{2}-phage from the *E. coli* O104:H4 outbreak isolates. Figure 6.4 shows the Mauve aligned complete Stx_{2}-phage sequences of a selection of Stx_{2}-bacteriophage that were closely related in Figures 6.1 and 6.2. There were common conserved regions among the strains, and Stx_{2}-phage such as O103:H2 strain 12009 and O157:H7 71074 had what appeared to be a very similar phage structure to the O104:H4 outbreak phage, despite not being as evolutionarily conserved at the nucleotide level, as demonstrated by the phylogenetic analyses of this study. Conversely, the O111:H- strain JB1-95 Stx_{2}-phage was grouped most-closely to the O104:H4 Stx_{2}-phage cluster in Figures 6.1, 6.2 and 6.5, but showed the absence of a phage region beginning at the 45 kb region of its phage that was conserved in all the Stx_{2}-phage except O111:H- strain 11128. The O111:H- strain 11128 Stx_{2}-phage was the least similar to the O104:H4 outbreak Stx_{-2} phage among those examined in Figure 6.4.

Figure 6.5 shows the maximum likelihood phylogram of the sequences depicted in Figure 6.4 after MAFFT alignment; the alignment file is available as Dataset S1. This ML tree showed the *E. coli* O111:H- strain JB1-95 Stx_{2}-bacteriophage to be most closely situated to the Stx_{2}-bacteriophage from the German *E. coli* O104:H4 outbreak isolates, and the O111:H- strain 11128 to be most distantly related among the strains in Figure 6.5. The Stx_{2}-phage from O111:H- strain JB1-95 was located on the tree between the O104:H4 outbreak cluster and a cluster of Stx_{2}-phage from *E. coli* O157:H7 lineage I/II strain 71074, O103:H2 strain 12009, and O111:NM strain OK1180.

Lastly, we compared the Stx_{2}-production by *E. coli* O104:H4 outbreak-related strain ON-2011 with that of two other outbreakassociated strains of *E. coli* O157:H7, EDL933 and Sakai. As shown in Figure 6.6, Stx_{2}-production of ON-2011 was significantly lower than that of the two O157:H7 strains ($P < 0.01$). However, after the addition of mitomycin C, Stx_{2}-production by ON-2011 was significantly greater than for both of the *E. coli* O157:H7 strains ($P < 0.01$).

Stx_{2}-mRNA levels were also compared between O104:H4 strain ON-2011 and O157:H7 strain EDL933 after induction by ciprofloxacin, norfloxacin and mitomycin C (Figure 6.7). All three treatments increased Stx_{2}-mRNA production in the strains, with the mitomycin C treatment causing the largest increase in Stx_{2}-mRNA production by both strains. The amount of Stx_{2}-mRNA produced by ON-2011 after mitomycin C treatment was significantly greater than that produced by EDL933 ($P < 0.05$).

6.5 Discussion

6.5.1 Pan-genomic Comparison

The *E. coli* O104:H4 outbreak strain is thought to be derived from an *E. coli* O104:H4 progenitor that recently acquired the Stx_{2}-phage [422]. Seven isolates of this outbreak strain were used in this study, but a detailed description of differences among these and other isolates can be found in the study by Grad et al. (2012) [427]. Stx_{p}hage are temperate lambdoid phage, which exist as an ordered set of interchangeable modules that have a high propensity to recombine into novel mosaic configurations in the laboratory [227]. As Stx_{2}-phage can be readily horizontally transferred, the evolutionary history of the phage and its bacterial host would not be expected to be highly concordant. We wished to identify the Stx_{2}-phage most closely related to the *E. coli* O104:H4 outbreak Stx_{2}-phage, by considering the bacteriophage sequence independently of that of the host bacteria.

Given the potential for high levels of recombination among Stx_2-bacteriophage, we initially examined the phylogenetic relationship of Stx_2-phage from 51 *E. coli* strains and *Shigella dysenteriae* using a network, where competing phylogenetic signals would be shown. We also chose to limit the Stx_2-phage pan-genome to loci present in at least three strains, as loci present in two or fewer are typically not informative phylogenetically [428].

Brzuszkiewicz et al. [340] had previously found that the VT2phi_272 O157:H7 bacteriophage from *E. coli* strain 71074 was closely related to the Stx_2-phage from the *E. coli* O104:H4 outbreak strain. This was particularly interesting to us, as strain 71074 is known to be from a clade of *E. coli* O157:H7 lineage I/II strains that has been termed hyper-virulent due to its more frequent association with severe human infections compared to other *E. coli* O157:H7 genotypes [71, 70]. However, it was unclear how many Stx_2-phage sequences had been examined in the Brzuszkiewicz et al. [340] study, so we repeated the analysis using 51 Stx_2-phage sequences and the outgroup *Shigella dysenteriae* phage sequence.

Both the neighbour-net and ML-tree examination of the pan-genomic data grouped the Stx_2-phage from the O104:H4 outbreak isolates as a cluster, with the common closest phage being from O111:H- strain JB1-95. While the phage from O157:H7 lineage I/II strains bear a structural similarity to the phage from the O104:H4 outbreak strain, our study suggests the possibility of a lateral gene transfer event from an STEC host more closely related to JB1-95 than any of the O157:H7 strains.

6.5.2 Evolutionary Concordance of Stx_2-bacteriophage and Host-genomes

Despite the expectation that Stx_2-phage would be highly heterogeneous, most were highly concordant with the phylogenetic trees of their bacterial hosts, suggesting that the host and phage had stably co-evolved for significant periods of time. This suggests that the lysogen offers an advantage to its host in the natural environment of the bacteria that is greater than the evolutionary cost of the bacterium continually copying the DNA of the foreign bacteriophage genome. The natural reservoir of most Stx$_{2\text{-}}$producing bacteria is the ruminant intestine [429]; therefore, the evolutionary forces acting to select for particular traits are those that confer an advantage in the ruminant host, or the 'source' habitat [178]. Humans are largely incidental hosts, with infection and clearance lasting only a few weeks and represent a 'sink' habitat. In this source / sink evolutionary model, the source habitat selects for long-term evolutionary change and the sink habitat allows proliferation over a limited time-scale, with eventual clearance.

In the case of Stx$_{,}$ it may aid in the colonization of the gut [270] and has also been suggested to defend the population against grazing protozoa that inhabit the bovine intestinal tract [222]. This resistance to predation may also facilitate increased environmental persistence and re-uptake by ruminants, allowing a single-clone to dominate in a herd, as has been shown in *E. coli* O157:H7 [242, 66]. The increased environmental persistence if associated with toxin production would increase the risk of human infection by increasing the probability of exposure. Once infected, the amount of toxin produced may also help the pathogen survive the human immune response, as Stx $_{\text{p}}$roduction has been shown to allow proliferation of *E. coli* O157:H7 within human macrophages [234]. Additionally, immunity to predatory bacteriophages in the rumen gut could also select for the persistence of lysogenic Stx$_{\text{-b}}$acteriophage in *E. coli* hosts.

We observed in the whole-genome tree of Figure 6.3 that serotypes invariably clustered together, suggesting that the discordance between the Stx$_{2\text{-}}$phage and *E. coli* host whole genome trees was caused by recent acquisition (horizontal transfer) of the Stx$_{2\text{-}}$phage. The evolution of pathogenicity in *E. coli* is thought to be driven by events such as phage-mediated horizontal gene transfer and our results suggest that once a phage has been acquired, it can be stably propagated along with the bacterial lineage [430]. This is most evident in the case of the O157:H7 strains, where the whole-genomic sequences in Figure 6.3 are clearly distributed among the three lineages, and the Stx$_{2\text{-}}$phage also group according to lineage (Figures 6.1 and 6.2), even in the case where a single strain has two integrated Stx$_{2\text{-}}$phage (Stx$_{2}$ and Stx$_{2c}$). The grouping of non-O157:H7 Stx$_{2\text{-}}$phage with O157:H7 Stx$_{2\text{-}}$phage on the same branch of the tree, as was the case for the O145:H2 strain and O157:H7 lineage I Stx$_{2\text{-}}$phage, likely indicates horizontal transfer of the phage between organisms. In the case of the O111 Stx_{2}-bacteriophage, their wider distribution among STEC strains also suggests at least one horizontal transfer event.

6.5.3 Stx_{2}-bacteriophage Sequence Alignment

E. coli O111:H- strain JB1-95 was found to be more closely related to the *E. coli* O104:H4 outbreak isolates than the previously reported Stx_{2}-phage from O157:H7 lineage I/II strain 71074 (VT2phi_272) and among the other Stx_{2}-phage sequences examined.

A plausible Stx_{2}-phage evolutionary scenario involves an STEC O111:H- like common ancestor to O111:H- JB1-95 and the *E. coli* O104:H4 outbreak isolates undergoing a lateral transfer event to the O104:H4 outbreak progenitor strain. This would account for the similar evolutionary histories of the Stx_{2}-phage, as well as account for the region missing from the JB1-95 Stx_{2}-phage sequence but present in all other closely related strains, which seems to indicate a loss of genetic information in the O111:H- branch. However, it should be noted that some Stx_{2}-phage sequences were incompletely recovered from the draft whole-genome sequence as a result of the Stx_{2}-phage sequence being distributed among multiple contigs or the absence of Stx_{2}-phage sequence from the draft-genome. Assembled genome sequences for all of these strains would be helpful in more precisely defining these relationships. The proposed evolutionary history of the *E. coli* O104:H4 bacterial ancestor giving rise to the 2011 *E. coli* O104:H4 Stx_{2}-containing strain and EAEC strain 55989 has recently been presented by Mellmann et al. [422].

Stx_{2}-phage are known to have an integration site preference in host bacteria [431]. Because this integration is dependent on the host genome, the site occupied by an integrated Stx_{2}-phage can be used to discriminate among members of the same species [356]. The fact that the Stx_{2}-bacteriophage in the O104:H4 outbreak strain is inserted in the *wrbA* locus and other *E. coli* strains within serotypes O111:H- and O157:H7 possess Stx_{2}-phage that are significantly different from those of the *E. coli* O104:H4 outbreak isolates inserted into other loci, leads us to conclude that the Stx_{2}-phage has been introduced multiple times within STEC through horizontal transfer.

The parallel evolution of this and other STEC virulence factors has been previously described

[76]. For example, evidence suggests the locus of enterocyte effacement (LEE) has been obtained through horizontal transfer multiple times in *E. coli*, giving rise to two distinct groupings of Stx$_{\text{-n}}$egative enteropathogenic *E. coli* (EPEC) and Stx$_{\text{-p}}$ositive enterohemorrhagic *E. coli* (EHEC) strains that share a LEE insertion site [432]. EPEC typically cause diarrhea in humans, while EHEC cause hemorrhagic colitis and HUS. These findings and the results of this study support the idea that multiple genomic backgrounds in *E. coli* are able to persist in the human gastrointestinal tract, and in some cases cause illness through a variety of distinct mechanisms; however, it is the Shiga toxin, and in particular Stx$_2$ that is responsible for severe forms of human disease such as HUS. It is now clear that Stx$_{2\text{-}}$phage are capable of integrating into diverse *E. coli* genomic backgrounds and of qualitatively transforming health risks associated with *E. coli* strains from other pathogroups such as EPEC and EAEC.

6.5.4 Stx$_{2\text{-}}$toxin Production

E. coli O157:H7 clade 8 lineage I/II strains are more frequently associated with HUS than other *E. coli* O157 genotypes and the incidence of human infection with clade 8 strains is thought to be increasing [70]. It is interesting to note that the Stx$_{2\text{-}}$phage from these 'hyper-virulent' *E. coli* O157 strains bear a striking structural similarity to the Stx$_{2\text{-}}$phage from the *E. coli* O104:H4 outbreak isolates. The *E. coli* O104:H4 outbreak isolates are also associated with high rates of HUS, suggesting that these phage share attributes that contribute to increased virulence through mechanisms such as increased expression of Stx$_{2.}$ We have previously shown that lineage I strains of *E. coli* O157:H7 produce significantly different amounts of Stx$_2$ *in vitro* than lineage II strains, and that within lineage I, strains from humans produce significantly more Stx$_2$ than strains from cattle [220]. In the current study, we found that an *E. coli* O104:H4 outbreak-related isolate (ON-2011) produced significantly more Stx$_2$ and Stx$_{2\text{-}}$mRNA than outbreak-related lineage I *E. coli* O157:H7 strains after mitomycin C induction. However, without mitomycin C added to the culture, O104:H4 strain ON-2011 produced very little toxin. Stx$_2$ is an extra-cellular toxin and has been shown to be released from STEC via two specific mechanisms [433]. Therefore, it is unlikely that the high level of Stx_2 produced by the mitomycin C-treated *E. coli* O104 strain was the result of differential toxin release from cell membranes by polymyxin B; however, this possibility cannot be excluded without further study.

A recent study examining the effect of antimicrobials on isolates from the German O104:H4 outbreak found that ciprofloxacin significantly increased both Stx_2 and Stx_{2}-mRNA production [434], which corroborates our current findings and suggests that the Stx_{2}-phage from the *E. coli* O104:H4 outbreak strains is induced and undergoes lytic conversion in vivo in a significant proportion of the bacterial population in the presence of both ciprofloxacin and mitomycin C.

While it can be argued that conditions in the gastrointestinal tract could induce a similar response to that obtained with mitomycin C, further study is required to determine if Stx_2 production by *E. coli* O104:H4 is greater or lesser than for *E. coli* O157 strains in vivo, particularly since other work has shown down-regulation of toxin-production by *E. coli* O157:H7 strains due to microbiota secreted factors present in the human gut [435].

It is also unknown whether the high-prevalence of serious human disease caused by the O104:H4 outbreak isolates and other Stx_{2}-producing bacteria is due simply to human exposure to high-numbers of the organism, or whether the production of high-levels of Stx_2 are responsible for the high-levels of serious human disease. Credence is given to the first hypothesis by the ‘supershedder’ phenomenon in cattle, where it has been reported that phage type (PT) 21/28 of *E. coli* O157:H7 is shed in significantly higher numbers by cattle than other PTs, and PT 21/28 is also the most commonly associated with serious human disease, suggesting that higher levels of human exposure to this PT account for its increased association with human disease [242]. Conversely, it may be that certain strains are more virulent in humans, simply by virtue of the amount of toxin or other virulence factors they produce. In theory, it would only take one such organism to establish and proliferate within the human intestine to cause severe human illness. Factors allowing the bacteria to withstand the hardships of the human gut and immune system, such as acid resistance and toxin-production could significantly influence disease outcomes. We have shown that in cultures containing mitomycin C, an O104:H4 strain produced significantly higher levels of Stx_2 than *E. coli* O157:H7 strains. It was also recently shown that O104:H4 outbreak isolates also had a higher number of cells survive pH 2.5 conditions than *E. coli* O157:H7 strain EDL933 [422].

It has also been suggested that the increased virulence of the *E. coli* O104:H4 outbreak strain could in part be related to the EAEC bacterial host genome, which has evolved adaptations for attachment and survival in the human intestine and that these adaptations have facilitated the systemic absorption of Stx$_{2,}$ which in turn increased the risk of developing HUS [416, 290]. If true, it is likely that the acquisition of any one of the many Stx$_{2}$-phage would have led to an increase in virulence of the *E. coli* O104:H4 host. However, the relative importance of Stx$_{2}$-phage heterogeneity and specific Stx$_{2}$-phage features compared to those of the bacterial host in the virulence of the *E. coli* O104:H4 outbreak strain awaits further study.

6.6 Conclusions

It is clear from the German *E. coli* O104:H4 outbreak that future novel combinations of bacteriophage and bacterial host are likely and that their impact on human health could be devastating. According to the source-sink model, attributes such as Stx $_{p}$roduction by STEC from ruminants is thought to provide a selective advantage in the animal reservoir, not in the 'dead end' human host. STEC virulence in humans is therefore not selected for but is an unintended consequence of accidental infection. As EAEC strains are thought to be human-restricted, the acquisition of a Stx$_{2}$-phage from an O111:H- like strain had to occur in an environment where both EAEC and STEC or their phage were concurrently present. This phenotypic jump could have been in the intestine of an animal or human harbouring both *E. coli* pathotypes, or an environment where both human and ruminant feces were present. In the case of the *E. coli* O104:H4 STEC strain, we do not know if the acquisition of this Stx$_{2}$-phage has provided a selective advantage for the bacterium in humans. While the infection was clearly foodborne, it would appear that human to human transmission has not been sustained and no new cases have been reported. The findings of this study suggest that efforts to control all *E. coli* pathogroups associated with enteric disease in both developed and developing countries as well as Stx$_{2}$-phage from both human and animal sources are warranted to prevent the re-occurrence of similar devastating outbreaks.

6.7 Acknowledgements

We would like to thank the Canadian Food Inspection Agency for allowing this research to be conducted at the Animal Diseases Research Institute.

Table 6.1: The Stx$_2$-bacteriophage and genomic sequences used in the current study.

Table 6.2: The Stx$_2$-bacteriophage and genomic sequences used in the current study.

GenBank ID	Bacterial Isolate	Bacteriophage	Source
ABHK	O157:H7 EC4206	Stx$_2$, Stx$_{2c}$	J. Craig Venter Institute
ABHM	O157:H7 EC4042	Stx$_2$, Stx$_{2c}$	J. Craig Venter Institute
ABHO	O157:H7 EC4196	Stx$_2$	J. Craig Venter Institute
ABHP	O157:H7 EC4113	Stx$_2$	J. Craig Venter Institute
ABHQ	O157:H7 EC4076	Stx$_2$	J. Craig Venter Institute
ABHR	O157:H7 EC4401	Stx$_2$	J. Craig Venter Institute
ABHS	O157:H7 EC4486	Stx$_2$	J. Craig Venter Institute
ABHU	O157:H7 EC869	Stx$_{2c}$	J. Craig Venter Institute
ABHW	O157:H7 EC508	Stx$_2$	J. Craig Venter Institute
ABJT	O157:H7 EC4024	Stx$_2$	J. Craig Venter Institute
ABKY	O157:H7 TW14588	Stx$_2$, Stx$_2$	J. Craig Venter Institute
ADUQ	O111:NM OK1180	Stx$_2$	[314]
AE005174.2	O157:H7 EDL933	Stx$_2$	[13]
AERP	O157:H7 1044	Stx$_2$	Virginia Bioinformatics Institute
AERQ	O157:H7 EC1212	Stx$_{2c}$	Virginia Bioinformatics Institute
AERR	O157:H7 1125 ECF	Stx$_2$	Los Alamos National Lab
AEZV	O111:H- JB1-95	Stx$_2$	J. Craig Venter Institute
AEZX	O121:H19 5.0959	Stx$_2$	J. Craig Venter Institute
AEZZ	O128:H2 9.0111	Stx$_2$	J. Craig Venter Institute
AF125520.1	n/a	Bacteriophage 933W	[365]
AFAA	O145:H2 4.0967	Stx$_2$	J. Craig Venter Institute
AFAB	O147:H- 2.3916	Stx$_{2e}$	J. Craig Venter Institute

Table 6.2: The Stx$_2$-bacteriophage and genomic sequences used in the current study.

AFAC	O153:H- 3.3884	Stx$_2$	J. Craig Venter Institute
AFDQ	O91:H21 B2F1	Stx$_2$	University of Maryland, IGS
AFDR	O73:H16 C165-02	Stx$_{2d}$	University of Maryland, IGS
AFDW	Ont:H12 EH250	Stx$_{2d}$	University of Maryland, IGS
AFEA	O139 S1191	Stx$_{2e}$	University of Maryland, IGS
AFOG	O104:H4 TY-2482	Stx$_2$	[436]
AFST	O104:H4 C227-11	Stx$_2$	[290]
AFWB	O104:H4 CS110	Stx$_2$	Health Protection Agency, UK
AFWC	O104:H4 CS70	Stx$_2$	Health Protection Agency, UK
AFWO	O104:H4 GOS1	Stx$_2$	[340]
AFWP	O104:H4 GOS2	Stx$_2$	[340]
AHZF	O104:H4 ON-2011	Stx$_2$	[425]
AP005154.1	n/a	Bacteriophage Stx$_2$_II	[437]
AP010958.1	O103:H2 12009	Stx$_2$	[314]
AP010960.1	O111:H- 11128	Stx$_2$	[314]
BA000007.2	O157:H7 Sakai	Stx$_2$	[12]
CP001164.1	O157:H7 EC4115	Stx$_{2,}$ Stx$_{2c}$	[438]
CP001368.1	O157:H7 TW14359	Stx$_{2,}$ Stx$_{2c}$	[389]
CP002890.1	O147 UMNF18	Stx$_{2e}$	[439]
FM180578.1	n/a	Bacteriophage 2851	[440]
HQ424691.1	O157:H7 71074	Stx$_2$	Public Health Agency of Canada
in-house	O157:H7 LRH6	Stx$_2$	Public Health Agency of Canada
in-house	O157:H7 EC970520	Stx$_{2c}$	Public Health Agency of Canada
NC_003525.1	n/a	Bacteriophage Stx$_2$_I	[24]
NC_007606.1	*Shigella dysenteriae* Sd197	Stx$_1$	[441]

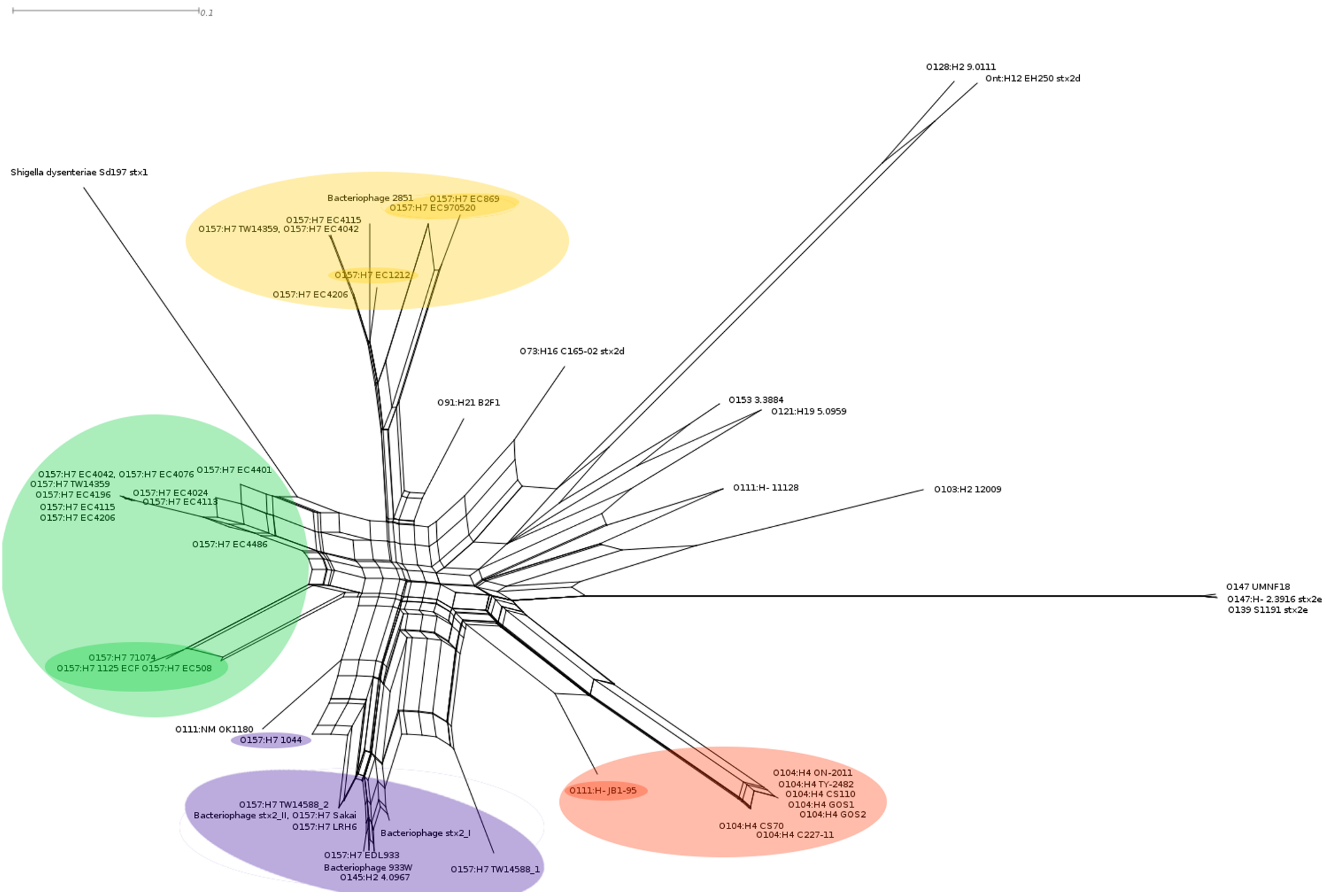


Figure 6.1: The neighbor-net visualization of the Stx$_2$ pan-genome among *E. coli* Stx$_2$-bacteriophage and a bacteriophage from *Shigella dysenteriae* (Table 6.2).

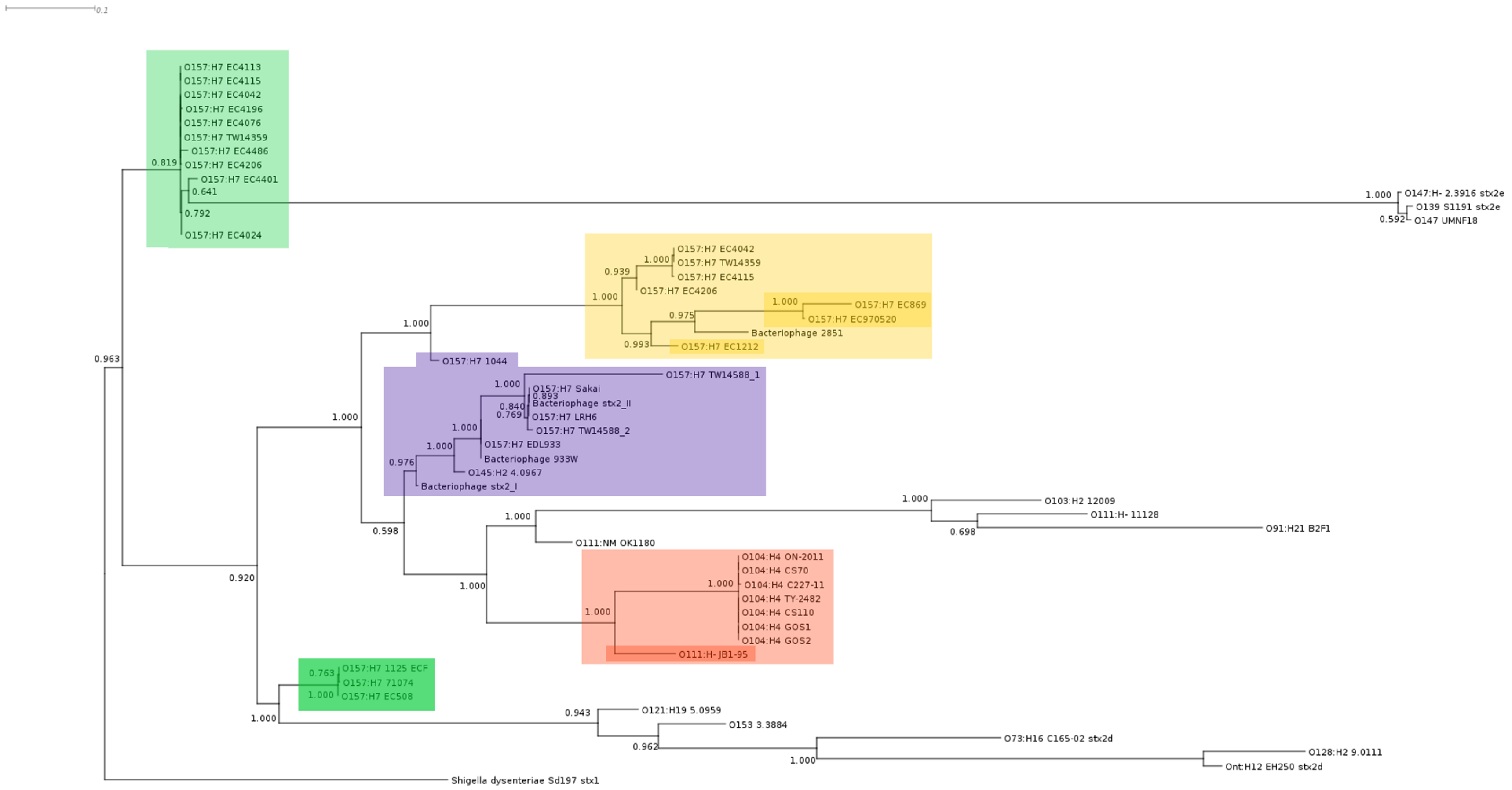


Figure 6.2: Maximum likelihood phylogram with aLRT branch-support values of the Stx$_2$ pan-genome data among *E. coli* Stx$_2$-bacteriophage and a bacteriophage from *Shigella dysenteriae* (Table 6.2).

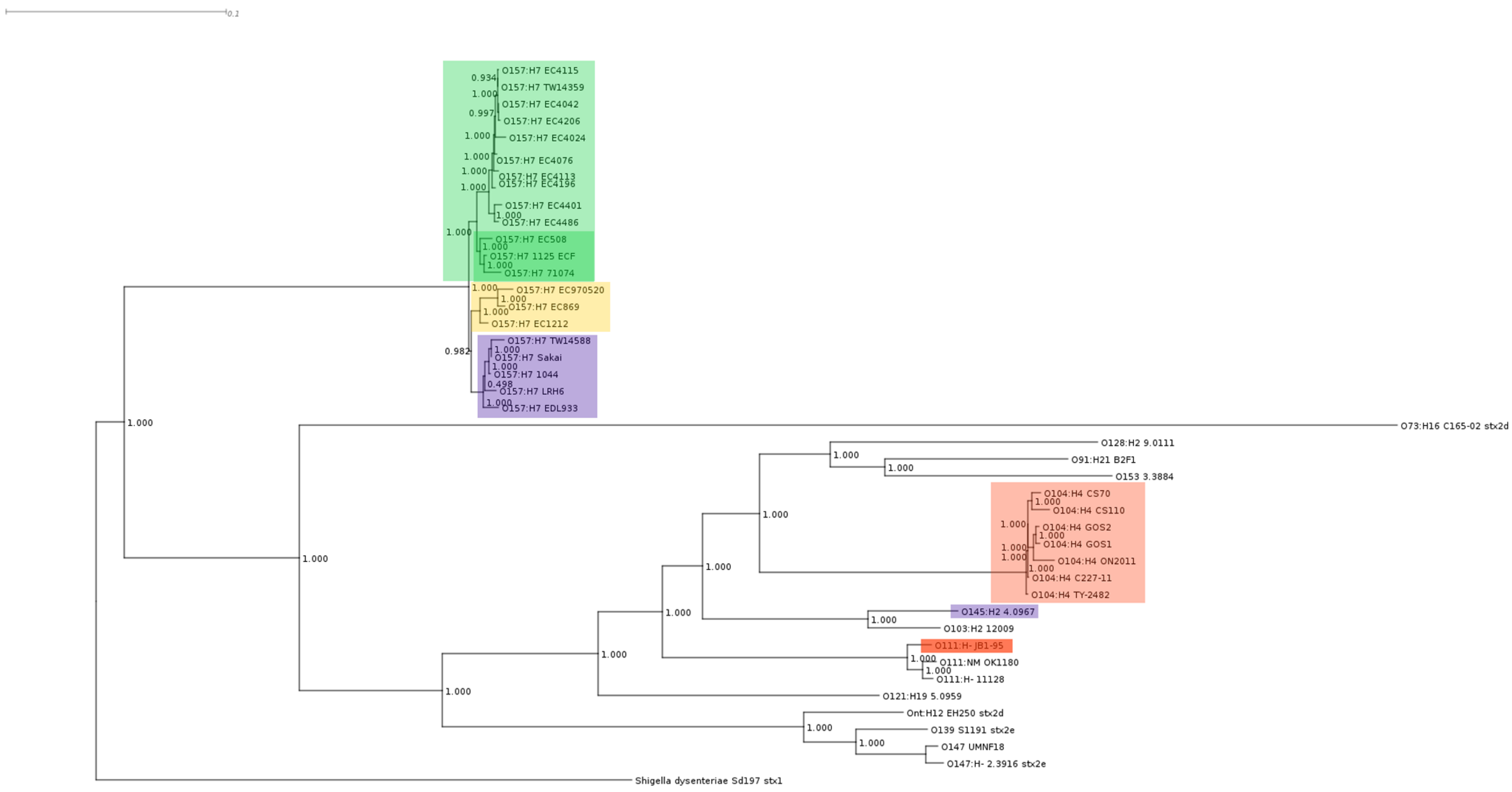


Figure 6.3: Maximum likelihood phylogram with aLRT branch-support values based on the complete pan-genome of the Stx_2-containing genomic sequences in Table 1.

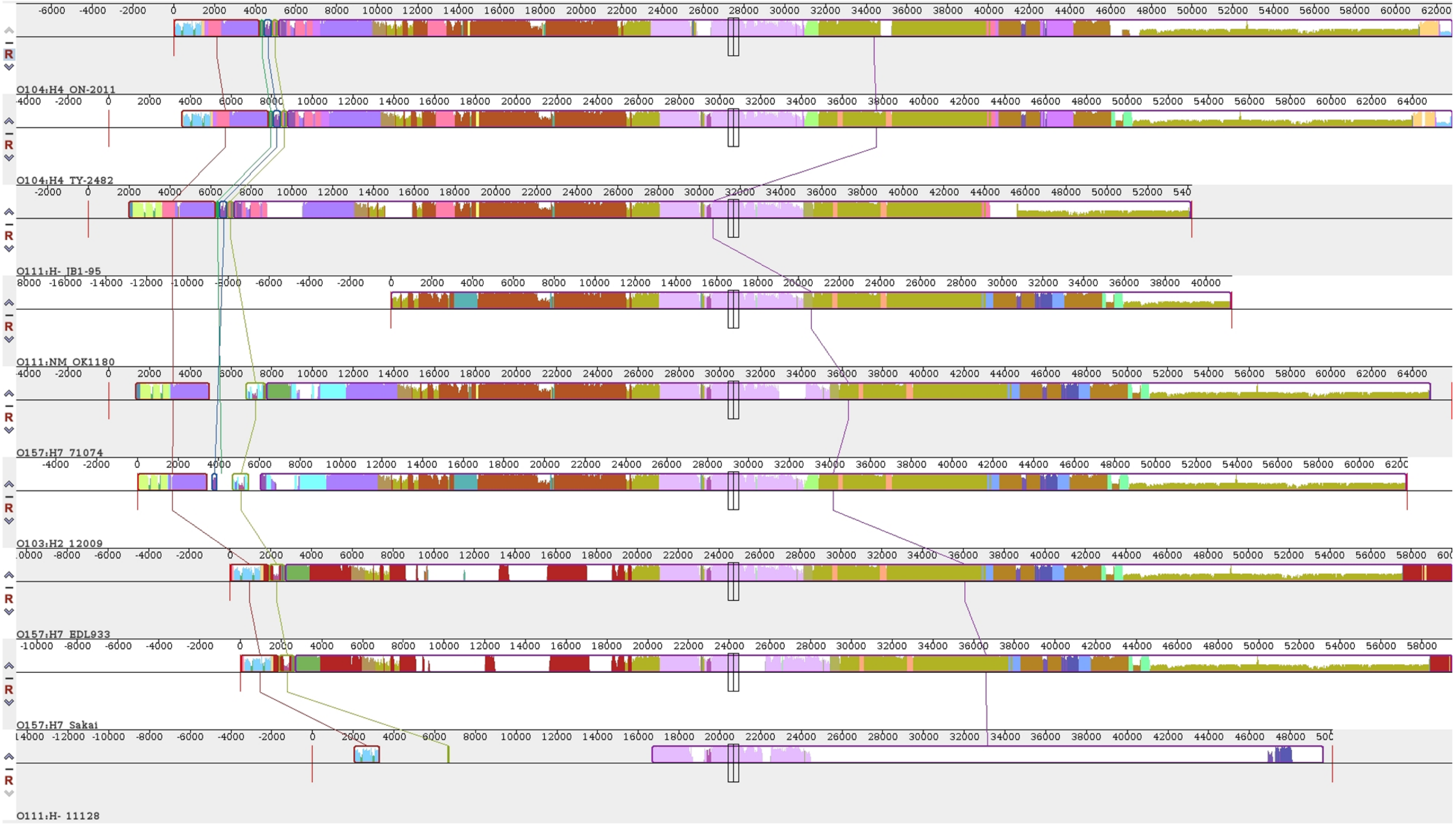


Figure 6.4: The progressiveMauve alignment of the Stx_2-bacteriophage closely related to the O104:H4 outbreak isolates [4]. Bounded boxes indicate similar sequence composition among sequences. The black box indicates the location of the stx_{2a}b gene cluster in each sequence.

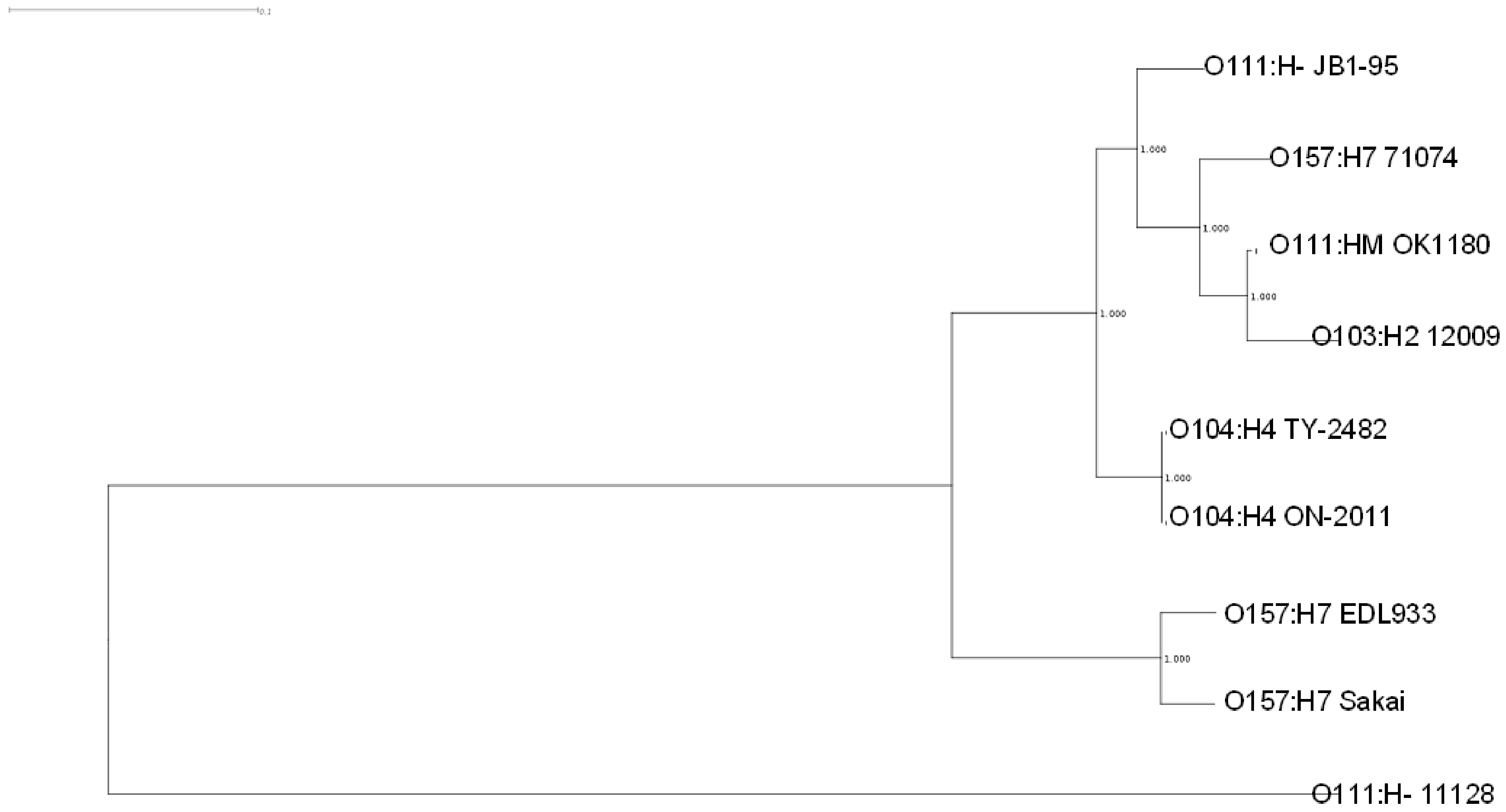


Figure 6.5: Maximum likelihood phylogram with aLRT branch-support values of the MAFFT alignment of the Stx$_2$-bacteriophage closely related to the O104:H4 outbreak isolates.

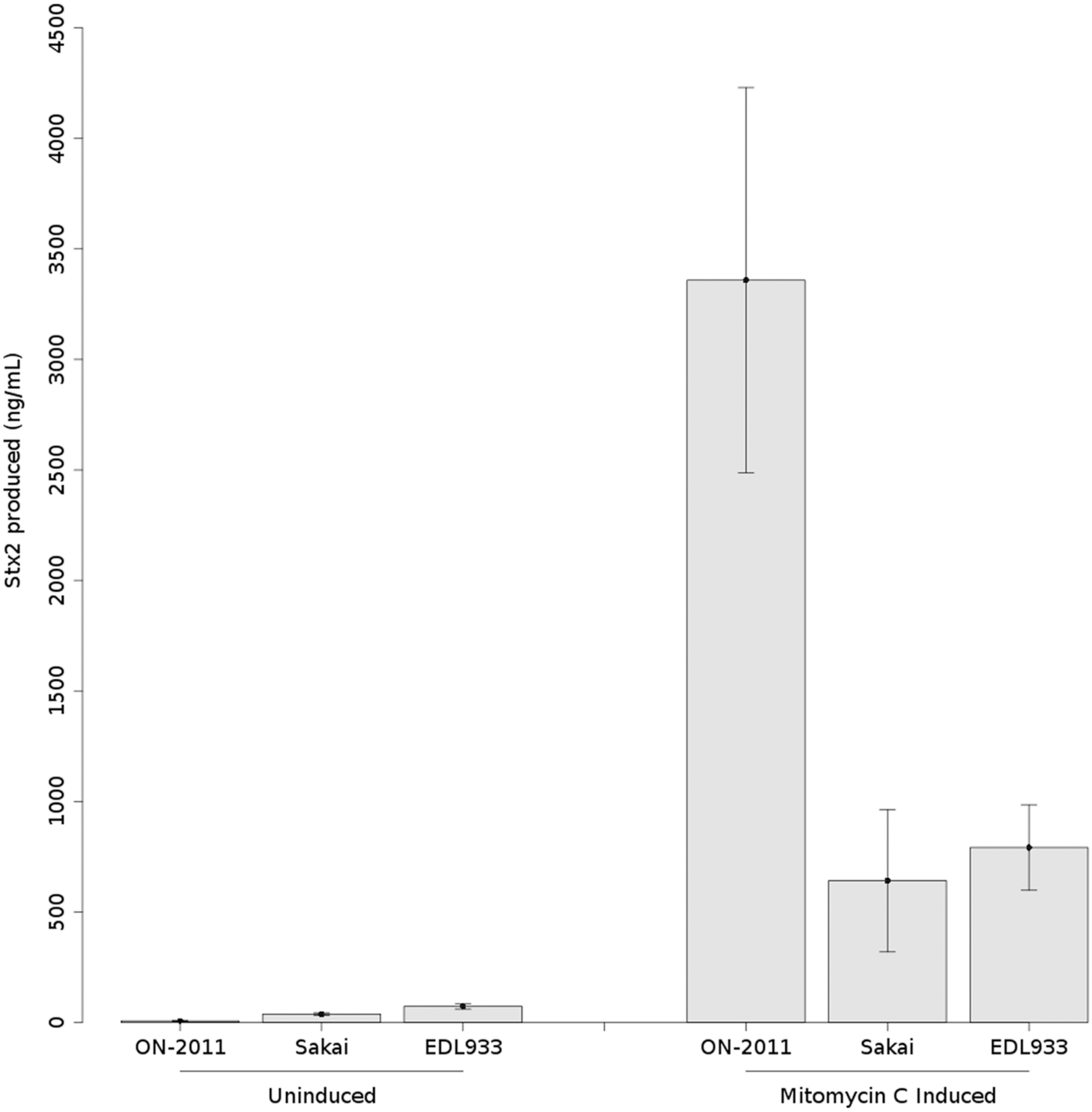


Figure 6.6: Stx_{2}-production by *E. coli* O104:H4 outbreak-related strain ON-2011 and *E. coli* O157:H7 strains EDL933 and Sakai in un-induced, and mitomycin C-induced states as measured by a Stx_{2}-specific ELISA. Error bars represent standard deviations from three independent replicates.

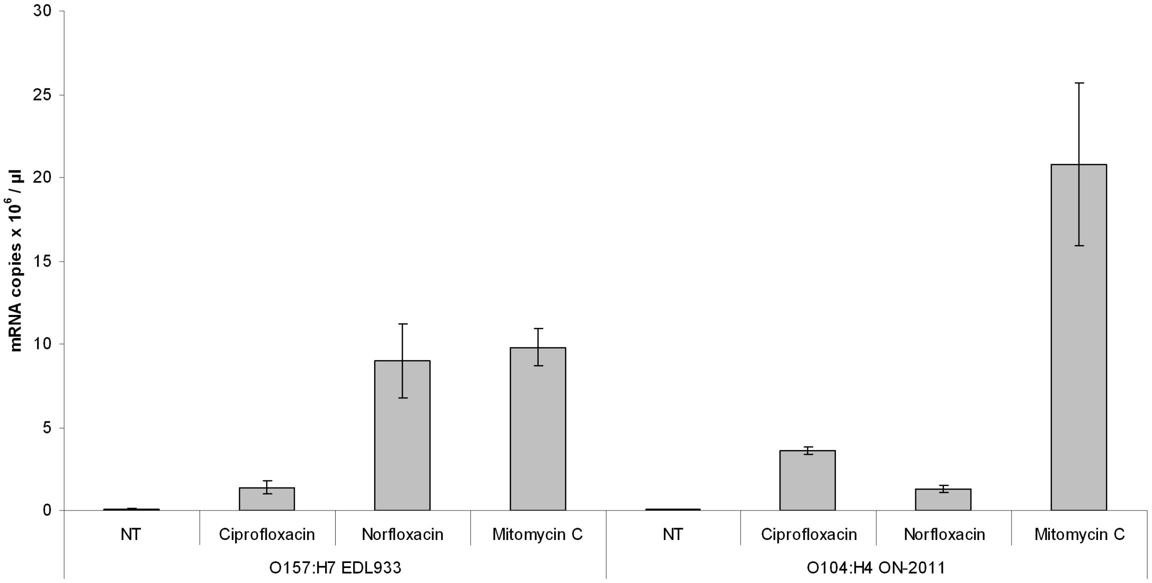


Figure 6.7: Stx$_2$-mRNA copy number differences between *E. coli* O104:H4 outbreak-related strain ON-2011 and *E. coli* O157:H7 strain EDL933 under ciprofloxacin, norfloxacin and mitomycin C treatments. NT = no treatment. Error bars represent standard deviations from three independent replicates.

Chapter 7

Thesis summary

7.1 Introduction

Understanding the evolution and population structure of bacterial pathogens allows us to answer fundamental questions related to human virulence. Infection by the same pathogenic species often results in illness of varying severity, often attributed to the differential carriage and expression of toxins and factors related to adherence, iron uptake and anti-microbial resistance [442]. The phylogenetic distribution of these traits within a population allows one to approach pathogenicity from an evolutionary perspective. This allows the tailoring of interventions such as vaccines to an appropriate lineage, the designing of rapid typing systems to representatively cover an entire population structure, and determining transmission dynamics of both bacteria within human populations, and mobile virulence factors within bacterial populations.

Prior to the advent of low-cost whole-genome sequencing (WGS), phylogeny and population structure of bacteria were inferred using methods ranging from biochemical utilization, banding patterns produced form genomic amplification or restriction digests, or the sequence of a small number of housekeeping genes. While all of these methods have proven useful in the past, the current price of WGS, when run in batch, has fallen from $100 / Mbp in 2008 to $0.10 / Mbp in 2012 Figure 7.1. Thus, while it may not have been feasible to sequence everything of interest a few years ago, benchtop-sequencers are now being targeted to those in the clinical setting, where culture-independent, real-time identification of bacterial pathogens may soon become commonplace [443, 444].

The dramatic drop in sequencing cost is reflected in the progression of this thesis, which started by analyzing molecular methods that sample whole-genome diversity, to designing software for whole-genome analyses of bacterial populations and finally using these tools to answer questions about bacterial populations in a pan-genome context. The pan-genome of a bacterial species consists of a core and an accessory gene pool. The accessory genome is thought to be an important source of genetic variability in bacterial populations and is gained through lateral gene transfer mechanisms such as bacteriophage, which allows subpopulations of bacteria to better adapt to specific niches.

My thesis was broadly constructed around the following objectives: 1) Develop tools for the analyses of bacterial genomes in a pan-genomic context; 2) Describe the phylogeny of *Eschericha coli* O157:H7 and other Shiga toxin-producing *E. coli* (STEC); and 3) Determine if the evolution of the the Shiga-toxin 2 producing bacteriophage parallels that of its bacterial STEC host.

7.2 Major Contributions

7.2.1 Design and Implementation of Pan-genomic Software

Low-cost and high-throughput sequencing platforms have created an exponential increase in genome sequence data and an opportunity to study the pan-genomes of many bacterial species. While originally applied to an entire species, any group of related strains can be said to contain a 'core' and 'accessory' set of genes. As such, tools that extract the new pieces of information from an extremely large pool of data and that can be used to determine the pan-genome and its distribution among strains will be invaluable in the study of genotypic and phenotypic traits in bacterial populations. Despite the myriad program options available, there is no comprehensive package for pan-genome analysis. Prior to the advent of the current generation of sequencing platforms, the number of genome sequences available for intra-species comparative genomic analysis and for the determination of 'accessory' genes was limiting; consequently, tools specifically designed for the analysis of the pan-genome were not available.

In order to fill the need, I created the pan-genome sequence analysis program, Panseq, that extracts novel DNA regions in one in one sequence of group of sequences with respect to another sequence or group of sequences, determines the core and accessory regions of sequences based on sequence identity and segmentation length parameters, creates files based on the core and accessory genome for use in phylogeny programs and determines the most discriminatory and variable set of loci from a dataset. The software allows for the core genome to be user-defined, where an investigator might wish to look for group-dominant (<100%), rather than group-specific (100%) genetic loci in collections of strains (present in many, rather than all of the strains). This feature helps in the analysis of strains that may have lost 'core' elements and are not phenotypically representative of the group, strains that may have lost genetic elements in the laboratory that are maintained by selective pressure in nature, and draft genome assemblies with incomplete sequence coverage. More importantly, it allows new sub-groups to be identified which share a subgroup-specific core that is distinct from the core possessed by the other members of the group of strains. These subgroups are not only likely to share a common ancestry, but to also be phenotypically divergent from other strains in the group, and to be exploiting a different ecological niche.

Using this software I was able to conduct pan-genomic studies using a number of bacterial species as test sets. Initially, I confirmed the identity of *Escherichia coli* O157:H7 and *E. coli* K-12 genomic islands. Following this, novel regions identified in 17 newly sequenced *Escherichia coli* O157:H7 strains were identified and their distribution was determined by PCR among 125 laboratory strains. Additionally the accessory genome and binary presence / absence data, and core genome and single nucleotide polymorphisms (SNPs) of six *L. monocytogenes* strains were identified by Panseq. The nucleotide core and binary accessory data were also used to construct maximum parsimony (MP) trees, which were compared to the MP tree generated by multi-locus sequence typing (MLST). The topology of the accessory and core trees was identical but differed from the tree produced using seven MLST loci.

In order to determine the effect of using a sequence identity cutoff on the size of the accessory genome, I examined groups of *L. monocytogenes*, *E. coli* O157:H7, *Clostridium difficile* and *C. jejuni* genomes with a fragmentation size of 500 bp (approximately half the size of an average gene) over a range of sequence identity cutoffs. The size of the estimated accessory genome increased as the sequence identity threshold was raised; however, this increase was observed to have two distinct phases: an initial linear growth in accessory genome size that was followed by an exponential increase in accessory genome size. The transition between the two stages occurred in the 80 - 90% sequence identity range for each species, suggesting that with values below 80% Panseq primarily identified accessory genome segments that were variably absent or present, whereas above this threshold Panseq identified core genome sequence heterogeneity (SNPs and small indels) among conserved segments .

I generated SNP-containing aligned concatenomes using the CAGF module of Panseq for 39 genomes of *Salmonella enterica* at a threshold of 39 genomes (the core genome) and at one genome (equivalent to the entire pan-genome). I also created aligned concatenomes based on the multi-locus sequence typing (MLST) scheme used to type *S. enterica* strains [3]. I found that both the phylogeny created based on the pan-genome and the phylogeny created based only on the core genome of all 39 *S. enterica* strains were largely concordant. The bootstrap values for the pan-genome trees were significantly higher than the bootstrap values of the core genome trees. Strains of the same *Salmonella* serovar tended to group together in both the core and pan-genome trees, and the overall tree architecture was very similar. Four strains found on branches bearing the lowest support values in the core-genome trees were found on topologically different branches with high branch support on the pan-genome trees. Contrasting this was the MLST tree, where although serovars tended to be clustered together, the overall tree architecture was much different; most notably the human host-restricted strain groups Typhi and Paratyphi A did not form a distinct branch in all of the MLST trees. The branch support successively increased between trees created from MLST, core-genome, and pan-genome loci.

Lastly, I used the Loci Selector module of Panseq to find the most variable and discriminatory combinations of four loci within a 100 loci set among 10 strains in 1s, compared to the 7min required to exhaustively search for all possible combinations. I was also able to identify the 20 most discriminatory loci from a 96 loci *E. coli* O157:H7 SNP dataset.

7.2.2 Population Structure and Phylogeny of E. coli O157:H7 and other STEC

Many approaches have been used to study the evolution, population structure and genetic diversity of *Escherichia coli* O157:H7; however, observations made with different genotyping systems are not easily relatable to each other. Three genetic lineages of *E. coli* O157:H7 designated I, II and I/II have been identified using octamer-based genome scanning and microarray comparative genomic hybridization (mCGH). Each lineage contains significant phenotypic differences, with lineage I strains being the most commonly associated with human infections. Similarly, a clade of hyper-virulent O157:H7 strains (clade 8), implicated in the 2006 spinach and lettuce outbreaks has been defined using single-nucleotide polymorphism (SNP) typing. I performed *in silico* comparison of six different genotyping approaches on the 19 *E. coli* genome sequences from 17 O157:H7 strains and single O145:NM and K12 MG1655 strains to provide an overall picture of diversity of the *E. coli* O157:H7 population, and to compare genotyping methods for O157:H7 strains.

In silico determination of lineage, Shiga-toxin bacteriophage integration site, comparative genomic fingerprint (CGF), mCGH profile, novel region distribution profile, SNP type and multi-locus variable number tandem repeat analysis type was performed and a supernetwork based on the combination of these methods was produced. This supernetwork showed three distinct clusters of strains that were O157:H7 lineage-specific, with the SNP-based hyper-virulent clade 8 synonymous with O157:H7 lineage I/II. Lineage I/II/clade 8 strains clustered closest on the supernetwork to *E. coli* K12 and *E. coli* O55:H7, O145:NM and sorbitol-fermenting O157 strains.

These results highlighted the similarities in relationships derived from multi-locus genome sampling methods and suggested that a ‘common genotyping language’ may be devised for population genetics and epidemiological studies. WGS has now become the genotyping method of choice for these types of studies. The first near real-time sequencing of bacterial pathogens during an outbreak was implemented for the 2008 Canadian *Listeria* outbreak related to processed meat, where two clinical strains were characterized [90]. More recently, WGS was used to determine the source of the Haitian Cholera outbreak [89], and the emergence of a novel human pathogen during the German *E. coli* O104:H4 outbreak [422].

In a later study detailed in Chapter 5, I was able to examine the whole-genome phylogeny of 42 STEC, based on publicly available genomic sequences. I found that all serotypes in the whole-genome tree formed discrete clusters, and that the tree was divided into two separate branches, with O157:H7 strains comprising one branch, and all other STEC strains the other. Within the large O157:H7 branch, three sub-groups that corresponded to genetic lineages I, I/II and II were observed. When put into the context of the entire *E. coli* population, using the same method as outlined in Chapter 5, we arrive at the picture in Figure 7.2, where there are four main 'trunks' in the neighbor-net depiction of the *E. coli* population, one containing O157:H7, O157:NM and O55:H7 strains (EHEC1 group), another of non-pathogenic *E. coli* (Cluster 1), a third containing extra-intestinal (ExPEC) and avian-pathogenic strains (APEC), and finally one comprised of non-O157 STEC. The EHEC1 trunk also contains O145:NM strains and *Shigella dysenteriae*, and the ExPEC / APEC trunk contains branches containing *E. coli* strains from urinary tract infections and pigeons. As can be seen, the four trunks are clearly delineated from one another, supporting previous results that show a clear distinction between O157:H7 strains and STEC of other serotypes.

7.2.3 The evolution of the Shiga-toxin 2 producing bacteriophage and its bacterial STEC host

Stx_2-bacteriophage are mobile, exist as an ordered set of interchangeable modules, and often contain 'passenger genes' inessential to phage function, which can account for large amounts of horizontal-gene transfer among bacteria. Therefore the Stx_2-bacteriophage evolutionary history would not be predicted to mimic that of their bacterial hosts. I analyzed the phylogeny and population structure of 52 Stx_2-bacteriophage from 42 STEC, and compared it to the aforementioned whole-genome STEC population structure.

Despite the expectation that Stx_2-phage would be highly heterogeneous, most were highly concordant with the core-genome phylogenetic trees of their bacterial hosts, suggesting that the host and phage had stably co-evolved for significant periods of time. This suggests that the lysogen offers an advantage to its bacterial host in the natural environment, namely the ruminant intestine and farm environment, which is greater than the evolutionary cost of the bacterium continually copying the DNA of the foreign bacteriophage genome.

The evolution of pathogenicity in *E. coli* is thought to be driven by events such as phage-mediated horizontal gene transfer and the results of this thesis suggest that once a phage has been acquired, it can be stably propagated along with the bacterial lineage [430]. This is most evident in the case of the O157:H7 strains, where the whole-genomic sequences are clearly distributed among the three lineages, and the Stx_2-phage also group according to lineage, even in the case where a single strain has two integrated Stx_2-phage (Stx_2 and Stx_{2c}).

I found that of the 42 whole-genomes examined, only *E. coli* serotype O111 strains, and an O145:H2 strain were placed differently on the whole-genome tree compared to the phylogeny of the Stx_2-phage, likely indicating horizontal acquisition of the phage from bacteria with different evolutionary histories.

During this study the *E. coli* O104:H4 outbreak in Germany had just ended, in which more than 4000 illnesses and 50 deaths were reported. This outbreak was unusual in that a novel Stx_2-producing serotype O104:H4 strain was implicated as the causative agent, and in which HGT is thought to have been responsible for bringing the Stx_2-bacteriophage into a human-adapted pathogen associated with mild gastrointestinal disease. I was able to include the WGS and Stx_2-bacteriophage from the outbreak strain into my analyses and found that among the *E. coli* Stx_2-phage sequences studied, that the Stx_2-bacteriophage from O111:H- strain JB1-95 was most closely related phylogenetically to the Stx_2-phage from the O104:H4 outbreak strain. This makes it clear that Stx2-phage are capable of integrating into diverse *E. coli* genomic backgrounds and of qualitatively transforming health risks associated with *E. coli* strains from other pathogroups such as enteropathogenic *E. coli* and enteroaggregative *E. coli*.

7.3 Conclusions and Future Directions

The number of genomic sequences of microorganisms being generated is increasing exponentially thanks to recent advances in sequencing technology. While the computing resources to store, transfer and analyze a small number of genomes currently exist, we are not able to effectively compare the tens of thousands of bacterial genomes anticipated to soon populate our databases and identify similarities and differences among them. Further, many of the computer programs are stand-alone and dedicated to specific isolated tasks rather than integrated into a logical workflow that meets end-user needs. In many ways the ‘future of biology’ depends on being able to effectively ask and answer questions by leveraging the vast amounts of data now being created. Part of my future goals are to focus on developing computing software to act as a service for those who would like to compare groups of thousands of sequenced bacterial genomes by applying novel algorithms to avoid the re-computation of data, and using methods that will allow newly sequenced bacteria to be easily compared to previously analyzed genomes, in near real time.

I would also like to work on the identification of STEC strain clusters associated with human disease and characterize them in a pan-genomics context. This would enable a more refined method of identification than serotype alone. The core and accessory genomic markers specific to these groups would also likely enhance the ability to detect strains with greater potential for causing human disease. These markers will also be useful for enhancing molecular subtyping schemes used for epidemiological and population studies relating to STEC, allowing greater public health emergency response (outbreak investigations), surveillance, source attribution and risk assessment. The accessory genome loci identified for clusters of human-related strains also offer potential targets for therapeutics such as vaccines.

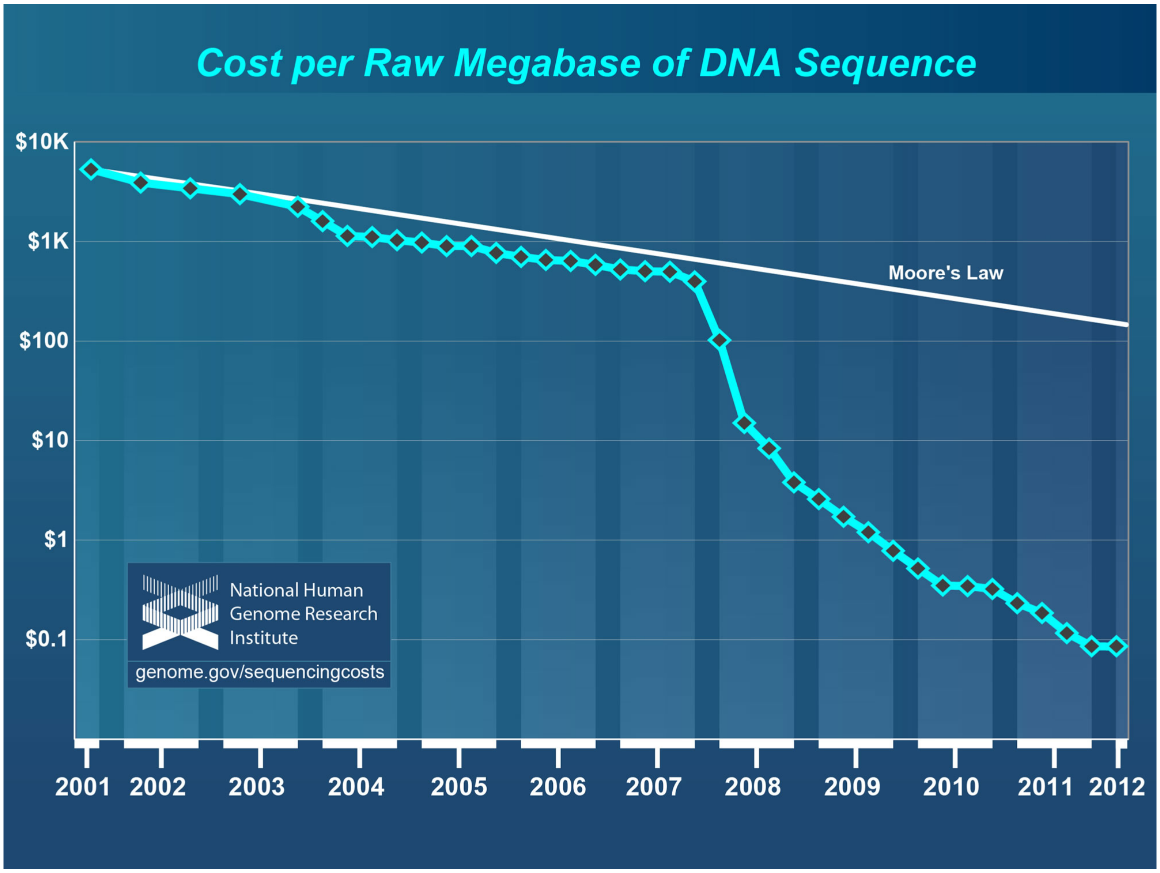


Figure 7.1: DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts

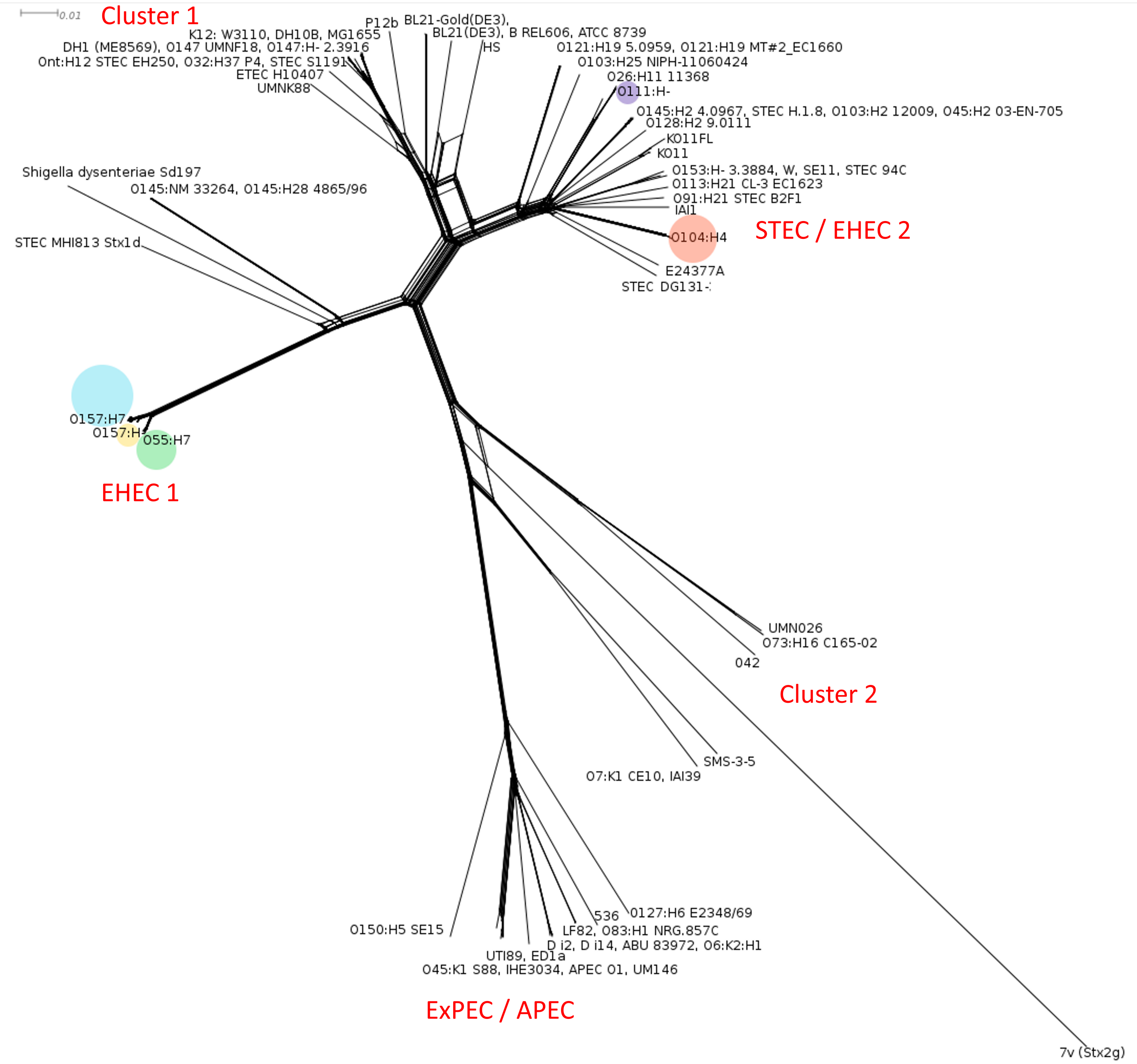

0.01
Cluster 1
P12b
BL21-Gold(DE3),
BL21(DE3), B REL606, ATCC 8739
K12: W3110, DH10B, MG1655
DH1 (ME8569), O147 UMNF18, O147:H- 2.3916
HS
Ont:H12 STEC EH250, O32:H37 P4, STEC S1191
ETEC H10407
UMNK88
O121:H19 5.0959, O121:H19 MT#2_EC1660
O103:H25 NIPH-11060424
O26:H11 11368
O111:H-
O145:H2 4.0967, STEC H.1.8, O103:H2 12009, O45:H2 03-EN-705
O128:H2 9.0111
KO11FL
KO11
O153:H- 3.3884, W, SE11, STEC 94C
O113:H21 CL-3 EC1623
O91:H21 STEC B2F1
IAI1
O104:H4
STEC / EHEC 2
E24377A
STEC DG131-3
Shigella dysenteriae Sd197
O145:NM 33264, O145:H28 4865/96
STEC MHI813 Stx1d
O157:H7
O157:H-
O55:H7
EHEC 1
UMN026
O73:H16 C165-02
042
Cluster 2
SMS-3-5
O7:K1 CE10, IAI39
O127:H6 E2348/69
536
O150:H5 SE15
LF82, O83:H1 NRG.857C
D i2, D i14, ABU 83972, O6:K2:H1
UTI89, ED1a
O45:K1 S88, IHE3034, APEC O1, UM146
ExPEC / APEC
7v (Stx2g)

Figure 7.2: Phylogenetic overview of the *E. coli* species

Bibliography

[1] Nightingale KK, Windham K, Wiedmann M: **Evolution and molecular phylogeny of *Listeria monocytogenes* isolated from human and animal listeriosis cases and foods**. *Journal of bacteriology* 2005, **187**(16):5537–5551. [PMID: 16077098].

[2] Felsenstein J: **PHYLIP- phylogeny inference package (version 3.2)**. *Cladistics* 1989, **5**:164–166.

[3] Kidgell C, Reichard U, Wain J, Linz B, Torpdahl M, Dougan G, Achtman M: ***Salmonella typhi*, the causative agent of typhoid fever, is approximately 50,000 years old**. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 2002, **2**:39–45, [http://www.ncbi.nlm.nih.gov/pubmed/12797999]. [PMID: 12797999].

[4] Darling AE, Mau B, Perna NT: **progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement**. *PloS One* 2010, **5**(6):e11147, [http://www.ncbi.nlm.nih.gov/pubmed/20593022]. [PMID: 20593022].

[5] Laing CR, Zhang Y, Thomas JE, Gannon VPJ: **Everything at once: comparative analysis of the genomes of bacterial pathogens**. *Veterinary microbiology* 2011, **153**(1-2):13–26. [PMID: 21764529].

[6] Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R: **The microbial pan-genome**. *Current Opinion in Genetics & Development* 2005, **15**(6):589–594, [http://www.ncbi.nlm.nih.gov/pubmed/16185861]. [PMID: 16185861].

[7] Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV: **PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, united States**. *Emerg Infect Dis* 2001, **7**(3):382–9, [11384513].

[8] Caetano-Anolls G, Gresshoff PM: *DNA markers: protocols, applications, and overviews*. Wiley-VCH 1997.

[9] Rademaker J, Hoste B, Louws F, Kersters K, Swings J, Vauterin L, Vauterin P, de Bruijn F: **Comparison of AFLP and rep-PCR genomic fingerprinting with DNA–DNA homology studies: xanthomonas as a model system**. *Int J Syst Evol Microbiol* 2000, **50**(2):665–677, [http://ijs.sgmjournals.org/cgi/content/abstract/50/2/665].

[10] Mira A, Martn-Cuadrado AB, D'Auria G, Rodrguez-Valera F: **The bacterial pan-genome:a new paradigm in microbiology**. *International Microbiology: The Official Journal of the Spanish Society for Microbiology* 2010, **13**(2):45–57, [http://www.ncbi.nlm.nih.gov/pubmed/20890839]. [PMID: 20890839].

[11] Alm RA, Ling LS, Moir DT, King BL, Brown ED, Doig PC, Smith DR, Noonan B, Guild BC, deJonge BL, Carmel G, Tummino PJ, Caruso A, Uria-Nickelsen M, Mills DM, Ives C, Gibson R, Merberg D, Mills SD, Jiang Q, Taylor DE, Vovis GF, Trust TJ: **Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori***. *Nature* 1999, **397**(6715):176–180, [http://www.ncbi.nlm.nih.gov/pubmed/9923682]. [PMID: 9923682].

[12] Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C, Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, Shinagawa H: **Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12**. *DNA Res* 2001, **8**:11–22, [http://dnaresearch.oxfordjournals.org/cgi/content/abstract/8/1/11].

[13] Perna NT, Plunkett G, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Posfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, Blattner FR: **Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7**. *Nature* 2001, **409**(6819):529–533, [http://dx.doi.org/10.1038/35054089].

[14] Beres SB, Musser JM: **Contribution of exogenous genetic elements to the group a streptococcus Metagenome**. *PLoS ONE* 2007, **2**(8):e800, [http://dx.plos.org/10.1371/journal.pone.0000800].

[15] Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM: **Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome"**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(39):13950–13955, [http://www.ncbi.nlm.nih.gov/pubmed/16172379]. [PMID: 16172379].

[16] McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P, Courtney L, Porwollik S, Ali J, Dante M, Du F, Hou S, Layman D, Leonard S, Nguyen C, Scott K, Holmes A, Grewal N, Mulvaney E, Ryan E, Sun H, Florea L, Miller W, Stoneking T, Nhan M, Waterston R, Wilson RK: **Complete genome sequence of *Salmonella enterica* serovar typhimurium LT2**. *Nature* 2001, **413**(6858):852–856, [http://www.ncbi.nlm.nih.gov/pubmed/11677609]. [PMID: 11677609].

[17] Whittam TS, Bumbaugh AC: **Inferences from whole-genome sequences of bacterial pathogens**. *Current Opinion in Genetics & Development* 2002, **12**(6):719–725, [http://www.sciencedirect.com/science/article/B6VS0-475R9SR-7/2/46f10ba47ad6272f6774d016f5fd3b6e].

[18] Langille M, Hsiao W, Brinkman F: **Evaluation of genomic island predictors using a comparative genomics approach**. *BMC Bioinformatics* 2008, **9**:329, [http://www.biomedcentral.com/1471-2105/9/329].

[19] Lee CA: **Pathogenicity islands and the evolution of bacterial pathogens**. *Infectious Agents and Disease* 1996, **5**:1–7, [http://www.ncbi.nlm.nih.gov/pubmed/8789594]. [PMID: 8789594].

[20] Read TD, Brunham RC, Shen C, Gill SR, Heidelberg JF, White O, Hickey EK, Peterson J, Utterback T, Berry K, Bass S, Linher K, Weidman J, Khouri H, Craven B, Bowman C, Dodson R, Gwinn M, Nelson W, DeBoy R, Kolonay J, McClarty G, Salzberg SL, Eisen J, Fraser CM: **Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39**. *Nucleic Acids Research* 2000, **28**(6):1397–1406, [http://www.ncbi.nlm.nih.gov/pubmed/10684935]. [PMID: 10684935].

[21] Read TD, Peterson SN, Tourasse N, Baillie LW, Paulsen IT, Nelson KE, Tettelin H, Fouts DE, Eisen JA, Gill SR, Holtzapple EK, Okstad OA, Helgason E, Rilstone J, Wu M, Kolonay JF, Beanan MJ, Dodson RJ, Brinkac LM, Gwinn M, DeBoy RT, Madpu R, Daugherty SC, Durkin AS, Haft DH, Nelson WC, Peterson JD, Pop M, Khouri HM, Radune D, Benton JL, Mahamoud Y, Jiang L, Hance IR, Weidman JF, Berry KJ, Plaut RD, Wolf AM, Watkins KL, Nierman WC, Hazen A, Cline R, Redmond C, Thwaite JE, White O, Salzberg SL, Thomason B, Friedlander AM, Koehler TM, Hanna PC, Kolst AB, Fraser CM: **The genome sequence of *Bacillus anthracis* ames and comparison to closely related bacteria**. *Nature* 2003, **423**(6935):81–86, [http://www.ncbi.nlm.nih.gov/pubmed/12721629]. [PMID: 12721629].

[22] Miao EA, Miller SI: **Bacteriophages in the evolution of pathogen-host interactions**. *Proceedings of the National Academy of Sciences of the United States of America* 1999, **96**(17):9452–4. [PMID: 10449711].

[23] Hooper S, Mavromatis K, Kyrpides N: **Microbial co-habitation and lateral gene transfer: what transposases can tell us**. *Genome Biology* 2009, **10**(4):R45, [http://genomebiology.com/2009/10/4/R45].

[24] Sato T, Shimizu T, Watarai M, Kobayashi M, Kano S, Hamabata T, Takeda Y, Yamasaki S: **Distinctiveness of the genomic sequence of shiga toxin 2-converting phage isolated from *Escherichia coli* O157:H7 okayama strain as compared to other shiga toxin 2-converting phages**. *Gene* 2003, **309**:35–48, [12727356].

[25] Waldor MK, Mekalanos JJ: **Lysogenic conversion by a filamentous phage encoding cholera toxin**. *Science (New York, N.Y.)* 1996, **272**(5270):1910–1914, [http://www.ncbi.nlm.nih.gov/pubmed/8658163]. [PMID: 8658163].

[26] Groman NB: **Conversion by corynephages and its role in the natural history of diphtheria**. *The Journal of Hygiene* 1984, **93**(3):405–417, [http://www.ncbi.nlm.nih.gov/pubmed/6439780]. [PMID: 6439780].

[27] Dahlberg C, Chao L: **Amelioration of the cost of conjugative plasmid carriage in eschericha coli K12**. *Genetics* 2003, **165**(4):1641–1649, [http://www.ncbi.nlm.nih.gov/pubmed/14704155]. [PMID: 14704155].

[28] Kung VL, Ozer EA, Hauser AR: **The accessory genome of *Pseudomonas aeruginosa***. *Microbiology and Molecular Biology Reviews: MMBR* 2010, **74**(4):621–641, [http://www.ncbi.nlm.nih.gov/pubmed/21119020]. [PMID: 21119020].

[29] Song Y, Roumagnac P, Weill FX, Wain J, Dolecek C, Mazzoni CJ, Holt KE, Achtman M: **A multiplex single nucleotide polymorphism typing assay for detecting mutations that**

result in decreased fluoroquinolone susceptibility in *Salmonella enterica* serovars typhi and paratyphi A. *Journal of Antimicrobial Chemotherapy* 2010, **65**(8):1631 –1641, [http://jac.oxfordjournals.org/content/65/8/1631.abstract].

[30] Olsen RJ, Sitkiewicz I, Ayeras AA, Gonulal VE, Cantu C, Beres SB, Green NM, Lei B, Humbird T, Greaver J, Chang E, Ragasa WP, Montgomery CA, Cartwright J, McGeer A, Low DE, Whitney AR, Cagle PT, Blasdel TL, DeLeo FR, Musser JM: **Decreased necrotizing fasciitis capacity caused by a single nucleotide mutation that alters a multiple gene virulence axis**. *Proceedings of the National Academy of Sciences* 2010, **107**(2):888 –893, [http://www.pnas.org/content/107/2/888.abstract].

[31] Fournier PE, El Karkouri K, Leroy Q, Robert C, Giumelli B, Renesto P, Socolovschi C, Parola P, Audic S, Raoult D: **Analysis of the *Rickettsia africae* genome reveals that virulence acquisition in rickettsia species may be explained by genome reduction**. *BMC Genomics* 2009, **10**:166, [http://www.ncbi.nlm.nih.gov/pubmed/19379498]. [PMID: 19379498].

[32] Rondini S, Kser M, Stinear T, Tessier M, Mangold C, Dernick G, Naegeli M, Portaels F, Certa U, Pluschke G: **Ongoing genome reduction in *Mycobacterium ulcerans***. *Emerging Infectious Diseases* 2007, **13**(7):1008–1015, [http://www.ncbi.nlm.nih.gov/pubmed/18214172]. [PMID: 18214172].

[33] Dziva F, Muhairwa AP, Bisgaard M, Christensen H: **Diagnostic and typing options for investigating diseases associated with *Pasteurella multocida***. *Veterinary Microbiology* 2008, **128**(1-2):1–22, [http://www.ncbi.nlm.nih.gov/pubmed/18061377]. [PMID: 18061377].

[34] Morgan MC, Boyette M, Goforth C, Sperry KV, Greene SR: **Comparison of the biolog OmniLog identification system and 16S ribosomal RNA gene sequencing for accuracy in identification of atypical bacteria of clinical origin**. *Journal of Microbiological Methods* 2009, **79**(3):336–343, [http://www.ncbi.nlm.nih.gov/pubmed/19837117]. [PMID: 19837117].

[35] Tannock GW, Fuller R, Smith SL, Hall MA: **Plasmid profiling of members of the family Enterobacteriaceae, lactobacilli, and bifidobacteria to study the transmission of bacteria from mother to infant.** *J. Clin. Microbiol.* 1990, **28**(6):1225–1228, [http://jcm.asm.org/cgi/content/abstract/28/6/1225].

[36] Grishchenko RI, Sakal NN, Zheltva AA: **[Method of colicin typing of Enterobacteriaceae]**. *Laboratornoe Delo* 1976, (8):500–501, [http://www.ncbi.nlm.nih.gov/pubmed/66413]. [PMID: 66413].

[37] Chirakadze I, Perets A, Ahmed R: **Phage typing**. *Methods in Molecular Biology (Clifton, N.J.)* 2009, **502**:293–305, [http://www.ncbi.nlm.nih.gov/pubmed/19082563]. [PMID: 19082563].

[38] Grif K, Karch H, Schneider C, Daschner FD, Beutin L, Cheasty T, Smith H, Rowe B, Dierich MP, Allerberger F: **Comparative study of five different techniques for epidemiological typing of *Escherichia coli* O157**. *Diagnostic Microbiology and Infectious Disease* 1998, **32**(3):165–76, [http://www.ncbi.nlm.nih.gov/pubmed/9884832]. [PMID: 9884832].

[39] Bezanson G, Khakhria R, Lacroix R: **Involvement of plasmids in determining bacteriophage sensitivity in *Salmonella typhimurium*: genetic and physical analysis of phagovar 204**. *Canadian Journal of Microbiology* 1982, **28**(9):993–1001, [http://www.ncbi.nlm.nih.gov/pubmed/6754045]. [PMID: 6754045].

[40] Kameyama L, Fernndez L, Caldern J, Ortiz-Rojas A, Patterson TA: **Characterization of wild lambdoid Bacteriophages: detection of a wide distribution of phage immunity groups and identification of a Nus-Dependent, nonlambdoid phage Group**. *Virology* 1999, **263**:100–111, [http://www.sciencedirect.com/science/article/B6WXR-45FK1RP-49/1/f7296b58737a6f09f1e7560e2ab42669].

[41] Bauer AW, Kirby WM, Sherris JC, Turck M: **Antibiotic susceptibility testing by a standardized single disk method**. *American Journal of Clinical Pathology* 1966, **45**(4):493–6, [http://www.ncbi.nlm.nih.gov/pubmed/5325707]. [PMID: 5325707].

[42] Fricke WF, Wright MS, Lindell AH, Harkins DM, Baker-Austin C, Ravel J, Stepanauskas R: **Insights into the environmental resistance gene pool from the genome sequence of the multidrug-resistant environmental isolate *Escherichia coli* SMS-3-5**. *Journal of Bacteriology* 2008, **190**(20):6779–94, [http://www.ncbi.nlm.nih.gov/pubmed/18708504]. [PMID: 18708504].

[43] Gautom RK: **Rapid pulsed-field gel electrophoresis protocol for typing of *Escherichia coli* O157:H7 and other gram-negative organisms in 1 day**. *Journal of Clinical Microbiology* 1997, **35**(11):2977–80, [http://www.ncbi.nlm.nih.gov/pubmed/9350772]. [PMID: 9350772].

[44] Halpin JL, Garrett NM, Ribot EM, Graves LM, Cooper KL: **Re-evaluation, optimization, and multilaboratory validation of the PulseNet-standardized pulsed-field gel electrophoresis protocol for *Listeria monocytogenes***. *Foodborne Pathogens and Disease* 2010, **7**(3):293–298, [http://www.ncbi.nlm.nih.gov/pubmed/19911934]. [PMID: 19911934].

[45] van Belkum A, van Leeuwen W, Kaufmann ME, Cookson B, Forey F, Etienne J, Goering R, Tenover F, Steward C, O'Brien F, Grubb W, Tassios P, Legakis N, Morvan A, El Solh N, de Ryck R, Struelens M, Salmenlinna S, Vuopio-Varkila J, Kooistra M, Talens A, Witte W, Verbrugh H: **Assessment of resolution and intercenter reproducibility of results of genotyping *Staphylococcus aureus* by pulsed-field gel electrophoresis of SmaI macrorestriction fragments: a multicenter study**. *Journal of Clinical Microbiology* 1998, **36**(6):1653–1659, [http://www.ncbi.nlm.nih.gov/pubmed/9620395]. [PMID: 9620395].

[46] Murchan S, Kaufmann ME, Deplano A, de Ryck R, Struelens M, Zinn CE, Fussing V, Salmenlinna S, Vuopio-Varkila J, El Solh N, Cuny C, Witte W, Tassios PT, Legakis N, van Leeuwen W, van Belkum A, Vindel A, Laconcha I, Garaizar J, Haeggman S, Olsson-Liljequist B, Ransjo U, Coombes G, Cookson B: **Harmonization of pulsed-field gel electrophoresis protocols for epidemiological typing of strains of methicillin-resistant *Staphylococcus aureus*: a single approach developed by consensus in 10 european laboratories and its application for tracing the spread of related strains**. *Journal of Clinical Microbiology* 2003, **41**(4):1574–85. [PMID: 12682148].

[47] Pei Y, Terajima J, Saito Y, Suzuki R, Takai N, Izumiya H, Morita-Ishihara T, Ohnishi M, Miura M, Iyoda S, Mitobe J, Wang B, Watanabe H: **Molecular characterization of enterohemorrhagic *Escherichia coli* O157:H7 isolates dispersed across japan by pulsed-field gel electrophoresis and multiple-locus variable-number tandem repeat analysis**. *Japanese Journal of Infectious Diseases* 2008, **61**:58–64. [PMID: 18219136].

[48] Holmes A, Edwards GF, Girvan EK, Hannant W, Danial J, Fitzgerald JR, Templeton KE: **Comparison of two multilocus variable-number tandem-repeat methods and pulsed-field gel electrophoresis for differentiating highly clonal methicillin-resistant *Staphylococcus aureus* isolates**. *Journal of Clinical Microbiology* 2010, **48**(10):3600–3607, [http://www.ncbi.nlm.nih.gov/pubmed/20702668]. [PMID: 20702668].

[49] Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M: **AFLP: a new technique for DNA fingerprinting**. *Nucleic Acids Research* 1995, **23**(21):4407–14, [http://www.ncbi.nlm.nih.gov/pubmed/7501463]. [PMID: 7501463].

[50] Heir E, Lindstedt BA, Vardund T, Wasteson Y, Kapperud G: **Genomic fingerprinting of shigatoxin-producing *Escherichia coli* (STEC) strains: comparison of pulsed-field gel electrophoresis (PFGE) and fluorescent amplified-fragment-length polymorphism (FAFLP)**. *Epidemiology and Infection* 2000, **125**(3):537–48. [PMID: 11218204].

[51] Lindstedt BA, Heir E, Vardund T, Kapperud G: **Fluorescent amplified-fragment length polymorphism genotyping of *Salmonella enterica* subsp. enterica serovars and comparison with pulsed-field gel electrophoresis typing**. *Journal of Clinical Microbiology* 2000, **38**(4):1623–1627, [http://www.ncbi.nlm.nih.gov/pubmed/10747153]. [PMID: 10747153].

[52] Fani R, Damiani G, Di Serio C, Gallori E, Grifoni A, Bazzicalupo M: **Use of random amplified polymorphic DNA (RAPD) for generating specific DNA probes for microorganisms**. *Molecular Ecology* 1993, **2**(4):243–250, [http://www.ncbi.nlm.nih.gov/pubmed/8167854]. [PMID: 8167854].

[53] de Bruijn FJ: **Use of repetitive (repetitive extragenic palindromic and enterobacterial repetitive intergeneric consensus) sequences and the polymerase chain reaction to fingerprint the genomes of *Rhizobium meliloti* isolates and other soil bacteria**. *Applied and Environmental Microbiology* 1992, **58**(7):2180–2187, [http://www.ncbi.nlm.nih.gov/pubmed/1637156]. [PMID: 1637156].

[54] Louws FJ, Fulbright DW, Stephens CT, de Bruijn FJ: **Specific genomic fingerprints of phytopathogenic xanthomonas and pseudomonas pathovars and strains generated with repetitive sequences and PCR**. *Applied and Environmental Microbiology* 1994, **60**(7):2286–2295, [http://www.ncbi.nlm.nih.gov/pubmed/8074510]. [PMID: 8074510].

[55] Grtler V, Stanisich VA: **New approaches to typing and identification of bacteria using the 16S-23S rDNA spacer region**. *Microbiology (Reading, England)* 1996, **142 (Pt 1)**:3–16, [http://www.ncbi.nlm.nih.gov/pubmed/8581168]. [PMID: 8581168].

[56] Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, Fujimoto E, Hoff M, Kumlin E: **Variable number of tandem repeat (VNTR) markers for human gene mapping**. *Science (New York, N.Y.)* 1987, **235**(4796):1616–1622, [http://www.ncbi.nlm.nih.gov/pubmed/3029872]. [PMID: 3029872].

[57] Fratamico PM, Bhunia AK, Smith JL: *Foodborne pathogens: microbiology and molecular biology*. Horizon Scientific Press 2005.

[58] Hyyti-Trees E, Smole SC, Fields PA, Swaminathan B, Ribot EM: **Second generation subtyping: a proposed PulseNet protocol for multiple-locus variable-number tandem repeat analysis of shiga toxin-producing *Escherichia coli* O157 (STEC O157)**. *Foodborne Pathogens and Disease* 2006, **3**:118–31. [PMID: 16602987].

[59] Cooley M, Carychao D, Crawford-Miksza L, Jay MT, Myers C, Rose C, Keys C, Farrar J, Mandrell RE: **Incidence and tracking of *Escherichia coli* O157:H7 in a major produce production region in California**. *PLoS ONE* 2007, **2**(11):e1159, [http://www.ncbi.nlm.nih.gov/pubmed/18174909]. [PMID: 18174909].

[60] Lindstedt BA, Tham W, Danielsson-Tham ML, Vardund T, Helmersson S, Kapperud G: **Multiple-locus variable-number tandem-repeats analysis of *Listeria monocytogenes* using multicolour capillary electrophoresis and comparison with pulsed-field gel electrophoresis typing**. *Journal of Microbiological Methods* 2008, **72**(2):141–148, [http://www.ncbi.nlm.nih.gov/pubmed/18096258]. [PMID: 18096258].

[61] Keim P, Price LB, Klevytska AM, Smith KL, Schupp JM, Okinaka R, Jackson PJ, Hugh-Jones ME: **Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis***. *Journal of Bacteriology* 2000, **182**(10):2928–2936, [http://www.ncbi.nlm.nih.gov/pubmed/10781564]. [PMID: 10781564].

[62] Zhang L, Srinivasan U, Marrs CF, Ghosh D, Gilsdorf JR, Foxman B: **Library on a slide for bacterial comparative genomics**. *BMC Microbiology* 2004, **4**:12, [http://www.ncbi.nlm.nih.gov/pubmed/15035675]. [PMID: 15035675].

[63] Hannon SJ, Taboada EN, Russell ML, Allan B, Waldner C, Wilson HL, Potter A, Babiuk L, Townsend HGG: **Genomics-based molecular epidemiology of *Campylobacter jejuni* isolates from feedlot cattle and from people in Alberta, Canada**. *Journal of Clinical Microbiology* 2009, **47**(2):410–420, [http://www.ncbi.nlm.nih.gov/pubmed/19036937]. [PMID: 19036937].

[64] Ogura Y, Ooka T, Asadulghani, Terajima J, Nougayrde J, Kurokawa K, Tashiro K, Tobe T, Nakayama K, Kuhara S, Oswald E, Watanabe H, Hayashi T: **Extensive genomic diversity and selective conservation of virulence-determinants in enterohemorrhagic *Escherichia coli* strains of O157 and non-O157 serotypes**. *Genome Biology* 2007, **8**(7):R138, [http://www.ncbi.nlm.nih.gov/pubmed/17711596]. [PMID: 17711596].

[65] Aguado-Urda M, Lpez-Campos GH, Fernndez-Garayzbal JF, Martn-Snchez F, Gibello A, Domnguez L, Blanco MM: **Analysis of the genome content of *Lactococcus garvieae* by genomic interspecies microarray hybridization**. *BMC Microbiology* 2010, **10**:79, [http://www.ncbi.nlm.nih.gov/pubmed/20233401]. [PMID: 20233401].

[66] Laing C, Pegg C, Yawney D, Ziebell K, Steele M, Johnson R, Thomas JE, Taboada EN, Zhang Y, Gannon VPJ: **Rapid determination of *Escherichia coli* O157:H7 lineage types and molecular subtypes by using comparative genomic Fingerprinting**. *Applied and Environmental Microbiology* 2008, **74**(21):6606–15, [http://www.ncbi.nlm.nih.gov/pubmed/18791027]. [PMID: 18791027].

[67] Taboada E, Mackinnon JM, Johnson J, Roberts MJ, Ross S, Mauro WO, Ratansi A, Yan J, Lorentz J, Thomas JE, Rahn K, Gannon VP: **The use of high-throughput comparative genomics-based molecular typing enhances cluster detection in epidemiological studies of campylobacter jejejuni.** *Campylobacter Helicobacter-Related Organisms 2007 Meet., 2 to 5 September 2007, Rotterdam, The Netherlands* 2007, **Zoonoses Public Health 54(Suppl. 1)**(20-0008).

[68] Cornelius AJ, Gilpin B, Carter P, Nicol C, On SLW: **Comparison of PCR binary typing (P-BIT), a new approach to epidemiological subtyping of *Campylobacter jejuni*, with serotyping, pulsed-field gel electrophoresis, and multilocus sequence typing methods**. *Applied and Environmental Microbiology* 2010, **76**(5):1533–1544, [http://www.ncbi.nlm.nih.gov/pubmed/20023103]. [PMID: 20023103].

[69] Taboada EN, Ross SL, Mutschall SK, Mackinnon JM, Roberts MJ, Buchanan CJ, Kruczkiewicz P, Jokinen CC, Thomas JE, Nash JHE, Gannon VPJ, Marshall B, Pollari F, Clark CG: **Development and validation of a comparative genomic fingerprinting method for high-resolution genotyping of *Campylobacter jejuni***. *Journal of clinical microbiology* 2012, **50**(3):788–797, [http://www.ncbi.nlm.nih.gov/pubmed/22170908]. [PMID: 22170908].

[70] Manning SD, Motiwala AS, Springman AC, Qi W, Lacher DW, Ouellette LM, Mladonicky JM, Somsel P, Rudrik JT, Dietrich SE, Zhang W, Swaminathan B, Alland D, Whittam TS: **Variation in virulence among clades of *Escherichia coli* O157:H7 associated with disease outbreaks**. *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**(12):4868–73. [PMID: 18332430].

[71] Laing C, Buchanan C, Taboada E, Zhang Y, Karmali M, Thomas J, Gannon V: **In silico genomic analyses reveal three distinct lineages of *Escherichia coli* O157:H7, one of which is associated with hyper-virulence**. *BMC Genomics* 2009, **10**:287, [http://www.ncbi.nlm.nih.gov/pubmed/19563677]. [PMID: 19563677].

[72] Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS, Pichon B, Baker S, Parry CM, Lambertsen LM, Shahinas D, Pillai DR, Mitchell TJ, Dougan G, Tomasz A, Klugman KP, Parkhill J, Hanage WP, Bentley SD: **Rapid pneumococcal evolution in response to clinical Interventions**. *Science* 2011, **331**(6016):430 –434, [http://www.sciencemag.org/content/331/6016/430.abstract].

[73] Woese CR: **Bacterial evolution**. *Microbiological Reviews* 1987, **51**(2):221–271, [http://www.ncbi.nlm.nih.gov/pubmed/2439888]. [PMID: 2439888].

[74] Pace NR: **Mapping the tree of life: progress and prospects**. *Microbiology and Molecular Biology Reviews: MMBR* 2009, **73**(4):565–576, [http://www.ncbi.nlm.nih.gov/pubmed/19946133]. [PMID: 19946133].

[75] Feil EJ, Smith JM, Enright MC, Spratt BG: **Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data**. *Genetics* 2000, **154**(4):1439–50, [http://www.ncbi.nlm.nih.gov/pubmed/10747043]. [PMID: 10747043].

[76] Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK, Whittam TS: **Parallel evolution of virulence in pathogenic *Escherichia coli***. *Nature* 2000, **406**(6791):64–7, [http://www.ncbi.nlm.nih.gov/pubmed/10894541]. [PMID: 10894541].

[77] Yan Y, Cui Y, Han H, Xiao X, Wong HC, Tan Y, Guo Z, Liu X, Yang R, Zhou D: **Extended MLST-based population genetics and phylogeny of *Vibrio parahaemolyticus* with high levels of recombination**. *International Journal of Food Microbiology* 2010, [http://www.ncbi.nlm.nih.gov/pubmed/21176856]. [PMID: 21176856].

[78] Chalmers G, Bruce HL, Hunter DB, Parreira VR, Kulkarni RR, Jiang YF, Prescott JF, Boerlin P: **Multilocus sequence typing analysis of *Clostridium perfringens* isolates from necrotic enteritis outbreaks in broiler chicken populations**. *Journal of Clinical Microbiology* 2008, **46**(12):3957–3964, [http://www.ncbi.nlm.nih.gov/pubmed/18945840]. [PMID: 18945840].

[79] Moodley Y, Linz B, Yamaoka Y, Windsor HM, Breurec S, Wu JY, Maady A, Bernhoft S, Thiberge JM, Phuanukoonnon S, Jobb G, Siba P, Graham DY, Marshall BJ, Achtman M: **The peopling of the pacific from a bacterial Perspective**. *Science* 2009, **323**(5913):527–530, [http://www.sciencemag.org/content/suppl/2009/01/22/323.5913.527.DC1].

[80] Konstantinidis KT, Ramette A, Tiedje JM: **Toward a more robust assessment of intraspecies diversity, using fewer genetic markers**. *Applied and Environmental Microbiology* 2006, **72**(11):7286–7293, [http://www.ncbi.nlm.nih.gov/pubmed/16980418]. [PMID: 16980418].

[81] Taboada EN, Mackinnon JM, Luebbert CC, Gannon VPJ, Nash JHE, Rahn K: **Comparative genomic assessment of Multi-locus sequence Typing: rapid accumulation of genomic heterogeneity among clonal isolates of *Campylobacter jejuni***. *BMC Evolutionary Biology* 2008, **8**:229, [http://www.ncbi.nlm.nih.gov/pubmed/18691421]. [PMID: 18691421].

[82] Noller AC, McEllistrem MC, Stine OC, Morris JG, Boxrud DJ, Dixon B, Harrison LH: **Multilocus sequence typing reveals a lack of diversity among *Escherichia coli* O157:H7 isolates that are distinct by pulsed-field gelelectrophoresis**. *Journal of Clinical Microbiology* 2003, **41**(2):675–9. [PMID: 12574266].

[83] Price E, Inman-Bamber J, Thiruvenkataswamy V, Huygens F, Giffard P: **Computer-aided identification of polymorphism sets diagnostic for groups of bacterial and viral genetic variants**. *BMC Bioinformatics* 2007, **8**:278, [http://www.biomedcentral.com/1471-2105/8/278].

[84] Frei UK, Wollenweber B, Lbberstedt T: **”PolyMin”: software for identification of the minimum number of polymorphisms required for haplotype and genotype differentiation**. *BMC Bioinformatics* 2009, **10**:176, [http://www.ncbi.nlm.nih.gov/pubmed/19515225]. [PMID: 19515225].

[85] Price EP, Huygens F, Giffard PM: **Fingerprinting of *Campylobacter jejuni* by using resolution-optimized binary gene targets derived from comparative genome hybridization studies**. *Applied and Environmental Microbiology* 2006, **72**(12):7793–7803, [http://www.ncbi.nlm.nih.gov/pubmed/16997982]. [PMID: 16997982].

[86] Gutacker MM, Smoot JC, Migliaccio CAL, Ricklefs SM, Hua S, Cousins DV, Graviss EA, Shashkina E, Kreiswirth BN, Musser JM: **Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains**. *Genetics* 2002, **162**(4):1533–1543, [http://www.ncbi.nlm.nih.gov/pubmed/12524330]. [PMID: 12524330].

[87] Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJM, Brinkman FSL, Brunham RC, Tang P: **Whole-genome sequencing and social-network analysis of a tuberculosis outbreak**. *The New England Journal of Medicine* 2011, **364**(8):730–739, [http://www.ncbi.nlm.nih.gov/pubmed/21345102]. [PMID: 21345102].

[88] Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, Edgeworth JD, de Lencastre H, Parkhill J, Peacock SJ, Bentley SD: **Evolution of MRSA during hospital transmission and intercontinental Spread**. *Science* 2010, **327**(5964):469 –474, [http://www.sciencemag.org/content/327/5964/469.abstract].

[89] Chin CS, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, Bullard J, Webster DR, Kasarskis A, Peluso P, Paxinos EE, Yamaichi Y, Calderwood SB, Mekalanos JJ, Schadt EE, Waldor MK: **The origin of the haitian cholera outbreak strain**. *The New England Journal of Medicine* 2011, **364**:33–42, [http://www.ncbi.nlm.nih.gov/pubmed/21142692]. [PMID: 21142692].

[90] Gilmour M, Graham M, Van Domselaar G, Tyler S, Kent H, Trout-Yakel K, Larios O, Allen V, Lee B, Nadon C: **High-throughput genome sequencing of two *Listeria monocytogenes* clinical isolates during a large foodborne outbreak**. *BMC Genomics* 2010, **11**:120, [http://www.biomedcentral.com/1471-2164/11/120].

[91] Leopold SR, Magrini V, Holt NJ, Shaikh N, Mardis ER, Cagno J, Ogura Y, Iguchi A, Hayashi T, Mellmann A, Karch H, Besser TE, Sawyer SA, Whittam TS, Tarr PI: **A precise reconstruction of the emergence and constrained radiations of *Escherichia coli* O157 portrayed by backbone concatenomic analysis**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, [http://www.ncbi.nlm.nih.gov/pubmed/19439656]. [PMID: 19439656].

[92] Leopold SR, Shaikh N, Tarr PI: **Further evidence of constrained radiation in the evolution of pathogenic *Escherichia coli* O157:H7**. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 2010, [http://www.ncbi.nlm.nih.gov/pubmed/20691811]. [PMID: 20691811].

[93] den Bakker HC, Cummings CA, Ferreira V, Vatta P, Orsi RH, Degoricija L, Barker M, Petrauskene O, Furtado MR, Wiedmann M: **Comparative genomics of the bacterial genus**

Listeria: genome evolution is characterized by limited gene acquisition and limited gene loss. *BMC Genomics* 2010, **11**:688, [http://www.ncbi.nlm.nih.gov/pubmed/21126366]. [PMID: 21126366].

[94] Xu Z, Chen X, Li L, Li T, Wang S, Chen H, Zhou R: **Comparative genomic characterization of *Actinobacillus pleuropneumoniae***. *Journal of Bacteriology* 2010, **192**(21):5625–5636, [http://www.ncbi.nlm.nih.gov/pubmed/20802045]. [PMID: 20802045].

[95] Lowder BV, Guinane CM, Ben Zakour NL, Weinert LA, Conway-Morris A, Cartwright RA, Simpson AJ, Rambaut A, Nbel U, Fitzgerald JR: **Recent human-to-poultry host jump, adaptation, and pandemic spread of *Staphylococcus aureus***. *Proceedings of the National Academy of Sciences* 2009, **106**(46):19545 –19550, [http://www.pnas.org/content/106/46/19545.abstract].

[96] Deng X, Phillippy AM, Li Z, Salzberg SL, Zhang W: **Probing the pan-genome of *Listeria monocytogenes*: new insights into intraspecific niche expansion and genomic diversification**. *BMC Genomics* 2010, **11**:500, [http://www.ncbi.nlm.nih.gov/pubmed/20846431]. [PMID: 20846431].

[97] Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nature Reviews. Genetics* 2009, **10**:57–63, [http://www.ncbi.nlm.nih.gov/pubmed/19015660]. [PMID: 19015660].

[98] Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proceedings of the National Academy of Sciences of the United States of America* 1977, **74**(12):5463–7, [http://www.ncbi.nlm.nih.gov/pubmed/271968]. [PMID: 271968].

[99] Maxam AM, Gilbert W: **A new method for sequencing DNA**. *Proceedings of the National Academy of Sciences of the United States of America* 1977, **74**(2):560–564, [http://www.ncbi.nlm.nih.gov/pubmed/265521]. [PMID: 265521].

[100] Droege M, Hill B: **The genome sequencer FLX System–longer reads, more applications, straight forward bioinformatics and more complete data sets**. *Journal of Biotechnology* 2008, **136**(1-2):3–10, [http://www.ncbi.nlm.nih.gov/pubmed/18616967]. [PMID: 18616967].

[101] Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ: **A large genome center's improvements to the illumina sequencing system**. *Nature Methods* 2008, **5**(12):1005–1010, [http://www.ncbi.nlm.nih.gov/pubmed/19034268]. [PMID: 19034268].

[102] Ondov BD, Varadarajan A, Passalacqua KD, Bergman NH: **Efficient mapping of applied biosystems SOLiD sequence data to a reference genome for functional genomic applications**. *Bioinformatics (Oxford, England)* 2008, **24**(23):2776–2777, [http://www.ncbi.nlm.nih.gov/pubmed/18842598]. [PMID: 18842598].

[103] Zhang J, Chiodini R, Badr A, Zhang G: **The impact of next-generation sequencing on genomics**. *Journal of genetics and genomics = Yi chuan xue bao* 2011, **38**(3):95–109. [PMID: 21477781 PMCID: 3076108].

[104] Glenn TC: **Field guide to next-generation DNA sequencers**. *Molecular ecology resources* 2011, **11**(5):759–769. [PMID: 21592312].

[105] Schadt EE, Turner S, Kasarskis A: **A window into third-generation sequencing**. *Human Molecular Genetics* 2010, **19**(R2):R227–240, [http://www.ncbi.nlm.nih.gov/pubmed/20858600]. [PMID: 20858600].

[106] Losada L, Varga JJ, Hostetler J, Radune D, Kim M, Durkin S, Schneewind O, Nierman WC: **Genome sequencing and analysis of yersina pestis KIM D27, an avirulent strain exempt from select agent Regulation**. *PLoS ONE* 2011, **6**(4):e19054, [http://dx.doi.org/10.1371/journal.pone.0019054].

[107] Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity**. *BMC Bioinformatics* 2004, **5**:113, [http://www.ncbi.nlm.nih.gov/pubmed/15318951]. [PMID: 15318951].

[108] Notredame C, Higgins DG, Heringa J: **T-Coffee: a novel method for fast and accurate multiple sequence alignment**. *Journal of Molecular Biology* 2000, **302**:205–217, [http://www.ncbi.nlm.nih.gov/pubmed/10964570]. [PMID: 10964570].

[109] Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic Acids Research* 1994, **22**(22):4673–4680, [http://www.ncbi.nlm.nih.gov/pubmed/7984417]. [PMID: 7984417].

[110] Pearson WR: **Rapid and sensitive sequence comparison with FASTP and FASTA**. *Methods in Enzymology* 1990, **183**:63–98, [http://www.ncbi.nlm.nih.gov/pubmed/2156132]. [PMID: 2156132].

[111] Altschul SF, Madden TL, Schffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Research* 1997, **25**(17):3389–402, [http://www.ncbi.nlm.nih.gov/pubmed/9254694]. [PMID: 9254694].

[112] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications**. *BMC Bioinformatics* 2009, **10**:421, [http://www.ncbi.nlm.nih.gov/pubmed/20003500]. [PMID: 20003500].

[113] Vouzis PD, Sahinidis NV: **GPU-BLAST: using graphics processors to accelerate protein sequence alignment**. *Bioinformatics (Oxford, England)* 2011, **27**(2):182–188, [http://www.ncbi.nlm.nih.gov/pubmed/21088027]. [PMID: 21088027].

[114] Kent WJ: **BLAT–the BLAST-like alignment tool**. *Genome Research* 2002, **12**(4):656–664, [http://www.ncbi.nlm.nih.gov/pubmed/11932250]. [PMID: 11932250].

[115] Lopez R, Silventoinen V, Robinson S, Kibria A, Gish W: **WU-Blast2 server at the european bioinformatics Institute**. *Nucleic Acids Research* 2003, **31**(13):3795–3798, [http://www.ncbi.nlm.nih.gov/pubmed/12824421]. [PMID: 12824421].

[116] Edgar RC: **Search and clustering orders of magnitude faster than BLAST**. *Bioinformatics (Oxford, England)* 2010, **26**(19):2460–2461, [http://www.ncbi.nlm.nih.gov/pubmed/20709691]. [PMID: 20709691].

[117] Li H, Durbin R: **Fast and accurate long-read alignment with BurrowsWheeler transform**. *Bioinformatics* 2010, **26**(5):589 –595, [http://bioinformatics.oxfordjournals.org/content/26/5/589.abstract].

[118] Chetouani F, Glaser P, Kunst F: **FindTarget: software for subtractive genome analysis**. *Microbiology (Reading, England)* 2001, **147**(Pt 10):2643–2649, [http://www.ncbi.nlm.nih.gov/pubmed/11577143]. [PMID: 11577143].

[119] Yao J, Lin H, Doddapaneni H, Civerolo EL: **nWayComp: a genome-wide sequence comparison tool for multiple strains/species of phylogenetically related microorganisms**. *In Silico Biology* 2007, **7**(2):195–200. [PMID: 17688445].

[120] Shao Y, He X, Harrison EM, Tai C, Ou HY, Rajakumar K, Deng Z: **mGenomeSubtractor: a web-based tool for parallel in silico subtractive hybridization analysis of multiple bacterial genomes**. *Nucleic Acids Research* 2010, **38**(Web Server issue):W194–200, [http://www.ncbi.nlm.nih.gov/pubmed/20435682]. [PMID: 20435682].

[121] Darling ACE, Mau B, Blattner FR, Perna NT: **Mauve: multiple alignment of conserved genomic sequence with rearrangements**. *Genome Research* 2004, **14**(7):1394–403, [http://www.ncbi.nlm.nih.gov/pubmed/15231754]. [PMID: 15231754].

[122] Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes**. *Genome Biology* 2004, **5**(2):R12, [http://www.ncbi.nlm.nih.gov/pubmed/14759262]. [PMID: 14759262].

[123] Schatz MC, Trapnell C, Delcher AL, Varshney A: **High-throughput sequence alignment using graphics processing Units**. *BMC Bioinformatics* 2007, **8**:474, [http://www.ncbi.nlm.nih.gov/pubmed/18070356]. [PMID: 18070356].

[124] Kryukov K, Saitou N: **MISHIMA–a new method for high speed multiple alignment of nucleotide sequences of bacterial genome scale data**. *BMC Bioinformatics* 2010, **11**:142, [http://www.ncbi.nlm.nih.gov/pubmed/20298584]. [PMID: 20298584].

[125] Angiuoli SV, Salzberg SL: **Mugsy: fast multiple alignment of closely related whole genomes**. *Bioinformatics (Oxford, England)* 2010, [http://www.ncbi.nlm.nih.gov/pubmed/21148543]. [PMID: 21148543].

[126] Rausch T, Emde AK, Weese D, Dring A, Notredame C, Reinert K: **Segment-based multiple sequence alignment**. *Bioinformatics* 2008, **24**(16):i187 –i192, [http://bioinformatics.oxfordjournals.org/content/24/16/i187.abstract].

[127] Laing C, Buchanan C, Taboada EN, Zhang Y, Kropinski A, Villegas A, Thomas JE, Gannon VPJ: **Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions**. *BMC Bioinformatics* 2010, **11**:461, [http://www.ncbi.nlm.nih.gov/pubmed/20843356]. [PMID: 20843356].

[128] Champion MD, Zeng Q, Nix EB, Nano FE, Keim P, Kodira CD, Borowsky M, Young S, Koehrsen M, Engels R, Pearson M, Howarth C, Larson L, White J, Alvarado L, Forsman M, Bearden SW, Sjstedt A, Titball R, Michell SL, Birren B, Galagan J: **Comparative genomic characterization of *Francisella tularensis* strains belonging to low and high virulence Subspecies**. *PLoS Pathog* 2009, **5**(5):e1000459, [http://dx.doi.org/10.1371/journal.ppat.1000459].

[129] Vasconcelos ATR, Ferreira HB, Bizarro CV, Bonatto SL, Carvalho MO, Pinto PM, Almeida DF, Almeida LGP, Almeida R, Alves-Filho L, Assuncao EN, Azevedo VAC, Bogo MR, Brigido MM, Brocchi M, Burity HA, Camargo AA, Camargo SS, Carepo MS, Carraro DM, de Mattos Cascardo JC, Castro LA, Cavalcanti G, Chemale G, Collevatti RG, Cunha CW, Dallagiovanna B, Dambros BP, Dellagostin OA, Falcao C, Fantinatti-Garboggini F, Felipe MSS, Fiorentin L, Franco GR, Freitas NSA, Frias D, Grangeiro TB, Grisard EC, Guimaraes CT, Hungria M, Jardim SN, Krieger MA, Laurino JP, Lima LFA, Lopes MI, Loreto ELS, Madeira HMF, Manfio GP, Maranhao AQ, Martinkovics CT, Medeiros SRB, Moreira MAM, Neiva M, Ramalho-Neto CE, Nicolas MF, Oliveira SC, Paixao RFC, Pedrosa FO, Pena SDJ, Pereira M, Pereira-Ferrari L, Piffer I, Pinto LS, Potrich DP, Salim ACM, Santos FR, Schmitt R, Schneider MPC, Schrank A, Schrank IS, Schuck AF, Seuanez HN, Silva DW, Silva R, Silva SC, Soares CMA, Souza KRL, Souza RC, Staats CC, Steffens MBR, Teixeira SMR, Urmenyi TP, Vainstein MH, Zuccherato LW, Simpson AJG, Zaha A: **Swine and poultry Pathogens: the complete genome sequences of two strains of *Mycoplasma hyopneumoniae* and a strain of *Mycoplasma synoviae***. *J. Bacteriol.* 2005, **187**(16):5568–5577, [http://jb.asm.org/cgi/content/abstract/187/16/5568].

[130] Tsolis RM, Seshadri R, Santos RL, Sangari FJ, Lobo JMG, de Jong MF, Ren Q, Myers G, Brinkac LM, Nelson WC, DeBoy RT, Angiuoli S, Khouri H, Dimitrov G, Robinson JR, Mulligan S, Walker RL, Elzer PE, Hassan KA, Paulsen IT: **Genome degradation in *Brucella ovis* corresponds with narrowing of its host range and tissue Tropism**. *PLoS ONE* 2009, **4**(5):e5519, [http://dx.doi.org/10.1371/journal.pone.0005519].

[131] Eppinger M, Worsham PL, Nikolich MP, Riley DR, Sebastian Y, Mou S, Achtman M, Lindler LE, Ravel J: **Genome sequence of the Deep-rooted *Yersinia pestis* strain angola reveals new insights into the evolution and pangenome of the plague Bacterium**. *J. Bacteriol.* 2010, **192**(6):1685–1699, [http://jb.asm.org/cgi/content/abstract/192/6/1685].

[132] Nash JH, Villegas A, Kropinski AM, Aguilar-Valenzuela R, Konczy P, Mascarenhas M, Ziebell K, Torres AG, Karmali MA, Coombes BK: **Genome sequence of adherent-invasive *Escherichia coli* and comparative genomic analysis with other E. coli pathotypes**. *BMC Genomics* 2010, **11**:667, [http://www.ncbi.nlm.nih.gov/pubmed/21108814]. [PMID: 21108814].

[133] Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, Goodhead I, Rance R, Baker S, Maskell DJ, Wain J, Dolecek C, Achtman M, Dougan G: **High-throughput sequencing provides insights into genome variation and evolution in salmonella Typhi**. *Nature genetics* 2008, **40**(8):987–993. [PMID: 18660809].

[134] Lewis T, Loman N, Bingle L, Jumaa P, Weinstock G, Mortiboy D, Pallen M: **High-throughput whole-genome sequencing to dissect the epidemiology of**

***Acinetobacter baumannii* isolates from a hospital outbreak**. *Journal of Hospital Infection* 2010, **75**:37–41, [http://www.sciencedirect.com/science/article/B6WJP-4YMHFJ3-8/2/e69b448a6d432519303c60f56e7a9449].

[135] Beres SB, Carroll RK, Shea PR, Sitkiewicz I, Martinez-Gutierrez JC, Low DE, McGeer A, Willey BM, Green K, Tyrrell GJ, Goldman TD, Feldgarden M, Birren BW, Fofanov Y, Boos J, Wheaton WD, Honisch C, Musser JM: **Molecular complexity of successive bacterial epidemics deconvoluted by comparative pathogenomics**. *Proceedings of the National Academy of Sciences* 2010, **107**(9):4371 –4376, [http://www.pnas.org/content/107/9/4371.abstract].

[136] Schrch AC, Kremer K, Kiers A, Daviena O, Boeree MJ, Siezen RJ, Smith NH, van Soolingen D: **The tempo and mode of molecular evolution of *Mycobacterium tuberculosis* at patient-to-patient scale**. *Infection, Genetics and Evolution* 2010, **10**:108–114, [http://www.sciencedirect.com/science/article/B6W8B-4XFPR1M-1/2/d1577a76eb28bcabc7d0a3e1329e2a8c].

[137] Kuroda M, Serizawa M, Okutani A, Sekizuka T, Banno S, Inoue S: **Genome-wide single nucleotide polymorphism typing method for identification of *Bacillus anthracis* species and strains among B. cereus group species**. *Journal of Clinical Microbiology* 2010, **48**(8):2821–2829, [http://www.ncbi.nlm.nih.gov/pubmed/20554827]. [PMID: 20554827].

[138] Pandya GA, Holmes MH, Petersen JM, Pradhan S, Karamycheva SA, Wolcott MJ, Molins C, Jones M, Schriefer ME, Fleischmann RD, Peterson SN: **Whole genome single nucleotide polymorphism based phylogeny of *Francisella tularensis* and its application to the development of a strain typing assay**. *BMC Microbiology* 2009, **9**:213, [http://www.ncbi.nlm.nih.gov/pubmed/19811647]. [PMID: 19811647].

[139] Sjdin A, Svensson K, Lindgren M, Forsman M, Larsson P: **Whole-genome sequencing reveals distinct mutational patterns in closely related laboratory and naturally propagated *Francisella tularensis* strains**. *PloS one* 2010, **5**(7):e11556. [PMID: 20657845].

[140] Morelli G, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, Wagner DM, Feldkamp M, Kusecek B, Vogler AJ, Li Y, Cui Y, Thomson NR, Jombart T, Leblois R, Lichtner P, Rahalison L, Petersen JM, Balloux F, Keim P, Wirth T, Ravel J, Yang R, Carniel E, Achtman M: ***Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity**. *Nature Genetics* 2010, **42**(12):1140–1143, [http://www.ncbi.nlm.nih.gov/pubmed/21037571]. [PMID: 21037571].

[141] Kennemann L, Didelot X, Aebischer T, Kuhn S, Drescher B, Droege M, Reinhardt R, Correa P, Meyer TF, Josenhans C, Falush D, Suerbaum S: ***Helicobacter pylori* genome evolution during human infection**. *Proceedings of the National Academy of Sciences* 2011, **108**(12):5033 –5038, [http://www.pnas.org/content/108/12/5033.abstract].

[142] Bettelheim KA: **The non-O157 shiga-toxigenic (verocytotoxigenic) *Escherichia coli*; under-rated pathogens**. *Critical Reviews in Microbiology* 2007, **33**:67–87, [http://www.ncbi.nlm.nih.gov/pubmed/17453930]. [PMID: 17453930].

[143] Bettelheim KA, Teoh-Chan CH, Chandler ME, O'Farrell SM, Rahamin L, Shaw EJ, Shooter RA: **Further studies of *Escherichia coli* in babies after normal delivery**. *The Journal of Hygiene* 1974, **73**(2):277–285, [http://www.ncbi.nlm.nih.gov/pubmed/4608224]. [PMID: 4608224].

[144] Majed NI, Bettelheim KA, Shooter RA, Moorhouse E: **The effect of travel on faecal *Escherichia coli* serotypes**. *The Journal of Hygiene* 1978, **81**(3):481–487, [http://www.ncbi.nlm.nih.gov/pubmed/366018]. [PMID: 366018].

[145] Clermont O, Bonacorsi S, Bingen E: **Rapid and simple determination of the *Escherichia coli* phylogenetic group**. *Applied and Environmental Microbiology* 2000, **66**(10):4555–4558, [http://www.ncbi.nlm.nih.gov/pubmed/11010916]. [PMID: 11010916].

[146] O'Reilly KM, Low JC, Denwood MJ, Gally DL, Evans J, Gunn GJ, Mellor DJ, Reid SWJ, Matthews L: **Associations between the presence of virulence determinants and the epidemiology and ecology of zoonotic *Escherichia coli***. *Applied and Environmental Microbiology* 2010, **76**(24):8110–8116, [http://www.ncbi.nlm.nih.gov/pubmed/20952647]. [PMID: 20952647].

[147] Mathusa EC, Chen Y, Enache E, Hontz L: **Non-O157 shiga toxin-producing *Escherichia coli* in foods**. *Journal of Food Protection* 2010, **73**(9):1721–1736, [http://www.ncbi.nlm.nih.gov/pubmed/20828483]. [PMID: 20828483].

[148] Madic J, Vingadassalon N, de Garam CP, Marault M, Scheutz F, Brugre H, Jamet E, Auvray F: **Detection of shiga toxin-producing *Escherichia coli* serotypes O26:H11, O103:H2, O111:H8, O145:H28, and O157:H7 in raw-milk cheeses by using multiplex real-time PCR**. *Applied and Environmental Microbiology* 2011, **77**(6):2035–2041, [http://www.ncbi.nlm.nih.gov/pubmed/21239543]. [PMID: 21239543].

[149] Berger CN, Sodha SV, Shaw RK, Griffin PM, Pink D, Hand P, Frankel G: **Fresh fruit and vegetables as vehicles for the transmission of human pathogens**. *Environmental Microbiology* 2010, **12**(9):2385–2397, [http://www.ncbi.nlm.nih.gov/pubmed/20636374]. [PMID: 20636374].

[150] Islam MA, Mondol AS, Azmi IJ, de Boer E, Beumer RR, Zwietering MH, Heuvelink AE, Talukder KA: **Occurrence and characterization of shiga toxin-producing *Escherichia coli* in raw meat, raw milk, and street vended juices in Bangladesh**. *Foodborne Pathogens and Disease* 2010, **7**(11):1381–1385, [http://www.ncbi.nlm.nih.gov/pubmed/20704491]. [PMID: 20704491].

[151] McCarthy TA, Barrett NL, Hadler JL, Salsbury B, Howard RT, Dingman DW, Brinkman CD, Bibb WF, Cartter ML: **Hemolytic-uremic syndrome and *Escherichia coli* O121 at a lake in Connecticut, 1999**. *Pediatrics* 2001, **108**(4):E59, [http://www.ncbi.nlm.nih.gov/pubmed/11581467]. [PMID: 11581467].

[152] Lienemann T, Pitknen T, Antikainen J, Mls E, Miettinen I, Haukka K, Vaara M, Siitonen A: **Shiga Toxin-producing *Escherichia coli* O100:H(-): stx (2e) in drinking water contaminated by waste water in Finland**. *Current Microbiology* 2011, **62**(4):1239–1244, [http://www.ncbi.nlm.nih.gov/pubmed/21188590]. [PMID: 21188590].

[153] Paton JC, Paton AW: **Pathogenesis and diagnosis of shiga Toxin-producing *Escherichia coli* Infections**. *Clin. Microbiol. Rev.* 1998, **11**(3):450–479, [http://cmr.asm.org/cgi/content/abstract/11/3/450].

[154] Paton AW, Ratcliff RM, Doyle RM, Seymour-Murray J, Davos D, Lanser JA, Paton JC: **Molecular microbiological investigation of an outbreak of hemolytic-uremic syndrome caused by dry fermented sausage contaminated with Shiga-like toxin-producing *Escherichia coli***. *Journal of Clinical Microbiology* 1996, **34**(7):1622–1627, [http://www.ncbi.nlm.nih.gov/pubmed/8784557]. [PMID: 8784557].

[155] Hussein HS: **Prevalence and pathogenicity of shiga toxin-producing *Escherichia coli* in beef cattle and their products**. *Journal of Animal Science* 2007, **85**(13 Suppl):E63–72, [http://www.ncbi.nlm.nih.gov/pubmed/17060419]. [PMID: 17060419].

[156] Nataro JP, Kaper JB: **Diarrheagenic *Escherichia coli***. *Clinical Microbiology Reviews* 1998, **11**:142–201, [http://www.ncbi.nlm.nih.gov/pubmed/9457432]. [PMID: 9457432].

[157] Karmali MA, Mascarenhas M, Shen S, Ziebell K, Johnson S, Reid-Smith R, Isaac-Renton J, Clark C, Rahn K, Kaper JB: **Association of genomic o island 122 of *Escherichia coli* EDL 933 with verocytotoxin-producing *Escherichia coli* seropathotypes that are linked to epidemic and/or serious disease**. *J Clin Microbiol* 2003, **41**(11):4930–40, [14605120].

[158] Sting R, Stermann M: **Duplex real-time PCR assays for rapid detection of virulence genes in E. coli isolated from post-weaning pigs and calves with diarrhoea**. *DTW. Deutsche Tierrztliche Wochenschrift* 2008, **115**(6):231–238, [http://www.ncbi.nlm.nih.gov/pubmed/18605375]. [PMID: 18605375].

[159] Arya G, Roy A, Choudhary V, Yadav MM, Joshi CG: **Serogroups, atypical biochemical characters, colicinogeny and antibiotic resistance pattern of shiga toxin-producing *Escherichia coli* isolated from diarrhoeic calves in Gujarat, India**. *Zoonoses and Public Health* 2008, **55**(2):89–98, [http://www.ncbi.nlm.nih.gov/pubmed/18234027]. [PMID: 18234027].

[160] Janke BH, Francis DH, Collins JE, Libal MC, Zeman DH, Johnson DD, Neiger RD: **Attaching and effacing *Escherichia coli* infection as a cause of diarrhea in young calves**. *Journal of the American Veterinary Medical Association* 1990, **196**(6):897–901. [PMID: 2179181].

[161] Bolton DJ: **Verocytotoxigenic (Shiga Toxin-Producing) *Escherichia coli*: virulence factors and pathogenicity in the farm to fork Paradigm**. *Foodborne Pathogens and Disease* 2011, **8**(3):357–365, [http://www.ncbi.nlm.nih.gov/pubmed/21114423]. [PMID: 21114423].

[162] Whittam TS, Wachsmuth IK, Wilson RA: **Genetic evidence of clonal descent of *Escherichia coli* O157:H7 associated with hemorrhagic colitis and hemolytic uremic syndrome**. *J Infect Dis* 1988, **157**(6):1124–33, [3286782].

[163] Wick LM, Qi W, Lacher DW, Whittam TS: **Evolution of genomic content in the stepwise emergence of *Escherichia coli* O157:H7**. *J. Bacteriol.* 2005, **187**(5):1783–1791, [http://jb.asm.org/cgi/content/abstract/187/5/1783].

[164] Kyle JL, Cummings CA, Parker CT, Quiones B, Vatta P, Newton E, Huynh S, Swimley M, Degoricija L, Barker M, Fontanoz S, Nguyen K, Patel R, Fang R, Tebbs R, Petrauskene O, Furtado M, Mandrell RE: ***Escherichia coli* serotype O55:H7 diversity supports parallel acquisition of bacteriophage at shiga toxin phage insertion sites during evolution of the O157:H7 lineage**. *Journal of bacteriology* 2012, **194**(8):1885–1896, [http://www.ncbi.nlm.nih.gov/pubmed/22328665]. [PMID: 22328665].

[165] Jenke C, Leopold SR, Weniger T, Rothgnger J, Harmsen D, Karch H, Mellmann A: **Identification of intermediate in evolutionary model of enterohemorrhagic *Escherichia coli* O157**. *Emerging infectious diseases* 2012, **18**(4):582–588, [http://www.ncbi.nlm.nih.gov/pubmed/22469031]. [PMID: 22469031].

[166] Wylie JL, Van Caeseele P, Gilmour MW, Sitter D, Guttek C, Giercke S: **Evaluation of a new chromogenic agar medium for detection of shiga Toxin-producing *Escherichia coli* (stec) and relative prevalence of O157 and non-O157 STEC, Manitoba, Canada**. *Journal of clinical microbiology* 2012. [PMID: 23175263].

[167] Brooks JT, Sowers EG, Wells JG, Greene KD, Griffin PM, Hoekstra RM, Strockbine NA: **Non-O157 shiga toxin-producing *Escherichia coli* infections in the united States, 1983-2002**. *The Journal of infectious diseases* 2005, **192**(8):1422–1429. [PMID: 16170761].

[168] Hermos CR, Janineh M, Han LL, McAdam AJ: **Shiga toxin-producing *Escherichia coli* in children: diagnosis and clinical manifestations of O157:H7 and non-O157:H7 infection**. *Journal of clinical microbiology* 2011, **49**(3):955–959. [PMID: 21177902].

[169] Rssmann H, Kothe E, Schmidt H, Franke S, Harmsen D, Caprioli A, Karch H: **Genotyping of Shiga-like toxin genes in non-O157 *Escherichia coli* strains associated with haemolytic uraemic syndrome**. *Journal of medical microbiology* 1995, **42**(6):404–410. [PMID: 7791204].

[170] Caprioli A, Luzzi I, Rosmini F, Resti C, Edefonti A, Perfumo F, Farina C, Goglio A, Gianviti A, Rizzoni G: **Community-wide outbreak of hemolytic-uremic syndrome associated with non-O157 verocytotoxin-producing *Escherichia coli***. *The Journal of infectious diseases* 1994, **169**:208–211. [PMID: 8277184].

[171] Huppertz HI, Busch D, Schmidt H, Aleksic S, Karch H: **Diarrhea in young children associated with *Escherichia coli* non-O157 organisms that produce Shiga-like toxin**. *The Journal of pediatrics* 1996, **128**(3):341–346. [PMID: 8774501].

[172] Caprioli A, Tozzi AE, Rizzoni G, Karch H: **Non-O157 shiga toxin-producing *Escherichia coli* infections in Europe**. *Emerging infectious diseases* 1997, **3**(4):578–579. [PMID: 9366613].

[173] Safety F, Inspection Service U: **Shiga Toxin-producing *Escherichia coli* in certain raw beef Products**. *FEDERAL REGISTER* 2012, **77**(105):31975–31981, [http://www.fsis.usda.gov/OPPDE/rdad/FRPubs/2010-0023FRN.htm]. [FR Doc No: 2012-13283].

[174] Lin A, Sultan O, Lau HK, Wong E, Hartman G, Lauzon CR: **O serogroup specific real time PCR assays for the detection and identification of nine clinically relevant non-O157 STECs**. *Food Microbiology* 2011, **28**(3):478–483, [http://www.ncbi.nlm.nih.gov/pubmed/21356454]. [PMID: 21356454].

[175] Fratamico PM, Bagi LK, Cray J William C, Narang N, Yan X, Medina M, Liu Y: **Detection by multiplex real-time polymerase chain reaction assays and isolation of shiga toxin-producing *Escherichia coli* serogroups O26, O45, O103, O111, O121, and O145 in ground beef**. *Foodborne pathogens and disease* 2011, **8**(5):601–607, [http://www.ncbi.nlm.nih.gov/pubmed/21214490]. [PMID: 21214490].

[176] Gill A, Martinez-Perez A, McIlwham S, Blais B: **Development of a method for the detection of verotoxin-producing *Escherichia coli* in food**. *Journal of food protection* 2012, **75**(5):827–837, [http://www.ncbi.nlm.nih.gov/pubmed/22564930]. [PMID: 22564930].

[177] Hegde NV, Jayarao BM, DebRoy C: **Rapid detection of the top six non-o157 shiga toxin-producing *Escherichia coli* O groups in ground beef by flow cytometry**. *Journal of clinical microbiology* 2012, **50**(6):2137–2139, [http://www.ncbi.nlm.nih.gov/pubmed/22493328]. [PMID: 22493328].

[178] Sokurenko EV, Gomulkiewicz R, Dykhuizen DE: **Source-sink dynamics of virulence evolution**. *Nature Reviews. Microbiology* 2006, **4**(7):548–555, [http://www.ncbi.nlm.nih.gov/pubmed/16778839]. [PMID: 16778839].

[179] Desmyter J, Melnick JL, Rawls WE: **Defectiveness of interferon production and of rubella virus interference in a line of african green monkey kidney cells (Vero)**. *Journal of Virology* 1968, **2**(10):955–961, [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC375423/]. [PMID: 4302013 PMCID: PMC375423].

[180] Greco KM, McDonough MA, Butterton JR: **Variation in the shiga toxin region of 20th-century epidemic and endemic *Shigella dysenteriae* 1 strains**. *The Journal of infectious diseases* 2004, **190**(2):330–334. [PMID: 15216469].

[181] Herold S, Karch H, Schmidt H: **Shiga toxin-encoding bacteriophages–genomes in motion**. *International journal of medical microbiology: IJMM* 2004, **294**(2-3):115–121. [PMID: 15493821].

[182] Grotiuz G, Sirok A, Gadea P, Varela G, Schelotto F: **Shiga toxin 2-producing *Acinetobacter haemolyticus* associated with a case of bloody Diarrhea**. *Journal of Clinical Microbiology* 2006, **44**(10):3838–3841, [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1594762/]. [PMID: 17021124 PMCID: PMC1594762].

[183] Lee JE, Reed J, Shields MS, Spiegel KM, Farrell LD, Sheridan PP: **Phylogenetic analysis of shiga toxin 1 and shiga toxin 2 genes associated with disease outbreaks**. *BMC Microbiology* 2007, **7**:109, [http://www.ncbi.nlm.nih.gov/pubmed/18053224]. [PMID: 18053224].

[184] Jackson MP, Newland JW, Holmes RK, O'Brien AD: **Nucleotide sequence analysis of the structural genes for Shiga-like toxin i encoded by bacteriophage 933J from *Escherichia coli***. *Microbial pathogenesis* 1987, **2**(2):147–153, [http://www.ncbi.nlm.nih.gov/pubmed/3333796]. [PMID: 3333796].

[185] Gyles CL: **Shiga toxin-producing *Escherichia coli*: an overview**. *Journal of Animal Science* 2007, **85**(13 Suppl):E45–62, [http://www.ncbi.nlm.nih.gov/pubmed/17085726]. [PMID: 17085726].

[186] Lentz EK, Leyva-Illades D, Lee MS, Cherla RP, Tesh VL: **Differential response of the human renal proximal tubular epithelial cell line HK-2 to shiga toxin types 1 and 2**. *Infection and immunity* 2011, **79**(9):3527–3540. [PMID: 21708996].

[187] Gallegos KM, Conrady DG, Karve SS, Gunasekera TS, Herr AB, Weiss AA: **Shiga toxin binding to glycolipids and glycans**. *PloS one* 2012, **7**(2):e30368. [PMID: 22348006].

[188] Blanco M, Blanco JE, Blanco J, Gonzalez EA, Mora A, Prado C, Fernndez L, Rio M, Ramos J, Alonso MP: **Prevalence and characteristics of *Escherichia coli* serotype O157:H7 and other verotoxin-producing E. coli in healthy cattle**. *Epidemiology and Infection* 1996, **117**(2):251–257, [http://www.ncbi.nlm.nih.gov/pubmed/8870622]. [PMID: 8870622].

[189] Beutin L, Geier D, Zimmermann S, Aleksic S, Gillespie HA, Whittam TS: **Epidemiological relatedness and clonal types of natural populations of *Escherichia coli* strains producing shiga toxins in separate populations of cattle and sheep**. *Applied and Environmental Microbiology* 1997, **63**(6):2175–2180, [http://www.ncbi.nlm.nih.gov/pubmed/9172336]. [PMID: 9172336].

[190] Scheutz F, Teel LD, Beutin L, Pirard D, Buvens G, Karch H, Mellmann A, Caprioli A, Tozzoli R, Morabito S, Strockbine NA, Melton-Celsa AR, Sanchez M, Persson S, O'Brien AD: **Multicenter evaluation of a sequence-based protocol for subtyping shiga toxins and standardizing stx nomenclature**. *Journal of clinical microbiology* 2012, **50**(9):2951–2963. [PMID: 22760050].

[191] Zheng J, Cui S, Teel LD, Zhao S, Singh R, O'Brien AD, Meng J: **Identification and characterization of shiga toxin type 2 variants in *Escherichia coli* isolates from Animals, Food, and Humans**. *Applied and Environmental Microbiology* 2008, **74**(18):5645–5652. [PMID: 18658282 PMCID: 2547040].

[192] Sonntag AK, Zenner E, Karch H, Bielaszewska M: **Pigeons as a possible reservoir of shiga toxin 2f-producing *Escherichia coli* pathogenic to humans**. *Berliner Und Mnchener Tierrztliche Wochenschrift* 2005, **118**(11-12):464–470, [http://www.ncbi.nlm.nih.gov/pubmed/16318270]. [PMID: 16318270].

[193] Gannon VP, Teerling C, Masri SA, Gyles CL: **Molecular cloning and nucleotide sequence of another variant of the *Escherichia coli* Shiga-like toxin II family**. *Journal of general microbiology* 1990, **136**(6):1125–1135. [PMID: 2200845].

[194] Prager R, Fruth A, Siewert U, Strutz U, Tschpe H: ***Escherichia coli* encoding shiga toxin 2f as an emerging human pathogen**. *International Journal of Medical Microbiology* 2009, **299**(5):343–353, [http://www.sciencedirect.com/science/article/pii/S1438422108001586].

[195] Prager R, Fruth A, Busch U, Tietze E: **Comparative analysis of virulence genes, genetic diversity, and phylogeny of shiga toxin 2g and heat-stable enterotoxin STIa encoding *Escherichia coli* isolates from humans, animals, and environmental sources**. *International Journal of Medical Microbiology: IJMM* 2011, **301**(3):181–191, [http://www.ncbi.nlm.nih.gov/pubmed/20728406]. [PMID: 20728406].

[196] Pruimboom-Brees IM, Morgan TW, Ackermann MR, Nystrom ED, Samuel JE, Cornick NA, Moon HW: **Cattle lack vascular receptors for *Escherichia coli* O157:H7 shiga toxins**. *Proceedings of the National Academy of Sciences* 2000, **97**(19):10325–10329, [http://www.pnas.org/content/97/19/10325].

[197] Nart P, Naylor SW, Huntley JF, McKendrick IJ, Gally DL, Low JC: **Responses of cattle to gastrointestinal colonization by *Escherichia coli* O157:H7**. *Infect. Immun.* 2008, **76**(11):5366–5372, [http://iai.asm.org/cgi/content/abstract/76/11/5366].

[198] Golan L, Gonen E, Yagel S, Rosenshine I, Shpigel NY: **Enterohemorrhagic *Escherichia coli* induce attaching and effacing lesions and hemorrhagic colitis in human and bovine intestinal xenograft models**. *Disease Models & Mechanisms* 2011, **4**:86–94, [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3014348/]. [PMID: 20959635 PMCID: PMC3014348].

[199] Sheng H, Lim JY, Knecht HJ, Li J, Hovde CJ: **Role of *Escherichia coli* O157:H7 virulence factors in colonization at the bovine terminal rectal mucosa**. *Infection and immunity* 2006, **74**(8):4685–4693. [PMID: 16861656].

[200] Frhlich J, Baljer G, Menge C: **Maternally and naturally acquired antibodies to shiga toxins in a cohort of calves shedding Shiga-toxigenic *Escherichia coli***. *Applied and environmental microbiology* 2009, **75**(11):3695–3704. [PMID: 19363081].

[201] Zumbrun SD, Hanson L, Sinclair JF, Freedy J, Melton-Celsa AR, Rodriguez-Canales J, Hanson JC, O'Brien AD: **Human intestinal tissue and cultured colonic cells contain globotriaosylceramide synthase mRNA and the alternate shiga toxin Receptor, Globotetraosylceramide**. *Infection and Immunity* 2010, [http://www.ncbi.nlm.nih.gov/pubmed/20732996]. [PMID: 20732996].

[202] Wahome PG, Robertus JD, Mantis NJ: **Small-molecule inhibitors of ricin and shiga toxins**. *Current topics in microbiology and immunology* 2012, **357**:179–207, [http://www.ncbi.nlm.nih.gov/pubmed/22006183]. [PMID: 22006183].

[203] Johannes L, Rmer W: **Shiga toxins from cell biology to biomedical applications**. *Nature Reviews Microbiology* 2010, **8**(2):105–116, [http://www.nature.com/nrmicro/journal/v8/n2/abs/nrmicro2279.html].

[204] Lauvrak SU, Torgersen ML, Sandvig K: **Efficient endosome-to-golgi transport of shiga toxin is dependent on dynamin and clathrin**. *Journal of cell science* 2004, **117**(Pt 11):2321–2331, [http://www.ncbi.nlm.nih.gov/pubmed/15126632]. [PMID: 15126632].

[205] Rmer W, Berland L, Chambon V, Gaus K, Windschiegl B, Tenza D, Aly MRE, Fraisier V, Florent JC, Perrais D, Lamaze C, Raposo G, Steinem C, Sens P, Bassereau P, Johannes

L: **Shiga toxin induces tubular membrane invaginations for its uptake into cells**. *Nature* 2007, **450**(7170):670–675, [http://www.ncbi.nlm.nih.gov/pubmed/18046403]. [PMID: 18046403].

[206] Sandvig K, Garred O, Prydz K, Kozlov JV, Hansen SH, van Deurs B: **Retrograde transport of endocytosed shiga toxin to the endoplasmic reticulum**. *Nature* 1992, **358**(6386):510–512, [http://www.ncbi.nlm.nih.gov/pubmed/1641040]. [PMID: 1641040].

[207] White J, Johannes L, Mallard F, Girod A, Grill S, Reinsch S, Keller P, Tzschaschel B, Echard A, Goud B, Stelzer EH: **Rab6 coordinates a novel golgi to ER retrograde transport pathway in live cells**. *The Journal of cell biology* 1999, **147**(4):743–760, [http://www.ncbi.nlm.nih.gov/pubmed/10562278]. [PMID: 10562278].

[208] Garred O, van Deurs B, Sandvig K: **Furin-induced cleavage and activation of shiga toxin**. *The Journal of biological chemistry* 1995, **270**(18):10817–10821, [http://www.ncbi.nlm.nih.gov/pubmed/7738018]. [PMID: 7738018].

[209] Yu M, Haslam DB: **Shiga toxin is transported from the endoplasmic reticulum following interaction with the luminal chaperone HEDJ/ERdj3**. *Infection and immunity* 2005, **73**(4):2524–2532, [http://www.ncbi.nlm.nih.gov/pubmed/15784599]. [PMID: 15784599].

[210] Aletrari MO, McKibbin C, Williams H, Pawar V, Pietroni P, Lord JM, Flitsch SL, Whitehead R, Swanton E, High S, Spooner RA: **Eeyarestatin 1 interferes with both retrograde and anterograde intracellular trafficking pathways**. *PloS one* 2011, **6**(7):e22713, [http://www.ncbi.nlm.nih.gov/pubmed/21799938]. [PMID: 21799938].

[211] Tam PJ, Lingwood CA: **Membrane cytosolic translocation of verotoxin A1 subunit in target cells**. *Microbiology (Reading, England)* 2007, **153**(Pt 8):2700–2710, [http://www.ncbi.nlm.nih.gov/pubmed/17660434]. [PMID: 17660434].

[212] Endo Y, Tsurugi K, Yutsudo T, Takeda Y, Ogasawara T, Igarashi K: **Site of action of a vero toxin (VT2) from *Escherichia coli* O157:H7 and of shiga toxin on eukaryotic ribosomes. RNA N-glycosidase activity of the toxins**. *European journal of biochemistry / FEBS***171**(1-2), [http://www.ncbi.nlm.nih.gov/pubmed/3276522].

[213] O'Brien AD, Tesh VL, Donohue-Rolfe A, Jackson MP, Olsnes S, Sandvig K, Lindberg AA, Keusch GT: **Shiga toxin: biochemistry, genetics, mode of action, and role in pathogenesis**. *Curr Top Microbiol Immunol* 1992, **180**:65–94, [1324134].

[214] Korcheva V, Wong J, Corless C, Iordanov M, Magun B: **Administration of ricin induces a severe inflammatory response via nonredundant stimulation of erk, jnk, and P38 MAPK and provides a mouse model of hemolytic uremic syndrome**. *The American journal of pathology* 2005, **166**:323–339. [PMID: 15632024].

[215] Jandhyala DM, Rogers TJ, Kane A, Paton AW, Paton JC, Thorpe CM: **Shiga toxin 2 and flagellin from Shiga-toxigenic *Escherichia coli* superinduce Interleukin-8 through synergistic effects on host Stress-activated protein kinase Activation**. *Infect. Immun.* 2010, **78**(7):2984–2994, [http://iai.asm.org/cgi/content/abstract/78/7/2984].

[216] O'Loughlin EV, Robins-Browne RM: **Effect of shiga toxin and Shiga-like toxins on eukaryotic cells**. *Microbes and Infection* 2001, **3**(6):493–507, [http://www.sciencedirect.com/science/article/B6VPN-4349FX5-7/2/ba4d7c5127e3468e252152b1e34c0311].

[217] Harrison LM, van Haaften WCE, Tesh VL: **Regulation of proinflammatory cytokine expression by shiga toxin 1 and/or lipopolysaccharides in the human monocytic cell line THP-1**. *Infect. Immun.* 2004, **72**(5):2618–2627, [http://iai.asm.org/cgi/content/abstract/72/5/2618].

[218] Morigi M, Galbusera M, Gastoldi S, Locatelli M, Buelli S, Pezzotta A, Pagani C, Noris M, Gobbi M, Stravalaci M, Rottoli D, Tedesco F, Remuzzi G, Zoja C: **Alternative pathway activation of complement by shiga toxin promotes exuberant C3a formation that triggers microvascular thrombosis**. *Journal of immunology (Baltimore, Md.: 1950)* 2011, **187**:172–180. [PMID: 21642543].

[219] Brigotti M, Tazzari PL, Ravanelli E, Carnicelli D, Rocchi L, Arfilli V, Scavia G, Minelli F, Ricci F, Pagliaro P, Ferretti AVS, Pecoraro C, Paglialonga F, Edefonti A, Procaccino MA, Tozzi AE, Caprioli A: **Clinical relevance of shiga toxin concentrations in the blood of patients with hemolytic uremic syndrome**. *The Pediatric infectious disease journal* 2011, **30**(6):486–490, [http://www.ncbi.nlm.nih.gov/pubmed/21164386]. [PMID: 21164386].

[220] Zhang Y, Laing C, Zhang Z, Hallewell J, You C, Ziebell K, Johnson RP, Kropinski AM, Thomas JE, Karmali M, Gannon VPJ: **Lineage and host source are both correlated with levels of shiga toxin 2 production by *Escherichia coli* O157:H7 strains**. *Applied and Environmental Microbiology* 2010, **76**(2):474–482, [http://www.ncbi.nlm.nih.gov/pubmed/19948861]. [PMID: 19948861].

[221] Eaton KA, Friedman DI, Francis GJ, Tyler JS, Young VB, Haeger J, Abu-Ali G, Whittam TS: **Pathogenesis of renal disease due to enterohemorrhagic *Escherichia coli* in germ-free mice**. *Infection and Immunity* 2008, **76**(7):3054–3063, [http://www.ncbi.nlm.nih.gov/pubmed/18443087]. [PMID: 18443087].

[222] Steinberg KM, Levin BR: **Grazing protozoa and the evolution of the *Escherichia coli* O157:H7 shiga toxin-encoding prophage**. *Proceedings. Biological Sciences / The Royal Society* 2007, **274**(1621):1921–1929, [http://www.ncbi.nlm.nih.gov/pubmed/17535798]. [PMID: 17535798].

[223] Lainhart W, Stolfa G, Koudelka GB: **Shiga toxin as a bacterial defense against a eukaryotic Predator, tetrahymena thermophila**. *Journal of Bacteriology* 2009, **191**(16):5116–5122, [http://jb.asm.org/content/191/16/5116].

[224] Ferens WA, Cobbold R, Hovde CJ: **Intestinal shiga Toxin-producing *Escherichia coli* bacteria mitigate bovine leukemia virus infection in experimentally infected Sheep**. *Infect. Immun.* 2006, **74**(5):2906–2916, [http://iai.asm.org/cgi/content/abstract/74/5/2906].

[225] Shaikh N, Tarr PI: ***Escherichia coli* O157:H7 shiga toxin-encoding bacteriophages: integrations, excisions, truncations, and evolutionary implications**. *J Bacteriol* 2003, **185**(12):3596–605, [12775697].

[226] Brussow H, Canchaya C, Hardt WD: **Phages and the evolution of bacterial Pathogens: from genomic rearrangements to lysogenic Conversion**. *Microbiol. Mol. Biol. Rev.* 2004, **68**(3):560–602, [http://mmbr.asm.org/cgi/content/abstract/68/3/560].

[227] Botstein D: **A theory of modular evolution for bacteriophages**. *Annals of the New York Academy of Sciences* 1980, **354**:484–490, [http://www.ncbi.nlm.nih.gov/pubmed/6452848]. [PMID: 6452848].

[228] Allison HE: **Stx-phages: drivers and mediators of the evolution of STEC and STEC-like pathogens**. *Future Microbiology* 2007, **2**:165–74, [http://www.ncbi.nlm.nih.gov/pubmed/17661653]. [PMID: 17661653].

[229] Li R, Harada T, Honjoh Ki, Miyamoto T: **Phylogenetic analysis and shiga toxin production profiling of shiga toxin-producing/enterohemorrhagic *Escherichia coli* clinical isolates**. *Microbial Pathogenesis* 2010, **49**(5):246–251, [http://www.sciencedirect.com/science/article/B6WN6-509W75G-4/2/ffcc8218f932ff1d806d61c46ff5d287].

[230] Tinsley CR, Bille E, Nassif X: **Bacteriophages and pathogenicity: more than just providing a toxin?** *Microbes and Infection* 2006, **8**(5):1365–1371, [http://www.sciencedirect.com/science/article/B6VPN-4JGJG2C-3/2/64998d48301c589d81f699524d13fce1].

[231] Wagner PL, Acheson DW, Waldor MK: **Human neutrophils and their products induce shiga toxin production by enterohemorrhagic *Escherichia coli***. *Infection and immunity* 2001, **69**(3):1934–1937, [http://www.ncbi.nlm.nih.gov/pubmed/11179378]. [PMID: 11179378].

[232] Zhang X, McDaniel AD, Wolf LE, Keusch GT, Waldor MK, Acheson DWK: **Quinolone antibiotics induce shiga toxin-encoding Bacteriophages, Toxin Production, and death in Mice**. *Journal of Infectious Diseases* 2000, **181**(2):664 –670, [http://jid.oxfordjournals.org/content/181/2/664.abstract].

[233] Merrikh H, Ferrazzoli AE, Bougdour A, Olivier-Mason A, Lovett ST: **A DNA damage response in *Escherichia coli* involving the alternative sigma factor, RpoS**. *Proceedings of the National Academy of Sciences* 2009, **106**(2):611–616, [http://www.pnas.org/content/106/2/611].

[234] Poirier K, Faucher SP, Bland M, Brousseau R, Gannon V, Martin C, Harel J, Daigle F: ***Escherichia coli* O157:H7 survives within human macrophages: global gene expression profile and involvement of the shiga toxins**. *Infection and Immunity* 2008, **76**(11):4814–4822, [http://www.ncbi.nlm.nih.gov/pubmed/18725421]. [PMID: 18725421].

[235] Bielaszewska M, Kck R, Friedrich AW, von Eiff C, Zimmerhackl LB, Karch H, Mellmann A: **Shiga toxin-mediated hemolytic uremic syndrome: time to change the diagnostic paradigm?** *PloS One* 2007, **2**(10):e1024, [http://www.ncbi.nlm.nih.gov/pubmed/17925872]. [PMID: 17925872].

[236] Garca-Aljaro C, Muniesa M, Jofre J, Blanch AR: **Genotypic and phenotypic diversity among induced, stx2-carrying bacteriophages from environmental *Escherichia coli* strains**. *Applied and Environmental Microbiology* 2009, **75**(2):329–336, [http://www.ncbi.nlm.nih.gov/pubmed/19011056]. [PMID: 19011056].

[237] Fogg PCM, Saunders JR, McCarthy AJ, Allison HE: **Cumulative effect of prophage burden on shiga toxin production in *Escherichia coli***. *Microbiology* 2012, **158**(2):488–497, [http://mic.sgmjournals.org/content/158/2/488].

[238] Watarai M, Sato T, Kobayashi M, Shimizu T, Yamasaki S, Tobe T, Sasakawa C, Takeda Y: **Identification and characterization of a newly isolated shiga toxin 2-converting phage from shiga Toxin-producing *Escherichia coli***. *Infect. Immun.* 1998, **66**(9):4100–4107, [http://iai.asm.org/cgi/content/abstract/66/9/4100].

[239] Smith DL, James CE, Sergeant MJ, Yaxian Y, Saunders JR, McCarthy AJ, Allison HE: **Short-tailed stx phages exploit the conserved YaeT protein to disseminate shiga toxin genes among Enterobacteria**. *J. Bacteriol.* 2007, **189**(20):7223–7233, [http://jb.asm.org/cgi/content/abstract/189/20/7223].

[240] Suh JK, Hovde CJ, Robertus JD: **Shiga toxin attacks bacterial ribosomes as effectively as eucaryotic ribosomes**. *Biochemistry* 1998, **37**(26):9394–9398, [http://www.ncbi.nlm.nih.gov/pubmed/9649321]. [PMID: 9649321].

[241] Gamage SD, Strasser JE, Chalk CL, Weiss AA: **Nonpathogenic *Escherichia coli* can contribute to the production of shiga toxin**. *Infection and immunity* 2003, **71**(6):3107–3115, [http://www.ncbi.nlm.nih.gov/pubmed/12761088]. [PMID: 12761088].

[242] Chase-Topping M, Gally D, Low C, Matthews L, Woolhouse M: **Super-shedding and the link between human infection and livestock carriage of *Escherichia coli* O157**. *Nature Reviews. Microbiology* 2008, **6**(12):904–912, [http://www.ncbi.nlm.nih.gov/pubmed/19008890]. [PMID: 19008890].

[243] Kaper JB: **The locus of enterocyte effacement pathogenicity island of shiga toxin-producing *Escherichia coli* O157:H7 and other attaching and effacing E. coli**. *Japanese Journal of Medical Science & Biology* 1998, **51 Suppl**:S101–107, [http://www.ncbi.nlm.nih.gov/pubmed/10211442]. [PMID: 10211442].

[244] Vidal JE, Navarro-Garca F: **EspC translocation into epithelial cells by enteropathogenic *Escherichia coli* requires a concerted participation of type v and III secretion systems**. *Cellular microbiology* 2008, **10**(10):1975–1986, [http://www.ncbi.nlm.nih.gov/pubmed/18547338]. [PMID: 18547338].

[245] Mellies JL, Elliott SJ, Sperandio V, Donnenberg MS, Kaper JB: **The per regulon of enteropathogenic *Escherichia coli* : identification of a regulatory cascade and a novel transcriptional activator, the locus of enterocyte effacement (LEE)-encoded regulator (Ler)**. *Molecular microbiology* 1999, **33**(2):296–306, [http://www.ncbi.nlm.nih.gov/pubmed/10411746]. [PMID: 10411746].

[246] Elliott SJ, Wainwright LA, McDaniel TK, Jarvis KG, Deng YK, Lai LC, McNamara BP, Donnenberg MS, Kaper JB: **The complete sequence of the locus of enterocyte effacement**

(LEE) from enteropathogenic *Escherichia coli* E2348/69. *Molecular microbiology* 1998, **28**:1–4, [http://www.ncbi.nlm.nih.gov/pubmed/9593291]. [PMID: 9593291].

[247] Nguyen Y, Sperandio V: **Enterohemorrhagic E. coli (EHEC) pathogenesis**. *Frontiers in Cellular and Infection Microbiology* 2012, **2**:90, [http://www.frontiersin.org/Cellular_and_Infection_Microbiology/10.3389/fcimb.2012.00090/full].

[248] Sekiya K, Ohishi M, Ogino T, Tamano K, Sasakawa C, Abe A: **Supermolecular structure of the enteropathogenic *Escherichia coli* type III secretion system and its direct interaction with the EspA-sheath-like structure**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(20):11638–11643, [http://www.ncbi.nlm.nih.gov/pubmed/11562461]. [PMID: 11562461].

[249] Wilson RK, Shaw RK, Daniell S, Knutton S, Frankel G: **Role of escf, a putative needle complex protein, in the type III protein translocation system of enteropathogenic *Escherichia coli***. *Cellular microbiology* 2001, **3**(11):753–762, [http://www.ncbi.nlm.nih.gov/pubmed/11696035]. [PMID: 11696035].

[250] Kenny B, DeVinney R, Stein M, Reinscheid DJ, Frey EA, Finlay BB: **Enteropathogenic E. coli (EPEC) transfers its receptor for intimate adherence into mammalian cells**. *Cell* 1997, **91**(4):511–520, [http://www.ncbi.nlm.nih.gov/pubmed/9390560]. [PMID: 9390560].

[251] Fitzhenry RJ, Pickard DJ, Hartland EL, Reece S, Dougan G, Phillips AD, Frankel G: **Intimin type influences the site of human intestinal mucosal colonisation by enterohaemorrhagic *Escherichia coli* O157:H7**. *Gut* 2002, **50**(2):180 –185, [http://gut.bmj.com/content/50/2/180.abstract].

[252] Torres AG, Zhou X, Kaper JB: **Adherence of diarrheagenic *Escherichia coli* strains to epithelial Cells**. *Infect. Immun.* 2005, **73**:18–29, [http://iai.asm.org].

[253] Beutin L, Marchs O, Bettelheim KA, Gleier K, Zimmermann S, Schmidt H, Oswald E: **HEp-2 cell adherence, actin aggregation, and intimin types of attaching and effacing *Escherichia coli* strains isolated from healthy infants in germany and Australia**. *Infection and Immunity* 2003, **71**(7):3995–4002, [http://www.ncbi.nlm.nih.gov/pubmed/12819087]. [PMID: 12819087].

[254] Bustamante VH, Santana FJ, Calva E, Puente JL: **Transcriptional regulation of type III secretion genes in enteropathogenic *Escherichia coli*: ler antagonizes H-NS-dependent repression**. *Molecular microbiology* 2001, **39**(3):664–678, [http://www.ncbi.nlm.nih.gov/pubmed/11169107]. [PMID: 11169107].

[255] Friedberg D, Umanski T, Fang Y, Rosenshine I: **Hierarchy in the expression of the locus of enterocyte effacement genes of enteropathogenic *Escherichia coli***. *Molecular microbiology* 1999, **34**(5):941–952, [http://www.ncbi.nlm.nih.gov/pubmed/10594820]. [PMID: 10594820].

[256] Grant AJ, Farris M, Alefounder P, Williams PH, Woodward MJ, O'Connor CD: **Co-ordination of pathogenicity island expression by the BipA GTPase in enteropathogenic**

***Escherichia coli* (EPEC)**. *Molecular microbiology* 2003, **48**(2):507–521, [http://www.ncbi.nlm.nih.gov/pubmed/12675808]. [PMID: 12675808].

[257] Goldberg MD, Johnson M, Hinton JC, Williams PH: **Role of the nucleoid-associated protein fis in the regulation of virulence properties of enteropathogenic *Escherichia coli***. *Molecular microbiology* 2001, **41**(3):549–559, [http://www.ncbi.nlm.nih.gov/pubmed/11532124]. [PMID: 11532124].

[258] Iyoda S, Watanabe H: **ClpXP protease controls expression of the type III protein secretion system through regulation of RpoS and GrlR levels in enterohemorrhagic *Escherichia coli***. *Journal of bacteriology* 2005, **187**(12):4086–4094, [http://www.ncbi.nlm.nih.gov/pubmed/15937171]. [PMID: 15937171].

[259] Sharma VK, Zuerner RL: **Role of hha and ler in transcriptional regulation of the esp operon of enterohemorrhagic *Escherichia coli* O157:H7**. *Journal of bacteriology* 2004, **186**(21):7290–7301, [http://www.ncbi.nlm.nih.gov/pubmed/15489441]. [PMID: 15489441].

[260] Iyoda S, Honda N, Saitoh T, Shimuta K, Terajima J, Watanabe H, Ohnishi M: **Coordinate control of the locus of enterocyte effacement and enterohemolysin genes by multiple common virulence regulators in enterohemorrhagic *Escherichia coli***. *Infection and immunity* 2011, **79**(11):4628–4637, [http://www.ncbi.nlm.nih.gov/pubmed/21844237]. [PMID: 21844237].

[261] Barba J, Bustamante VH, Flores-Valdez MA, Deng W, Finlay BB, Puente JL: **A positive regulatory loop controls expression of the locus of enterocyte effacement-encoded regulators ler and GrlA**. *Journal of bacteriology* 2005, **187**(23):7918–7930, [http://www.ncbi.nlm.nih.gov/pubmed/16291665]. [PMID: 16291665].

[262] Brito PH, Rocha EPC, Xavier KB, Gordo I: **Natural genome diversity of AI-2 quorum sensing in *Escherichia coli*: conserved signal production but labile signal Reception**. *Genome biology and evolution* 2013, **5**:16–30. [PMID: 23246794].

[263] Sircili MP, Walters M, Trabulsi LR, Sperandio V: **Modulation of enteropathogenic *Escherichia coli* virulence by quorum sensing**. *Infection and immunity* 2004, **72**(4):2329–2337, [http://www.ncbi.nlm.nih.gov/pubmed/15039358]. [PMID: 15039358].

[264] Njoroge J, Sperandio V: **Enterohemorrhagic *Escherichia coli* virulence regulation by two bacterial adrenergic kinases, QseC and QseE**. *Infection and immunity* 2012, **80**(2):688–703, [http://www.ncbi.nlm.nih.gov/pubmed/22144490]. [PMID: 22144490].

[265] Flockhart AF, Tree JJ, Xu X, Karpiyevich M, McAteer SP, Rosenblum R, Shaw DJ, Low CJ, Best A, Gannon V, Laing C, Murphy KC, Leong JM, Schneiders T, La Ragione R, Gally DL: **Identification of a novel prophage regulator in *Escherichia coli* controlling the expression of type III secretion**. *Molecular microbiology* 2012, **83**:208–223, [http://www.ncbi.nlm.nih.gov/pubmed/22111928]. [PMID: 22111928].

[266] Gomez-Duarte OG, Kaper JB: **A plasmid-encoded regulatory region activates chromosomal eaeA expression in enteropathogenic *Escherichia coli***. *Infect Immun* 1995, **63**(5):1767–76, [7729884].

[267] Iyoda S, Watanabe H: **Positive effects of multiple pch genes on expression of the locus of enterocyte effacement genes and adherence of enterohaemorrhagic *Escherichia coli* O157 : H7 to HEp-2 cells**. *Microbiology (Reading, England)* 2004, **150**(Pt 7):2357–2571, [http://www.ncbi.nlm.nih.gov/pubmed/15256577]. [PMID: 15256577].

[268] Honda N, Iyoda S, Yamamoto S, Terajima J, Watanabe H: **LrhA positively controls the expression of the locus of enterocyte effacement genes in enterohemorrhagic *Escherichia coli* by differential regulation of their master regulators PchA and PchB**. *Molecular microbiology* 2009, **74**(6):1393–1341, [http://www.ncbi.nlm.nih.gov/pubmed/19889091]. [PMID: 19889091].

[269] Lehnen D, Blumer C, Polen T, Wackwitz B, Wendisch VF, Unden G: **LrhA as a new transcriptional key regulator of flagella, motility and chemotaxis genes in *Escherichia coli***. *Molecular microbiology* 2002, **45**(2):521–532, [http://www.ncbi.nlm.nih.gov/pubmed/12123461]. [PMID: 12123461].

[270] Robinson CM, Sinclair JF, Smith MJ, O'Brien AD: **Shiga toxin of enterohemorrhagic *Escherichia coli* type O157:H7 promotes intestinal colonization**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(25):9667–9672, [http://www.ncbi.nlm.nih.gov/pubmed/16766659]. [PMID: 16766659].

[271] Xu X, McAteer SP, Tree JJ, Shaw DJ, Wolfson EBK, Beatson SA, Roe AJ, Allison LJ, Chase-Topping ME, Mahajan A, Tozzoli R, Woolhouse MEJ, Morabito S, Gally DL: **Lysogeny with shiga toxin 2-encoding bacteriophages represses type III secretion in enterohemorrhagic *Escherichia coli***. *PLoS Pathog* 2012, **8**(5):e1002672, [http://dx.doi.org/10.1371/journal.ppat.1002672].

[272] Bhatt S, Romeo T, Kalman D: **Honing the message: post-transcriptional and post-translational control in attaching and effacing pathogens**. *Trends in microbiology* 2011, **19**(5):217–224, [http://www.ncbi.nlm.nih.gov/pubmed/21333542]. [PMID: 21333542].

[273] Bhatt S, Edwards AN, Nguyen HTT, Merlin D, Romeo T, Kalman D: **The RNA binding protein CsrA is a pleiotropic regulator of the locus of enterocyte effacement pathogenicity island of enteropathogenic *Escherichia coli***. *Infection and immunity* 2009, **77**(9):3552–3568, [http://www.ncbi.nlm.nih.gov/pubmed/19581394]. [PMID: 19581394].

[274] Hansen AM, Kaper JB: **Hfq affects the expression of the LEE pathogenicity island in enterohaemorrhagic *Escherichia coli***. *Molecular microbiology* 2009, **73**(3):446–465, [http://www.ncbi.nlm.nih.gov/pubmed/19570135]. [PMID: 19570135].

[275] Lodato PB, Kaper JB: **Post-transcriptional processing of the LEE4 operon in enterohaemorrhagic *Escherichia coli***. *Molecular microbiology* 2009, **71**(2):273–290, [http://www.ncbi.nlm.nih.gov/pubmed/19019141]. [PMID: 19019141].

[276] Sledjeski DD, Whitman C, Zhang A: **Hfq is necessary for regulation by the untranslated RNA DsrA**. *Journal of bacteriology* 2001, **183**(6):1997–2005, [http://www.ncbi.nlm.nih.gov/pubmed/11222598]. [PMID: 11222598].

[277] Laaberki MH, Janabi N, Oswald E, Repoila F: **Concert of regulators to switch on LEE expression in enterohemorrhagic *Escherichia coli* O157:H7: interplay between Ler, grla, HNS and RpoS**. *International journal of medical microbiology: IJMM* 2006, **296**(4-5):197–210, [http://www.ncbi.nlm.nih.gov/pubmed/16618552]. [PMID: 16618552].

[278] Karmali MA: **Host and pathogen determinants of verocytotoxin-producing escherichia coli-associated hemolytic uremic syndrome**. *Kidney International. Supplement* 2009, (112):S4–7, [http://www.ncbi.nlm.nih.gov/pubmed/19180132]. [PMID: 19180132].

[279] Tobe T, Beatson SA, Taniguchi H, Abe H, Bailey CM, Fivian A, Younis R, Matthews S, Marches O, Frankel G, Hayashi T, Pallen MJ: **An extensive repertoire of type III secretion effectors in *Escherichia coli* O157 and the role of lambdoid phages in their dissemination**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(40):14941–6, [http://www.ncbi.nlm.nih.gov/pubmed/16990433]. [PMID: 16990433].

[280] Parsot C: **Shigella type III secretion effectors: how, where, when, for what purposes?** *Current opinion in microbiology* 2009, **12**:110–116, [http://www.ncbi.nlm.nih.gov/pubmed/19157960]. [PMID: 19157960].

[281] Coombes BK, Wickham ME, Mascarenhas M, Gruenheid S, Finlay BB, Karmali MA: **Molecular analysis as an aid to assess the public health risk of non-O157 shiga toxin-producing *Escherichia coli* strains**. *Applied and Environmental Microbiology* 2008, **74**(7):2153–2160, [http://www.ncbi.nlm.nih.gov/pubmed/18245257]. [PMID: 18245257].

[282] Wu B, Skarina T, Yee A, Jobin MC, Dileo R, Semesi A, Fares C, Lemak A, Coombes BK, Arrowsmith CH, Singer AU, Savchenko A: **NleG type 3 effectors from enterohaemorrhagic *Escherichia coli* are U-box E3 ubiquitin ligases**. *PLoS pathogens* 2010, **6**(6):e1000960, [http://www.ncbi.nlm.nih.gov/pubmed/20585566]. [PMID: 20585566].

[283] Sham HP, Shames SR, Croxen MA, Ma C, Chan JM, Khan MA, Wickham ME, Deng W, Finlay BB, Vallance BA: **Attaching and effacing bacterial effector NleC suppresses epithelial inflammatory responses by inhibiting NF-B and p38 mitogen-activated protein kinase activation**. *Infection and immunity* 2011, **79**(9):3552–3562, [http://www.ncbi.nlm.nih.gov/pubmed/21746856]. [PMID: 21746856].

[284] Royan SV, Jones RM, Koutsouris A, Roxas JL, Falzari K, Weflen AW, Kim A, Bellmeyer A, Turner JR, Neish AS, Rhee KJ, Viswanathan VK, Hecht GA: **Enteropathogenic E. coli non-LEE encoded effectors NleH1 and NleH2 attenuate NF-B activation**. *Molecular microbiology* 2010, **78**(5):1232–1245, [http://www.ncbi.nlm.nih.gov/pubmed/21091507]. [PMID: 21091507].

[285] Thanabalasuriar A, Koutsouris A, Weflen A, Mimee M, Hecht G, Gruenheid S: **The bacterial virulence factor NleA is required for the disruption of intestinal tight junctions by enteropathogenic *Escherichia coli***. *Cellular microbiology* 2010, **12**:31–41, [http://www.ncbi.nlm.nih.gov/pubmed/19712078]. [PMID: 19712078].

[286] Schwidder M, Hensel M, Schmidt H: **Regulation of nleA in shiga toxin-producing *Escherichia coli* O84:H4 strain 4795/97**. *Journal of bacteriology* 2011, **193**(4):832–841, [http://www.ncbi.nlm.nih.gov/pubmed/21131485]. [PMID: 21131485].

[287] Blattner FR, Plunkett r G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y: **The complete genome sequence of *Escherichia coli* K-12**. *Science (New York, N.Y.)* 1997, **277**(5331):1453–1462. [PMID: 9278503].

[288] Neidhardt FC, B B: *Escherichia coli and Salmonella: cellular and molecular Biology*. American Society for Microbiology 1996.

[289] Lukjancenko O, Wassenaar TM, Ussery DW: **Comparison of 61 sequenced *Escherichia coli* genomes**. *Microbial Ecology* 2010, **60**(4):708–720, [http://www.ncbi.nlm.nih.gov/pubmed/20623278]. [PMID: 20623278].

[290] Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin CS, Iliopoulos D, Klammer A, Peluso P, Lee L, Kislyuk AO, Bullard J, Kasarskis A, Wang S, Eid J, Rank D, Redman JC, Steyert SR, Frimodt-Mller J, Struve C, Petersen AM, Krogfelt KA, Nataro JP, Schadt EE, Waldor MK: **Origins of the E. coli strain causing an outbreak of Hemolytic-uremic syndrome in Germany**. *The New England Journal of Medicine* 2011, [http://www.ncbi.nlm.nih.gov/pubmed/21793740]. [PMID: 21793740].

[291] Pupo GM, Karaolis DK, Lan R, Reeves PR: **Evolutionary relationships among pathogenic and nonpathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and mdh sequence studies**. *Infect Immun* 1997, **65**(7):2685–92, [9199437].

[292] Newton HJ, Sloan J, Bulach DM, Seemann T, Allison CC, Tauschek M, Robins-Browne RM, Paton JC, Whittam TS, Paton AW, Hartland EL: **Shiga toxin-producing *Escherichia coli* strains negative for locus of enterocyte effacement**. *Emerging Infectious Diseases* 2009, **15**(3):372–380, [http://www.ncbi.nlm.nih.gov/pubmed/19239748]. [PMID: 19239748].

[293] Mellmann A, Fruth A, Friedrich AW, Wieler LH, Harmsen D, Werber D, Middendorf B, Bielaszewska M, Karch H: **Phylogeny and disease association of shiga toxin-producing *Escherichia coli* O91**. *Emerging Infectious Diseases* 2009, **15**(9):1474–1477, [http://www.ncbi.nlm.nih.gov/pubmed/19788818]. [PMID: 19788818].

[294] Tarr CL, Nelson AM, Beutin L, Olsen KEP, Whittam TS: **Molecular characterization reveals similar virulence gene content in unrelated clonal groups of *Escherichia coli* of serogroup O174 (OX3)**. *Journal of Bacteriology* 2008, **190**(4):1344–1349, [http://www.ncbi.nlm.nih.gov/pubmed/18083801]. [PMID: 18083801].

[295] Vogt SL, Nevesinjac AZ, Humphries RM, Donnenberg MS, Armstrong GD, Raivio TL: **The cpx envelope stress response both facilitates and inhibits elaboration of the enteropathogenic *Escherichia coli* bundle-forming pilus**. *Molecular Microbiology* 2010, **76**(5):1095–1110, [http://www.ncbi.nlm.nih.gov/pubmed/20444097]. [PMID: 20444097].

[296] Cookson AL, Bennett J, Thomson-Carter F, Attwood GT: **Molecular subtyping and genetic analysis of the enterohemolysin gene (ehxA) from shiga toxin-producing *Escherichia coli* and atypical enteropathogenic E. coli**. *Applied and Environmental Microbiology* 2007, **73**(20):6360–6369, [http://www.ncbi.nlm.nih.gov/pubmed/17720842]. [PMID: 17720842].

[297] Cookson AL, Cao M, Bennett J, Nicol C, Thomson-Carter F, Attwood GT: **Relationship between virulence gene profiles of atypical enteropathogenic *Escherichia coli* and shiga toxin-producing E. coli isolates from cattle and sheep in new Zealand**. *Applied and Environmental Microbiology* 2010, **76**(11):3744–3747, [http://www.ncbi.nlm.nih.gov/pubmed/20400570]. [PMID: 20400570].

[298] Paton AW, Paton JC: **Multiplex PCR for direct detection of shiga toxigenic *Escherichia coli* strains producing the novel subtilase Cytotoxin**. *J. Clin. Microbiol.* 2005, **43**(6):2944–2947, [http://jcm.asm.org/cgi/content/abstract/43/6/2944].

[299] Law D: **Virulence factors of *Escherichia coli* O157 and other shiga toxinproducing E. coli**. *Journal of Applied Microbiology* 2000, **88**(5):729–745, [http://onlinelibrary.wiley.com/doi/10.1046/j.1365-2672.2000.01031.x/abstract].

[300] Large TM, Walk ST, Whittam TS: **Variation in acid resistance among shiga toxin-producing clones of pathogenic *Escherichia coli***. *Appl Environ Microbiol* 2005, **71**(5):2493–500, [15870339].

[301] Coldewey SM, Hartmann M, Schmidt DS, Engelking U, Ukena SN, Gunzer F: **Impact of the rpoS genotype for acid resistance patterns of pathogenic and probiotic *Escherichia coli***. *BMC Microbiology* 2007, **7**:21, [http://www.ncbi.nlm.nih.gov/pubmed/17386106]. [PMID: 17386106].

[302] Bhagwat AA, Tan J, Sharma M, Kothary M, Low S, Tall BD, Bhagwat M: **Functional heterogeneity of RpoS in stress tolerance of enterohemorrhagic *Escherichia coli* Strains**. *Applied and Environmental Microbiology* 2006, **72**(7):4978–4986. [PMID: 16820496 PMCID: 1489321].

[303] Orth D, Grif K, Dierich MP, Wrzner R: **Prevalence, structure and expression of urease genes in shiga toxin-producing *Escherichia coli* from humans and the environment**. *International Journal of Hygiene and Environmental Health* 2006, **209**(6):513–520, [http://www.sciencedirect.com/science/article/B7GVY-4KJ0SYW-1/2/71d11f6da3194c9cb7f0855b6e23f56d].

[304] Yin X, Wheatcroft R, Chambers JR, Liu B, Zhu J, Gyles CL: **Contributions of o island 48 to adherence of enterohemorrhagic *Escherichia coli* O157:H7 to epithelial cells in vitro and in ligated pig ileal Loops**. *Appl. Environ. Microbiol.* 2009, **75**(18):5779–5786, [http://aem.asm.org/cgi/content/abstract/75/18/5779].

[305] Toma C, Martnez Espinosa E, Song T, Miliwebsky E, Chinen I, Iyoda S, Iwanaga M, Rivas M: **Distribution of putative adhesins in different seropathotypes of shiga Toxin-producing *Escherichia coli***. *Journal of Clinical Microbiology* 2004, **42**(11):4937–4946. [PMID: 15528677 PMCID: 525252].

[306] Tarr PI, Bilge SS, Vary JCJ, Jelacic S, Habeeb RL, Ward TR, Baylor MR, Besser TE: **Iha: a novel *Escherichia coli* O157:H7 adherence-conferring molecule encoded on a recently acquired chromosomal island of conserved structure**. *Infect Immun* 2000, **68**(3):1400–7, [10678953].

[307] Nicholls L, Grant TH, Robins-Browne RM: **Identification of a novel genetic locus that is required for in vitro adhesion of a clinical isolate of enterohaemorrhagic *Escherichia coli* to epithelial cells**. *Molecular Microbiology* 2000, **35**(2):275–288, [http://www.ncbi.nlm.nih.gov/pubmed/10652089]. [PMID: 10652089].

[308] Lu Y, Iyoda S, Satou H, Satou H, Itoh K, Saitoh T, Watanabe H: **A new immunoglobulin-binding protein, eibg, is responsible for the chain-like adhesion phenotype of locus of enterocyte effacement-negative, shiga toxin-producing *Escherichia coli***. *Infect Immun* 2006, **74**(10):5747–55, [16988252].

[309] Miller VL, Beer KB, Loomis WP, Olson JA, Miller SI: **An unusual pagC::TnphoA mutation leads to an invasion- and virulence-defective phenotype in Salmonellae**. *Infection and Immunity* 1992, **60**(9):3763–3770, [http://www.ncbi.nlm.nih.gov/pubmed/1323535]. [PMID: 1323535].

[310] Okeke IN, Scaletsky IC, Soars EH, Macfarlane LR, Torres AG: **Molecular epidemiology of the iron utilization genes of enteroaggregative *Escherichia coli***. *J Clin Microbiol* 2004, **42**:36–44, [14715729].

[311] Karch H, Schubert S, Zhang D, Zhang W, Schmidt H, Olschlager T, Hacker J: **A genomic island, termed High-pathogenicity Island, Is present in certain Non-O157 shiga Toxin-producing *Escherichia coli* clonal Lineages**. *Infect. Immun.* 1999, **67**(11):5994–6001, [http://iai.asm.org/cgi/content/abstract/67/11/5994].

[312] Afset JE, Bruant G, Brousseau R, Harel J, Anderssen E, Bevanger L, Bergh K: **Identification of virulence genes linked with diarrhea due to atypical enteropathogenic *Escherichia coli* by DNA microarray analysis and PCR**. *Journal of Clinical Microbiology* 2006, **44**(10):3703–3711, [http://www.ncbi.nlm.nih.gov/pubmed/17021100]. [PMID: 17021100].

[313] Zhang W, Zhao M, Ruesch L, Omot A, Francis D: **Prevalence of virulence genes in *Escherichia coli* strains recently isolated from young pigs with diarrhea in the US**. *Veterinary Microbiology* 2007, **123**(1-3):145–152, [http://www.ncbi.nlm.nih.gov/pubmed/17368762]. [PMID: 17368762].

[314] Ogura Y, Ooka T, Iguchi A, Toh H, Asadulghani M, Oshima K, Kodama T, Abe H, Nakayama K, Kurokawa K, Tobe T, Hattori M, Hayashi T: **Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic *Escherichia coli*** 2009, **106**(42):17939–17944. [PMID: 19815525 PMCID: 2764950].

[315] Souza MRSM, Klassen G, Toni FD, Rigo LU, Henkes C, Pigatto CP, Dalagassa CdB, Fadel-Picheth CMT: **Biochemical properties, enterohaemolysin production and plasmid carriage of shiga toxin-producing *Escherichia coli* strains**. *Memrias Do Instituto Oswaldo Cruz* 2010, **105**(3):318–321, [http://www.ncbi.nlm.nih.gov/pubmed/20512247]. [PMID: 20512247].

[316] Fratamico PM, Yan X, Caprioli A, Esposito G, Needleman DS, Pepe T, Tozzoli R, Cortesi ML, Morabito S: **The complete DNA sequence and analysis of the virulence plasmid and of five additional plasmids carried by shiga toxin-producing *Escherichia coli* O26:H11 strain H30**. *International Journal of Medical Microbiology: IJMM* 2011, **301**(3):192–203, [http://www.ncbi.nlm.nih.gov/pubmed/21212019]. [PMID: 21212019].

[317] Beutin L, Montenegro MA, Orskov I, Orskov F, Prada J, Zimmermann S, Stephan R: **Close association of verotoxin (Shiga-like toxin) production with enterohemolysin production in strains of *Escherichia coli***. *J Clin Microbiol* 1989, **27**(11):2559–64, [2681256].

[318] von Mffling T, Smaijlovic M, Nowak B, Sammet K, Blte M, Klein G: **Preliminary study of certain serotypes, genetic and antimicrobial resistance profiles of verotoxigenic *Escherichia coli* (VTEC) isolated in bosnia and germany from cattle or pigs and their products**. *International Journal of Food Microbiology* 2007, **117**(2):185–191, [http://www.ncbi.nlm.nih.gov/pubmed/17011059]. [PMID: 17011059].

[319] Li H, Granat A, Stewart V, Gillespie JR: **rpos, H-ns, and DsrA influence EHEC hemolysin operon (ehxCABD) transcription in *Escherichia coli* O157:H7 strain EDL933**. *FEMS microbiology letters* 2008, **285**(2):257–262. [PMID: 18616595].

[320] Brockmeyer J, Bielaszewska M, Fruth A, Bonn ML, Mellmann A, Humpf HU, Karch H: **Subtypes of the plasmid-encoded serine protease EspP in shiga toxin-producing *Escherichia coli*: distribution, secretion, and proteolytic activity**. *Applied and Environmental Microbiology* 2007, **73**(20):6351–6359, [http://www.ncbi.nlm.nih.gov/pubmed/17704265]. [PMID: 17704265].

[321] Tozzoli R, Caprioli A, Cappannella S, Michelacci V, Marziano ML, Morabito S: **Production of the subtilase AB5 cytotoxin by shiga toxin-negative *Escherichia coli***. *Journal of Clinical Microbiology* 2010, **48**:178–183, [http://www.ncbi.nlm.nih.gov/pubmed/19940059]. [PMID: 19940059].

[322] Uhlich GA, Chen CY, Cottrell BJ, Irwin PL, Phillips JG: **Peroxide resistance in *Escherichia coli* serotype O157 : H7 biofilms is regulated by both RpoS-dependent and -independent mechanisms**. *Microbiology (Reading, England)* 2012, **158**(Pt 9):2225–2234, [http://www.ncbi.nlm.nih.gov/pubmed/22700652]. [PMID: 22700652].

[323] Uhlich GA: **KatP contributes to OxyR-regulated hydrogen peroxide resistance in *Escherichia coli* serotype O157 : H7**. *Microbiology (Reading, England)* 2009, **155**(Pt 11):3589–3598, [http://www.ncbi.nlm.nih.gov/pubmed/19713239]. [PMID: 19713239].

[324] Yoon JW, Lim JY, Park YH, Hovde CJ: **Involvement of the *Escherichia coli* O157:H7(pO157) ecf operon and lipid a myristoyl transferase activity in bacterial survival in the bovine gastrointestinal tract and bacterial persistence in farm water troughs**. *Infection and Immunity* 2005, **73**(4):2367–2378, [http://www.ncbi.nlm.nih.gov/pubmed/15784583]. [PMID: 15784583].

[325] Grys TE, Walters LL, Welch RA: **Characterization of the StcE protease activity of *Escherichia coli* O157:H7**. *Journal of bacteriology* 2006, **188**(13):4646–4653, [http://www.ncbi.nlm.nih.gov/pubmed/16788173]. [PMID: 16788173].

[326] Grys TE, Siegel MB, Lathem WW, Welch RA: **The StcE protease contributes to intimate adherence of enterohemorrhagic *Escherichia coli* O157:H7 to host cells**. *Infection and immunity* 2005, **73**(3):1295–1303, [http://www.ncbi.nlm.nih.gov/pubmed/15731026]. [PMID: 15731026].

[327] Tatsuno I, Horie M, Abe H, Miki T, Makino K, Shinagawa H, Taguchi H, Kamiya S, Hayashi T, Sasakawa C: **toxB gene on pO157 of enterohemorrhagic *Escherichia coli* O157:H7 is required for full epithelial cell adherence phenotype**. *Infect Immun* 2001, **69**(11):6660–9, [11598035].

[328] White DG, Zhao S, Simjee S, Wagner DD, McDermott PF: **Antimicrobial resistance of foodborne pathogens**. *Microbes and Infection / Institut Pasteur* 2002, **4**(4):405–412, [http://www.ncbi.nlm.nih.gov/pubmed/11932191]. [PMID: 11932191].

[329] White TA, Kell DB: **Comparative genomic assessment of novel Broad-spectrum targets for antibacterial Drugs**. *Comparative and Functional Genomics* 2004, **5**(4):304–327, [http://www.hindawi.com/GetArticle.aspx?doi=10.1002/cfg.411].

[330] Bradford PA, Petersen PJ, Fingerman IM, White DG: **Characterization of expanded-spectrum cephalosporin resistance in E. coli isolates associated with bovine calf diarrhoeal disease**. *The Journal of Antimicrobial Chemotherapy* 1999, **44**(5):607–610, [http://www.ncbi.nlm.nih.gov/pubmed/10552976]. [PMID: 10552976].

[331] Mirzaagha P, Louie M, Sharma R, Yanke LJ, Topp E, McAllister TA: **Distribution and characterization of ampicillin- and tetracycline-resistant *Escherichia coli* from feedlot cattle fed subtherapeutic antimicrobials**. *BMC microbiology* 2011, **11**:78, [http://www.ncbi.nlm.nih.gov/pubmed/21504594]. [PMID: 21504594].

[332] Seto K, Taguchi M, Kobayashi K, Kozaki S: **Biochemical and molecular characterization of minor serogroups of shiga toxin-producing *Escherichia coli* isolated from humans in osaka prefecture**. *The Journal of Veterinary Medical Science / the Japanese Society of Veterinary Science* 2007, **69**(12):1215–1222, [http://www.ncbi.nlm.nih.gov/pubmed/18176015]. [PMID: 18176015].

[333] Gow SP, Waldner CL: **Antimicrobial resistance and virulence factors stx1, stx2, and eae in generic *Escherichia coli* isolates from calves in western canadian cow-calf herds**. *Microbial Drug Resistance (Larchmont, N.Y.)* 2009, **15**:61–67, [http://www.ncbi.nlm.nih.gov/pubmed/19216645]. [PMID: 19216645].

[334] Garcia-Aljaro C, Moreno E, Andreu A, Prats G, Blanch AR: **Phylogroups, virulence determinants and antimicrobial resistance in stx2 gene-carrying *Escherichia coli* isolated from aquatic environments**. *Research in Microbiology* 2009, **160**(8):585–591, [http://www.sciencedirect.com/science/article/B6VN3-4X49YG4-2/2/f789a64fc805bfa374ec8c37ba8ebdb0].

[335] Erickson MC, Doyle MP: **Food as a vehicle for transmission of shiga toxin-producing *Escherichia coli***. *Journal of Food Protection* 2007, **70**(10):2426–49, [http://www.ncbi.nlm.nih.gov/pubmed/17969631]. [PMID: 17969631].

[336] Bettelheim KA: **Enterohaemorrhagic *Escherichia coli* O157:H7: a red herring?** *J Med Microbiol* 2001, **50**(2):201–2, [11211230].

[337] Sansonetti PJ: **To be or not to be a pathogen: that is the mucosally relevant question**. *Mucosal Immunology* 2011, **4**:8–14, [http://www.ncbi.nlm.nih.gov/pubmed/21150896]. [PMID: 21150896].

[338] Puttamreddy S, Minion FC: **Linkage between cellular adherence and biofilm formation in *Escherichia coli* O157:H7 EDL933**. *FEMS microbiology letters* 2011, **315**:46–53. [PMID: 21166710].

[339] Ingle DJ, Clermont O, Skurnik D, Denamur E, Walk ST, Gordon DM: **Biofilm formation by and thermal niche and virulence characteristics of escherichia spp**. *Applied and environmental microbiology* 2011, **77**(8):2695–2700. [PMID: 21335385].

[340] Brzuszkiewicz E, Thrmer A, Schuldes J, Leimbach A, Liesegang H, Meyer FD, Boelter J, Petersen H, Gottschalk G, Daniel R: **Genome sequence analyses of two isolates from the recent *Escherichia coli* outbreak in germany reveal the emergence of a new pathotype: Entero-Aggregative-haemorrhagic *Escherichia coli* (EAHEC)**. *Archives of Microbiology* 2011, [http://www.ncbi.nlm.nih.gov/pubmed/21713444]. [PMID: 21713444].

[341] Al Safadi R, Abu-Ali GS, Sloup RE, Rudrik JT, Waters CM, Eaton KA, Manning SD: **Correlation between in vivo biofilm formation and virulence gene expression in *Escherichia coli* O104:H4**. *PloS one* 2012, **7**(7):e41628. [PMID: 22848550].

[342] Ehrlich GD, Ahmed A, Earl J, Hiller NL, Costerton JW, Stoodley P, Post JC, DeMeo P, Hu FZ: **The distributed genome hypothesis as a rubric for understanding evolution in situ during chronic bacterial biofilm infectious processes**. *FEMS Immunology and Medical Microbiology* 2010, **59**(3):269–279, [http://www.ncbi.nlm.nih.gov/pubmed/20618850]. [PMID: 20618850].

[343] Berry ED, Wells JE: ***Escherichia coli* O157:H7 recent advances in research on Occurrence, Transmission, and control in cattle and the production Environment**. *Advances in Food and Nutrition Research* 2010, **60**:67–117, [http://www.ncbi.nlm.nih.gov/pubmed/20691954]. [PMID: 20691954].

[344] Niu YD, McAllister TA, Xu Y, Johnson RP, Stephens TP, Stanford K: **Prevalence and impact of bacteriophages on the presence of *Escherichia coli* O157:H7 in feedlot cattle and their environment**. *Applied and environmental microbiology* 2009, **75**(5):1271–1278, [http://www.ncbi.nlm.nih.gov/pubmed/19139243]. [PMID: 19139243].

[345] Labrie SJ, Samson JE, Moineau S: **Bacteriophage resistance mechanisms**. *Nature Reviews Microbiology* 2010, **8**(5):317–327, [http://www.nature.com/nrmicro/journal/v8/n5/full/nrmicro2315.html].

[346] Kashiwagi A, Yomo T: **Ongoing phenotypic and genomic changes in experimental coevolution of RNA bacteriophage Q and *Escherichia coli***. *PLoS Genetics* 2011, **7**(8), [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3150450/]. [PMID: 21829387 PMCID: PMC3150450].

[347] Dennehy JJ: **What can phages tell us about Host-pathogen Coevolution?** *International Journal of Evolutionary Biology* 2012, **2012**, [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3506893/]. [PMID: 23213618 PMCID: PMC3506893].

[348] Bondy-Denomy J, Pawluk A, Maxwell KL, Davidson AR: **Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system**. *Nature* 2012. [PMID: 23242138].

[349] Short FL, Blower TR, Salmond GPC: **A promiscuous antitoxin of bacteriophage T4 ensures successful viral replication**. *Molecular microbiology* 2012, **83**(4):665–668. [PMID: 22283468].

[350] Siegler RL: **The hemolytic uremic syndrome**. *Pediatric Clinics of North America* 1995, **42**(6):1505–29. [PMID: 8614598].

[351] Yoon JW, Hovde CJ: **All blood, no stool: enterohemorrhagic *Escherichia coli* O157:H7 infection**. *Journal of Veterinary Science (Suwn-Si, Korea)* 2008, **9**(3):219–31, [http://www.ncbi.nlm.nih.gov/pubmed/18716441]. [PMID: 18716441].

[352] Ohnishi M, Kurokawa K, Hayashi T: **Diversification of *Escherichia coli* genomes: are bacteriophages the major contributors?** *Trends in Microbiology* 2001, **9**(10):481–5, [http://www.ncbi.nlm.nih.gov/pubmed/11597449]. [PMID: 11597449].

[353] Kim J, Nietfeldt J, Benson AK: **Octamer-based genome scanning distinguishes a unique subpopulation of *Escherichia coli* O157:H7 strains in cattle**. *Proceedings of the National Academy of Sciences of the United States of America* 1999, **96**(23):13288–93, [http://www.ncbi.nlm.nih.gov/pubmed/10557313]. [PMID: 10557313].

[354] Zhang Y, Laing C, Steele M, Ziebell K, Johnson R, Benson A, Taboada E, Gannon V: **Genome evolution in major *Escherichia coli* O157:H7 lineages**. *BMC Genomics* 2007, **8**:121, [http://www.biomedcentral.com/1471-2164/8/121].

[355] Yang Z, Kovar J, Kim J, Nietfeldt J, Smith DR, Moxley RA, Olson ME, Fey PD, Benson AK: **Identification of common subpopulations of non-sorbitol-fermenting, beta-glucuronidase-negative *Escherichia coli* O157:H7 from bovine production environments and human clinical samples**. *Appl Environ Microbiol* 2004, **70**(11):6846–54, [15528552].

[356] Besser TE, Shaikh N, Holt NJ, Tarr PI, Konkel ME, Malik-Kale P, Walsh CW, Whittam TS, Bono JL: **Greater diversity of shiga Toxin-Encoding bacteriophage insertion sites among *Escherichia coli* O157:H7 isolates from cattle than in those from Humans**. *Appl. Environ. Microbiol.* 2007, **73**(3):671–679, [http://aem.asm.org/cgi/content/abstract/73/3/671].

[357] Feng PCH, Monday SR, Lacher DW, Allison L, Siitonen A, Keys C, Eklund M, Nagano H, Karch H, Keen J, Whittam TS: **Genetic diversity among clonal lineages within *Escherichia coli* O157:H7 stepwise evolutionary model**. *Emerging Infectious Diseases* 2007, **13**(11):1701–6. [PMID: 18217554].

[358] Maddison DR, Swofford DL, Maddison WP: **NEXUS: an extensible file format for systematic information**. *Systematic Biology* 1997, **46**(4):590–621, [http://www.ncbi.nlm.nih.gov/pubmed/11975335]. [PMID: 11975335].

[359] Wilgenbusch JC, Swofford D: **Inferring evolutionary trees with PAUP***. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al* 2003, **Chapter 6**:Unit 6.4, [http://www.ncbi.nlm.nih.gov/pubmed/18428704]. [PMID: 18428704].

[360] Riordan JT, Viswanath SB, Manning SD, Whittam TS: **Genetic differentiation of *Escherichia coli* O157:H7 clades associated with human disease by Real-time PCR**. *J. Clin. Microbiol.* 2008, **46**(6):2070–2073, [http://jcm.asm.org/cgi/content/abstract/46/6/2070].

[361] Huson DH, Bryant D: **Application of phylogenetic networks in evolutionary studies**. *Molecular Biology and Evolution* 2006, **23**(2):254–67, [http://www.ncbi.nlm.nih.gov/pubmed/16221896]. [PMID: 16221896].

[362] Abe H, Miyahara A, Oshima T, Tashiro K, Ogura Y, Kuhara S, Ogasawara N, Hayashi T, Tobe T: **Global regulation by horizontally transferred regulators establishes the pathogenicity of *Escherichia coli***. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes* 2008, **15**:25–38, [http://www.ncbi.nlm.nih.gov/pubmed/18222925]. [PMID: 18222925].

[363] Hacker J, Kaper JB: **Pathogenicity islands and the evolution of microbes**. *Annual Review of Microbiology* 2000, **54**:641–79. [PMID: 11018140].

[364] O'Brien A, Newland J, Miller S, Holmes R, Smith H, Formal S: **Shiga-like toxin-converting phages from *Escherichia coli* strains that cause hemorrhagic colitis or infantile diarrhea**. *Science* 1984, **226**(4675):694–696, [http://www.sciencemag.org/cgi/content/abstract/226/4675/694].

[365] Plunkett G, Rose DJ, Durfee TJ, Blattner FR: **Sequence of shiga toxin 2 phage 933W from *Escherichia coli* O157:H7: shiga toxin as a phage late-gene product**. *Journal of Bacteriology* 1999, **181**(6):1767–78, [http://www.ncbi.nlm.nih.gov/pubmed/10074068]. [PMID: 10074068].

[366] Neely MN, Friedman DI: **Functional and genetic analysis of regulatory regions of coliphage H-19B: location of shiga-like toxin and lysis genes suggest a role for phage functions in toxin release**. *Molecular Microbiology* 1998, **28**(6):1255–67, [http://www.ncbi.nlm.nih.gov/pubmed/9680214]. [PMID: 9680214].

[367] McDaniel TK, Jarvis KG, Donnenberg MS, Kaper JB: **A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens**. *Proceedings of the National Academy of Sciences of the United States of America* 1995, **92**(5):1664–8, [http://www.ncbi.nlm.nih.gov/pubmed/7878036]. [PMID: 7878036].

[368] Kudva IT, Evans PS, Perna NT, Barrett TJ, Ausubel FM, Blattner FR, Calderwood SB: **Strains of *Escherichia coli* O157:H7 differ primarily by insertions or deletions, not single-nucleotide polymorphisms**. *J Bacteriol* 2002, **184**(7):1873–9, [11889093].

[369] Kelly BG, Vespermann A, Bolton DJ: **The role of horizontal gene transfer in the evolution of selected foodborne bacterial pathogens**. *Food and Chemical Toxicology:*

An International Journal Published for the British Industrial Biological Research Association 2009, **47**(5):951–968, [http://www.ncbi.nlm.nih.gov/pubmed/18420329]. [PMID: 18420329].

[370] Ziebell K, Steele M, Zhang Y, Benson A, Taboada EN, Laing C, McEwen S, Ciebin B, Johnson R, Gannon V: **Genotypic characterization and prevalence of virulence factors among canadian *Escherichia coli* O157:H7 strains**. *Applied and Environmental Microbiology* 2008, **74**(14):4314–23, [http://www.ncbi.nlm.nih.gov/pubmed/18487402]. [PMID: 18487402].

[371] Steele M, Ziebell K, Zhang Y, Benson A, Konczy P, Johnson R, Gannon V: **Identification of *Escherichia coli* O157:H7 genomic regions conserved in strains with a genotype associated with human infection**. *Applied and Environmental Microbiology* 2007, **73**:22–31. [PMID: 17056689].

[372] Dowd S, Ishizaki H: **Microarray based comparison of two *Escherichia coli* O157:H7 lineages**. *BMC Microbiology* 2006, **6**:30, [http://www.biomedcentral.com/1471-2180/6/30].

[373] Friedrich AW, Bielaszewska M, Zhang W, Pulz M, Kuczius T, Ammon A, Karch H: ***Escherichia coli* harboring shiga toxin 2 gene variants: frequency and association with clinical symptoms**. *The Journal of Infectious Diseases* 2002, **185**:74–84, [http://www.ncbi.nlm.nih.gov/pubmed/11756984]. [PMID: 11756984].

[374] Dagerhamn J, Blomberg C, Browall S, Sjstrm K, Morfeldt E, Henriques-Normark B: **Determination of accessory gene patterns predicts the same relatedness among strains of *Streptococcus pneumoniae* as sequencing of housekeeping genes does and represents a novel approach in molecular epidemiology**. *Journal of Clinical Microbiology* 2008, **46**(3):863–8, [http://www.ncbi.nlm.nih.gov/pubmed/18160453]. [PMID: 18160453].

[375] Laing C, Villegas A, Taboada EN, Kropinski A, Thomas JE, Gannon VPJ: **Identification of *Salmonella enterica* species- and subgroup-specific genomic regions using panseq 2.0**. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 2011, [http://www.ncbi.nlm.nih.gov/pubmed/22001825]. [PMID: 22001825].

[376] Ansorge WJ: **Next-generation DNA sequencing techniques**. *New Biotechnology* 2009, **25**(4):195–203, [http://www.sciencedirect.com/science/article/B8JG4-4VHSDPY-1/2/7026ea66c50167c1a02237749a2e9228].

[377] Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores**. *Genome Research* 2008, **18**(11):1851–1858, [http://www.ncbi.nlm.nih.gov/pubmed/18714091]. [PMID: 18714091].

[378] MacLean D, Jones JDG, Studholme DJ: **Application of 'next-generation' sequencing technologies to microbial genetics**. *Nature Reviews. Microbiology* 2009, **7**(4):287–296, [http://www.ncbi.nlm.nih.gov/pubmed/19287448]. [PMID: 19287448].

[379] Stiens M, Becker A, Bekel T, Gdde V, Goesmann A, Niehaus K, Schneiker-Bekel S, Selbitschka W, Weidner S, Schlter A, Phler A: **Comparative genomic hybridisation and ultrafast pyrosequencing revealed remarkable differences between the sino*Rhizobium meliloti* genomes of the model strain Rm1021 and the field isolate SM11**. *Journal of Biotechnology* 2008, **136**(1-2):31–37, [http://www.ncbi.nlm.nih.gov/pubmed/18562031]. [PMID: 18562031].

[380] Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison**. *Proceedings of the National Academy of Sciences of the United States of America* 1988, **85**(8):2444–2448. [PMID: 3162770].

[381] Abbott JC, Aanensen DM, Bentley SD: **WebACT: an online genome comparison suite**. *Methods in molecular biology (Clifton, N.J.)* 2007, **395**:57–74. [PMID: 17993667].

[382] Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko A, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, John Wilbur W, Yaschenko E, Ye J: **Database resources of the national center for biotechnology Information**. *Nucleic acids research* 2010, **38**(Database issue):D5–16. [PMID: 19910364].

[383] Chiapello H, Gendrault A, Caron C, Blum J, Petit MA, El Karoui M: **MOSAIC: an online database dedicated to the comparative genomics of bacterial strains at the intra-species level**. *BMC bioinformatics* 2008, **9**:498. [PMID: 19038022].

[384] Fong C, Rohmer L, Radey M, Wasnick M, Brittnacher MJ: **PSAT: a web tool to compare genomic neighborhoods of multiple prokaryotic genomes**. *BMC bioinformatics* 2008, **9**:170. [PMID: 18366802].

[385] Ou HY, He X, Harrison EM, Kulasekara BR, Thani AB, Kadioglu A, Lory S, Hinton JCD, Barer MR, Deng Z, Rajakumar K: **MobilomeFINDER: web-based tools for in silico and experimental discovery of bacterial genomic islands**. *Nucleic acids research* 2007, **35**(Web Server issue):W97–W104. [PMID: 17537813].

[386] Langille MGI, Brinkman FSL: **IslandViewer: an integrated interface for computational identification and visualization of genomic islands**. *Bioinformatics (Oxford, England)* 2009, **25**(5):664–665. [PMID: 19151094].

[387] Steele M, Ziebell K, Zhang Y, Benson A, Johnson R, Laing C, Taboada E, Gannon V: **Genomic regions conserved in lineage II *Escherichia coli* O157:H7 strains**. *Applied and Environmental Microbiology* 2009, **75**(10):3271–3280, [http://www.ncbi.nlm.nih.gov/pubmed/19329668]. [PMID: 19329668].

[388] Taboada EN, Luebbert CC, Nash JHE: **Studying bacterial genome dynamics using microarray-based comparative genomic hybridization**. *Methods in Molecular Biology (Clifton, N.J.)* 2007, **396**:223–53, [http://www.ncbi.nlm.nih.gov/pubmed/18025696]. [PMID: 18025696].

[389] Kulasekara BR, Jacobs M, Zhou Y, Wu Z, Sims E, Saenphimmachak C, Rohmer L, Ritchie JM, Radey M, McKevitt M, Freeman TL, Hayden H, Haugen E, Gillett W, Fong C, Chang J, Beskhlebnaya V, Waldor MK, Samadpour M, Whittam TS, Kaul R, Brittnacher M, Miller SI: **Analysis of the genome of the *Escherichia coli* O157:H7 2006 spinach-associated outbreak isolate indicates candidate genes that may enhance virulence**. *Infection and Immunity* 2009, [http://www.ncbi.nlm.nih.gov/pubmed/19564389]. [PMID: 19564389].

[390] Willenbrock H, Hallin P, Wassenaar T, Ussery D: **Characterization of probiotic *Escherichia coli* isolates with a novel pan-genome microarray**. *Genome Biology* 2007, **8**(12):R267, [http://genomebiology.com/2007/8/12/R267].

[391] Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehvslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E: **The bioperl toolkit: perl modules for the life sciences**. *Genome research* 2002, **12**(10):1611–1618. [PMID: 12368254].

[392] Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O: **New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0**. *Systematic Biology* 2010, **59**(3):307–321, [http://www.ncbi.nlm.nih.gov/pubmed/20525638]. [PMID: 20525638].

[393] Goloboff PA, Farris JS, Nixon KC: **tnt, a free program for phylogenetic analysis**. *Cladistics* 2008, **24**(5):774–786, [http://onlinelibrary.wiley.com/doi/10.1111/j.1096-0031.2008.00217.x/abstract].

[394] Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R: **Dendroscope: an interactive viewer for large phylogenetic trees**. *BMC Bioinformatics* 2007, **8**:460, [http://www.ncbi.nlm.nih.gov/pubmed/18034891]. [PMID: 18034891].

[395] Consortium U: **The universal protein resource (UniProt) in 2010**. *Nucleic acids research* 2010, **38**(Database issue):D142–148. [PMID: 19843607].

[396] Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, Churcher C, Mungall KL, Bentley SD, Holden MT, Sebaihia M, Baker S, Basham D, Brooks K, Chillingworth T, Connerton P, Cronin A, Davis P, Davies RM, Dowd L, White N, Farrar J, Feltwell T, Hamlin N, Haque A, Hien TT, Holroyd S, Jagels K, Krogh A, Larsen TS, Leather S, Moule S, O'Gaora P, Parry C, Quail M, Rutherford K, Simmonds M, Skelton J, Stevens K, Whitehead S, Barrell BG: **Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar typhi CT18**. *Nature* 2001, **413**(6858):848–52, [11677608].

[397] Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MTG, Prentice MB, Sebaihia M, James KD, Churcher C, Mungall KL, Baker S, Basham D, Bentley SD, Brooks K, Cerdeno-Tarraga AM, Chillingworth T, Cronin A, Davies RM, Davis P, Dougan G, Feltwell T, Hamlin N, Holroyd S, Jagels K, Karlyshev AV, Leather S, Moule S, Oyston PCF, Quail M, Rutherford K, Simmonds M, Skelton J, Stevens K, Whitehead S, Barrell BG: **Genome sequence of *Yersinia pestis*, the causative agent of plague**. *Nature* 2001, **413**(6855):523–527, [http://dx.doi.org/10.1038/35097083].

[398] Dingle KE, Colles FM, Wareing DR, Ure R, Fox AJ, Bolton FE, Bootsma HJ, Willems RJ, Urwin R, Maiden MC: **Multilocus sequence typing system for *Campylobacter jejuni***. *Journal of clinical microbiology* 2001, **39**:14–23. [PMID: 11136741].

[399] Best EL, Fox AJ, Frost JA, Bolton FJ: **Real-time single-nucleotide polymorphism profiling using taqman technology for rapid recognition of *Campylobacter jejuni* clonal complexes**. *Journal of medical microbiology* 2005, **54**(Pt 10):919–925. [PMID: 16157544].

[400] Ward TJ, Ducey TF, Usgaard T, Dunn KA, Bielawski JP: **Multilocus genotyping assays for single nucleotide polymorphism-based subtyping of *Listeria monocytogenes* isolates**. *Applied and environmental microbiology* 2008, **74**(24):7629–7642. [PMID: 18931295].

[401] Bronowski C, Winstanley C: **Identification and distribution of accessory genome DNA sequences from an invasive african isolate of salmonella Heidelberg**. *FEMS Microbiology Letters* 2009, **298**:29–36, [http://www.ncbi.nlm.nih.gov/pubmed/19594621]. [PMID: 19594621].

[402] Hallin M, De Mendona R, Denis O, Lefort A, El Garch F, Butaye P, Hermans K, Struelens MJ: **Diversity of accessory genome of human and livestock-associated ST398 methicillin resistant *Staphylococcus aureus* strains**. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 2011, **11**(2):290–299, [http://www.ncbi.nlm.nih.gov/pubmed/21145988]. [PMID: 21145988].

[403] Vos M: **A species concept for bacteria based on adaptive divergence**. *Trends in Microbiology* 2011, **19**:1–7, [http://www.sciencedirect.com/science/article/B6TD0-51F6HJY-1/2/afebadd43b84c8cd40242e8730475924].

[404] McClelland M, Sanderson KE, Clifton SW, Latreille P, Porwollik S, Sabo A, Meyer R, Bieri T, Ozersky P, McLellan M, Harkins CR, Wang C, Nguyen C, Berghoff A, Elliott G, Kohlberg S, Strong C, Du F, Carter J, Kremizki C, Layman D, Leonard S, Sun H, Fulton L, Nash W, Miner T, Minx P, Delehaunty K, Fronick C, Magrini V, Nhan M, Warren W, Florea L, Spieth J, Wilson RK: **Comparison of genome degradation in paratyphi a and Typhi, human-restricted serovars of *Salmonella enterica* that cause typhoid**. *Nat Genet* 2004, **36**(12):1268–1274, [http://dx.doi.org/10.1038/ng1470].

[405] Chiu CH, Tang P, Chu C, Hu S, Bao Q, Yu J, Chou YY, Wang HS, Lee YS: **The genome sequence of *Salmonella enterica* serovar Choleraesuis, a highly invasive and resistant zoonotic pathogen**. *Nucleic Acids Research* 2005, **33**(5):1690–1698, [http://www.ncbi.nlm.nih.gov/pubmed/15781495]. [PMID: 15781495].

[406] Kim HJ, Park SH, Kim HY: **Comparison of *Salmonella enterica* serovar typhimurium LT2 and non-LT2 salmonella genomic sequences, and genotyping of salmonellae by using PCR**. *Applied and Environmental Microbiology* 2006, **72**(9):6142–6151, [http://www.ncbi.nlm.nih.gov/pubmed/16957240]. [PMID: 16957240].

[407] Thomson NR, Clayton DJ, Windhorst D, Vernikos G, Davidson S, Churcher C, Quail MA, Stevens M, Jones MA, Watson M, Barron A, Layton A, Pickard D, Kingsley RA, Bignell A, Clark L, Harris B, Ormond D, Abdellah Z, Brooks K, Cherevach I, Chillingworth T, Woodward J, Norberczak H, Lord A, Arrowsmith C, Jagels K, Moule S, Mungall K, Sanders

M, Whitehead S, Chabalgoity JA, Maskell D, Humphrey T, Roberts M, Barrow PA, Dougan G, Parkhill J: **Comparative genome analysis of *Salmonella enteritidis* PT4 and salmonella gallinarum 287/91 provides insights into evolutionary and host adaptation pathways**. *Genome Research* 2008, **18**(10):1624–1637, [http://www.ncbi.nlm.nih.gov/pubmed/18583645]. [PMID: 18583645].

[408] Lan R, Reeves PR, Octavia S: **Population structure, origins and evolution of major *Salmonella enterica* clones**. *Infection, Genetics and Evolution* 2009, **9**(5):996–1005, [http://www.sciencedirect.com/science/article/B6W8B-4W4JDSG-1/2/ab62ddd15efc85b9eca625b1cf75b7dc].

[409] Falush D, Torpdahl M, Didelot X, Conrad DF, Wilson DJ, Achtman M: **Mismatch induced speciation in Salmonella: model and data**. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2006, **361**(1475):2045 –2053, [http://rstb.royalsocietypublishing.org/content/361/1475/2045.abstract].

[410] Didelot X, Bowden R, Street T, Golubchik T, Spencer C, McVean G, Sangal V, Anjum MF, Achtman M, Falush D, Donnelly P: **Recombination and population structure in *Salmonella enterica***. *PLoS Genet* 2011, **7**(7):e1002191, [http://dx.doi.org/10.1371/journal.pgen.1002191].

[411] Bonaventura MPD, Lee EK, DeSalle R, Planet PJ: **A whole-genome phylogeny of the family Pasteurellaceae**. *Molecular Phylogenetics and Evolution* 2010, **54**(3):950–956, [http://www.sciencedirect.com/science/article/B6WNH-4X0XF8P-3/2/3d75ea317789a67532147864ebd8a0c1].

[412] Philippe H, Douady CJ: **Horizontal gene transfer and phylogenetics**. *Current Opinion in Microbiology* 2003, **6**(5):498–505, [http://www.ncbi.nlm.nih.gov/pubmed/14572543]. [PMID: 14572543].

[413] Laing CR, Zhang Y, Gilmour MW, Allen V, Johnson R, Thomas JE, Gannon VPJ: **A comparison of Shiga-toxin 2 bacteriophage from classical enterohemorrhagic *Escherichia coli* serotypes and the german E. coli O104:H4 outbreak strain**. *PloS one* 2012, **7**(5):e37362. [PMID: 22649523].

[414] Blaser MJ: **Deconstructing a lethal foodborne epidemic**. *The New England Journal of Medicine* 2011, **365**(19):1835–1836, [http://www.ncbi.nlm.nih.gov/pubmed/22029755]. [PMID: 22029755].

[415] Buchholz U, Bernard H, Werber D, Bhmer MM, Remschmidt C, Wilking H, Deler Y, an der Heiden M, Adlhoch C, Dreesman J, Ehlers J, Ethelberg S, Faber M, Frank C, Fricke G, Greiner M, Hhle M, Ivarsson S, Jark U, Kirchner M, Koch J, Krause G, Luber P, Rosner B, Stark K, Khne M: **German outbreak of *Escherichia coli* O104:H4 associated with sprouts**. *The New England Journal of Medicine* 2011, **365**(19):1763–1770, [http://www.ncbi.nlm.nih.gov/pubmed/22029753]. [PMID: 22029753].

[416] Bielaszewska M, Mellmann A, Zhang W, Kck R, Fruth A, Bauwens A, Peters G, Karch H: **Characterisation of the *Escherichia coli* strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: a microbiological study**. *The Lancet*

Infectious Diseases 2011, [http://www.ncbi.nlm.nih.gov/pubmed/21703928]. [PMID: 21703928].

[417] Mossoro C, Glaziou P, Yassibanda S, Lan NTP, Bekondi C, Minssart P, Bernier C, Le Bougunec C, Germani Y: **Chronic diarrhea, hemorrhagic colitis, and hemolytic-uremic syndrome associated with HEp-2 adherent *Escherichia coli* in adults infected with human immunodeficiency virus in Bangui, central african Republic**. *Journal of Clinical Microbiology* 2002, **40**(8):3086–3088, [http://www.ncbi.nlm.nih.gov/pubmed/12149388]. [PMID: 12149388].

[418] Denamur E: **The 2011 shiga toxin-producing *Escherichia coli* O104:H4 german outbreak: a lesson in genomic plasticity**. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 2011, **17**(8):1124–1125, [http://www.ncbi.nlm.nih.gov/pubmed/21781204]. [PMID: 21781204].

[419] Scheutz F, Moller Nielsen E, Frimodt-Moller J, Boisen N, Morabito S, Tozzoli R, Nataro J, Caprioli A: **Characteristics of the enteroaggregative shiga toxin/verotoxin-producing *Escherichia coli* O104:H4 strain causing the outbreak of haemolytic uraemic syndrome in Germany, may to june 2011**. *Euro Surveillance: Bulletin Europen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* 2011, **16**(24), [http://www.ncbi.nlm.nih.gov/pubmed/21699770]. [PMID: 21699770].

[420] Askar M, Faber MS, Frank C, Bernard H, Gilsdorf A, Fruth A, Prager R, Hohle M, Suess T, Wadl M, Krause G, Stark K, Werber D: **Update on the ongoing outbreak of haemolytic uraemic syndrome due to shiga toxin-producing *Escherichia coli* (STEC) serotype O104, Germany, may 2011**. *Euro Surveillance: Bulletin Europen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* 2011, **16**(22), [http://www.ncbi.nlm.nih.gov/pubmed/21663710]. [PMID: 21663710].

[421] Frank C, Werber D, Cramer JP, Askar M, Faber M, an der Heiden M, Bernard H, Fruth A, Prager R, Spode A, Wadl M, Zoufaly A, Jordan S, Kemper MJ, Follin P, Mller L, King LA, Rosner B, Buchholz U, Stark K, Krause G: **Epidemic profile of Shiga-toxin-producing *Escherichia coli* O104:H4 outbreak in Germany**. *The New England Journal of Medicine* 2011, **365**(19):1771–1780, [http://www.ncbi.nlm.nih.gov/pubmed/21696328]. [PMID: 21696328].

[422] Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W, McLaughlin SF, Henkhaus JK, Leopold B, Bielaszewska M, Prager R, Brzoska PM, Moore RL, Guenther S, Rothberg JM, Karch H: **Prospective genomic characterization of the german enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing Technology**. *PLoS ONE* 2011, **6**(7):e22751, [http://dx.doi.org/10.1371/journal.pone.0022751].

[423] Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS: **PHaST: A fast phage search Tool**. *Nucleic Acids Research* 2011, **39**(Web Server issue):W347–352, [http://www.ncbi.nlm.nih.gov/pubmed/21672955]. [PMID: 21672955].

[424] Katoh K, Kuma Ki, Toh H, Miyata T: **MAFFT version 5: improvement in accuracy of multiple sequence alignment**. *Nucleic Acids Research* 2005, **33**(2):511–518, [http://www.ncbi.nlm.nih.gov/pubmed/15661851]. [PMID: 15661851].

[425] Alexander DC, Hao W, Gilmour MW, Zittermann S, Sarabia A, Melano RG, Peralta A, Lombos M, Warren K, Amatnieks Y, Virey E, Ma JH, Jamieson FB, Low DE, Allen VG: ***Escherichia coli* O104:H4 infections and international travel**. *Emerging Infectious Diseases* 2012, **18**(3):473–476, [http://www.ncbi.nlm.nih.gov/pubmed/22377016]. [PMID: 22377016].

[426] Vareille M, de Sablet T, Hindr T, Martin C, Gobert AP: **Nitric oxide inhibits Shiga-toxin synthesis by enterohemorrhagic *Escherichia coli***. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(24):10199–10204, [http://www.ncbi.nlm.nih.gov/pubmed/17537918]. [PMID: 17537918].

[427] Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, Fitzgerald M, Godfrey P, Haas BJ, Murphy CI, Russ C, Sykes S, Walker BJ, Wortman JR, Young S, Zeng Q, Abouelleil A, Bochicchio J, Chauvin S, Desmet T, Gujja S, McCowan C, Montmayeur A, Steelman S, Frimodt-Mller J, Petersen AM, Struve C, Krogfelt KA, Bingen E, Weill FX, Lander ES, Nusbaum C, Birren BW, Hung DT, Hanage WP: **Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011**. *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109**(8):3065–3070, [http://www.ncbi.nlm.nih.gov/pubmed/22315421]. [PMID: 22315421].

[428] Wiens JJ, Morrill MC: **Missing data in phylogenetic Analysis: reconciling results from simulations and empirical Data**. *Systematic Biology* 2011, [http://sysbio.oxfordjournals.org/content/early/2011/03/27/sysbio.syr025.short].

[429] Boerlin P: **Evolution of virulence factors in Shiga-toxin-producing *Escherichia coli***. *Cell Mol Life Sci* 1999, **56**(9-10):735–41, [11212333].

[430] Donnenberg MS, Whittam TS: **Pathogenesis and evolution of virulence in enteropathogenic and enterohemorrhagic *Escherichia coli***. *Journal of Clinical Investigation* 2001, **107**(5):539–548, [http://www.jci.org/articles/view/12404/version/1].

[431] Serra-Moreno R, Jofre J, Muniesa M: **Insertion site occupancy by stx2 bacteriophages depends on the locus availability of the host strain chromosome**. *Journal of Bacteriology* 2007, **189**(18):6645–6654, [http://www.ncbi.nlm.nih.gov/pubmed/17644594]. [PMID: 17644594].

[432] Wieler LH, McDaniel TK, Whittam TS, Kaper JB: **Insertion site of the locus of enterocyte effacement in enteropathogenic and enterohemorrhagic *Escherichia coli* differs in relation to the clonal phylogeny of the strains**. *FEMS Microbiology Letters* 1997, **156**:49–53, [http://www.ncbi.nlm.nih.gov/pubmed/9368360]. [PMID: 9368360].

[433] Shimizu T, Ohta Y, Noda M: **Shiga toxin 2 is specifically released from bacterial cells by two different mechanisms**. *Infection and Immunity* 2009, **77**(7):2813–2823, [http://www.ncbi.nlm.nih.gov/pubmed/19380474]. [PMID: 19380474].

[434] Bielaszewska M, Idelevich EA, Zhang W, Bauwens A, Schaumburg F, Mellmann A, Peters G, Karch H: **Epidemic *Escherichia coli* O104:H4: effects of antibiotics on shiga toxin 2 production and bacteriophage induction**. *Antimicrobial Agents and Chemotherapy* 2012, [http://www.ncbi.nlm.nih.gov/pubmed/22391549]. [PMID: 22391549].

[435] de Sablet T, Chassard C, Bernalier-Donadille A, Vareille M, Gobert AP, Martin C: **Human microbiota-secreted factors inhibit shiga toxin synthesis by enterohemorrhagic *Escherichia coli* O157:H7**. *Infection and Immunity* 2009, **77**(2):783–790, [http://www.ncbi.nlm.nih.gov/pubmed/19064636]. [PMID: 19064636].

[436] Rohde H, Qin J, Cui Y, Li D, Loman NJ, Hentschke M, Chen W, Pu F, Peng Y, Li J, Xi F, Li S, Li Y, Zhang Z, Yang X, Zhao M, Wang P, Guan Y, Cen Z, Zhao X, Christner M, Kobbe R, Loos S, Oh J, Yang L, Danchin A, Gao GF, Song Y, Li Y, Yang H, Wang J, Xu J, Pallen MJ, Wang J, Aepfelbacher M, Yang R: **Open-source genomic analysis of Shiga-Toxin-producing E. coli O104:H4**. *The New England Journal of Medicine* 2011, [http://www.ncbi.nlm.nih.gov/pubmed/21793736]. [PMID: 21793736].

[437] Sato T, Shimizu T, Watarai M, Kobayashi M, Kano S, Hamabata T, Takeda Y, Yamasaki S: **Genome analysis of a novel shiga toxin 1 (Stx1)-converting phage which is closely related to Stx2-converting phages but not to other Stx1-converting phages**. *Journal of Bacteriology* 2003, **185**(13):3966–3971, [http://www.ncbi.nlm.nih.gov/pubmed/12813092]. [PMID: 12813092].

[438] Eppinger M, Mammel MK, Leclerc JE, Ravel J, Cebula TA: **Genomic anatomy of *Escherichia coli* O157:H7 outbreaks**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**(50):20142–20147, [http://www.ncbi.nlm.nih.gov/pubmed/22135463]. [PMID: 22135463].

[439] Shepard SM, Danzeisen JL, Isaacson RE, Seemann T, Achtman M, Johnson TJ: **Genome sequences and phylogenetic analysis of K88- and F18-positive porcine enterotoxigenic *Escherichia coli***. *Journal of Bacteriology* 2012, **194**(2):395–405, [http://www.ncbi.nlm.nih.gov/pubmed/22081385]. [PMID: 22081385].

[440] Strauch E, Hammerl JA, Konietzny A, Schneiker-Bekel S, Arnold W, Goesmann A, Phler A, Beutin L: **Bacteriophage 2851 is a prototype phage for dissemination of the shiga toxin variant gene 2c in E. coli O157:H7**. *Infection and Immunity* 2008, [http://www.ncbi.nlm.nih.gov/pubmed/18824528]. [PMID: 18824528].

[441] Yang F, Yang J, Zhang X, Chen L, Jiang Y, Yan Y, Tang X, Wang J, Xiong Z, Dong J, Xue Y, Zhu Y, Xu X, Sun L, Chen S, Nie H, Peng J, Xu J, Wang Y, Yuan Z, Wen Y, Yao Z, Shen Y, Qiang B, Hou Y, Yu J, Jin Q: **Genome dynamics and diversity of shigella species, the etiologic agents of bacillary dysentery**. *Nucleic Acids Res* 2005, **33**(19):6445–58, [16275786].

[442] Wilson DJ: **Insights from genomics into bacterial pathogen Populations**. *PLoS Pathogens* 2012, **8**(9):e1002874, [http://dx.plos.org/10.1371/journal.ppat.1002874].

[443] Pallen MJ, Loman NJ, Penn CW: **High-throughput sequencing and clinical microbiology: progress, opportunities and challenges**. *Current Opinion in Micro-*

biology 2010, **13**(5):625–631, [http://www.sciencedirect.com/science/article/B6VS2-511HN2H-1/2/d4a96e82fe99c89d0c410509b48b2c34].

[444] Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW: **Transforming clinical microbiology with bacterial genome sequencing**. *Nature Reviews Genetics* 2012, **13**(9):601–612, [http://www.nature.com/nrg/journal/v13/n9/abs/nrg3226.html].