Norwegian University
of Life Sciences

Master's Thesis 2018 60 ECTS
Faculty of Chemistry, Biotechnology and Food Science

In silico exploration of *stx2a*-positive (STEC) and *stx*-negative *Escherichia coli* (STEC-LST)

Abira Sivanesan
Chemistry and Biotechnology, Bioinformatics

ACKNOWLEDGEMENTS

This master's thesis has been written as a part of the Engineering Programme in Chemistry and Biotechnology (Bioinformatics as programme variant) at Norwegian University of Life Science (NMBU). It was a collaboration between the Faculty of Chemistry, Biotechnology and Food Science (KBM) at NMBU and the Bacteriological Department at Norwegian Institute of Public Health (NIPH).

I would like to thank my supervisors at NIPH and NMBU. I wish to express my deepest thanks and sincere appreciation to Jon Bohlin for his mentorship and Lin C. T. Brandal for her support. Truly, no better motivation exists than the enthusiasm you spread. My gratitude also goes to Lars Snipen for his guidance. The three of you have been great sources of knowledge and inspiration to this thesis. I would also like to thank Ole B. Brynildsrud for his bioinformatical support.

I would like to extend my thanks to my family and friends. To my parents, thank you for the endless encouragement and support. To my sister, Arabi, thank you so much for the strength and motivation. To Erlend, I am sincerely grateful for your patience and understanding. Thank you for always being positive and optimistic. And to my friends, thank you for watching over me and keeping me happy.

Stavanger, December 2018

Abira Sivanesan

SAMMENDRAG

Escherichia coli er en harmløs bakterie som naturlig sameksisterer med mennesker og dyr i tarmen. *E. coli*-genomet er dynamisk og kan skape store interne endringer (rekombinasjon), overføre plasmider (konjugasjon), tilegne seg DNA (transformasjon) og virus eller fager (transduksjon). Dette betyr at det finnes mange varianter av *E. coli* og at bakterien er vanskelig å analysere siden rekombinasjon og genetiske forandringer skjer fra generasjon til generasjon. Noen endringer er til fordel, som inkorporering av plasmider med antibiotikaresistens. Andre endringer kan være til skade, som integrasjon av en bakteriofag eller toksiner.

Det siste konseptet er av interesse når det omhandler shigatoksiner (Stx) produsert av shigatoksingener (*stx*). Toksinet observeres å bli overført mellom *E. coli* bakterier gjennom bakteriofager. Fagen er kartlagt ganske godt, men det er fortsatt karakteristisk informasjon som må avdekkes. Siden *stx* er veldig smittsomme for mennesker og dyr (som ofte er asymptomatiske), er det et behov for å forstå hvordan fagene overfører toksinene til *E. coli*.

I dette studiet er det av interesse å undersøke *E. coli* bakterier som er like. Noen *E. coli* i samme serogruppe med identisk multi-locus VNTR analyse-profil (MLVA-profil) produserer Stx, mens andre avstår. Siden *stx* er overført ved *stx*-fager, er det et behov for å analysere om det er rester fra fagen som kan ha forsvunnet fra genomet, eller andre varierende genetiske områder med spesifikke DNA-elementer hvor fager lett kan integreres.

Det er forskjellige metoder for tilnærming av problemet. Fenotypisk identifisering kan bli utført i et laboratorium ved ekstraksjon av DNA, kloning og amplifisering, og PCR gelelektroforese sammenlignet med en referanse. Analysemetoder som benytter databehandling kan være Sanger sekvensering og read-mapping ved bruk av referansegenom. Når det ikke finnes en referanse, anvendes neste generasjons sekvensering (NGS). I begge tilfeller skjer amplifisering digitalt, og sammenstillingen visualiseres ved identitetsrate og distanser som fylogenetiske trær. NGS-metoden er benyttet for dette studiet.

Målet med dette studiet er å forstå samsvarende *E. coli* med og uten *stx*, gitt at serogruppen og MLVA-profilen er identisk. Meningen bak dette er å kunne gi råd når det gjelder infeksjonskontroll for pasienter som er bærere av *stx*-negativ *E. coli* (atypisk EPEC / aEPEC). Er det en ekte aEPEC eller kan det være STEC-LST (*stx*-produserende E. coli – lost *stx*) som potensielt kan konvertere til STEC ved en *stx*-faginduksjon? Barn med aEPEC kan gå i barnehagen etter at diaré har opphørt, men barn med en potensiell STEC må holdes tilbake til tre kontrolltester viser negativt resultat. Dette vil derfor påvirke konsekvensene for den enkelte pasient og deres familie.

ABSTRACT

Escherichia coli is a harmless bacterium that naturally coexists with humans and animals in their intestinal tracts. The *E. coli* genome is dynamic and can make larger internal changes (recombination) like exchange plasmids (conjugation), add DNA (transformation) and virus or phages (transduction). This means that *E. coli* have many variants and are hard to analyze since recombination and genomic changes can happen from generation to generation. Some changes are beneficial, while others can be harmful. An example of a benefit could be incorporation of plasmids with antibiotic resistance genes, a disadvantage could be insertion of phages with toxins.

The last concept is of special interest regarding shiga toxins (Stx) produced by shiga toxin genes (*stx*). The toxin seems to be transferred between *E. coli* bacteria through bacteriophages. The phages are characterized relatively well, but there is still a lot of information left to reveal. Since *stx* is very infectious to humans and animals (often asymptomatic), there is a need to understand how the phages transfer these toxins to *E. coli*.

In this study there is an interest in researching *E. coli* bacteria that are similar. Some *E. coli* within the same serotype with identical multilocus VNTR analysis (MLVA) profile produce Stx, while others do not. Since *stx* is transferred through *stx*-phages, there exists a need to examine if any traces are left from the phage (that could be missing from the genome) or various genetic regions with certain DNA patterns where the phages easily integrate.

There are different methods to approach this problem. Phenotypic identification can be executed in a laboratory by extracting DNA, doing clonal amplification and PCR gel electrophoresis (compared to a ladder or reference). Sanger sequencing and computational methods like read-mapping use a reference genome. When there is no reference, next generation sequencing (NGS) can be applied. In both cases amplification happens digitally, and comparisons are visualized by identity rate or distances (f. ex. phylogenetic tree). The NGS method is chosen for this study.

The goal of this study is to understand similar *E. coli* with and without *stx*, given identical serotypes and MLVA-profiles. The reason for this, is to be able to give advice regarding infection prevention and control for a patient carrying *stx*-negative *E. coli* (atypical EPEC, aEPEC). Is it a true aEPEC or could it be a STEC-LST (*stx*-producing E. coli – lost *stx*) that potentially could be very dangerous if the *stx*-phage is recycled? Children with aEPEC can go back to kindergarten after diarrhea have ceased, but children with a potential STEC must be held home until three control tests show negative results. This will have larger consequences for the individual patient and their family.

TABLE OF CONTENTS

1 INTRODUCTION

1.1 *Escherichia coli*

Escherichia coli (*E. coli*) is a harmless bacterium that is part of a healthy human microbiota. It naturally lives in the intestinal tract and has simple requirements to survive. *E. coli* helps with absorption of nutrients and production of vitamin K. The bacterium is an environmentally versatile and adaptable species (Didelot *et al*., 2012). The genome is dynamic and can have internal changes (recombination), exchange plasmids (conjugation), add or delete DNA (transformation) or a virus (transduction). The genome can adapt traits from other organisms that coexist in the surroundings. This concept is called plasticity (Gordo *et al*., 2014). The high level of plasticity in *E. coli* leads to variety between individual bacteria and cause diversity – both within the genome (genotype) and in expression of genes (phenotype) (Gordo *et al*., 2014, Leimbach *et al*., 2013). These changes occur from generation to generation, and for *E. coli* the generation time is short (minimum 20 min). Some changes can be beneficial like incorporation of antibiotic resistance genes, while other traits can be harmful for example an insertion of a phage with toxins. Transduction by virus can cause anything from non-immediate effect (lysogeny) to virion (virus parts) production and cell death (lysis) (Kruger & Lucchesi, 2015). For an infected person, this could mean anything from asymptomatic reaction to gastrointestinal infections and severe illness.

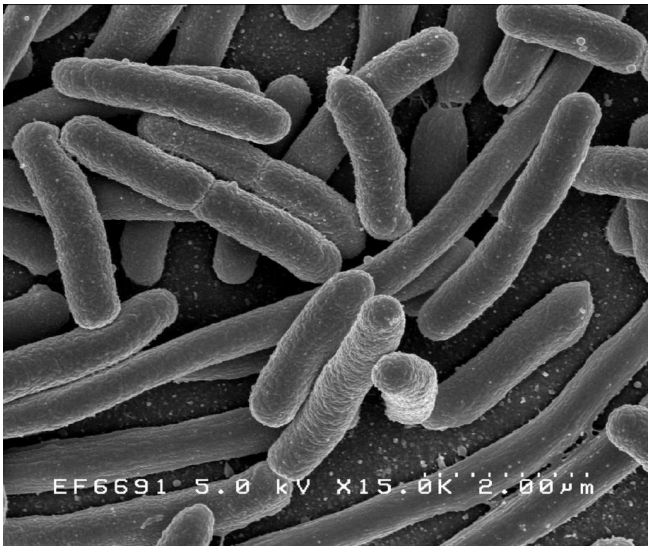

Figure 1.1: Micrograph of *E. coli.* "*Escherichia coli*: Scanning electron micrograph of *Escherichia coli*, grown in culture and adhered to a cover slip" by Rocky Mountain Laboratories, NIAID, NIH – NIAID (https://commons.wikimedia.org/w/index.php?curid=104228)

Humans and animals are at risk of infection due to the adaptability of *E. coli* to the environment. There are several pathogenic (infectious) variants of *E. coli*. Some *E. coli* are strictly pathogenic, but others are opportunistic causing infections when introduced to other organs or tissues (Ussery *et al*., 2009). Some are harmless to certain species but infectious to others. Because of the many variants of *E. coli* and a fast-changing rate of mutation and recombination, it is difficult to analyze these bacteria efficiently. It has been challenging to classify *E. coli*, especially the pathogenic variants (Rasko *et al*., 2011). New emerging hybrid *E. coli* strains are observed and described; for instance, the pathogenic *E. coli* strain causing the severe outbreak in Germany 2011 (L'Abée-Lund *et al*., 2012). Depending on the species and strain, bacteria may be categorized as strictly pathogenic, opportunistic, commensal or non-pathogenic (Center for Disease and Control, 2014). *E. coli* can possess the properties of all the above-mentioned categories (Leimbach *et al*., 2013). Multilocus sequence typing (MLST) and multiple-locus variable number tandem repeat analysis (MLVA) are traditional genotyping methods to track outbreaks of *E. coli* (Maiden *et al*., 1998). MLVA detects specific variable number tandem repeat (VNTR) regions in the genome (Nadon *et al*., 2013). Core genome MLST (cgMLST) is "gene-by-gene" approach to detect allelic differences in the genome and can be used to make phylogenetic groupings or trees (Center for Disease and Control, 2016, National Institute for Public Health and the Environment, 2016).

E. coli classification happens by phylogenetic groups (phylogroups), serotypes and pathogenic profiles (pathotypes) (Didelot *et al*., 2012). There are five major phylogroups (A, B1, B2, C, D), two minor (E and F) and an eighth group named *Escherichia* cryptic clade I (Clermont *et al.*, 2013). Classification by serotyping is done by detecting combination of bodily or somatic (O), tail or flagellar (H), and protective or capsular (K) polysaccharide antigens presented on the surface of the bacterium (Stenutz *et al*., 2006). The abbreviation for serotype is OH-type, for the somatic and flagellar antigens. Most serotypes are commensal, but certain are pathogenic (Didelot *et al*., 2012). Pathotyping is grouping by virulence or pathogenicity and is divided into extraintestinal (ExPEC) and diarrheagenic *E. coli* (DEC). ExPEC usually infect the urinary tract, the blood stream (sepsis), or the membranes of the brain (meninges) (Stenutz *et al*., 2006). DEC is divided into six main pathotypes: shiga toxin-producing (STEC) or enterohemorrhagic (EHEC), enterotoxigenic (ETEC), typical and atypical enteropathogenic (tEPEC / aEPEC), enteroaggregative (EAEC),

enteroinvasive (EIEC), and diffusely adherent (DAEC) *E. coli* (Center for Disease and Control, 2014). Infection severity by DEC depends on virulence and pathogenicity of the specific strain.

STEC, also known as verocytotoxin-producing *E. coli* (VTEC), are characterized by their ability to produce Shiga toxin(s) (Stx) (Scheutz, 2014). All STEC strains are able to produce Stx as their main virulence factor and harbour one or more of the Stx-encoding genes (*stx*) (Brandal *et al*., 2015). Apart from the main virulence factor Stx, the individual STEC strain can carry other virulence factors or toxins, including phage related genes or genes located on pathogenicity islands or virulence plasmids (Mellmann *et al*., 2009). The pathogenicity island locus of enterocyte effacement (LEE), includes the *eae* gene encoding intimin, which is a large outer membrane protein, necessary for attachment and effacing (AE) lesion formation (Brandal *et al*., 2015, Scheutz, 2014, Mellmann *et al*., 2009). Other important virulence genes are enterohemolysin (*ehxA*) and cytolethal distending toxin B (*cdtB*), located on virulence plasmids (Franzin & Sircili, 2015, Clements *et al*., 2012).

A STEC infection can vary from asymptomatic carriage to life-threatening illness: hemorrhagic uremic syndrome (HUS) (Didelot *et al*., 2012). The primary infection site of STEC is the colon, and disease symptoms can vary from watery diarrhea to haemorrhagic colitis (HC) including bloody diarrhea, stomach cramps and abdominal pain, and can progress to HUS (Krüger *et al*., 2015). Stx is the key virulence factor for development of HUS. It binds to the Gb3 receptor located on eukaryotic cells in the kidney and brain (Obrig & Karpman, 2012, Obrig, 2010). This might lead to dehydration, hypertension, anemia, renal failure and in a few occasions this infection has a deadly outcome (L'Abée-Lund & Wasteson, 2015). Antibiotic treatment during a STEC infection is contradictory due to studies showing induction of Stx-production after antibiotic treatment, thus increasing the risk for HUS (Bielaszewska *et al*., 2012). Diarrhea-associated HUS is treated symptomatically by supportive care of controlling the fluid and electrolyte balance.

Ruminants (bovine and ovine) are asymptomatic STEC carriers and the main reservoir for STEC. Humans are infected through ingestion of fecal contaminated food or water, through direct contact with the animals, or by person-to-person spread (Krüger *et al*., 2015). The

infectious dose of STEC as low as 100-1000 bacteria are enough for development of an infection (L'Abée-Lund & Wasteson, 2015). This makes STEC highly infectious.

STEC are classified into five seropathotypes (SPT) based on a gradient ranging of pathogenicity from A-E, where A is "high risk" and E is "minimal risk" (Scheutz, 2014). The highly pathogenic STEC serotypes are O157:H7, O26:H11, O145:H25, O103:H25, O111:H8, O121:H19, O177:H25, O145:H28 and their non-motile (NM) derivatives (Delannoy *et al*., 2013). STEC O157:H7 and O157:NM serotypes are the most virulent of the strains having SPT-A classification (Scheutz, 2014). In persons infected with non-sorbitol-fermenting (NSF) O157, HC ensues after 1-3 days in approximately 90% of the cases. Children, elderly and immunocompromised patients are most prone for severe complications such as HUS. Several studies have found low age (≤ 5 years), the presence of *stx2* and *eae* as risk factors for HUS (Brandal *et al*., 2015, Scheutz, 2014, Friedrich *et al*., 2002, Naseer *et al*., 2017). NSF STEC O157 are most commonly involved in outbreaks worldwide. However, the sorbitol-fermenting (SF) STEC O157 is emerging as an important pathogen and has been the cause of several HUS outbreaks in Europe (Bieleszewska *et al*., 2006, Byrne *et al*., 2018).

There have been examples in which *eae*-positive *E. coli* without *stx* have been isolated from cases with HC and HUS. STEC and HUS-associated *eae*-positive *E. coli* with identical serotypes have been compared in studies and show phenotypical similarity and same virulence factors excluding *stx*-genes (Ferduous *et al*., 2015, Mellmann *et al*., 2009, Byrne *et al*., 2018). These bacteria are named STEC – lost shiga toxin (STEC-LST) and should be differentiated from "true" aEPEC as aEPEC is less harmful. However, distinguishing STEC-LST from aEPEC in the laboratory is difficult and few genetic markers exist (Bugarel *et al*., 2011). Studies have shown that STEC-LST have the ability to recycle *stx* during an infection due to phage induction. A theory is that these bacteria live in a dynamic environment where *stx*-phages are gained and lost (Mellmann *et al*., 2009, Haugum, 2014). The mechanisms behind this recycling event are still under investigation.

1.2 Bacteriophages

Bacteriophages (phages) are viruses that are attuned to bacteria. The viral particles (virions) consist of the genome containing either DNA or RNA (double or single stranded), a protein coat (capsid) which protects the genetic material, and in some cases a lipid envelope surrounding the capsid (Koonin & Starokadomskyy, 2016). Phages are a hundred times smaller than bacteria and their shapes range from simple helices to more complex structures. Viruses are classified by morphology; size, shape, chemical composition, structure of genome, and mode of replication (Gelderblom, 1996). They can carry pathogenic genes and cause viral infections by transferring genetic material (transduction) into a bacterial genome. Some phages are attuned to infect one bacterial species while others have a broader range of potential hosts. The lambda and T4-phages have the ability to infect *E. coli* (Muniesa & Schmidt, 2014, Willey *et al.*, 2014). Studying virulent phages are necessary when discussing alternative treatments for antibiotic-resistant bacteria (Cisek *et al.*, 2017).

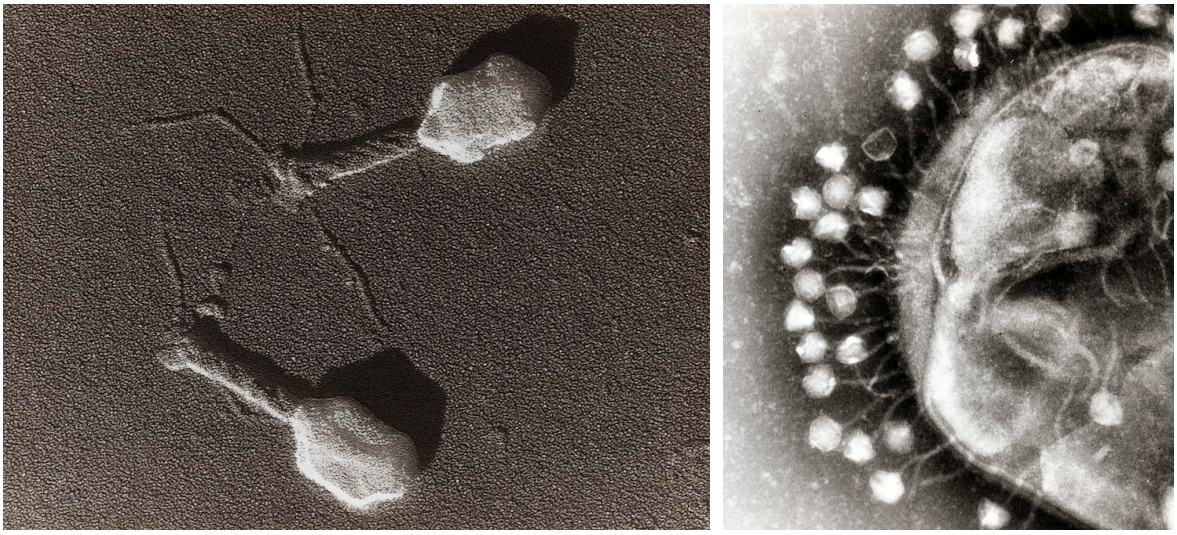

Figure 1.2: Bacteriophages. "Bacteriophage: a virus that feeds on bacteria" (left) by AFADadcADSasd (https://upload.wikimedia.org/wikipedia/commons/e/eb/Bacteriophage.jpg) and "Transmission electron micrograph of multiple bacteriophages attached to a bacterial cell wall" (right) by Dr. Graham Beards (https://commons.wikimedia.org/wiki/File:Phage.jpg)

Bacteriophages may be in a virulent or temperate infection mode (Krüger & Lucchesi, 2015). While temperate phages can either cause cell death (lysis) or remain within the host like the lambda phage, a virulent phage exploits and destroys the host like the T4-phage. If not

induced to a lytic cycle, the genes adhere to the lysogenic cycle and become a part of the bacterial genome – called a prophage (Willey *et al*., 2014). The bacteria carrying the prophage are called lysogens or lysogenic bacteria (Casjens, 2003). If the pathogenic gene does not extinguish the host by lysis, they may even benefit each other by higher nutritional extraction, energy saving or hardened exterior membrane for better survival (Krüger & Lucchesi, 2015). The lysogen might be triggered to convert by imbalances in the environment such as stress and shock from nutritional devastation, dehydration, UV irradiation, growth deprivation, antibiotics or another infection (Muniesa & Schmidt, 2014). This conversion is called prophage induction where virions are synthesized, and the host enters the lytic cycle. Prophage induction leads to free phages that can infect new bacteria (Krüger & Lucchesi, 2015).

Prophage DNA contributes to large varieties within bacterial species as up to 10-20% of a bacterial genome may consist of prophage DNA (Casjens, 2003). For preserving the genomic variation and at the same time prevent lysis, the prophage DNA often needs to be repressed to incorporate well with the bacterial host genome. This often happens through point mutations or recombination in the prophage region. Genomic mutations are biased towards richness in AT-content and cause lower entropy because of less rigidity in AT-bonds (Bohlin *et al*., 2014). Areas containing bacteriophages are more AT-rich compared to the rest of the host genome often caused by these high mutation rates and adaptabilities. This makes it easier to detect areas with prophage DNA. Although the integrated viral genes often are defective and do not produce virions any longer, the importance of transduction as a method for gene transfer is depicted through the vast number of bacterial genomes with prophage DNA (Casjens, 2003). Phage-mediated gene transfer is therefore an important study as it leads to the evolution of pathogenic bacteria.

Although specific bacteria can be infected by various phages, phages are selective toward their hosts. Within *E. coli*, certain viruses target a specific phylogenetic group of the host (Gamage *et al.,* 2004, Muniesa & Schmidt, 2014). Phages DNA can cause beneficial lysogenic effects to bacteria, like incorporations of anti-microbial resistance (AMR) genes. These genes are usually a great cost to maintain and the bacteria are easier outmatched by competitive microbes. But as long as they are resistant to antibiotics, they might be a threat to human

health. When combating bacterial or viral infections and other illnesses, CRISPR and phage therapy are currently hot topics. Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) is a gene-editing tool that targets and edits specific genetic codes at precise locations (Hsu *et al*., 2014). The CRISPR-Cas9 genome editing tool was adapted from a naturally occurring bacterial editing system (Gupta & Musunuru, 2014). The CRISPR-associated protein 9 (Cas9) enzyme cuts the DNA at the targeted location (Hsu *et al*., 2014, Jiang *et al*., 2015). This enzyme is used most often but can be exchanged by other enzymes (for example Cpf1). After cutting, the DNA repair system inserts and deletes sequences as needed (Genetics Home Reference, 2018). CRISPR can prevent and treat human disease by altering the genome and is currently utilized for single-locus treatment which targets one region at a time. There might be developed a method to eliminate prophages and toxins completely through CRISPR. Phage therapy is treatment using lytic bacteriophages to target pathogenic bacterial infections. Phage-mediated therapy can induce phages to interact with lysogenic bacteria and result in host lysis (Nagel, 2018). Cell death during early phage infection leads to limited spread of virions, which at least halts illness development. There might be a potential to evolve phage therapy from only treating complications to targeting severe diseases, in the future (Górski *et al*., 2018). Phages are present in the intestinal tract in high concentrations, thus might phage therapy be an efficient treatment method to eliminate STEC infections (Nagel, 2018).

stx-phages carry *stx* and are related to temperate lambdoid bacteriophages (Huang *et al.*, 1987, Scheutz, 2014). They have similar phage cycle regulations. In lysogenic state, the DNA of the *stx*-phage is integrated into the STEC chromosome and the expression of the majority of *stx*-phage genes is inhibited. Even if most of the lysogens are stable, external instabilities can spontaneously induce phage production by expression of the *stx*-genes (Krüger *et al.*, 2015, Scheutz, 2014). In this case, the bacterial host will lyse. Expression or repression of *stx* in STEC is highly dependent on prophage induction, although transcription can be driven by own promotors under certain conditions. A higher level of spontaneous induction occurs in *stx*-phages in comparison to non *stx*-phages (Krüger *et al.*, 2015). Stx causes death of eukaryotic cells like human leukocytes (Steinberg *et al.*, 2007).

stx-prophages can convert the host strains to pathogenic bacteria. Prophages are integrated into the genome by attaching to specific insertion sites (Krüger & Lucchesi, 2015). The phages use one insertion site depending on the host strain, and when unavailable the phage integrates into a secondary site. There are several integration sites for *stx*-phages: *wrbA, yehV, sbcB, argW, yecE, potC, prfC, serU, ssrA, yciD, yecD, yjbM, ynfH, Z2577* (Krüger & Lucchesi, 2015). Five target sites have been described specifically for SFO157: *wrbA*, which codes for a NADH: quinone oxidoreductase; *yehV*, which codes for a transcriptional regulator; *sbcB*, which produces an exonuclease; *yecE*, whose function remains unknown; and *Z2577*, which encodes an oxidoreductase (Sierra-Moreno *et al.*, 2007, Scheutz, 2014, L'Abée-Lund *et al.,* 2012).

Other phages in the host genome can regulate expression of *stx*-phages. Strains harboring several *stx*-phages affect Stx production more in comparison to strains with only one *stx*-phage (Krüger & Lucchesi, 2015). There is also a co-regulation between *stx*-phages and the bacterial host. Internal bacterial factors affect phage induction in *stx*-phages (Brandal *et al.*, 2015). *stx*-phage lysogeny increase rigidity by higher acid tolerance and motility in the host. Since *stx*-phages are infectious to humans and animals (often asymptomatic), there is a need to understand how the phages transfer these toxins to *E. coli* (Scheutz *et al.,* 2012). The genomes of *stx*-phages are characterized relatively well, but still there is some information left to reveal. Some genes in the phage genomes are still to be uncovered, and as of now many annotate as encoding "hypothetical proteins". Therefore, virulence factors and disease severity are difficult to determine for the *stx*-phages. *stx*-phages can persist longer than their hosts especially in aquatic environments and can tolerate exposure to disinfectants and maintain their infectivity under food-processing conditions. Therefore, may transmission occur in water, food and biofilms (Krüger *et al*., 2015). Antibiotics are contradicted during a STEC infection due to induced *stx*-phage induction and increased Stx production, although certain antibiotics have shown to eliminate STEC without triggering the lytic cycle (Krüger & Lucchesi, 2015). Anti-induction strategies, phage induction repression studies and amino acid-starvation have been tested to diminish the efficiency of phage formation. Fasting and providing minerals and citrate have helped in managing STEC infections, but variability in *stx*-phages cause difficulty in treatment (Krüger & Lucchesi, 2015).

1.3 *stx* genes

The *stx*-genes are transferred between *E. coli* bacteria through bacteriophages or *stx*-phages. Stx shares a similarity in structure and activity to *Shigella dysenteriae* toxin. It inhibits protein synthesis by inactivating the 60S ribosomal subunits (part of eukaryotic ribosome) and the toxins are released when the bacteriophage induces lysis (Krüger *et al.*, 2015). The *stx* genes are classified in two major types: *stx1* and *stx2* – subtyped into *stx1a*, *stx1c* and *stx1d*, and *stx2a*, *stx2b*, *stx2c*, *stx2d*, *stx2e*, *stx2f* and *stx2g* (Scheutz *et al*., 2012). In humans STEC with *stx2a* and/or *stx2d* are associated with high virulence, whereas STEC with *stx1*, *stx2b*, *stx2e*, *stx2f* or *stx2g* is associated with low virulence (Krüger & Lucchesi, 2015, Brandal *et al*., 2015). The *stx*-genes are encoded in the late region of lambdoid prophages where they are located downstream of the promoter $p_{R'}$ and late terminator $t_{R'}$. *stx* have their own promotors, but induction of the prophage and transcription from $p_{R'}$ is important for the expression of the *stx*-genes, as well as release of Stx from the bacteria. The anti-terminator activity of the Q protein, encoded by *q*, is necessary for read-through of the late terminator ($t_{R'}$) and activation of $p_{R'}$. It has been proposed that the expression of Stx2 might be influenced by the *q*-gene and in O157:H7 STEC q_{933} is associated with higher *stx2a* production than q_{21} (Haugum *et al.*, 2012, Olavesen *et al.*, 2016).

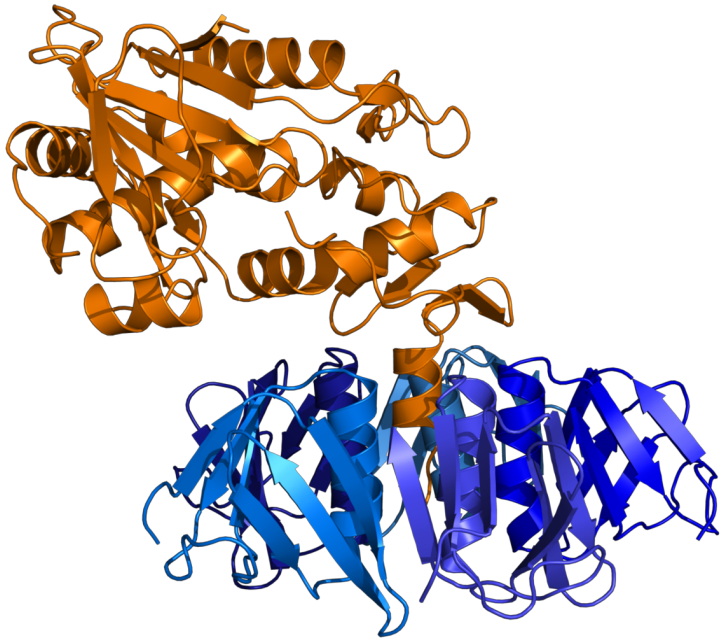

Figure 1.3: Protein structure of shiga toxin. "Ribbon diagram of Shiga toxin (Stx) from *Shigella dysenteriae*" by Fraser *et al*. (2004) (https://www.rcsb.org/structure/1R4Q). Subunit A1 and A2 (orange), subunit B (dark blue), and volumetric surface (light blue).

1.4 Biotechnology

Several decades ago researchers either studied pathogens inside a host or outside on a petri dish, called *in vivo* and *in vitro* study techniques respectively. The biological field have evolved making biotechnology the future when resolving problems regarding medical biology. Bioinformatics is the fusion of biology and computer science, where biological data is often rapidly analyzed with computer technology that can process information in bulk simultaneously, called *in silico* research. *In silico* research through bioinformatics and computational biology has the ability to characterize genomes precisely and faster. Optimization is ideal when conducting an experiment with genome construction that lead to enhanced understanding of gene contents and their functions. Through phylogenetics and taxonomical graphs organisms can be classified, which upon a medical practitioner can determine a diagnosis for the patient accurately.

The biological data has to be collected in a laboratory before it is analyzed *in silico*. Phylogenetic identification happens in a laboratory through methods like MALDI-TOF, biochemistry and serology. Genotypic identification is utilized to determining virulence genes, serotype, antibiotic resistance genes, *et cetera*. by extracting DNA through PCR-based technology and capillary electrophoreses, real-time PCR, Sanger sequencing and NGS. Sanger sequencing and read-mapping use a reference genome to execute *in silico* analysis. When there is no reference, next generation sequencing (NGS) can be applied. In NGS, amplification happens digitally, and comparisons are visualized by identity rate or distances (f. ex. phylogenetic tree).

The evolution of bioinformatics started with Sanger sequencing as first generation sequencing. In the beginning, bioinformatics was directed towards Sanger sequencing to duplicate longer strains of DNA. This was beneficial because the strains were held intact, but the sequencing efficiency was low. Later on, massive paralleled sequencing was introduced as second generation sequencing with Illumina sequencers leading the way. Illumina sequencing technologies (NGS methods) cannot produce the complete sequence of a chromosome in one continuous strain. Instead, they generate large numbers of reads (short sequences) ranging from tens to thousands of consecutive bases sampled from different parts of the genome. Genome assembly software combines the reads into larger regions

called contigs (Gurevich *et al.*, 2013). Single molecule real time (SMRT) sequencing with one template DNA has become third generation sequencing with the Oxford Nanopore sequencer method. SMRT sequences only one DNA strand as its template similar to Sanger sequencing but is parallelized to save time. In this study, Illumina sequencing was applied. With NGS the efficiency is higher due to multiple sequencing of smaller reads in lager parallels. When fusing reads, there could be erroneous overlapping because of frequent Single Nucleotide Polymorphisms (SNPs) (Karki *et al*., 2015). The errors are corrected with an error corrector when overlapping reads into contigs. NGS is applied on big genomes (>1000 base pairs), and is usually advanced and upgraded quickly to fit new modern technology.

NGS is executed with Illumina sequencing technology to determine the nucleotide base pairs in a chain of DNA, called DNA sequencing (Ussery *et al*., 2009). Purified DNA is cut into smaller pieces and added different adapters to each side. The DNA is put on a flow cell where the adapters attach to the reverse-complementary adapters, and folds over into a bridge-like shape (Ussery *et al*., 2009). A primer attaches to the bend and a polymerase synthesizes the reverse-complementary strand producing a reverse stand to the forward strand. The strands release and each form a new bridge, repeating the process. This is performed multiple times creating clonal copies of both strands simultaneously. The result is a cluster of clones containing thousands of copies of the same DNA strand. The flow cells are arranged like a chip and the nucleotides added via polymerase are loaded with fluorescent tags (Zvelebil & Baum, 2007). The wavelength of the fluorescent light is signaled to a detector that determines the base added. This is recorded for every spot on the chip, converted into sequences, and then the data is analyzed on a computer. This method is completed in massive parallel and can be used for whole genomes, regions, transcriptomes, metagenomics, nucleic acid-protein interaction, and other analysis (Zvelebil & Baum, 2007).

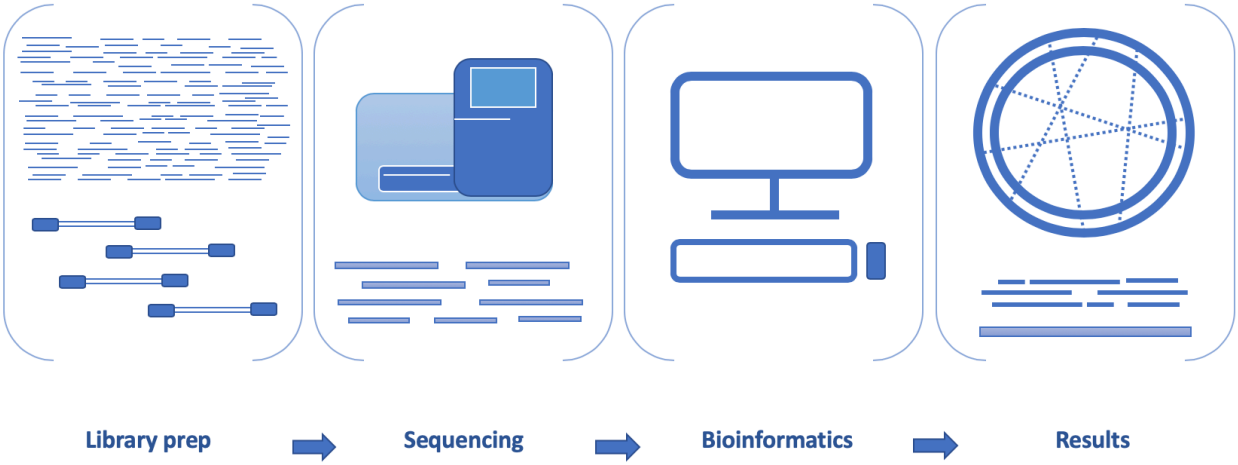

Figure 1.4: General overview. Rough steps in analyzing *E. coli* data from laboratory testing to *in silico* methods.

Biotechnology evolve fast with the change in technology and computer science. The usual way of analyzing biological information on a computer was *in situ*, where a reference genome was used (Sanger). Now, there exist elaborated databases and *in silico* researches, where the whole genome sequence is formed by itself through specific algorithms. This is called *de novo* assembly.

De novo assembly performs methods with metrics that do not need a reference to assemble a genome (Sohn & Nam, 2018). It is similar to attempt to read a whole-genome sequence at once. This is difficult to do with the limitations of modern technology. The key is to approach it like a puzzle with million pieces and use the fragments to assemble the genome. This is achieved by matching regions of shared reads that fit together. *De novo* assembly methods use k-mer counting to create the most probable sequence, where k is the best fitted number (Sohn & Nam, 2018). k represents how many nucleotides a read contains, meaning smaller ks produce many k-mers. The k-mers are then used to produce a de Bruijin graph, which is the main building chart for the assembly. In this graph the most probable sequence of overlapping k-mers is shown to a specific prefix and suffix of (k-1)-mer (Sohn & Nam, 2018). De Bruijin graphs have two types of classification of expressing nodes and edges: Hamiltonian and Eulerian. Hamiltonian de Bruijin graphs expresses the k-mer as a node and

the overlapping suffix or prefix as an edge, and Eulerian the other way around where the overlap is a node and the k-mer is an edge (Sohn & Nam, 2018). The Eulerian de Bruijin graph-based assembler generally performs better in terms of assembly results. By first assembling the contigs and scaffolds and then filling the gaps, the build goes from nucleotide bases to contigs and scaffolds to possibly an error-free whole genome assembly.

The correction of sequencing errors is of critical importance, because it can obstruct the process of assembly and introduce faulty constructs (Pevzner *et al*., 2001). There are certain challenges in *de novo* assembly because human genomes consist of non-randomly repeated elements and topological complexity that cause interrelatedness between genes. These can be long and short interspersed nuclear elements (LINEs and SINEs, respectively), long terminal repeats (LTRs) and simple tandem repeats (STRs), and cause misarrangements and gaps (Sohn & Nam, 2018). Therefore, is the overlapping of short reads (about 50-300 nt for Illumina) enabling large genome assemblies. Erroneous sequencing regions can be limited by directly aligning reads with each other and correct by consensus. Highly repetitive structures in the De Bruijin graph can create ambiguous paths and gaps in the resulting genome assembly, which can be solved by longer reads or larger read depth (Sohn & Nam, 2018). Hybrid or long-read-only methods use overlapping, gap closers, and error correction through SMRT long reads, scaffolding and contig assembly sometimes assisted by short reads to correct for sequencing errors and generate the assembly rather than relying on de Bruijin graph methods (Sohn & Nam, 2018). Quality evaluation is executed by QUAST with the broken contigs. It applies the N50 method. N50 is similar to a mean of length where greater weight is given to the longer contigs and is the minimum contig length to cover 50% of the genome. It is a measure to describe the quality of the assembled genome that is fragmented in contigs of different lengths.

Another way of assembly is to use reference-based read-mapping. This is usually performed by Bowtie2 and SAMtools, or Burrows-Wheeler Aligner (BWA). A previously assembled genome is used as a reference or template. Sequenced reads are then independently aligned against the reference and placed at the most likely position (Langmead, 2010). As this method is too slow for many reads on a big reference, it is sped up with reference indexes (Langmead *et al*., 2009). These indexes are fast lookup tables for subsequences in the

reference. All possible alignment positions are found from the index, called seeds (Langmead & Salzberg, 2012). Every seed is evaluated and extended into full alignments. Insertions and deletions increase the complexity of the alignment. Optimal alignment is executed through dynamic programming like “Smith-Waterman” algorithms. The output is a BAM alignment file (Langmead *et al*., 2009). The method is used for smaller projects. It is also used after de novo assembly as variant calling to detect alignment differences from a database reference genome.

After sequencing, different samples can be aligned to look for similarities. “Smith-Waterman” and “Needleman-Wunsch” both perform alignments of nucleotide or protein sequences (Langmead *et al*., 2009). The results form a scorings matrix to find the optimal alignment by giving punishments for mismatches, gaps and substitutions. “Smith-Waterman” performs local alignments by selecting a part of the sequence that align optimally. “Needleman-Wunsch” performs global alignments regarding the optimal match of the whole sequence. CLUSTAL, Muscle and MAFFT are some examples of alignment tools.

Clustering and mapping can be done by distance measures from the alignment data. The distances are usually measured as Euclidian or Hamiltonian units and then a distance or dissimilarity matrix is created (Zvelebil & Baum, 2007). Phylogenetic trees, hierarchical clusters and heatmaps can illustrate relatedness and matches within and between organisms by utilizing the distance matrix to make nodes and edges as mentioned above. Neighbor-joining for phylogenetic trees, maximum likelihood with Bayesian inference, and unweighted pair group method with arithmetic mean (UPGMA) and its weighted counterpart (WPGMA) for hierarchical clustering are statistical methods or algorithms to set up a model and visualize the sample data (Ussery *et al*., 2009). With shortened processing time due to modern technology, likelihood methods are used frequently because of the optimization.

These statistical algorithms can also determine the build of a model or function to predict future outcomes by separating the data into training data and testing data. These algorithms can be prone to overfitting or underfitting of the model. This usually happens when the data is too complicated or not complicated enough – too many or too little observations. Overfitting happens when the model fits the current data too well and fails to fit predicted

observations reliably. Underfitting occurs when the model cannot capture the correct structure because of missing data. In this way, Occam's razor or the law of parsimony – a principle for problem solving indicating the simplest solution to be the correct one – will usually give something closer to the correctly specified model (Ussery *et al*., 2009). By a standard, all models are usually incorrect, but some are more useful than others. This learning algorithm method, with training and testing data, is called machine learning and is currently used for creation of artificial intelligence for example. Biologists would like to use similar methods to be able to classify cellular images, make genomic connections and advance drug discovery, to mention a few things.

De novo assembly and read-mapping is done through several different programs specified for each part of the procedure in chapter 2, materials and methods.

1.5 Aim of study

STEC affects children under 5 years old, elderly and patients with severe autoimmune diseases with serious complications like HC and HUS. In Norway strict guidelines for prevention and control of cases with high virulent STEC infection belonging to risk groups are implemented (Folkehelseinstituttet, 2010). Although, limited knowledge is available considering differentiating a STEC-LST from a "true" aEPEC. A person infected with an *eae*-positive *E. coli* defined as atypical EPEC can return to kindergarten and work 48 hours after cessation of diarrhea without any need of follow-up measures. On the contrary, a case with an *eae*-positive *E. coli* with identical serotype and MLVA-profile to a STEC isolated from a HUS case, might reflect a STEC-LST, and thus require follow-up guidelines comparable with a high-virulent STEC. Therefore, it is important to understand if the infection stems from a "true" aEPEC or a STEC-LST to predict a proper diagnosis. Strict control measures have huge socioeconomic impact on cases and their families and should only be recommended if needed.

The aim of this study was to compare STEC and *stx*-negative but *eae*-positive *E. coli* with identical serotype and similar MLVA-profile. Computational biology and bioinformatics were applied in order to see how similar these strains were. The goal was to find characteristics substantiating the classification of the *eae*-positive *E. coli* as STEC-LST (not a "true" aEPEC). The bioinformatics analyses were performed on 22 *eae*-positive *E. coli* isolates, with main focus on serotype SFO157:H7. Five additional serotypes were analyzed: O145:H25, O26:H11, O103:H25, O145:H28, and O177:H25, all associated with severe *E. coli* infection. Genomic examination was performed on *stx*-phage traces and DNA-patterns in certain genetic regions where phages easily integrate. Additionally, in *E. coli* SFO157, specific integration sites as *yecE*, *wrbA*, *yehV*, and *sbcB* were examined.

2 MATERIALS AND METHODS

2.1 Research methodology

This study uses quantitative methods in a natural-scientific research analysis on biological data through computational framework. Biostatistical and bioinformatical programs are utilized as tools to assemble, annotate, and align sequence reads made from the *E. coli* isolates. The *E. coli* strains were sent from clinical microbiological laboratories throughout Norway to the National Reference Laboratory (NRL) at Norwegian Institute for Public Health (NIPH) for verification and further characterization (serotype both phenotypically and molecularly, PCR for *stx1*, *stx2*, *eae* and *ehxA*, PCR for *stx*-subtyping, and MLVA).

In this study Next-Generation Sequencing (NGS) is applied through *de novo* assembly. Reference-based genome sequencing has also been utilized through read-mapping. Most commands for programs were run through the UNIX system (OS X on a macbook, 2016).

2.2 Material

An *in silico* research of STEC was conducted with isolates of 22 *E. coli* cultures analyzed at Norwegian medical facilities. “Pairs” of isolates with identical serotype and MLVA-profile (including single locus variant [SLV]), with and without the *stx2a* gene, were selected. Six different serotypes, associated with severe clinic of cases infected in Norway, were included. The main focus was on sorbitol-fermenting O157:H7 (SFO157). SFO157 was chosen because of identical MLVA-profiles, closest relatedness in cgMLST, the SPT-A rating from seropathotype classification, and interesting clinical outcomes (Scheutz, 2014). An overview of the *E. coli* isolates is presented in Table 2.1. When the data was collected at NIPH, the MLVA- and cgMLST-profiles were produced as described under section 2.3.5 Alignment, and phylogenetic trees were provided. The virulence genes including *stx2a* were also identified. The general explanation of the approach can be viewed in Figure 2.2.

Table 2.1: Overview of *E. coli*. Serotype, virulence genes, MLVA profile, year of isolation and source are presented according to the specified *E. coli* (EC) isolate number.

Isolate	Serotype	Virulence gene	MLVA profile	Year	Source
EC1	O145:H25	*stx2a, eae, ehxA*	5-3-0-8-4-1-1-16-9-15	2009	Human
EC2	O145:H25	*stx2a, eae, ehxA*	5-3-0-8-4-1-1-16-9-15	2009	Human
EC3	O145:H25	*eae, ehxA*	5-3-0-8-4-1-1-16-9-**16**	2009	Human
EC4	O26:H11	*stx2a, eae, ehxA*	6-0-0-8-3-2-1-6-25-11	2014	Human
EC5	O26:H11	*eae, ehxA*	6-0-0-8-3-2-1-6-**27**-11	2011	Human
EC6	O26:H11	*eae*	6-0-0-8-3-2-1-6-**26**-11	2009	Ovine
EC7	O26:H11	*stx2a, eae, ehxA*	6-0-0-8-3-4-1-6-15-15	2015	Human
EC8	O26:H11	*eae, ehxA*	6-0-0-8-3-4-1-6-15-**14**	2010	Human
EC9	O103:H25	*stx2a, eae, ehxA*	7-3-0-5-0-7-1-16-9-12	2006	Human
EC10	O103:H25	*eae, ehxA*	7-3-0-5-0-7-1-16-9-12	2007	Ovine
EC11	O103:H25	*eae*	7-3-0-5-0-7-1-16-9-12	2006	Mutton
EC12	O103:H25	*stx2a, eae*	7-3-0-5-0-7-1-16-9-12	2006	Human
EC13	SFO157	*stx2a, eae, ehxA*	11-0-23-0-6-4-0	2009	Human
EC14	SFO157	*eae, ehxA*	11-0-23-0-6-4-0	2009	Human
EC15	SFO157	*eae, ehxA*	4-23-3-6-3-5-3	2008	Human
*EC16	SFO157	*stx2a, eae, ehxA*	4-23-3-6-3-5-3	2008	Human
*EC17	SFO157	*eae, ehxA*	4-23-3-6-3-5-3	2008	Human
EC18	O145:H28	*stx2a, eae, ehxA*	7-3-0-8-3-2-1-35-0-0	2013	Human
EC19	O145:H28	*eae, ehxA*	7-3-0-8-3-2-1-35-0-0	2013	Human
EC20	O177:H25	*stx2a, eae, ehxA*	5-3-0-15-4-1-1-16-11-15	2013	Human
EC21	O177:H25	*stx2a, eae, ehxA*	5-3-0-15-4-1-1-16-11-**13**	2013	Human
EC22	O177:H25	*eae, ehxA*	5-3-0-15-4-1-1-16-11-**10**	2013	Human

*SFO157:H7 – EC16 and EC17 are from the same incident.

The strains were sequenced using Illumina paired-end sequencing technology (NGS method). The raw read data consisted of four FASTQ files per isolate; the paired forward and reverse reads and the unpaired forward and reverse reads. For further processing only the paired FASTQ files were used.

2.3 Method

First, Illumina sequencing was applied. Then, the datasets were trimmed, combined, and assembled using different approaches. FASTQC was used to check the quality of the NGS reads before trimming. *De novo* assembly and multiple sequence alignment was performed on the data after assembling and error correction. Thereafter, phages were collected, and all the different data were compared to each other.

Specific and specialized biotechnological tools were used to process and analyze the *E. coli* strains. The methods can be divided into seven sections: 1. Illumina sequencing, 2. Preprocessing, 3. Genome assembly, 4. Genome annotation, 5. Alignment, 6. Phage detection and 7. Dataset comparison. An overview of all the steps included in the analysis is provided in Figure 2.1.

2.3.1 Illumina sequencing

Illumina sequencing was completed at NIPH and fastq-files were handed down along with the Nextera adapter sequences. Illumina paired end sequencing technology is sequencing DNA fragments from both ends. The fragments library size was about 600-900 bp, and the read pairs size was 250 bp on each end. The theory behind Illumina sequencing technology is explained in chapter 1, Introduction, under section 1.4 Biotechnology.

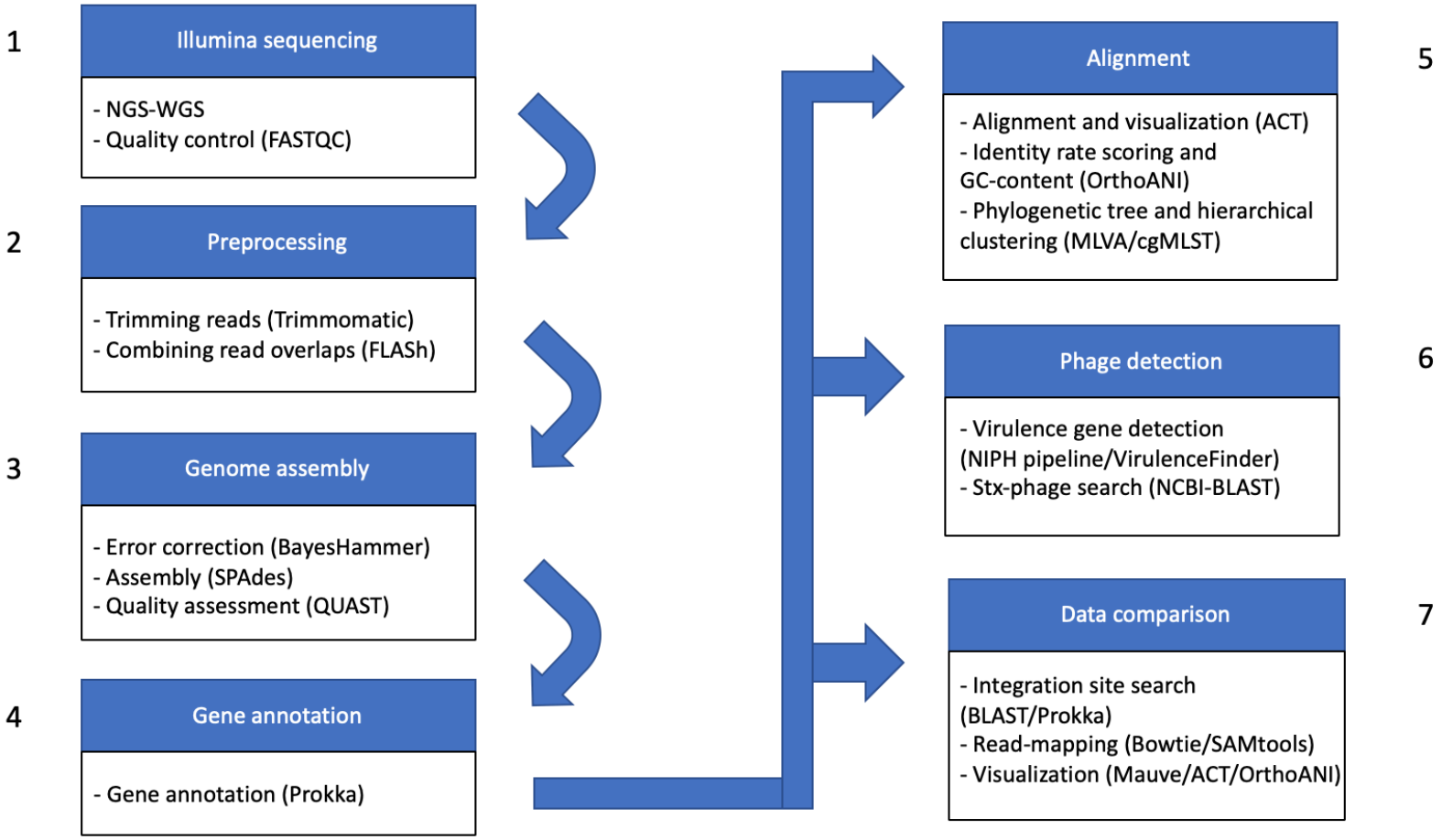

Figure 2.1: Overview of methods. Seven parts explaining the procedure of thesis. First four parts executed consecutively. 5, 6 and 7 completed more or less simultaneously.

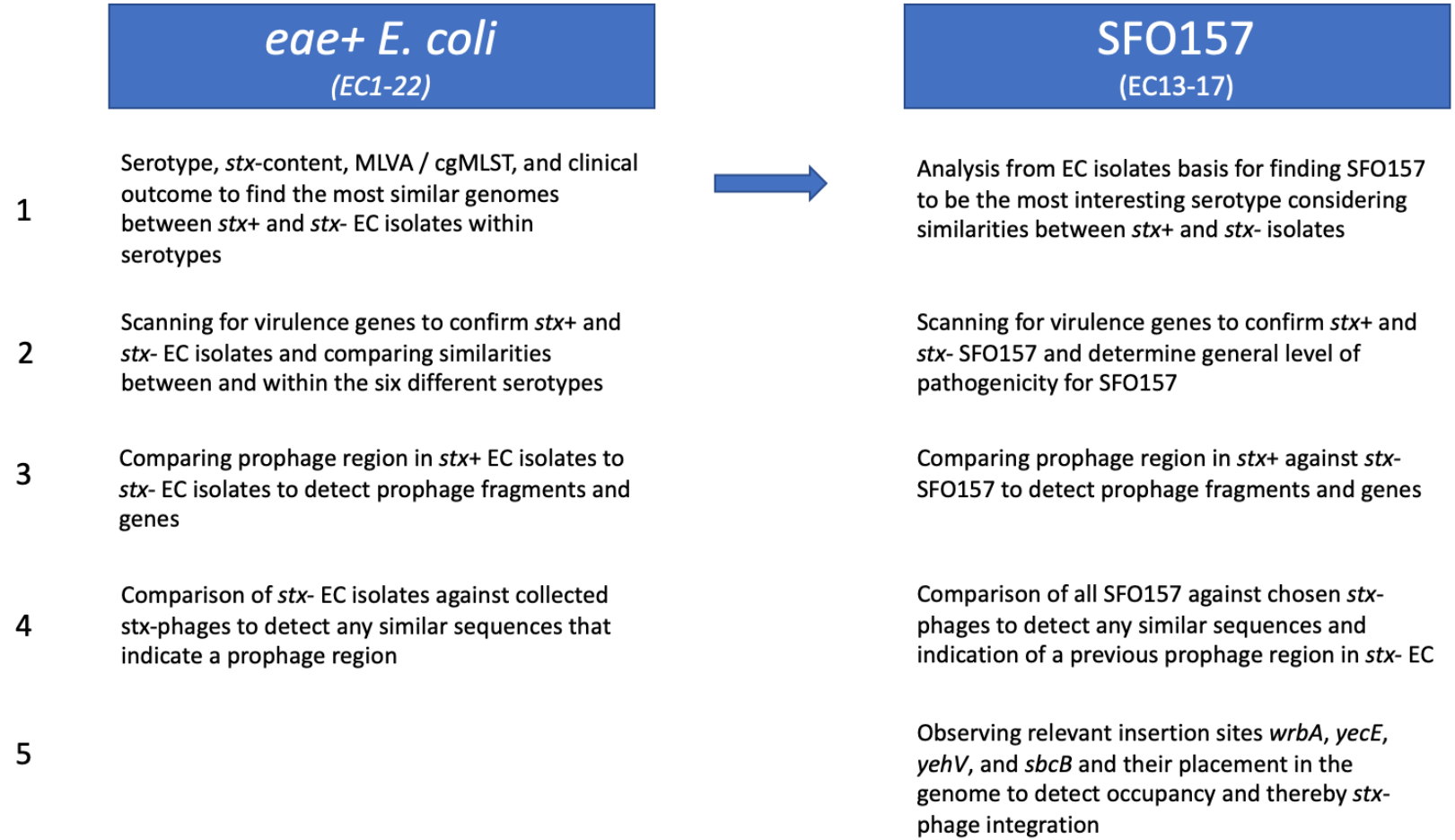

Figure 2.2: Comparison of the analysis process of *eae+ E. coli* and SFO157. The chart explains why certain comparisons and analyses were executed.

2.3.2 Preprocessing

After FASTQC application, the first step in processing the raw data was trimming the reads in order to remove adapter sequences and low-quality read ends. Removing unwanted parts of the reads is beneficial for downstream analysis like genome assembly. The command line tool Trimmomatic (version 0.36) was used to perform this task. Trimmomatic has been developed for processing Illumina NGS data, and is able to correctly handle paired-end data (Ussery *et al*., 2009, Bolger *et al.*, 2014). Trimmomatic is a read trimming tool and filtering for Illumina NGS data before assembly, and enters the palindrome mode to remove adapter sequences, filters low quality reads below a certain threshold, and drops reads below a certain length of bases (Bolger *et al.*, 2014). In palindrome mode forward and reverse reads are globally aligned to score the overlapping region. The program combines length threshold, error rate and coverage factor to trim at a peak of a combined score (Bolger *et al.*, 2014). Trimmomatic has several other commandoes to perform sliding window trimming, cut bases off start or end of a read, cut read to a specific length, and show quality scores in phred-33 or phred-64. This kind of preprocessing is important as the N50 contig size increases for *de novo* assembly; the contigs are larger (Bolger *et al*., 2014).

Using Trimmomatic, the adapter sequences (Nextera) were removed using the *ILLUMINACLIP* command, while providing a FASTA file containing the DNA sequences of the used adapters. The maximum amount of mismatches per seed was set to 2, simple clip threshold to 10 and palindrome clip threshold to 30. After adapter removal the leading and trailing bases with a quality score below 3 were trimmed off, and finally, the reads were scanned with a sliding window of 3 bases wide and trimmed if the average base quality dropped below 15. The remaining reads smaller than 36 bases were also removed from the dataset. The output gave four FASTQ files per isolate, containing the trimmed forward and reverse reads for the paired and unpaired set. The two paired FASTQ files were used for further analysis.

2.3.3 Genome assembly

After trimming, the paired reads were combined where the reads overlap. The Fast Length Adjustment of Short reads (FLASh, version 1.2.11), is an accurate tool that merges paired-end reads form fragments that are shorter than twice the length of reads (Magoc & Salzberg, 2011). FLASh found the overlapping reads in the two corresponding paired end reads and merged them into one single long read. The average read length that must be identical in both reads to overlap was set to 200 bp. The merged read is written to an unpaired FASTQ file and the remaining reads are written to standard paired end FASTQ files. The extended length of reads has a significant positive impact on improvement of genome assemblies. FLASh was used to combine the reads into sets of longer reads before error correction was performed. Error correction is often used to improve the quality of the read data by correction of nucleotides that are expected to be sequencing errors (Magoc & Salzberg, 2011). There are different kinds of error correction methods available. The used error corrector was BayesHammer through SPAdes. BayesHammer corrects the data through k-mer counting, as described under.

When assembling a genome, usually the reads are merged into contigs and then scaffolds by comparing to a reference genome. *De novo* assembly does not use a reference but assembles the genome by k-mer counting. All read data have been assembled using St. Petersburg genome assembler (SPAdes, version 3.11.1), which runs BayesHammer by default. SPAdes is a Eulerian de Bruijin graph assemblers designed for single-cell sequencing (SCS), which uses paired de Bruijin graph and is kind of “double-layered” (Bankevich *et al.*, 2012). The k-mers from short reads build the inner graph and assemble contigs, then the “paired k-mers” (k-bimers) with large insert size build the outer graph for repeat resolving or scaffolding (Sohn *et al*., 2018). First stage is graph simplification through multisized de Bruijin graphs to remove bubbles or bulges, chimeric reads and make distance histograms from bireads (Bankevich *et al.*, 2012). The choice of k affects the graph construction, where small values collapse reads together making the graph tangled and large values ignore read overlaps and make the graph fragmented (Sohn *et al*., 2018). Therefore, it is important to vary k through multisized de Bruijin graphs to fit low-coverage and high-coverage regions. Then the k-mer distances are estimated by joint analysis of the distance histograms and the paths in the assembly graph. Afterwards the paired assembly graph and read-mapping of

contigs are constructed (Bankevich *et al*., 2012). BayesHAMMER is a tool for error correction in single-cell sequencing (SCS) and other methods, based on Multiple Diceplacement Amplification (MDA) technology employed through SPAdes (Nikolenko *et al*., 2013). Programs like QUAKE and the HAMMER algorithm uses a clustering technique to select an error-free central k-mer in each connected component of the Hamming graph. The Hamming graph describes the distance relation of two vectors with the number of coordinates where they differ (Nikolenko *et al*., 2013). This method might be oversimplified because it's more reasonable to assume two central k-mers rather than one, and the programs produce poor results on non-uniform coverage. Thus, the BayesHAMMER does not rely on uniform coverage, subclusters beyond the HAMMER algorithm by read quality, and introduces Bayesian (BIC) penalties for extra parameters leading to overfitted models (Nikolenko *et al*., 2013). The algorithms are heavily parallelized and significantly sped up, making BayesHAMMER an easier error corrector to use compared to other tools that might need more input to run (Nikolenko *et al.,* 2013).

Although, SPAdes is designed to process single-cell data, it performs well on standard multi-cell bacterial datasets (Bankevich *et al*., 2012). Mismatch and short indels were prevented using the careful option. A folder containing the k-mer data and the contigs data was made for each dataset. The contigs.fasta files were used for further analysis. Quality Assesment Tool (QUAST, version 4.6.1) was used on all of the contigs data. QUAST is a quality assessment tool for evaluating and comparing genome assemblies with or without a reference genome (Gurevich *et al.*, 2013). When a new species does not have a finished reference genome, being able to assess the quality of an assembly without a reference helps a lot. QUAST evaluates information about the contigs such as unalignment, ambiguous mapping, misassembly or correctness, though it does not distinguish between SNPs and single-nucleotide errors (Gurevich *et al.*, 2013). QUAST is also highly parallelized to make it faster. Quast uses N50-metrics and builds convenient plots like coverage, sequence depth and GC-content. After the quality control, the genes can be annotated.

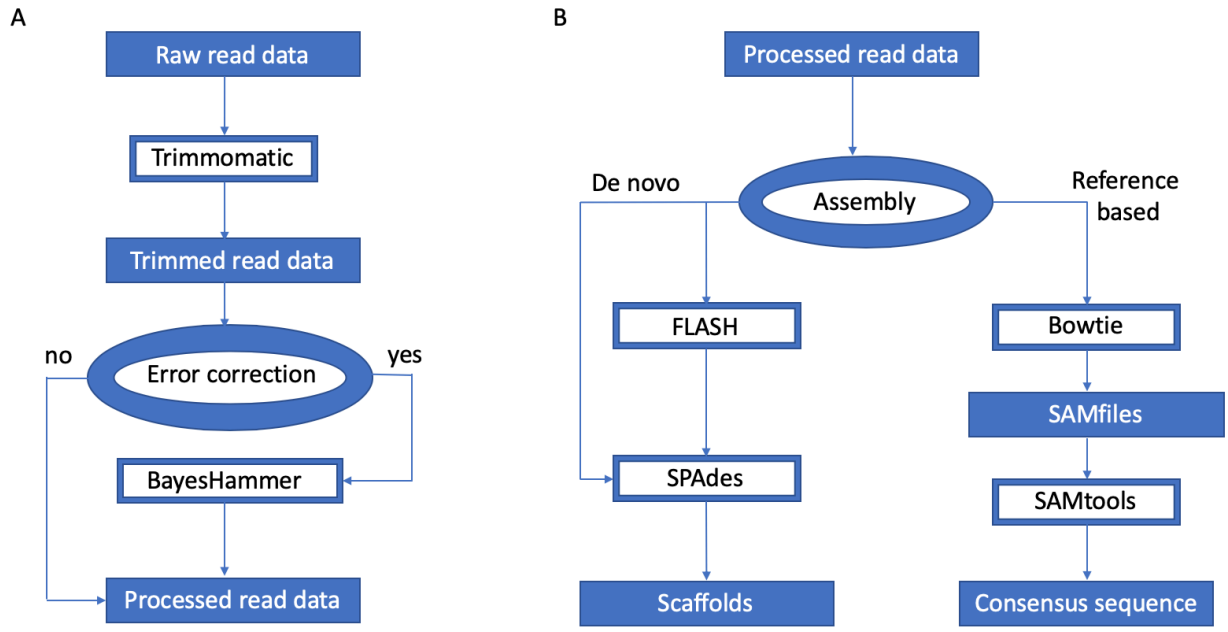

Figure 2.3: Flow chart of method. Going from raw read data to scaffolds and consensus sequences through preprocessing, error correction and genome assembly. Both *de novo* method and reference-based sequencing shown.

2.3.4 Gene annotation

Gene annotation is applied to identify genes in the assembled genome and mark them with the acquired information. Hereby, it is easier to find the position of the specific genes to look into. After finding the *stx*-gene in the genomes, multiple alignment tools are used to compare the different *E. coli* strains to each other and also to different *E. coli*-related phages and *stx*-phages.

Gene annotation is performed on the contigs data by Prokka. Prokka is designed for rapid automatic annotation of prokaryotic genomes by searching through databases like Basic Local Alignment Search Tool (NCBI-BLAST, version 2.7.1+) and applying Hidden Markov Models (HMM), and finding similar sequences to the contigs file (Ussery *et al.*, 2009, Sohn & Nam, 2018). The program matches the sequences and marks them with the appropriate gene names. Prokka outputs Genbank files, among other things, that have genomic information. Prokka was run with default settings. After Prokka was run, the *stx*-gene was searched for in all datasets. The *E. coli* datasets were run through an NIPH in-house pipeline checking the serotyping and finding virulence genes including *stx*-genes. The pipeline used

NCBI-BLAST to run an overall check of the *E. coli* isolates against the Center for Genomic Epidemology (CGE) VirulenceFinder database at Technical University of Denmark (DTU) to search for the serotype and virulence factors. The database contains known *stx* and virulence genes.

2.3.5 Alignment

Artemis Comparison Tool (ACT, version 16.0.0) visualizes and compares sequences and Prokka files (.ffn), Genbank files (.gb) as well as fasta files are compatible with the program. ACT was used to align the different annotated genomes to each other and look at the placement of the *stx2* on each set of data. Identity rates between isolate genomes were noted from Orthologous Average Nucleotide Identity (OrthoANI) tool for all *E. coli*. OrthoANI determines similarities obtained by mimicking DNA-DNA hybridization (Lee *et al.,* 2016). It calculates nucleotide identities of orthologous fragment pairs of two genome sequences, by using BLASTn (Lee *et al.,* 2016).

Core genome MLST (cgMLST) was performed in Ridom SeqSphere+ (version 5.1.0, Ridom GmbH, Germany) (Maiden *et al*., 1998). Briefly, raw sequence reads were trimmed until an average base quality of 30 was reached in a window of 20 bases, and *de novo* assembly was performed using Velvet (version 1.1.04) with default settings. The SeqSphere+ integrated "*E. coli*/*Shigella* cgMLST scheme v1" from Enterobase was used (Enterobase, 2018). With allele calling procedure with minimum accepted BLAST identity of 80%, no BLASTp search, frame-shift detection turned on and independent SeqSphere+ allele numbering nomenclature, the allelic profiles of the isolates were visualized as a minimum spanning tree (MST) and a neighbor-joining (NJ) phylogenetic tree using the parameter "pairwise ignoring missing values". Phylogenetic trees or clustering from cgMLST-profiles and the MLVA-profiles were utilized to look at relatedness.

Data	Illumina sequencing	Preprocessing	Genome assembly	Gene annotation	Alignment	Phage detection	Data comparison
- SFO157 included in EC isolates - Prophages are 2-4 contigs from EC isolates - *stx*-phage attained from NCBI-BLAST	- NGS-WGS - Quality control (FASTQC)	- Trimming reads (Trimmomatic) - Combining read overlaps (FLASh)	- Error correction (BayesHammer) - Assembly (SPAdes) - Quality assessment (QUAST)	- Gene annotation (Prokka)	- Alignment and visualization (ACT) - Identity rate scoring and GC-content (OrthoANI) - Phylogenetic tree and hierarchical clustering (MLVA/cgMLST)	- Virulence gene detection (NIPH pipeline /VirulenceFinder) - *stx*-phage search (NCBI-BLAST)	- Integration site search (BLAST/Prokka) - Read-mapping (Bowtie/SAMtools) - Visualization (Mauve/ACT/ OrthoANI)
EC isolates	✓	✓	✓	✓	✓	Only VG detection	Only visualization in OrthoANI
SFO157	✓	✓	✓	✓	✓	Only VG detection	✓
Prophage	✓	✓	✓	✓	✗	Basis for *stx*-phage search	Only visualization in OrthoANI to SFO157 and ACT to EC isolates
stx-phage	✗	✗	✗	✗	✗	Attained from *stx*-phage search	Only visualization in OrthoANI and Mauve

Figure 2.4: Work-flow progress chart. Visualization of *E. coli*, SFO157, prophage and *stx*-phage data processing at any given part of the analyzing process.

2.3.6 Phage detection

The shiga toxin containing (*stx+*) contigs from the *E. coli* with the *stx2a*-genes were saved as prophages. These prophage files consist of additional one to three contigs surrounding *stx* with a total of 50-60 kbp. 50-60 kbp sequence length was chosen because the average length of characterized *stx*-phages was within this range. The prophage data was run in BLAST, and phage setting was chosen to "tailed phages". Different relevant *stx2*-phages were hit and collected as whole genome FASTA files and Genbank files (.gb) from NCBI's database for eventual viewing. The prophage data was also run in BLAST and OrthoANI to detect any similarities within and between the other contigs.

The FASTA files of the *stx*-phages were run through OrthoANI to observe alignment against the *E. coli* isolates, the prophage files, and each of the other *stx*-phage files. OrthoANI compares sequences locally and gives a general identity rating score and GC content in percentages. *stx*-converting phages from NCBI-BLAST were aligned with the *E. coli* isolates in Artemis to visualize the alignment. The goal was to observe the prophage region of approximately 50 kbp and detect any phage fragments. The scorings were observed in OrthoANI. The phages were then compared to the *stx*-negative isolates specifically. For SFO157 there were run many cross checking with BLAST, OrthoANI, ACT and Mauve (version 2) to match specific phages to isolates. Mauve is a visualization tool like ACT and computes with .ffn and .fasta. It aligns specifically selected genes and contigs and easily compares certain areas of the genome, like the prophage.

During phage detection an in-house NIPH pipeline for serotyping and finding virulence genes was run and several virulence genes were found. The genes of main interest were *stx2a*, *eae*, *ehxA*, and *cdtB*. *nleB* was identified in all the isolates. *stx2a*, *eae* and *ehxA* were found through NGS methods at first, *cdtB* and *nleB* were only identified through the pipeline. In the Prokka files the different integration seats for phages in general were searched for, like *intA* and *intS*. Specific insertion seats were run on BLAST and compared to the serogroup SFO157. *yecE*, *wrbA* and *sbcB* were collected from "*Escherichia coli* str. K-12 substr. MG1655, complete genome" at Genbank, and *yehV* from "*Shigella dysenteriae* Sd197". They were compared to the SFO157 isolates with special interest of the *stx*-negative isolates within the

serotype. Detecting occupancy of the insertion sites within *stx*-positive *E. coli*, lead to some information about where the prophage could have integrated.

2.3.7 Data comparison

Bowtie2 (version 2.3.4.1) and SAMtools (version 1.8), together with bcftools (version 1.8) and seqtk, were reference based tools used to map certain *stx*-phages found in the BLAST search to all of the *E. coli* datasets to compare similarities and matching sequences. These reference-based mappings were visualized in Mauve. Comparison of location and preservation of *stx*-phage was tried detected in the STEC and then the non-carrying *E. coli*.

The prophage data of SFO157 was run in BLAST to detect matches with other contigs, phages and insertion sites specifically for the serotype. The genes around the prophage area, 50 kbp surrounding the *stx2a*-gene, were inspected with Prokka annotation for any relevant and known *stx*-phage genes. If any were close enough, they were checked with Mauve how close by nucleotide base placement with relevance to *stx2a*.

In total, during alignment, phage detection and data comparison, the *E. coli* isolates were aligned with ACT, OrthoANI, phylogenetic tree and hierarchical clustering, phages were collected through BLAST and tried detected through Mauve, ACT, OrthoANI and Prokka, and *E. coli* were compared to each other as isolates and prophage data, *stx*-phages and insertion sites to find matches to reveal STEC and STEC – lost shiga toxin (STEC-LST).

3 RESULTS

3.1 Comparison of *E. coli* strains with identical serotype and MLVA-type

This study attempts to observe if *eae*-positive *E. coli*, with identical serotypes and MLVA-profiles, are *stx2a*-positive STEC that lost the *stx*-prophage and have a vacant insertion site, or whether phage fragments are available within the genome and only the *stx*-gene is missing. STEC is *stx*-positive (*stx*+) *E. coli*, and *stx*-negative (*stx*-) *E. coli* means *eae*-positive *E. coli* without *stx*. The sequence data of 22 different *E. coli* isolates were analyzed to compare STEC and *stx*-negative *E. coli*. The goal was to find characteristics substantiating the classification of the *eae*-positive *E. coli* as STEC-LST.

General hypotheses were that the bacteriophage is lost from the genome, or the *stx*-gene goes missing because of recombination and mutations during infection, digestion or isolation of strains. *In silico* methods were used to preprocess, assemble, and compare the *E. coli* to each other and to *stx*-phages. Six serotypes were analyzed, O145:H25, O26:H11, O103:H25, SFO157:H7, O145:H28, and O177:H25. Table 3.1 shows an overview of the *E. coli* strains, their corresponding serotypes, virulence genes, MLVA-profile, year of isolation, source of the outbreak, and *stx*-gene and -nucleotide placement. The main focus of this thesis is on serotype sorbitol-fermenting O157:H7 (SFO157); EC13 to EC17, and the *stx*-variant *stx2a*.

Traditional extraction methods in laboratories showed half of the *E. coli* isolates containing *stx2a*. In addition, did gene annotation show the placements of the *stx2a*-genes on every STEC strain with genetic placement (gene no.) and nucleic placement (locus). The *stx2a* subtyping was also confirmed with NCBI-BLAST search. All *E. Coli* strains carried the associated pathogenic *eae*-gene, and all isolates except for EC6, EC11 and EC12 contain *ehxA*. All SFO157 contained *nleB* and only EC13 and EC14 contained *cdtb*, which were some of the virulence genes observed with the in-house NIPH pipeline.

Table 3.1: Overview of *E. coli* data. Extended table from Table 2.1 (chapter 2, materials and methods). Additional info is *stx* content, and for the *stx*-positive EC – gene and nucleotide placement.

Isolate	Serotype	Virulence gene	MLVA profile	Year	Source	*stx*	*Stx2a*-gene placement	*stx2a* nucleotide placement
EC1	O145:H25	*stx2a, eae, ehxA*	5-3-0-8-4-1-1-16-9-15	2009	Human	+	5102	4 531 866
EC2	O145:H25	*stx2a, eae, ehxA*	5-3-0-8-4-1-1-16-9-15	2009	Human	+	481	3 942 051
EC3	O145:H25	*eae, ehxA*	5-3-0-8-4-1-1-16-9-**16**	2009	Human	-		
EC4	O26:H11	*stx2a, eae, ehxA*	6-0-0-8-3-2-1-6-25-11	2014	Human	+	4737	4 304 366
EC5	O26:H11	*eae, ehxA*	6-0-0-8-3-2-1-6-**27**-11	2011	Human	-		
EC6	O26:H11	*eae*	6-0-0-8-3-2-1-6-**26**-11	2009	Ovine	-		
EC7	O26:H11	*stx2a, eae, ehxA*	6-0-0-8-3-4-1-6-15-15	2015	Human	+	4029	3 646 730
EC8	O26:H11	*eae, ehxA*	6-0-0-8-3-4-1-6-15-**14**	2010	Human	-		
EC9	O103:H25	*stx2a, eae, ehxA*	7-3-0-5-0-7-1-16-9-12	2006	Human	+	1571	1 442 612
EC10	O103:H25	*eae, ehxA*	7-3-0-5-0-7-1-16-9-12	2007	Ovine	-		
EC11	O103:H25	*eae*	7-3-0-5-0-7-1-16-9-12	2006	Mutton	-		
EC12	O103:H25	*stx2a, eae*	7-3-0-5-0-7-1-16-9-12	2006	Human	+	4632	4 253 874
EC13	SFO157	*stx2a, eae, ehxA*	11-0-23-0-6-4-0	2009	Human	+	1881	1 666 962
EC14	SFO157	*eae, ehxA*	11-0-23-0-6-4-0	2009	Human	-		
EC15	SFO157	*eae, ehxA*	4-23-3-6-3-5-3	2008	Human	-		
EC16	SFO157	*stx2a, eae, ehxA*	4-23-3-6-3-5-3	2008	Human	+	4581	4 299 956
EC17	SFO157	*eae, ehxA*	4-23-3-6-3-5-3	2008	Human	-		
EC18	O145:H28	*stx2a, eae, ehxA*	7-3-0-8-3-2-1-35-0-0	2013	Human	+	524	454 502
EC19	O145:H28	*eae, ehxA*	7-3-0-8-3-2-1-35-0-0	2013	Human	-		
EC20	O177:H25	*stx2a, eae, ehxA*	5-3-0-15-4-1-1-16-11-15	2013	Human	+	5134	4 541 779
EC21	O177:H25	*stx2a, eae, ehxA*	5-3-0-15-4-1-1-16-11-**13**	2013	Human	+	4441	4 102 151
EC22	O177:H25	*eae, ehxA*	5-3-0-15-4-1-1-16-11-**10**	2013	Human	-		

*SFO157:H7 – EC16 and EC17 are from the same incident.

Observing the MLVA-profiles in Table 3.1, differences within serotypes are found at one specific locus for "pairs" of isolates. In serotype O145:H25, EC1 and EC2 have identical MLVA-profiles but *stx*- EC3 differs with one repeat on the last VNTR locus. In O26:H11, the situation is similar with one repeat at the VNTR locus for EC8 compared to EC7. The repeat is 25 in EC4 but 27 and 26 in *stx*- EC5 and EC6, respectively. Within O103:H25, all the profiles are identical. The same is true for EC13 and EC14, and EC15-17 within SFO157. The two O145:H28 isolates also have identical profiles. And, O177:H25 have the most differing MLVA-profiles within the serotype.

Table 3.2: Average nucleotide identity rates in OrthoANI (%). Similarities visualized by colors between *stx+* and *stx*- isolates. STEC are presented as rows, and *eae*-positive *E. coli* are presented in columns.

STEC → / ***eae+* EC**	**1**	**2**	**4**	**7**	**9**	**12**	**13**	**16**	**18**	**20**	**21**
3		x				x				x	x
5		x	x	x	x						x
6		x	x	x	x	x					x
8		x	x	x	x						
10		x			x	x					x
11		x			x	x					x
14						x	x	x			
15						x	x	x			
17						x	x	x			
19						x	x		x		
22	x	x				x				x	x

x 100.0% x 99.99% x 99.90% - 99.98% x 99.00% - 99.89% blank = 94.99% - 98.99%

Table 3.2 shows average nucleotide identity scorings in OrthoANI between the *stx- E. coli* and *stx+ E coli*. Isolates from the same serotypes scored well in local alignment, although there were some identity rate variations in global alignments when applying BLAST. Different serotypes had higher percentage of identity where the same flagellar serotype (H25) was observed; isolates EC1-EC3, EC9-EC12 and EC20-EC22 had ≥99% similarities. EC2 seemed to have very high identity with EC22, and EC1 and EC3 with EC20-EC22. Observing the zoomed visualization of the Mauve comparisons, isolates from the same serotypes aligned well, excluding *stx2a*. The MST in Figure 3.1 and neighbor-joining (NJ) tree in Figure 3.2 confirm the serotypes with flagellar gene H25 (green and brown) to cluster very well together and show relatedness in H25 and H11 (beige and yellow). H28 (white and neon yellow) and H7 (red) were clustered further apart from the other flagellar antigens. All of the somatic (O) types showed close relatedness within the same serogroup, except for O145 that was split in two and clustered further apart regarding the flagellar gene – O145:H25 (brown) and O145:H28 (white and neon yellow). The color corresponding the specific isolate and serotype for Figure 3.1 and 3.2 is found in Table 3.3.

Table 3.3: Serotyping for *E. coli*. The corresponding colors for EC serotypes for Figure 3.1 and 3.2 are shown in the table.

EC	Color code	Serotype
1		O145:H25
2		O145:H25
3		O145:H25
4		O26:H11
5		O26:H11
6		O26:H11
7		O26:H11
8		O26:H11
9		O103:H25
10		O103:H25
11		O103:H25
12		O103:H25
13		SFO157
14		SFO157
15		SFO157
16		SFO157
17		SFO157
18		O145:H28
19		O145:H28
20		O177:H25
21		O177:H25
22		O177:H25

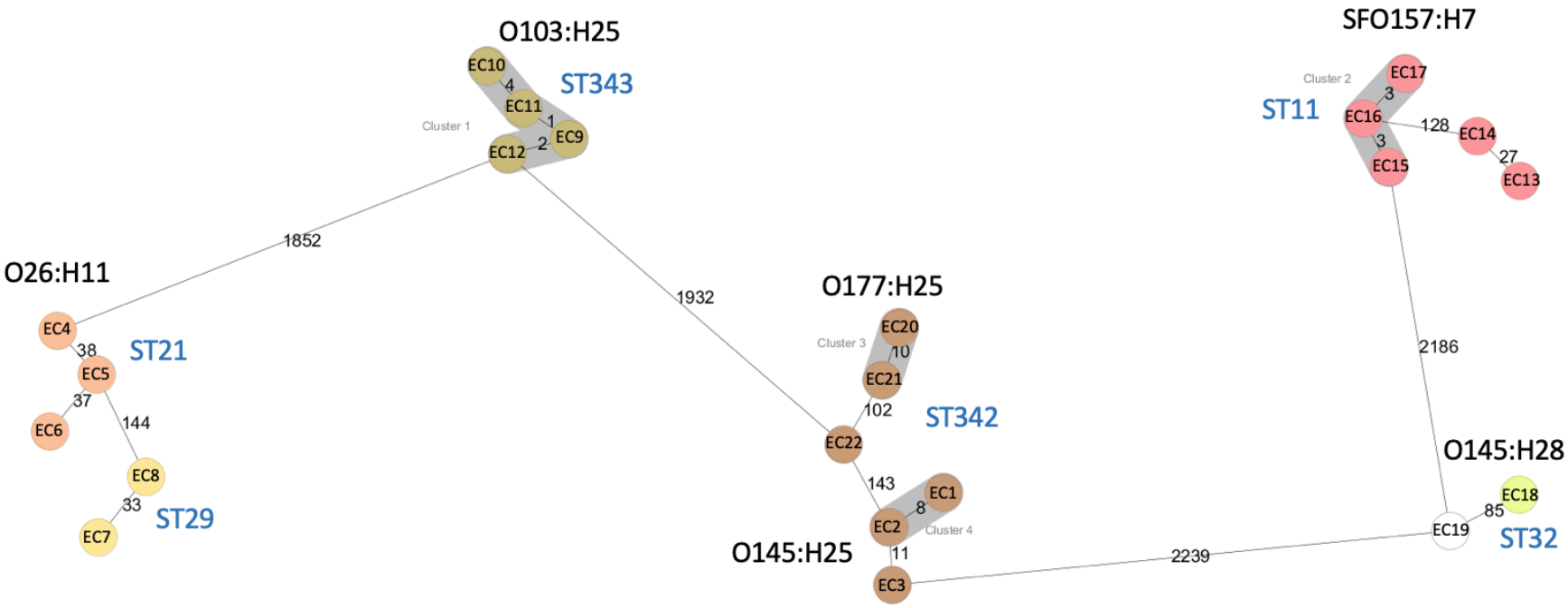

Figure 3.1: Phylogenetic tree by *E. coli* cgMLST scheme v1.0 run in Ridom SeqSphere+. Minimum spanning tree (MST) for 22 samples based on 2528 genes (pairwise ignoring missing values) (*E. coli* Warwick (7 genes), *E. coli* Pasteur (8 genes) and *E. coli* cgMLST (2513 genes)).

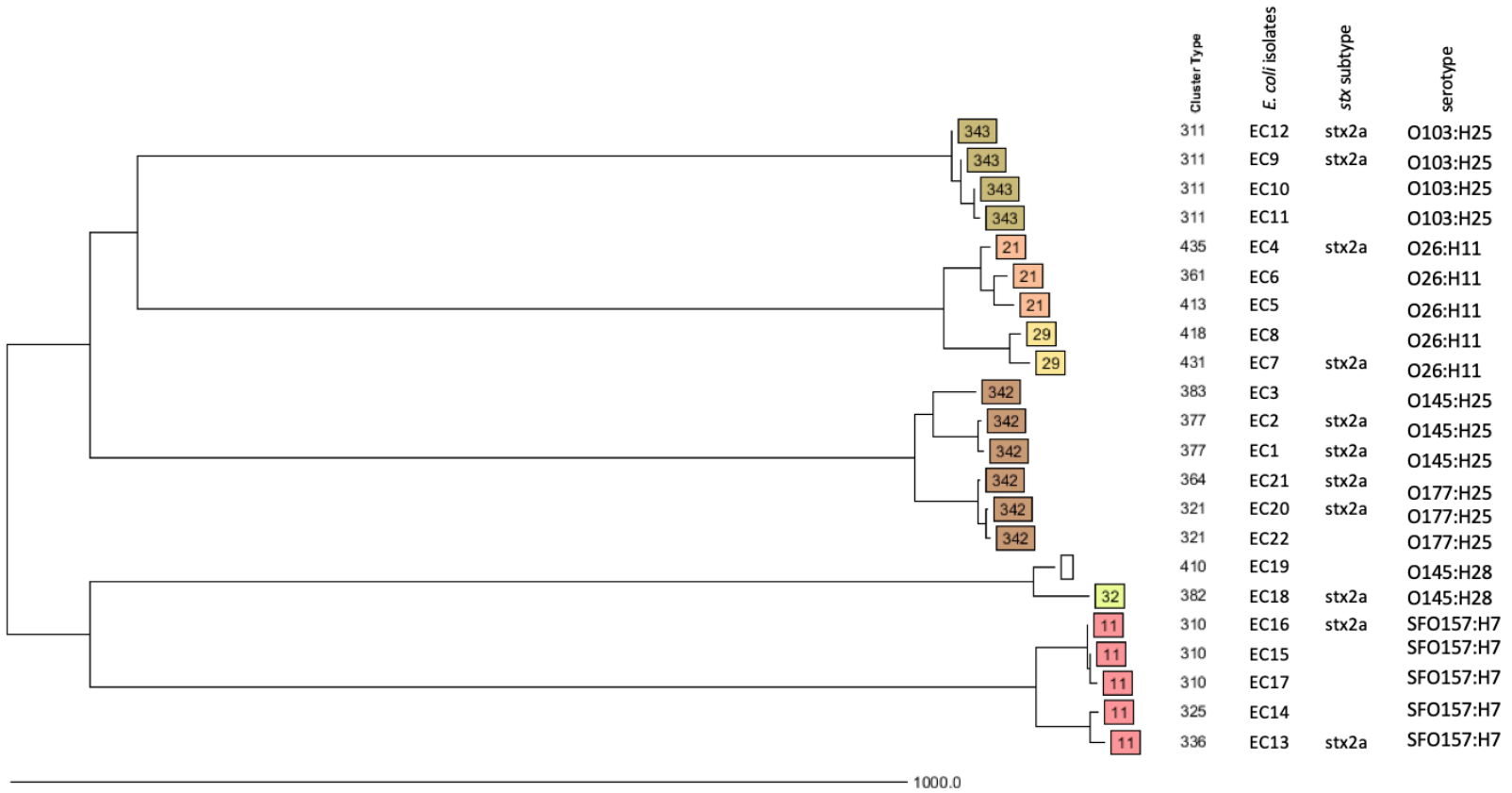

Figure 3.2: Neighbor-joining (NJ) tree by *E. coli* cgMLST scheme v1.0 run in Ridom SeqSphere+. Produced by 22 samples based on 2528 genes, same as for Figure 3.1.

3.2 Similarities within SFO157

In 2008, two cases within the same family were infected with STEC SFO157. One case had a severe clinical outcome whereas the other case was asymptomatic. The case with severe clinical outcome (EC15) carried an *eae*-positive *E. coli* with no evidence of *stx*. From the case without symptoms, two isolates were available – one containing *stx2a* (EC16) and the other isolate missing the *stx*-gene (EC17). All three isolates showed identical MLVA-profile, but with cgMLST there were some allelic differences. There were three allelic differences between the isolates EC16 and EC17 (isolated from the same case), and three allelic differences between the isolates EC16 and EC15 from the two different cases. During 2009, two other occasions happened within a close geographical area and a short time period apart. Both isolates, EC13 and EC14, shared identical MLVA-profiles indicating the same bacteria, but the cgMLST showed 27 differing alleles. The red circles in the MST in Figure 3.1 represent SFO157. The allelic differences from the cgMLST can be observed as the vectors (lines) connecting the red circles. The relatedness can also be viewed as red squares in the NJ-tree in Figure 3.2.

Table 3.4: Identity rate between SFO157 in OrthoANI. All SFO157 isolates compared with colors indicating percentage level of match. EC13 and EC16 are STEC. EC14, EC15 and EC17 are *eae+ E.coli*.

Isolates	13	14	15	16	17
13	x	x	x	x	x
14	x	x	x	x	x
15	x	x	x	x	x
16	x	x	x	x	x
17	x	x	x	x	x

x 100.0% x 99.90% - 99.99% x 99.87% - 99.89%

The SFO157 (EC13 to EC17) comparisons are presented in Table 3.4 that shows percentages of matches. Observing the table, EC13 has higher identity to EC14 and EC15 than EC16 and EC17. EC14 is closest related to EC13, whereas EC15 is similar to EC16. EC16 and EC17 are closest related to each other and EC15. Summing up, the table shows a higher identity rate between EC13 and EC14, and between EC15 to EC17. The isolates within the same grouping has identical year of outbreak and MLVA-profiles, which can be viewed in Table 3.1.

3.3 Prophage data

The prophages of the different *E. coli* strains were estimated to be around 50-60 kbp, calculated roughly from averaging the collected *stx2a*-phages. The prophage data consists of two to four contigs surrounding the *stx*-gene and contain most of the supposed *stx*-bacteriophage, if there were any, – so about 50 kbp depending on the genomic region it was collected from. The prophage data was also collected from EC13 and EC16 in SFO157.

Even though the genomes of the different *E. coli* strains were highly identical, the prophages might vary. The prophage contained an insertion site and integration seat to attach to the genome. It had regulatory genes, the Q anti-terminator with q_{933} and q_{21}, the $p_{R'}$ promotor and $t_{R'}$ terminator, *stx* with own promotor and terminator, the lysis genes, and structural genes for the phage – like capsid, tail and lom encoding genes. Figure 3 illustrates the assumed prophage and the expected genomic content: A. shows the elements of the whole prophage genome and B. visualizes the replication, Q anti-terminator, the transcription of *stx* and the lysis genes.

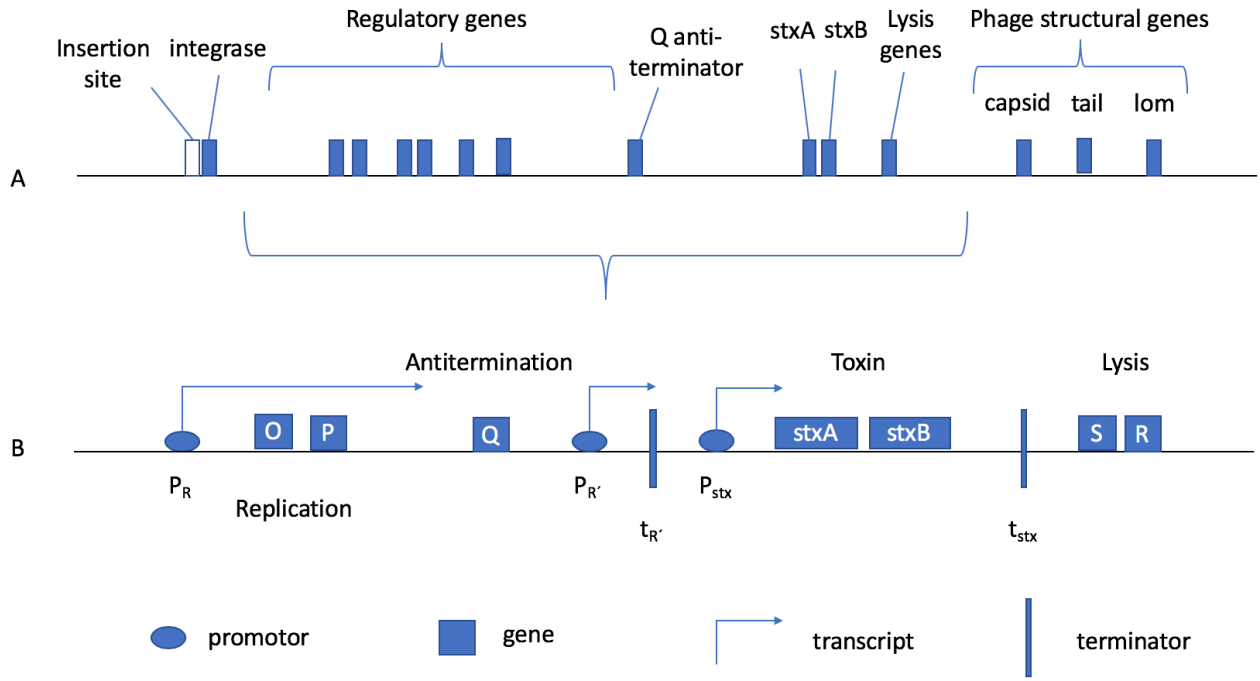

Figure 3.3: Visualization of the presumed prophage. A. shows prophage as a whole with insertion site, regulatory genes, Q-gene, toxins, and phage structural genes. B. shows, in detail, the specific area between the insertion sites and the phage structural genes.

wrbA, *yecE*, *yehV*, and *sbcB* are integration sites in the SFO157-genome where the *stx2a*-phages integrate. q_{933} and q_{21} is usually found within SFO157. Prokka annotation gave some information in the assumed prophage region, along with "hypothetical proteins". Gene names that Prokka annotated in the prophage area were related to replication, recombination and transcription, flagella and pilin, outermembrane protein, tRNA, anti-adaptor, ribose, modification methylase, endonucleases, RNA polymerase, transporter, toxins (*ccdB* and *ccdA* for EC16), sporulating inhibitors, virulence regulator, prophage tail and sheath, and shiga toxin. BLAST search supported these findings.

Observations in Mauve and ACT showed DNA methylase, tRNA, outer membrane lipoprotein (lom), and prophage tail and tail sheath genes scattered throughout the genome. The genomic loci of these genes did not cohere well with the prophage area. Figure 3.4 illustrates an *E. coli* O157:H7 *stx2*-phage and its non-motile variants (Haugum *et al.*, 2012).

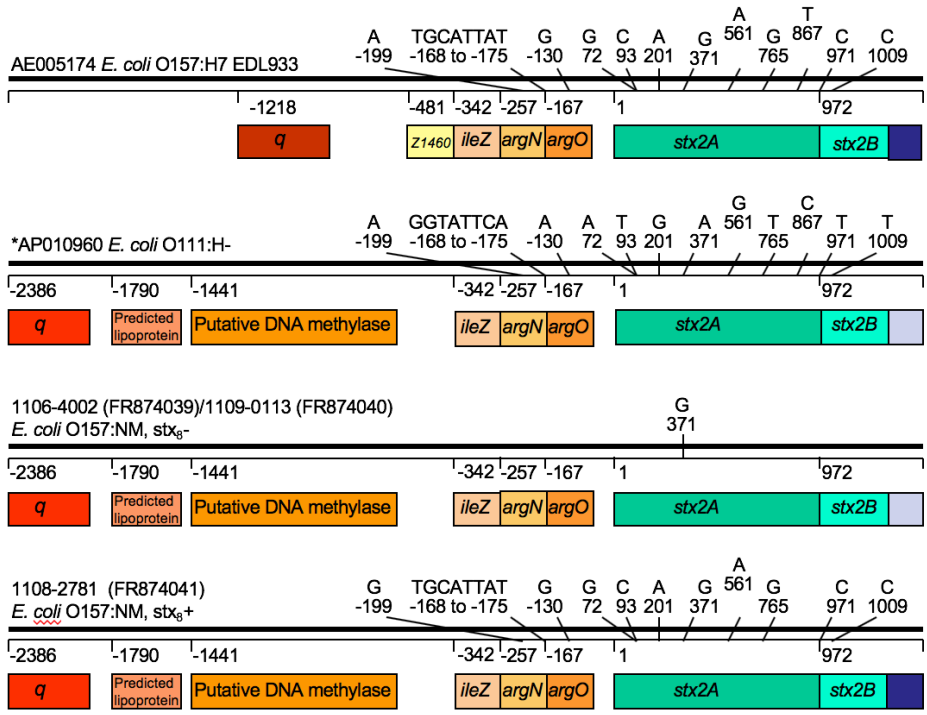

Figure 2. Comparison of the structure of the *stx* genes and upstream genes in the *E. coli* strains O157:H7 EDL933 (AE005174), O111:H- 11128 (AP010960), O157:NM strains 1106-4002 (FR874039), 1109-0113 (FR874040), and 1108-2781 (FR874041). Additionally, approximately 500 bp downstream of the *stx* genes are shown. The **numbers indicates** base positions or gene start positions (translational start for *stx2A*; +1).
*Reversed complement of AP010960.

Figure 3.4: Visualization of O157 prophage. Illustration from Haugum *et al.* (2012). O157-specific example of the relevant prophages and their build-up.

The prophages of EC13 (pro-13) and EC16 (pro-16) were also compared to all of the SFO157 contigs files in BLAST. Pro-13 (node 91) showed 100% sequence identity to a specific node in each EC contig with length 13190, but pro-16 (node 26) showed some differences. Pro-16 was most similar to EC15 with 94% sequence identity to one specific node (19). EC13, EC14 and EC17 had parts of the pro-16 scattered around the genome with maximum 20% identity rate within one sequence.

In Figure 3.5, Mauve shows alignments of the isolates within SFO157. The Prokka annotations of EC13 to EC17 are aligned from top to bottom. The visualization showed some genes to be similar in several isolates close to the prophage. The *stx+* EC13 and EC16 aligned with some gaps and a total of 5000-6000 bp right around *stx2a* (outlined rectangle) and some genes downstream for EC13 and upstream for EC16 consecutively. This is visualized by

the colors pink, orange, purple and blue about 1667000 to 1672500 in EC13 and 4295000 to 4301000 in EC16. A 2000 bp sequence (purple) upstream for *stx2a* in EC13 at 1664500 to 1666500 matches with *stx*- EC15 at 4544000 to 4546000. In EC17 at 2920000 to 2925000 (yellow and blue), a 5000 bp sequence align relatively well with EC16 ca. 3000 bp downstream for *stx2a* at ca. 4303500 to 4307000 with a gap. The genetic positions of the matching genes are reversed and somewhat in order in the isolates.

QUAST analyses of background guanine-cytosine (GC) content for all data sets showed around 50.5% GC at 30000-32000 window frame. GC content might indicate variation in selection, mutational bias and biased recombination-associated DNA repair. In Mauve, observations on nucleotide base level show regions around the prophage susceptible to mutations and a little richer in AT, although not significant enough.

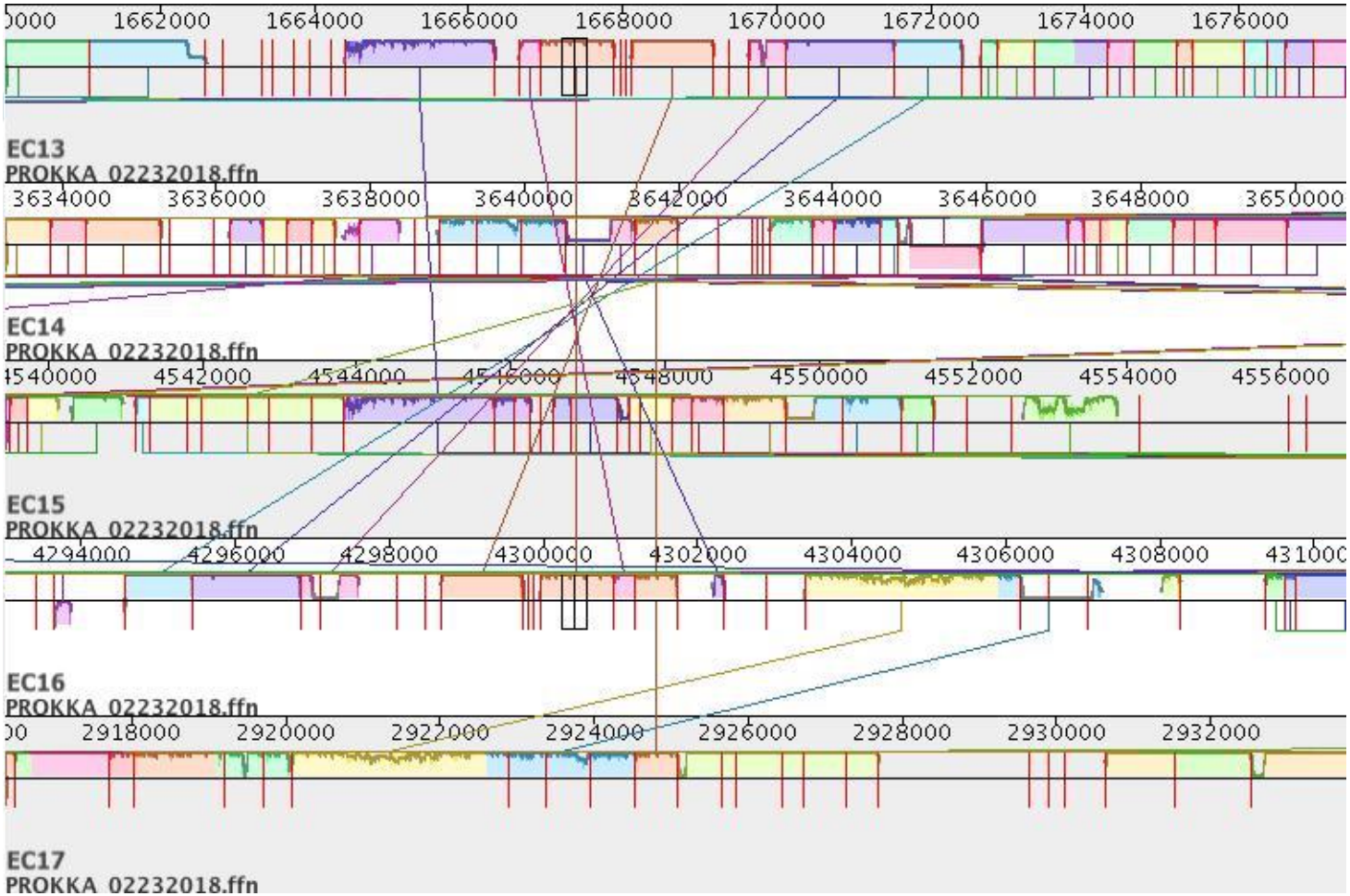

Figure 3.5: Mauve alignment of all SFO157 isolates. Isolates EC13 to EC17 is visualized top down in the suspected prophage area. Specific genes align with the other isolates and is observed by the crossing lines. The blocks of colors show differing genes. Stx2a is marked with a black outlined rectangle visual in EC13 and EC16.

3.4 Insertion sites

There are several insertion sites for *stx*-phages. *wrbA*, *yecE*, *yehV*, and *sbcB* were chosen for SFO157 since *stx2a*-phages insert at these sites. Table 3.5 shows percentage level of the insertion sites found in the *E. coli* strains, which can determine occupancy of integration seat and lead to phage detection. *stx2a* content and clinical outcome were also added to the table to determine if probable phage insertion coheres with severity of illness.

Table 3.5: SFO157 match with insertion sites. Percentage level BLAST results of specific insertion sites found in SFO157. Clinical outcome and *stx2a* profile added.

EC	13	14	15	16	17
stx2a	+	-	-	+	-
Clinical outcome	severe	unknown	severe	asymptomatic	asymptomatic
serogroup	SFO157				
wrbA	100	100	100	100	100
yecE	**81**	100	100	100	100
yehV	**93**	**93**	**93**	**93**	**93**
sbcB	100	100	100	100	**82**

All the SFO157 insertion sites were collected through NCBI-BLAST, since they were not found in the Prokka results. The details of the insertion sites are specified in chapter 2, materials and methods. Although none of the insertion sites for SFO157 are close enough to the prophage area to be proven as integrated seats, some interesting information was collected by observing occupancy of insertion sites. *yehV* shows occupancy of 7% in all SFO157, and *sbcB* is 82% conserved in EC17 and *yecE* is 81% conserved in EC13. The only insertion site found with Prokka of the relevant sites for STEC, was *prfC.* Although Prokka did not annotate the specific insertion sites for SFO157, two “prophage integrase” (*intA* and *intS*) were noted as possible insertion sites and can be viewed in Table 3.6.

Table 3.6: Gene placement of *intA* and *intS* in SFO157. The sites in the left columns of *intA* and *intS* indicates how many sites were found within each EC. The placements in the right columns are gene placement number from Prokka annotation.

	intA		***intS***	
EC	**Sites**	**Placement**	**Sites**	**Placement**
13	3	**1743**, 3037, 3350	5	469, **2016**, 3021, 3755, 4755
14	3	1616, 3596, 3879	4	498, 1514, 2660, 3497
15	5	444, 788, 3808, 3816, 4972	7	681, 1896, 2257, 3490, 3749, 4444, 4775
16	5	515, 830, 838, 2017, **5018**	7	246, 414, 2554, 2768, 2910, **3949**, **4819**
17	4	1922, 1930, 2959, 4256	6	375, 459, 733, 1684, 2930, 2979

The genes in bold were viewed in ACT to detect any closeness to *stx2a*. As listed in Table 3.1, *stx2a* can be found at gene placement 1881 and 4581 for EC13 and EC16, respectively. In Figure 3.6 *intA* genes are marked with blue, *intS* with orange and *stx2a* with black. *intA* and *intS* are about 110 kbp afar from *stx2a* in EC13, and in EC16 *intA* is almost 290 kbp apart from *stx2a* and *intS* is 160 kbp apart. This means none of them are within the prophage insertion area of 50 kbp. Checking the Prokka annotation, none of these genes contain anything in particular indicating prophage genes.

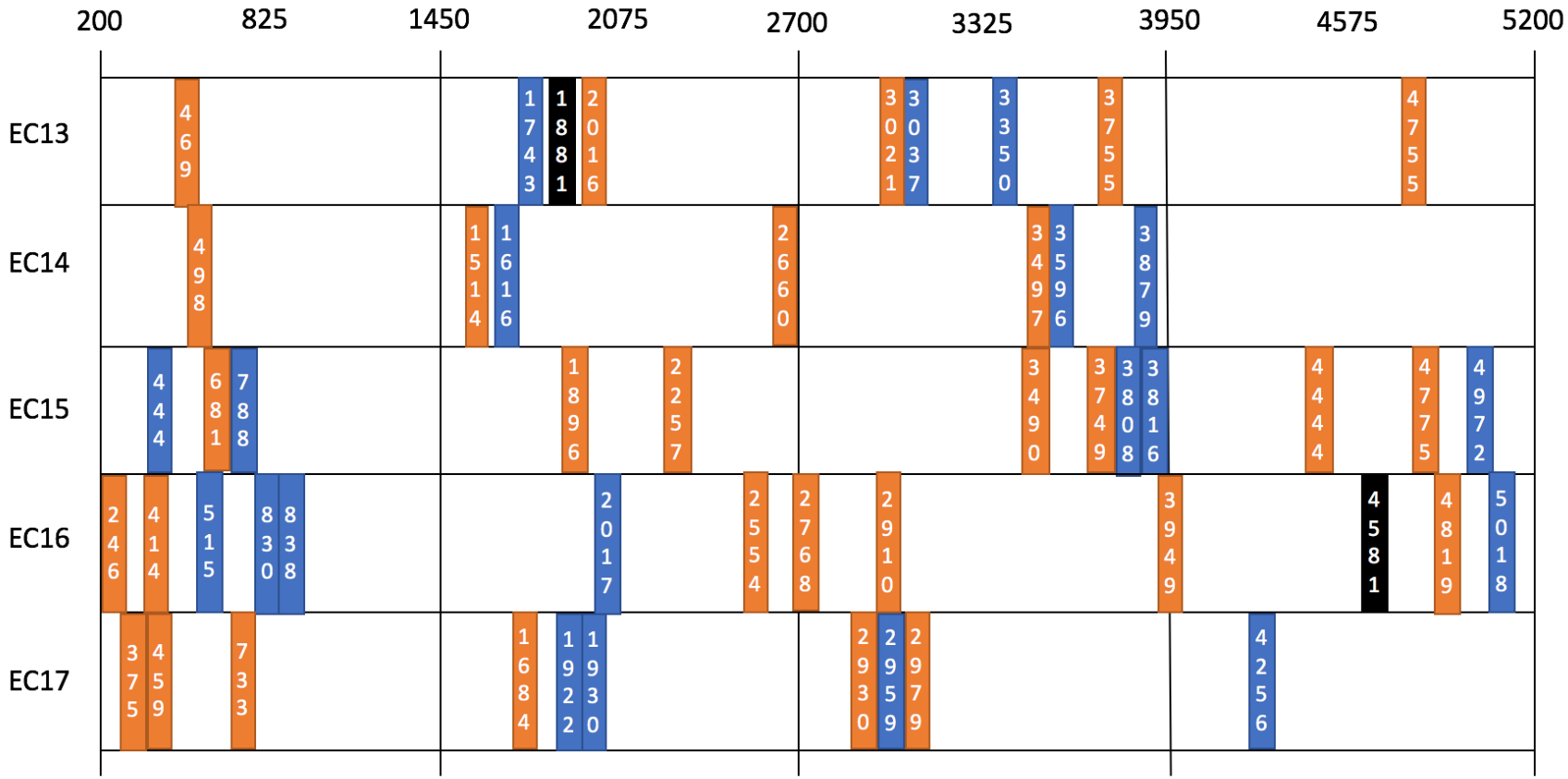

Figure 3.6: Gene placement of *intA* and *intS* in SFO157. Viewed together with Table 3.6. *intA* genes are filled with blue color and *intS* genes are marked in orange. *stx2a* is marked with black. Numbering at the very top is gene placement in the genome.

3.5 *stx2a*-phages

In Table 3.7, the *stx2a*-phages and the *stx- E. coli* isolates are organized. These observations show that the local alignment is highly identical and that EC10 and EC11 contains parts from several phages. The other isolates, however, do only match with some of the phages. When run through BLAST specifically for serogroup SFO157 (EC13-EC17) few of the phages align well, which is coherent with Table 3.7. In BLAST EC13-EC17 aligned with a percentage level of 3% to 14% similarities with phages 1717, 1447, F349, P27 and WGPS (2, 4, 6 and 8) shown in Table 3.8.1. EC13 and EC16 aligned well with phages VTB46 and VTB60 (99.0-99.7%), but as all of the *stx- E. coli* showed 0% alignment, these were cut.

Table 3.7: Phage detection. Nucleotide identity rate between *stx*- isolates and *stx2a*-phages with percentage level of best matches (varying sequence of overlap automatically chosen by OrthoANI).

Isolates / **Phage**	**3**	**5**	**6**	**8**	**10**	**11**	**14**	**15**	**17**	**19**	**22**
1717	x		x	x							x
1447											x
933W					x	x					
TL-2011c					x	x					
8624					x	x					
86				x		x					
24B	x				x	x					
P22					x	x					
P27	x			x	x	x			x		
P32					x	x					
I					x	x					
II					x	x					
F349	x							x	x	x	x
F403					x	x					
F422					x	x					
F451				x	x	x					
F723					x	x					
F765					x	x					
WGPS2											
WGPS4	x		x	x				x			x
WGPS6	x			x	x						
WGPS8	x				x	x					x
WGPS9				x	x	x					

x 98.00% - 98.64% x 95.00% - 97.99% blank = 83.09% - 94.99%

Comparing the *stx2a*-phage to the *E. coli* isolates in Mauve, fragments of the phages were scattered throughout the genomes. In some of the *stx+* isolates certain genetic sequences of the phage could be found within a 50 kb distance around *stx2a*. In EC13, this was observed with phages 1447 and WGPS2, and in EC16 with phages 1717, P27 and WGPS (6 and 8). The nucleotide placements of *stx2a* in EC13 was 1666926 and 4299956 in EC16 as shown in Table 3.1, and the phages placement can be observed in Table 3.8.2. Phages 1447 and WGPS2 were just outside of the specified region of 50 kb in EC16. When comparing the *stx+* and *stx-E. coli* around the *stx2a*-gene, there were certain similarities. In some cases, lengths of 5000-15000 nucleotide bases around *stx2a* lined up, although with some SNPs. Determining the genes content is difficult, even with Prokka annotation and NCBI-BLAST search.

Table 3.8.1: Alignment of relevant phages to SFO157. Percentage level alignment rate in BLAST between *stx2a*-phages relevant to SFO157 isolates collected from Table 3.7.

Phages	1717	1447	F349	P27	WGPS2	WGPS4	WGPS6	WGPS8
13	8	14	9	6	14	9	9	10
14	8	10	9	3	9	9	9	10
15	10	11	11	5	11	11	11	12
16	6	7	6	7	7	6	7	7
17	6	7	7	4	7	7	7	7

Table 3.8.2: Phage detection within prophage region. Results from search for relevant phages to SFO157 in a 50 kbp region of *stx2a*. Nucleotide base placement shown.

Phages	1717	1447	F349	P27	WGPS2	WGPS4	WGPS6	WGPS8
13		1698000			1694000			
16	4150000-4430000	4480000 (maybe)		4150000-4430000	4430000 (maybe)		4150000-4430000	4150000-4430000

4 DISCUSSION

In this study the focus was to explore *eae*-positive *E. coli* with and without *stx* to determine if the *stx*-negative *E. coli* were STEC-LST or "true" aEPEC. The phylogenetic similarities (MLVA- and cgMLST-profiles) of "pairs" (*stx*-positive and *stx*-negative isolates) within identical serotypes determined relatedness and allelic differences. The observation of virulence genes (*stx*, *eae*, *ehxA*, *cdtB* and *nleB*) revealed the pathogenic profiles of the *E. coli* isolates. The *stx*-subtype discovered within the *stx*-positive isolates was *stx2a*.

Regions where the *stx*-negative *E. coli* had high identity rates with the *stx* or prophage area in STEC were of special interest. The *stx*-gene is carried by a *stx*-phage, therefore could the area containing *stx* within the STEC be a prophage. If larger parts of the prophage were found within the *stx*-negative *E. coli*, there might have been previous content of a *stx*-carrying phage. Therefore, could the amount of phage-related gene content within the *stx*-negative *E. coli* indicate whether just the *stx*-gene or larger fragments of the prophage was missing from the genome. Observations of insertion site occupancy in *stx*-negative isolates compared to STEC helped reveal traces of prophage integration.

While many serotypes were considered, the focus was particularly on SFO157. SFO157 was chosen as main serotype because of identical MLVA-profiles, closest relatedness in cgMLST (three allelic differences), the SPT-A rating from seropathotype classification ("high risk"), and interesting clinical outcomes. The SFO157 *stx*-negative *E. coli* cause severe clinical outcomes that resemble STEC infections. Therefore, within this serotype, the *stx*-negative *E. coli* could indicate STEC – lost shiga toxin (STEC-LST). *In silico* methods were applied to explore this.

4.1 *E. coli* data and SFO157 isolate similarities

STEC with *stx2a*-subtyping and *stx*-negative *E. coli* were identified through several methods; PCR, NGS and NCBI-BLAST. Exactly half of the isolates had *stx2a* and the other half were *stx*-negative, with an almost even distribution of *stx*-positive and *stx*-negative *E. coli* within all of the six serotypes. The *E. coli* receipt from the in-house NIPH pipeline determined virulence genes, the most important being *stx2a*, *eae*, *ehxA* and *cdtB*. These were compared to each other to reveal particular similarities and dissimilarities.

Generally, the *E. coli* isolates showed overall relatedness from the OrthoANI. It was observed that most of the serogroups (O-types) had high identity rates (>99%) of average nucleotide sequences when compared. This indicates that many parts of the *E. coli* have similar sequences, which would be natural for bacteria within the same serotype. Observing Table 3.1, all of the serotypes are very similar although O177:H25 and O26:H11 are the most varying within serotype. O103:H25 (EC9-12), SFO157 (EC13-17) and O145:H28 (EC18-19) were the serotypes that had identical MLVA-profiles. The *E. coli* sharing the same flagellar antigen (H-type), especially H25, scored high matching rates. O177:H25 (EC20-22) and O145:H25 (EC1-3) were clustered within the same serotype (brown). Figure 3.1 and 3.2 shows the MST and NJ-tree from cgMLST-analyses, where O103:H25 and SFO157:H7 have the least allelic differences within the serotype. Within SFO157, EC15-17 consistently have three allelic differences between the various isolates. This might indicate similar origin of infection or the bacteria stemming from identical culture.

For the isolates collected from ovine (O26:H11 EC6 and O103:H25 EC10) and mutton (O103:H25 EC11) there were no *stx*-content, but they contained *eae*-gene encoding intimin. The intimin gene is associated with highly pathogenic *E. coli* but could not correctly differentiate a “true” aEPEC from STEC-LST (Ferduous *et al.,* 2015, Mellmann *et al*., 2009, Brandal *et al*., 2015, Scheutz, 2014). Although ovine and bovine are one of the main carriers of STEC, they are asymptomatic, and it is difficult to determine the circumstances of infection (Krüger *et al*., 2015, Bieleszewska *et al*., 2006). The MLVA-profile of EC10 and EC11 were exactly the same as EC9 and EC12, the *stx2a*-positive isolates within the O103:H25 serotype, which could indicate a possible STEC. NGS-data shows 99-100% similarities within

this serotype. The cgMLST scheme supports this and shows only one allelic difference between EC11 (*stx*-negative) and EC9 (*stx*-positive).

SFO157:H7 is classified as a "high risk" *E. coli*, and the clinical outcomes of the serotype are interesting. All the isolates were collected from humans within two years – 2008 and 2009. All of the SFO157 were *eae*- and *ehxA*-positive. EC13 and EC14 also contained *cdtB*. Only EC13 and EC16 carried *stx2a*. EC13 and EC14 had identical MLVA-profiles, and so did EC15-EC17. EC15 (*stx*-negative) caused a severe clinical outcome that resembled STEC infections, while EC16 (*stx*-positive) lead to asymptomatic outcome. Due to the severity of infection, there might be a possibility that EC15 could have previously contained *stx* (Bieleszewska *et al*., 2006, Mellmann *et al*., 2009, Haugum, 2014, Byrne *et al*., 2018). All of the isolates with *stx2a* in this study caused severe illness, except for EC16.

EC16 and EC17 came from the same incident, but EC17 did not contain *stx*. Either EC16 and EC17 could be from different cultures even though they are collected from the same incident, or the *stx*-gene have disappeared from EC17 during infection or isolation. In Mauve, EC17 was observed to contain phage-related genes. EC17 could have been a STEC and selectively lost *stx*. Carrying a highly pathogenic toxin is costly for the bacteria, as *stx* can induce the phage cycle and lead to bacterial lysis (Krüger & Lucchesi, 2015, Ferduous *et al.,* 2015). It is a major disadvantage for the survival of the host. Another supporting factor of EC15-17 stemming from the same bacteria, is the close relation between the geographical outbreak and timeline. The likeliness of separate infections is lower when the bacteria are closely related. As mentioned before, the MLVA-profiles are identical and the MST show only three allelic differences between the isolates. Therefore, the *stx*-negative *E. coli* isolates could be considered to stem from the same culture as the *stx*-positive, but with missing *stx*, resulting in EC15 and EC17 possibly being STEC-LST (Byrne *et al*., 2018).

Although EC13 and EC14 had the exact same MLVA-profile, they clustered further apart from each other than isolates EC15-17. The MST and NJ-tree showed 27 allelic differences, and this might be caused by EC13 containing a *stx2a*-prophage not found in EC14. EC13 and EC14 shows 100% similarities in Table 3.2 and 99% in Table 3.4. Also, the Mauve comparison, Figure 3.5, shows a smaller sequence that could be a prophage fragment corresponding in

EC13 (*stx*-positive) and EC14 (*stx*-negative). The clinical outcome of EC14 could have helped determine whether the infectious source was *stx* or not, but it is unknown. It is a challenge to determine whether *stx2a* had disappeared from the EC14 or if it never initially contained *stx*. There might not have been a *stx*-phage infection in EC14.

4.2 Prophage data

stx-phages carry *stx*. Through a *stx*-phage infection the *E. coli* genome attains the prophage containing *stx*. In determining where the prophage area could be in the genome of the *stx*-negative isolates, naturally the *stx2a*-positive *E. coli* would be templates within their serotypes. The prophage data matched the *stx*-negative isolates by ≥95% average nucleotide identity when analyzing local alignments. For SFO157, EC13 and EC16 contain *stx2a*, while EC14, EC15 and EC17 were *stx*-negative isolates. Here, local alignment of isolates did not output identity rates lower than 99.85%. The OrthoANI results first and foremost indicate a general level of similarity in *E. coli* prophage and the contigs data of *stx*-negative *E. coli* within the same serotype. There are some genes that match up consecutively in certain areas of the prophage data with the *stx*-negative *E. coli*. The contents of these genes were checked with gene annotation. Prokka mostly annotated “hypothetical proteins”, but other genes surrounding the prophage area were discovered to relate to tRNAs, anti-adaptor, ribose, DNA-replication, flagella or pilin, modification methylases, *et cetera* (Krüger & Lucchesi, 2015). All these genes, including the “hypothetical protein” could imply phage-related areas. Since phage genes are not annotated well yet, the specificities of the Prokka findings are not great. The reason is that these genes are not properly explored or categorized yet.

The Mauve alignment (Figure 3.5) illustrates the potential prophage positions. In EC13 a 5000-6000 bp sequence upstream of *stx2a* is reciprocated in EC16 downstream from the *stx2a* in this genome. The genes are inverted but placed in a somewhat right order. This could indicate a partly conserved prophage – although this is 10-15% of what would be expected in *stx*-positive *E. coli* if they contain a fully preserved prophage (50-60 kbp). Separate regions in EC13 and EC16 further apart from the *stx2a*, within the assumed prophage area, are not found in the *stx*-negative *E. coli* genomes. Although some genes in

the *stx*-negative *E. coli* match with the *stx*-positive *E. coli*, they are considered insignificant due to distanced placement from *stx2a*, non-matching sequence length or larger gaps.

Although *E. coli* have higher plasticity than many bacteria and are prone to mutations (Gordo *et al*., 2014, Leimbach *et al*., 2013), the observed GC-contents of the *E. coli* do not indicate higher levels of mutation. As the standard amount of GC content in *E. coli* is 50.4% to 50.8%, a level of 50.5% is within the limits (Mann & Chen, 2010). If a much lower GC-content was observed, phage-prone mutations could be discussed due to AT-richness. As the observed level is normal for *E. coli* there might not be any significant signs of larger prophage content within the isolates.

4.3 Insertion sites

wrbA, yehV, sbcB, argW, yecE, potC, prfC, serU, ssrA, yciD, yecD, yjbM, ynfH, Z2577 are specific insertion sites for *stx*-phages. *wrbA*, *yecE*, *yehV*, and *sbcB* were selected for SFO157 (Friedrich *et al.*, 2007, Sierra-Moreno *et al*., 2007). Table 3.5 show sequence matches with the SFO157 isolates. Observing the contigs, none of the insertion sites seem to be located close enough to the prophage area within the genomes. This could mean that comparison errors occurred when the insertion sites were matched with the isolates, or that the selected *stx*-phages do not integrate at these sites. Other options could be that the prophage does not exist within the *E. coli* or that it is not characterized in the databases.

If the insertion sites have 100% conserved DNA, there are no insertions at these sites. For all SFO157, *wrbA* and *yehV* are preserved 100% and 93% respectively, on a specific contig. The high identity rates could indicate absence of phage integration at these sites. In EC14-16 *yecE* and *sbcB* is preserved 100%. For EC13 *yecE* is 81% conserved and for EC17 *sbcB* is 82% conserved. There could have been phage integration at *yecE* in EC13 because of occupancy and findings of phage fragments. In EC17, there could be a potential prophage integration at *sbcB*. All of these sites are relatively misplaced to the *stx2a*-gene and indicate no prophage integration, although STEC (EC13 and EC17) should have at least one relevant insertion site occupancy close to stx2a. Another oddity is that the *stx*-positive EC16 did not show any difference from *stx*-negative EC14 and EC15, meaning no insertion site occupancy other than

yehV (93%). *yehV* might be an insertion site for the *stx*-phage in EC16, or there might be another insertion site that is not yet observed. If *yehV* is the insertion site for EC16, this might also be true for EC14 and EC15, and there could have been a prophage integration here at some point. This means that EC14 and EC15 could be STEC-LST.

Prokka annotates possibly relevant insertion sites as "prophage integrase" A and S, *intA* and *intS*. Figure 3.6 visualizes insertion sites to *stx2a*-placement, but none are located within a 50-60 kbp region of the prophage area. The unidentified sites, *intA* and *intS*, might not be probable insertion sites. Descriptive or qualitative information could help determine the specificities of *intA* and *intS*, which is not attainable. Determining what effect *intA* and *intS* have and whether they are occupied or not is difficult. Although *intA* and *intS* are not located close enough to *stx2a* to be viable as the specific insertion sites, they might become interesting if new *stx*-phages are recognized and integrate here. In that case, the "prophage integrases" need to be identified correctly.

Observing the insertion site data, important information could have been missed during the *stx*-phage selection. There might have been more than one *stx*-phage integration (relevant for SFO157). There might be possibilities of the relevant *stx*-phage not being discovered and characterized yet. The DNA sequence of this putative new phage might be larger or different from what is expected in discovered *stx*-phages. The insertion site placements could potentially be more relevant with findings of new *stx*-phages.

4.4 *stx2a*-phages

In the phage detection phase, *stx2a*-phages were attained through BLAST. In Table 3.7, the *stx*-negative *E. coli* EC10 and EC11 matched with the greatest number of phages with >83% identity. EC14 had no hits and EC15 and EC17 had two matches, each between 83-98% identity. Low levels of local alignment match indicate no infection from the collected *stx*-phages in *stx2a*-negative SFO157 isolates (Krüger & Lucchesi, 2015, Casjens, 2003). In Table 3.8.1 and 3.8.2, the phages with hits were examined more thoroughly through BLAST. Only fragments of the phages were similar to the SFO157, and the genetic composition where the phage coheres with the genome were difficult to identify. In the SFO157 isolates, OrthoANI comparisons with *stx*-phages showed a low percentage in identity rating (<15%), which coheres with the prophage detection in the Mauve alignment. In the *stx*-positive isolates, EC13 and EC16, any genes matching with the *stx*-phages within 50kpb of *stx2a* in the genome were noted. These sequences roughly indicated preserved amount of the prophage.

The BLAST search with *stx*-phages against SFO157 showed interesting findings. The phages match with contigs from the *E. coli* sequences, especially the *stx*-positive. Here, sequences in each end of the phage could be found within the *E. coli* (including *stx*) mostly conserved. However, the sequences in the center did not match. Many *stx*-phages have similar prophage encoding genes for replication, recombination and regulation genes with insertion sites, integration seats, methylases and tRNAs, among other genes, in one end. Then, in the other end there are genes encoding Q, Stx, lysis, tail and sheath of the prophage (Haugum *et al.*, 2012, Olavesen *et al*., 2016). The mismatches in the center of the phage could mean that parts of the prophage have been replaced by indels (insertion and deletion), several point mutations or transposons. It could mean that BLAST did not identify the phages completely correctly, and a larger phage sequences might fit better within the prophage area. Another possibility could be that the SFO157 could contain two or more *stx2a*-phages (Bielaszewska *et al*., 2008). This would explain the insertion site placements not matching the observed *stx2a*-area, and the gaps and mismatch in fragment order in Figure 3.5.

Prokka could not annotate these regions properly, and it is uncertain which genes might be undiscovered. For EC13 and EC16 certain genes indicating a prophage might be detected, such as tRNA, modification methylase, anti-adaptor, transcriptional repressor (*lexA*),

virulence regulator (*virB*), prophage tail and sheath, virulence genes (*stx2a*) and many "hypothetical proteins". Unknown phage genes are often annotated to be hypothetical proteins, and this could indicate fragments of a previous prophage. If the three tRNA located downstream *stx2a* are *ileZ*, *argN* and *argO* like in Figure 3.4, it could indicate a *stx*-prophage (Haugum *et al*., 2012). If Prokka annotated DNA methylase as "modification methylase" and the annotated anti-adaptor genes are *q*-genes, this will support the *stx*-prophage theory. Virulence genes detected by pipeline programs (*nleB*, *cdtB*, *eae*, *ehxA*, *stx2a*) also show high pathogenicity. All SFO157 contain *nleB* which is associated with virulent EHEC and EPEC, and *cdtB* is a potential virulence factor that may have been attained from phage transduction (Karch & Bielaszewska, 2001; Janka *et al*., 2003, Bugarel *et al.,* 2011). Prophage tail and sheath genes, together with the virulence factors, imply that a prophage most likely have infected the bacteria.

4.5 Future perspectives and limitations

In silico methods of analyses introduce new ways to identify and compare STEC, aEPEC and potential STEC-LST. For further exploration, suggestions could be Oxford Nanopore sequencing of "pairs" of *eae*-positive *E. coli* with and without *stx*, with main focus on SFO157:H7.

In her thesis, Larsen (2017) established that aEPEC isolates were susceptible to *stx2a*-phage incorporation through lysogenic infection if all insertion sites were available. She suggests *in silico* research with whole genome comparison and phylogenetic study of aEPEC vs. STEC. Human, ovine, and bovine isolates were used to evaluate relations between pathogens by searching for insertion sites, virulence genes and relevant genomic constituents. This, to identify factors essential for *stx2a*-phage susceptibility, and compare *stx2a*-phages within STEC. Scheffer (2017) suggests a method of sequence analysis by trimming reads with Trimmomatic, executing *de novo* assembly in SPAdes with BayesHammer as error correction, reference-based assembly with BWA-MEM (Bowtie2 and SAMtool chosen for this study), annotating with Prokka, and then comparing the results. This was attempted in this thesis.

Limitations of this study would be the small sample size and computational and human errors. The focus of this study was to compare *eae*-positive *E. coli* with and without *stx2a*. The sample size was chosen to be smaller to explore qualitative information in depth. Five samples from one serotype, SFO157, were chosen from 22 *E. coli* samples total. A larger sample size could possibly reveal other necessary information and is highly recommended to determine statistical significance in comparisons. In this study, a large sample size could not be provided considering the research limitation of Norwegian isolates. As noted, the programs can be erroneous and running error correctors can be helpful. Some programs will function better with updates and new information with time.

Currently, the *stx*-phages are not yet fully characterized. There is a lot of information in the database of NCBI-BLAST that is truly relevant. Due to the *E. coli* evolution caused by high recombination rate, the prophages could change and adapt relatively fast. This means that the phages evolve and could be different from what has been previously observed. A lack in knowledge about gene existence and function cause trouble in determining phenotypical traits. This is evident in Prokka for example, as the genes are not optimally annotated. NCBI-BLAST was used as additional annotator but was not completely successful either. Programs output what researchers have already discovered, collected and categorized. Thus, they can produce erroneous information by not having an optimal reference structure. High through-put in programs that sequence and assemble genomes could cause poor sequence quality. Extended time to rerun the programs, could allow better processing of the genomic data.

5 CONCLUSION

The *in silico* data analysis show relatedness by serotypes in all the *E. coli* data, with higher similarity rates within similar serogroups (O-type) and flagellar antigens (H25). The *stx*-positive *E. coli* share some potential phage-related genes in addition to *stx2a*, *eae* and *ehxA* (except for EC12). SFO157 (EC13 to EC17) was chosen out of the six serotypes to be investigated more thoroughly. The serotype had interesting combination of *stx* content compared to the clinical outcome – asymptomatic to severe illness. The isolates matched in MLVA- and cgMLST-profiling and had corresponding genes matching within the prophage area. EC15 to EC17 were closely related and concluded to stem from the same culture being STEC and STEC-LST. The "high risk" pathogenic profile of SFO157 has led to many necessary researches of the bacteria. Though many studies have already been conducted of SFO157, further analyses can support and improve the attained information about STEC-LST.

In Mauve, *stx*-positive SFO157 EC13 and EC16 match with ca. 5-6 kbp (ca. 10-15%) to each other within the prophage area of the genomes. EC17 match better with EC16 at the prophage region (with gaps) than EC14 and EC15 match with any of the other SFO157 isolates here. Prokka annotates genes that portray phage-related genes and "hypothetical proteins" in the area surrounding *stx2a*. The annotation implies that many phage-genes are not yet categorized properly, but the findings point to a probable phage-region. Although the *stx*-positive *E. coli* have indications of prophage genes in these areas, the annotation of these genes is also not very specific. It is difficult to determine what the prophage contains.

EC17, which is *stx*-negative, had occupancy at insertion site *sbcB* (82% conserved), whereas *stx*-positive EC13 had occupancy at *yecE* (81% conserved). EC14 to EC16 had a possible occupancy at *yehV* (93% conserved). The insertion site occupancy could explain *stx*-phage susceptibility at these sites, even though the sites are not close enough to the prophage region (50-60 kbp). This means that there could have been a previous phage infection, and that the prophage was selectively lost from the genome through recombination. The putative new phage is also believed to contain a larger genome and might be a recombined *stx*-phage.

Of all the collected *stx2a*-phages from BLAST, only eight showed relevant alignment match. In ACT, EC16 and EC13, match with a few *stx2a*-phage genes around the *stx2a* region within the prophage. In BLAST comparisons, the start and end of the phages showed high identity, but a larger middle part of the prophage was deleted and inserted (indel or transposon) with a non-matching sequence. One reason is that the *stx*-positive *E. coli* might have lost some of the *stx*-phage during infection *in vivo* or isolation *in vitro*. Another reason could be that an unidentified *stx*-phage has caused the infection. A third possibility could be that the SFO157 could contain two or more *stx2a*-phages. This would explain the insertion site placements being observed far from the *stx2a*-area.

There is no doubt that investigating the topic further is important. In this thesis the aim was to find characteristics substantiating the classification of the *eae*-positive *E. coli* as STEC-LST. The similarities within STEC and the *eae*-positive *E. coli* could reflect a STEC-LST. A patient suffering from HUS can immediately be identified as STEC infected. The challenge is when the patient is asymptomatic or has a milder symptom than HC or HUS. The severe clinical outcome, occupancy of insertion sites and findings of *stx2a*-phage parts within the *E. coli* indicates high-risk bacteria. What could determine a STEC-LST is the close relatedness of these *eae*-positive *E. coli* and STEC, especially within serotype SFO157. Continuous lab-testing specifically for STEC by virulence genes (*stx*, *eae*, *ehxA*, *cdtB*, *nleB*) would always be an efficient identification method at first. To further test the *E. coli* by *in silico* methods could be a better way of identifying STEC-LST. By comparing previous STEC infections worldwide and monitoring the evolution of the bacteria, identification of STEC and STEC-LST will become faster and more efficient.

The conclusion of this thesis would be that STEC could lose *stx* selectively and become STEC-LST. When suspecting STEC, an elaborate follow-up routine is needed at the medical facilities. The evolution of STEC could be affected by the changes in the attuned phages, and vice versa. Therefore, the putative new phage is of interest to further research phage susceptibility in STEC. To do so, *in silico* methods are applicable to efficiently understand STEC and *stx*-phages. Further work and analyses are needed to determine more specific and clear results. Increasing the sample size could help with more reliable and viable results. The

databases could also be updated to improve reference structure to help characterize *stx*-phage genes more completely, as relevant information is obtained.

REFERENCES

AFADadcADSasd (2017). “Bacteriophage: a virus that feeds on bacteria”. Downloaded from https://upload.wikimedia.org/wikipedia/commons/e/eb/Bacteriophage.jpg

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., Wishart, D. (2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Research, 44*, 16-21. doi: 10.1093/nar/gkw387

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., … Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology, 19(5),* 455-477. doi: 10.1089/cmb.2012.0021

Beards, G. (2008). “Transmission electron micrograph of multiple bacteriophages attached to a bacterial cell wall”. Dowloaded from https://commons.wikimedia.org/wiki/File:Phage.jpg

Bielaszewska, M., Prager, R., Zhang, W., Friedrich, A. W., Mellmann, A., Tschäpe, H., Karch, H. (2006). Chromosomal Dynamism in Progeny of Outbreak-Related Sorbitol-Fermenting Enterohemorrhagic *Escherichia coli* O157:NM. *Applied and Environmental Microbiology, 72(3)*, 1900-1909. doi: 10.1128/AEM.72.3.1900-1909.2006

Bielaszewska, M., Middendorf, B., Köck, R., Friedrich, A. W., Furth, A., Karch, H., … Mellmann, A. (2008). Shiga Toxin-Negative Attaching and Effacing *Escherichia coli*: Distinct Clinical Associations with Beneficial Phylogeny and Virulence Traits and Inferred In-Host Pathogen Evolution. *Clinical Infectious Diseases, 47, 208-217*. doi: 10.1086/589245

Bielaszewska M., Idelevich E. A., Zhang W., Bauwens A., Schaumburg F., Mellmann A., Peters G., Karch H. (2012). Effects of Antibiotics on Shiga Toxin 2 Production and Bacteriophage Induction by Epidemic *Escherichia coli* O104:H4 Strain. *Antimicrobial Agents Chemotherapy, 56(6), 3277–3282*. doi: 10.1128/AAC.06315-11

Bohlin, J., Brynildsrud, O. B., Sekse, C., Snipen, L. (2014). An evolutionary analysis of genome expansion and pathogenicity in *Escherichia coli*. *BMC Genomics, 15(882).* doi: 10. 1186/1471-2164-15-882

Bohlin, J. (2018). Bioinformatical and biological information – conversations, mails and notes. Ås: Norwegian University of Life Science & Oslo: Norwegian Institution of Public Health (NIPH).

Bolger, A. M., Lohse, M., Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 40(15),* 2114-2120. doi: 10.1093/bioinformatics/btu170

Brandal L. T., Wester A. L., Lange H., Lobersli I., Lindstedt B. A., Vold L., Kapperud G. (2015). Shiga toxin-producing *Escherichia coli* infections in Norway, 1992-2012: characterization of isolates and identification of risk factors for haemolytic uremic syndrome. *BMC Infectious Diseases, 15(324)*. doi: 10.1186/s12879-015-1017-6

Bugarel M., Beutin L., Scheutz F., Loukiadis E., Fach P. (2011). Identification of Genetic Markers for Differentiation of Shiga Toxin-Producing, Enteropathogenic, and Avirulent Strains of *Escherichia coli* O26. *Applied and Environmental Microbiology, 77(7), 2275-2281*. doi: 10.1128/AEM.02832-10

Byrne, L., Dallman, T. J., Adams, N., Mikhail, A. F. W., McCarthy, N., Jenkins C. (2018). Highly Pathogenic Clone of Shiga Toxin–Producing *Escherichia coli* O157:H7, England and Wales. *Emerging Infectious Diseases, 24(12), 2303-2308*. doi: 10.3201/eid2412.180409

Casjens, S. (2003). Prophages and bacterial genomics: what have we learned so far?*. Molecular Microbiology, 49(2)*, 277-300. doi: 10.1046/j.1365-2958.2003.03580.x

Center of Disease Control and Prevention. (2014). Dowloaded from https://www.cdc.gov/ecoli/general/index.html

Center of Disease Control and Prevention. (2016). Dowloaded from https://www.cdc.gov/pulsenet/pathogens/mlva.html

Cisek, A. A., Dabrowska, I., Gregorczyk, K. P., Wyzewski, Z. (2017). Phage Therapy in Bacterial Infections Treatment: One Hundred Years After the Discovery of Bacteriophages*. Current Microbiology, 74(2), 277-283*. doi: 10.1007/s00284-016-1166-x.

Clements, A., Young, J. C., Constantinou, N. & Frankel, G. (2012). Infection strategies of enteric pathogenic *Escherichia coli*. *Gut Microbes, 3(2), 71-87*. doi: 10.4161/gmic.19182

Clermont, O., Christenson, J. K., Denamur, E., Gordon, D. M. (2013). The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environmental Microbiology Reports, 5(1),* 58-65. doi: 10.1111/1758-2229.12019

Delannoy, S., Beutin, L., Fach, P. (2013). Discrimination of enterohemorrhagic *Escherichia coli* (EHEC) from non-EHEC strains based on detection of various combinations of type III effector genes. *Journal of Clinical Microbiology, 51(10)*, 3257-3262. doi 10.1128/JCM.01471-13

Delannoy, S., Mariani-Kurkdjian, P., Webb, H. E., Bonacorsi, S., Fach, P. (2017). The Mobilome; A Major Contributor to *Escherichia coli stx2*-Positive O26:H11 Strains Intra-Serotype Diversity. *Frontiers in Microbiology, 8(1625),* 1-17. doi: 10.3389/fmicb.2017.01625

Didelot, X., Méric, G., Falush, D., Darling, A. E. (2012). Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. *BMC Genomics, 13(256).* doi: 10.1186/1471-2164-13-256

Eichhorn, I., Heidemanns, K., Semmler, T., Kinnemann, B., Mellmann, A., Harmsen, D., … Wieler, L. H. (2015). Highly Cirulent Non-O157 Enterohemorrhagic *Escherichia coli* (EHEC) Serotypes Reflect Similar Phylogenetic Lineages, Providing New Insights into the Evolution of EHEC. *Applied and Environmental Microbiology, 81(20),* 7041-7047. doi: 10.1128/AEM.01921-15

Eichhorn, I., Semmler, T., Mellmann, A., Pickard, D., Anjum, M. F., Furth, A., … Wieler, L. H. (2018). Microevolution of epidemiological highly relevant non-O157 enterohemorrhagic *Escherichia coli* of serogroups O26 and O111. *International Journal of Medical Microbiology.* doi: 10.1016/j.ijmm.2018.08.003

Eichhorn, I., Heidemanns, K., Ulrich, R. G., Schmidt, H., Semmler, T., Fruth, A., … Wieler, L. H. (2018). Lysogenic conversion of atypical enteropathogenic *Escherichia coli* (aEPEC)
from human, murine, and bovine origin with bacteriophage Φ3538
Δ*stx2*::cat proves their enterohemorrhagic *E. coli* (EHEC) progeny. *International Journal of Medical Microbiology.* doi: 10.1016/j.ijmm.2018.06.005

Enterobase version 1.1.2. (2018). Downloaded from https://enterobase.warwick.ac.uk/species/index/e.coli

Feng, Y., Zhang, Y., Ying, C., Wang, D., Du, C. (2015). Nanopore-based Fourth-generation DNA Sequencing Technology. *Genomics Proteomics Bioinformatics, 13,* 4-16. doi: 10.1016/j.gpb.2015.01.009

Ferdous, M., Zhou, K., Mellmann, A., Morabito, S., Croughs, P. D., de Boer, R. F., … Friedrich, A. W. (2015). Is Shiga Toxin-Negative *Escherichia coli* O157:H7 Enteropathogenic or Enterohemorrhagic Eschericia coli? Comprehensive Molecular Analysis Using Whole-Genome Sequencing. *Journal of Clinical Microbiology, 53(11)*, 3530-3538. doi: 10.1128/JCM.01899-15

Folkehelseintituttet (2010, 24. february). *E. coli*-enteritt (inkludert EHEC-infeksjon og HUS) - veileder for helsepersonell. Downloaded from https://www.fhi.no/nettpub/smittevernveilederen/sykdommer-a-a/e.-coli-enteritt-inkludert-ehec-inf/

Franzin, F. M. & Sircili, M. P. (2015). Locus of enterocyte effacement: a pathogenicity island involved in the virulence of enteropathogenic and enterohemorragic Escherichia coli subjected to a complex network of gene regulation. *BioMed Research International, 2015, 534738*. doi: 10.1155/2015/534738

Fraser, M. E., Fujinaga, M., Cherney, M.M., Melton-Celsa, A. R., Twiddy, E. M., O'Brien, A. D., James, M. N. G. (2004). "Structure of shiga toxin type 2 (Stx2) from *Escherichia coli* O157:H7". *Journal of Biological Chemistry, 279(26), 27511–27517*. doi: 10.1074/jbc.M401939200.

Friedrich, A. W., Zhang, W., Bielaszewska, M., Mellmann, A., Köck, R., Fruth, A., … Karch, H. (2007). Prevalence, Virulence Profiles, and Clinical Significance of Shiga Toxin-Negative Variants of Enterohemorrhagic *Escherichia coli* O157 Infection in Humans. *Clinical Infectious Diseases, 2007(45)*, 39-45. doi: 10.1086/518573

Friedrich A. W., Bielaszewska M., Zhang W. L., Pulz M., Kuczius T., Ammon A., Karch H. (2002) *Escherichia coli* harboring Shiga toxin 2 gene variants: frequency and association with clinical symptoms. *Journal of Infectious Diseases, 185(1), 74-84*. doi: 10.1086/338115

Gamage, S. D., Patton, A. K., Hanson, J. F. & Weiss, A. A. (2004). Diversity and host range of Shiga toxin-encoding phage. *Infection and Immunology, 72(12), 7131-7139.* doi: 10.1128/IAI.72.12.7131-7139.2004

Gelderblom, H. R. (1996). Medical Microbiology (4th edition). Chapter 41: Structure and Classification of Viruses. Galveston (TX): University of Texas Medical Branch at Galveston.

Genetics Home Reference (2018). Downloaded from https://ghr.nlm.nih.gov/primer/genomicresearch/genomeediting

Gordo, I., Demengeot, J. & Xavier, K. (2014). *Escherichia coli* adaptation to the gut environment: a constant fight for survival. *Future Microbiology, 9(11),* 1235-1238. doi: 10.2217/fmb.14.86.

Górski A., Jończyk-Matysiak E., Międzybrodzki R., Weber-Dąbrowska B., Łusiak-Szelachowska M., Bagińska N., Borysowski J., … Węgrzyn G. (2018). Phage Therapy: Beyond Antibacterial Action. *Frontiers in Medicine, 5, 146*. doi: 10.3389/fmed.2018.00146

Gupta R. M. & Musunuru K. (2014). Expanding the genetic editing tool kit: ZFNs, TALENs, and CRISPR-Cas9. *Journal of Clinical Investigation, 124(10), 4154-4161*. doi: 10.1172/JCI72992.

Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics Application Note, 29(8),* 1072-1075. doi: 10.1093/bioinformatics/btt086

Haugum, K., Lindstedt, B.-A., Løbersli, I., Kapperud, G., Brandal, L. T. (2012). Identification of the anti-terminator $q_{O111:H-}$ gene in Norwegian sorbitol-fermenting *Escherichia coli* O157:NM. *Federation of European Microbiological Societies, 329*, 102-110. doi: 10.1111/j.1574-6968.2012.02505.x

Haugum, K. (2014). Studies of genetic characteristics in Shiga toxin-producing *Escherichia coli* (STEC) from patients with and without haemolytic uremic syndrome (HUS) in Norway. (Ph. D. Thesis). Trondheim: Norwegian University of Science and Technology (NTNU).

Hsu P. D., Lander E. S., Zhang F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. *Cell Press, 157(6), 1262-1278*. doi:10.1016/j.cell.2014.05.010.

Huang, A., Friesen, J. & Brunton, J. L. (1987). Characterization of a bacteriophage that carries the genes for production of Shiga-like toxin 1 in *Escherichia coli*. *Journal of Bacteriology, 169(9)*, 4308-4312.

Janka, A., Bielaszewska, M., Dobrindt, U., Greune, L., Schmidt, M. A., Karch, H. (2003). Cytolethal distending toxin gene cluster in enterohemorrhagic *Escherichia coli* O157:H- and O157:H7: characterization and evolutionary considerations. *Infection and Immunity, 71(6), 3634–3638*. doi: 10.1128/IAI.71.6.3634-3638.2003

Jiang, Y., Chen, B., Duan, C., Sun, B., Yang, J., Yang, S. (2015). Multigene editing in the *Escherichia coli* genome via the CRISPR-Cas9 system. *Applied Environmental Microbiology, 81, 2506-2514*. doi: 10.1128/AEM.04023-14

Karch H. & Bielaszewska M. (2001). Sorbitol-Fermenting Shiga Toxin-Producing *Escherichia coli* O157:H− Strains: Epidemiology, Phenotypic and Molecular Characteristics, and Microbiological Diagnosis. *Journal of Clinical Mircrobiology, 39(6), 2043-2049*. doi: 10.1128%2FJCM.39.6.2043-2049.2001

Karki, R., Pandya, D., Elston, R. C., Ferlini, C. (2015). Defining «mutation» and «polymorphism» in the era of personal genomics. *BMC Medical Genomics, 8(37)*. doi: 10.1186/s12920-015-0115-z

Koonin E. V. & Starokadomskyy P. (2016). Are viruses alive? The replicator paradigm sheds decisive light on an old but misguided question. *Studies in History and Philosophy of Biological and Biomedical Sciences, 59, 125-134*. doi: 10.1016/j.shpsc.2016.02.016

Krüger, A. & Lucchesi, P. M. A. (2015). Shiga toxins and *stx* phages: highly diverse entities. *Microbiology, 161,* 451-462. doi: 10.1099/mic.0.000003

L'Abée-Lund, T. M., Jørgensen, H. J., O'Sullivan, K., Bohlin, J., Ligård, G., Granum, P. E., Lindbäck, T. (2012). The Highly Virulent 2006 Norwegian EHEC O103:H25 Outbreak Strain Is Related to the 2011 German O104:H4 Outbreak Strain. *Plos One, 7(3),* 1-11. doi: 10.1371/journal.pone.0031413

L'Abée-Lund & Wasteson, Y. (2015). Chapter 7: *Escherichia coli*. *In:* GRANUM, P. E. (ed.) *Matforgiftning*. 4. edition ed. -: Cappelen Damm Akademisk.

Langmead, B. (2010). Aligning short sequencing reads with Bowtie. *Current Protocols in Bioinformatics. Chapter: Unit–11.7*. doi: 10.1002/0471250953.bi1107s32

Langmead, B. & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9(4), 357–359*. doi: 10.1038/nmeth.1923

Langmead, B., Trapnell, C., Pop, M., Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology, 10(3).* doi: 10.1186/gb-2009-10-3-r25

Larsen, R. K. L. (2017). *Shiga toxin-producing E. Coli O26 – transduction of STX bacteriophages.* (Masterthesis). Ås: Norwegian University of Life Science (NMBU).

Lee, I., Kim, Y. O., Park, S-C., Chun, J. (2016). OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *International Journal of Systematic and Evolutionary Microbiology, 66, 1100–1103.* doi: 10.1099/ijsem.0.000760

Leimbach, A., Hacker, J., & Dobrindt, U. (2013). *E. coli* as an all-rounder: the thin line between commensalism and pathogenicity. *Current Topics in Microbiology Immunology, 358,* 3-32. doi: 10.1007/82_2012_303.

Magoc T. & Salzberg S. (2011). FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics 27(21), 2957-2963*. doi:10.1093/bioinformatics/btr507

Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q. … Spratt, B. G. (1998). Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America, 95(6), 3140–3145*.

Mann, S. & Chen, Y-P. P. (2010). Bacterial genomic G + C composition-electing environmental adaptation. *Genomics, 95(1), 7-15.* doi: 10.1016/j.ygeno.2009.09.002

Mellmann, A., Lu, S., Karch, H., Xu, J.-g., Harmsen, D., Schmidt, M. A., Bielaszewska, M. (2008). Recycling of Shiga Toxin 2 Genes in Sorbitol-Fermenting Enterohemorrhagic *Escherichia coli* O157:NM. *Applied and Environmental Microbiology, 74(1),* 67-72. doi: 10.1128/AEM.01906-07

Mellmann, A., Bielaszewska, M., Karch, H. (2009). Intrahost Genome Alterations in Enterohemorrhagic *Escherichia coli*. *Gastroenterology, 2009(136)*, 1925-1938. doi: 10.1053/j.gastro.2008.12.072

Milligan, I. (2014, 20. september). Programming Historian: Introduction to the Bash Command Line [Blogpost]. Dowloaded from https://programminghistorian.org/en/lessons/intro-to-bash.

Muniesa, M., De Simon, M., Prats, G., Ferrer, D., Panella, H. & Jofre, J. (2003). Shiga toxin 2-converting bacteriophages associated with clonal variability in *Escherichia coli* O157:H7 strains of human origin isolated from a single outbreak. *Infection and Immunology, 71(8), 4554-4562*. doi: 10.1128/IAI.71.8.4554-4562.2003

Nadon, C. A., Trees, E., Ng, K., Nielsen, E. M., Reimer, A., Maxwell, N., Kubota, K. A., Gerner-Smidt, P. and the MLVA Harmonization Working Group5 (2013). Development and application of MLVA methods as a tool for inter-laboratory surveillance. *Euro Surveillance, 18(35), 20565*.

Nagel, T. E. (2018). Delivering Phage Products to Combat Antibiotic Resistance in Developing Contries: Lessons Learned from the HIV/AIDS Epidemic in Africa. *Viruses, 10(7), 345*. doi: 10.3390/v10070345

Naseer U., Lobersli I., Hindrum M., Bruvik T., Brandal L. T. (2017). Virulence factors of Shiga toxin-producing *Escherichia coli* and the risk of developing haemolytic uraemic syndrome in Norway, 1992-2013. *European Journal of Clinical Microbiological Infectious Diseases, 36, 1613-1620*. doi: 10.1007/s10096-017-2974-z

National Institute for Public Health and the Environment (Ministry of Health, Welfare and Sport). Downloaded from https://www.mlva.net/

Nikolenko, S. I., Korobeynikov, A. I., Alekseyev, M. A. (2013). BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics, 14(57).* doi: 10.1186/1471-2164-14-S1-S7

Obrig T. G., (2010). *Escherichia coli* Shiga Toxin Mechanisms of Action in Renal Disease. Toxins, 2(12), 2769-2794. doi: 10.3390/toxins2122769

Obrig T. G. & Karpman D. (2012). Shiga Toxin Pathogenesis: Kidney Complications and Renal Failure. *Current Topics in Microbiology and Immunology. 357, 105-136*. doi: 10.1007/82_2011_172

Olavesen, K. K., Linstedt, B.-A., Løbersli, I., Brandal, L. T. (2016). Expression of Shiga toxin 2 (Stx2) in highly virulent Stx-producing *Escherichia coli* (STEC) carrying different anti-terminator (q) genes. *Microbial Pathogenesis (2016)*. doi: 10.1016/j.micpath.2016.05.010

Pevzner, P. A., Tang, H., Waterman M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America, 98(17), 9748–9753*. doi: 10.1073/pnas.171285098

Rasko, D. A., Webster, D. R., Sahl, J. W., Bashir, A., Boisen, N., Scheutz, F., … Waldor, M. K. (2011). Origins of the *E. coli* Strain causing an Outbreak of Hemolytic-Uremic Syndrome in Germany. *The New England Journal of Medicine, 365(8)*, 709-717. doi: 10.1056/NEJMoa1106920

Recktenwald, J. & Schmidt, H. (2002). The Nucleotide Sequence of Shiga Toxin (Stx) 2e-Encoding Phage ϕP27 Is Not Related to Other Stx Phage Genomes, but the Modular Genetic Structure Is Conserved. *Infection and Immunity, 70(4)*, 1896-1908. doi: 10.1128/IAI.70.4.1896-1908.2002

Rocky Mountain Laboratories, NIAID, NIH (2005). "*Escherichia coli*: Scanning electron micrograph of *Escherichia coli*, grown in culture and adhered to a cover slip". Dowloaded from https://commons.wikimedia.org/w/index.php?curid=104228

Scheffer, L. (2017). *Best Practices for 3D Protein Modelling.* (Thesis). Oslo: Norwegian Institute of Public Health (NIPH).

Scheutz, F., Teel, L. D., Beutin, L., Piérard, D., Buvens, G., Karch, H., … O'Brien, A.D. (2012). Multicenter evaluation of a sequence-based protocol for subtyping Shiga toxins and standardizing Stx nomenclature. *Journal of Clinical Microbiology, 50(9)*, 2951-2963. doi: 10.1128/JCM.00860-12

Scheutz, F. (2014). Taxonomy Meets Public Health: The Case of Shiga Toxin-Producing *Escherichia coli*. *Microbioloy Spectrum, 2(3)*. doi: 10.1128/microbiolspec.EHEC-0019-2013.

Senthakumaran, T., Brandal, L. T., Lindstedt, B.-A., Jørgensen, S. B., Charnock, C., Tunsjø, H. S. (2018). Implications of *stx* loss for clinical diagnostics of Shiga toxin-producing *Escherichia coli*. *European Journal of Clinical Microbiology & Infectious Diseases.* Germany: Springer Nature. doi: 10.1007/s10096-018-3384-6

Sierra-Moreno, R., Jofre, J., Muniesa, M. (2007). Insertion Site Occupancy by *stx2* Bacteriophages Depends on the Locus Availability of the Host Strain Chromosome. *Journal of Bacteriology, 189(18)*, 6645-6654. doi: 10.1128/JB.00466-07

Snipen, L., & Hvidsten, T. R., (2013-2018). BIN210, BIN310, BIN315 - Notes from lectures. Ås: Norwegian University of Life Science (NMBU).

Song, L., Florea, L., Langmead, B. (2014). Lighter: fast and memory-efficient sequencing error correction without counting. *Genome Biology, 15(509).* doi: 10.1186/s13059-014-0509-9

Sohn, J.-i. & Nam, J.-W. (2018). The present and future of *de novo* whole-genome assembly. *Briefings in Bioinformatics, 19(1),* 23-40. doi: 10.1093/bib/bbw096

Steinberg, K. M., & Levin, B. R. (2007). Grazing protozoa and the evolution of the *Escherichia coli* O157:H7 Shiga toxin-encoding prophage. *Proceedings of the Royal Society B*. doi: 10.1098/rspb.2007.0245

Stenutz, R., Weintraub, A. & Widmalm, G. (2006). The structures of *Escherichia coli* O-polysaccharide antigens. *FEMS Microbiology Reviews, 30(3)*, 382-403. doi: 10.1111/j.1574-6976.2006.00016.x

Tietze, E., Dabrowski, P. W., Prager, R., Radonic, A., Fruth, A., Auraß, P., … Flieger, A. (2015). Comparative Genomic Analysis of Two Novel Sporadic Shiga Toxin-Producing *Escherichia coli* O104:H4 Strains Isolated 2011 in Germany. *Plos One, 10(4).* doi: 10.1371/journal.pone.0122074

Ussery, D. W., Wassenaar, T. M., Borini, S. (2009). *Computing for Comparative Microbial Genomics.* London: Springer.

Willey, J. M., Sherwood, L. & Woolverton, C. J. (2014). 6.4 Types of Viral Infections. Prescott's microbiology. Ninth edition. New York, NY: McGraw-Hill.

Zvelebil, M. J. & Baum, J. O. (2007). *Understanding Bioinformatics*. USA: Garland Science.

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway