Degree project

UNIVERSITY OF SKÖVDE

NEXT-GENERATION DIAGNOSTICS OF *ESCHERICHIA COLI* FROM COMMUNITY-ONSET SEPSIS PATIENTS IN SWEDEN: STUDYING THE BIODIVERSITY OF *ESCHERICHIA COLI* GENOMES

Master Degree Project (120 credits) in Systems Biology

Second Cycle 45 credits

Spring term Year 2021

Student: Nada Mahmoud

Supervisor: Helena Enroth, University of Skövde

Examiner: Sanja Jurcevic, University of Skövde

Abstract

Escherichia coli (E. coli) is among the gram-negative bacteria that can cause several infections including sepsis. Confirmed sepsis patients must show a sequential organ failure assessment (SOFA) score of ≥2 with verified infection. Understanding the genotypic characteristics of *E. coli* clinical isolates from sepsis patients can help directing treatment strategies, tracking antibiotic resistance, and monitoring acquired virulence factors that can contribute to the severity of sepsis. The isolates included in this thesis were collected from confirmed sepsis patients (SOFA 2-3) during a prospective observational study that was conducted in Sweden. The aim was to study the biodiversity in *E. coli* clinical isolates using whole genome sequencing (WGS) paired-end reads that were produced by the next-generation sequencer Illumina. To perform the WGS-based analysis, two bioinformatics pipelines were used. The first is the in-house developed pipeline and the second is the 1928Diagnostics *E. coli* pipeline. The obtained *in silico* results were compared with the phenotypic findings for species identification and the *in vitro* predictions of antibiotic resistance. Species identification by the bioinformatics pipelines matched the phenotypic method, except for three isolates that were highly contaminated with other species. Both pipelines predicted the exact multi locus sequence types, which revealed that the most common sequence types (STs) were ST73(17%), ST95(9%), and ST131(6%). The phenotype of the isolates resulted in 5% resistant to at least one of the assessed antibiotics. The 1928Diagnostics predicted 28% of the isolates were resistant to at least one class of the tested antibiotic classes, while the in-house pipeline predicted 33% of the isolates to be resistant. Out of the predicted resistant isolates, 52% coded for multi-drug resistance. The in-house pipeline reported virulence genes. The common reported genes were coding for iron reuptake, adhesins, cell outer membrane and increased serum survival. It was observed that the isolates that belonged to ST73 and ST95 showed a more susceptible antibiotic profile than isolates that belonged to ST131, those harbored the highest mean of virulence genes. In conclusion, the present study provided an evidence of the usefulness of the WGS-based analysis to study the biodiversity in *E. coli*. The obtained results are valuable for surveillances, tracking antibiotic resistance and identifying virulence factors, but with a limited use in clinical settings.

Popular scientific summary

The word sepsis is derived from the ancient Greek verb "sepo", It means literally "I rot". Thousands of years ago, several writings from the ancient world reported the death of a previously healthy person after an injury. As reported, the injury was followed by a series of painful and stressful events, then death within a few days to a few weeks. Those ill-fated events were a mix between fever, sweating, unstable heart beats, difficulties to breath, hallucination and in some cases loss of consciousness. It was believed that there was a dangerous matter that enters the body through the wound, poison the blood and make the body rotten. It was also reported that a lot of people died in this way. Thousands of years later and with all the advancements the humans could reach, thousands of people still die every year from sepsis. So, what is the definition of sepsis according to the modern medical science? It is simply a condition that is caused by the overreaction of the immune system in response to an infection with a microorganism, it can be a bacteria, a fungus, a parasite, or a virus. So, what happens after the body get infected? Well, normally our immune system will react reasonably, and fight those microorganisms by the white blood cells, which are programmed to identify many of the microorganisms; However, sometimes our immune system will not identify the infecting microorganisms, but it will develop a memory to remember them the next time we get infected. In case of sepsis, the immune system will overreact to an infection. It can be an infected wound, a urinary tract infection or even food poisoning, that can cause inflammation in the body. The blood vessels will be affected and will start to leak, which in turn will drop the blood pressure and decrease the oxygen being carried to the body organs. Due to the low oxygenation, organs such as the heart, kidneys, lungs, and the brain will not function properly. Do not worry, not every infection will develop into sepsis! Anyway, it is still not well understood how and why the immune system overreact to some infections. Probably because some microorganisms can escape the primary location of the infection and produce toxic matters into the blood. How can sepsis be diagnosed? In 2016 a scale to evaluate the degree of dysfunction of the body organs, besides a confirmed infection was suggested. It is called the SOFA score and it ranges from zero to four and up. Patients with a score two and above are considered confirmed sepsis patients. Then how can it be treated and reduce the severe outcomes? By identifying the causing microorganism and then the use of a suitable antimicrobial. One of the most common microorganisms that can cause sepsis is the bacteria called *E. coli*. There are a good and a bad *E. coli*. The good ones normally live in our digestive tract. They are not harmful if they just stay there as they actually benefit our bodies. Yet, in some cases the *E. coli* may become dangerous. They can exchange toxic materials with other harmful bacteria, or other bad *E. coli.* Moreover, the *E. coli*, and other bacteria can develop mechanisms to overcome the antibiotics so they can survive and continue harming the body even with the administration of antibiotics. Those mechanisms are known as antibiotic resistance. There are many antibiotics that were produced to try to defeat *E. coli*, but they are smart enough to understand how the antibiotics work and develop a resistance against them. To understand how *E. coli* can do this, a detailed look at their genetic material was needed. Everything that happens in a living cell, is simply a translation to what was written (coded) on the genetic material which is called the DNA. In the present study, an advanced technology called next-generation sequencing was used to read (sequence) the full genome in extracted DNA from *E. coli* isolates that was collected from sepsis patients. To understand the information obtained from sequencing of the isolates, a series of bioinformatics computational tools were used. The used tools could identify *E. coli* from other species, identify genes responsible for antibiotic resistance, and genes responsible for making *E. coli* virulent (bad) instead of being commensal (good). The obtained information can assist in research, surveillances, and new drug strategies.

Table of Contents

Abbreviations 1

Introduction 2

Definitions of sepsis 2

Sepsis in Sweden 2

Detection of the causative microorganism and their resistance pattern 3

Escherichia coli 3

Next-generation sequencing (NGS) 5

Aim 7

Materials and Methods 8

Collection of isolates 8

Species identification (phenotypic methods) 8

Antibiotic susceptibility testing 8

DNA extraction 9

Library preparation, sequencing and NGS data generation 9

Merging of raw data files 9

Quality evaluation 10

Pre-processing (trimming) 10

Assembly and evaluation of quality 10

Genotypic analysis (Center for Genomic Epidemiology) 11

Species identification (genotypic method) 11

1928Diagnostics genotypic analysis 11

Statistical analysis 11

Ethical approval 12

Results 13

1928Diagnostics platform 13

Quality evaluation, trimming, and assembly 13

Species identification, MLST, and phylogroup 14

Antibiotic resistance genes 15

Virulence genes 16

SNP analysis (1928Diagnostics) and CSIphylogeny (CGE) 17

Discussion 18

Conclusion 22

Ethical aspects, gender perspectives, and impact on the society 23
Future perspectives 23
Acknowledgments 24
References 25
Appendix 1: Freely available tools for NGS analysis 35
Appendix 2: Material and Methods 36
Merging Fastq files 36
Quality evaluation of the reads: FastQC 36
Fastq files trimming: Trimmomatic 40
Assembly: Unicycler 40
Quast Results 41
Appendix 3: Species identification 44
Appendix 4: Antibiotic resistance predictions 46
Appendix 5: Virulence genes 50
Appendix 6: SNP and CSIphylogeny trees 52

Abbreviations

Aminoglycosides Modifying Enzymes AME

Antibiotic Susceptibility Testing AST

Average Nucleotide Identity ANI

Center for Genomic Epidemiology CGE

Core Genome Multi-Locus Sequence Typing cgMLST

European Committee on Antimicrobial Susceptibility Testing EUCAST

Extended-Spectrum Beta-lactamase ESBL

Extra-intestinal Pathogenic *Escherichia Coli* ExPEC

Guanine – Cytosine content GC content

Macrolides, Lincosamide, and Streptogramin MLS

Matrix-assisted laser desorption ionization-time of flight mass spectrometry MALDI-TOF MS

Multi-Locus Sequence Typing MLST

National Center for Biotechnology Information NCBI

Next Generation Sequencing NGS

Quality Assessment Tool for Genome Assembly QUAST

Sequence Type ST

Sequencing by Synthesis SBS

Sequential Organ Failure Assessment score SOFA

Single Nucleotide Polymorphism SNP

Systemic Inflammatory Response Syndrome SIRS

The comprehensive Antibiotic Resistance genes Database CARD

Virulence Factor VF

Whole Genome Sequencing WGS

Introduction

Definitions of sepsis

A reliable definition of sepsis is a systemic response to infection with the presence of some degree of organ dysfunction (Vincent et al., 2013). Several definitions of sepsis were proposed in the last few decades (Bone et al., 1992; levy et al., 2003), yet the diagnosis of sepsis remained challenging. The most applied definition was the sepsis-2 definition, which defined sepsis as Systemic Inflammatory Response Syndrome (SIRS) plus infection (levy et al., 2003). The patient must show two or more SIRS criteria to be diagnosed with SIRS. The SIRS criteria were established as follows: tachycardia (heart rate > 90 beats/minute), tachypnea (respiratory rate > 20 breaths/minute), fever or hypothermia (body temperature > 38 ℃ or < 36 ℃) and leukocytosis, leukopenia (white blood cells > 1200/ mm^3, < 4000 / mm^3) (levy et al., 2003; Singer et al., 2016; Vincent et al., 2013). After around two decades, it was realized that the sepsis-2 definition was unhelpful, because the change in heart rate, body temperature, and blood counts are signs of inflammation or trauma and are not necessarily signs for sepsis. It was also understood that SIRS is not a marker for sepsis (Singer et al., 2016; Torsvik et al., 2016; Vincent et al., 2013). In 2016 the sepsis-3 definition was introduced, and the Society of Critical Care Medicine and the European Society of Intensive Care Medicine conference suggested the Sequential Organ Failure Assessment score (SOFA) to evaluate the degree of organ dysfunction. Sepsis-3 was defined as an infection with the possibility of organ failure and a SOFA score ≥ 2 (Rhodes et al., 2017; Singer et al., 2016). The SOFA score was developed to describe as accurately as possible the degree of an organ or multi-organ failure or dysfunction over time in an individual patient or group of patients (Vincent et al., 1996). The score ranges from 0-4 and up to evaluate each of the following six systems: respiratory, cardiovascular, hepatic, coagulation, renal, and neurological systems, with each increase in the score, indicates a further deterioration in the organ function (Vincent et al., 1996). The sepsis-3 definition is the most used to diagnose sepsis today.

Sepsis in Sweden

The diagnosis of sepsis is not standardized, and it may differ between different regions in the world and according to the studied population (Ljungström et al., 2019; Mariansdatter et al., 2016; Singer et al., 2016). In a retrospective study that was carried out over a year in Sweden, it was found that the incidence rate of sepsis was 780/100000 according to sepsis-3 definition (Mellhammar et al., 2016). However, in the Skaraborg Sepsis study that was performed in 2011-2012, the incidence rate of community-onset sepsis in the Skaraborg area in southern Sweden was 838/100000 according to the sepsis-3 definition (Ljungström et al., 2017; Ljungström et al., 2019).

Sepsis requires fast and reliable diagnosis to save patients from further deterioration of their organs that can reach multi organs failure, septic shock, and death (Angus et al., 2001). The fast and accurate identification of the causative microorganism, guide the rapid initiation of the suitable antimicrobial therapy. The quick action is effective to manage the life-threatening infection that causes sepsis or septic shock (Ferrer et al., 2014; Seymour et al., 2017; Zhang et al., 2015). Many types of pathogens can lead to sepsis bacteria, virus, fungi, or parasites (Fan et al., 2016). However, gram-negative bacteria are among the most observed causes of sepsis (Guo et al., 2015; How et al., 2015; Ljungström et al., 2019).

Detection of the causative microorganism and their resistance pattern

The gold standard for detecting microorganisms in collected samples is conventional laboratory culture methods. For example, blood cultures were thought to be a sensitive and reliable method to detect bacteremia (Coburn et al., 2012), where bacteremia is defined as a true-positive culture. Though, it is common not to rely on cultures results due to the following reasons (Coburn et al., 2012; Phua et al., 2013; Vincent et al., 2015;): (i) They are time consuming, a culture might take up to five days to recognize the microorganisms in a sample and their resistance pattern, (ii) Cultures can produce either false negatives due to the empirical antibiotic administration or administration of antibiotic before diagnosis of sepsis, or false positives due to a contaminated sample, (iii) cultures generate a huge amount of waste because different types of microorganisms need different types of media to grow in. Besides, other commonly used phenotypic methods proved occasionally inaccurate identification of the causative microorganisms for many diseases. For example, the protein signature-based matrix-assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS) has misidentified the causative microorganisms on several occasions including samples that were collected from sepsis patients (Kawahara-Matsumizu et al., 2018; Poonawala et al., 2018; Saxenborn, 2018).

Antibiotic resistance is a major problem that was dangerously spread globally. New antibiotic resistances are emerging and causing treatment failure in several common infections (Frieri et al., 2017). Antibiotic resistance is defined as the ability of a certain bacteria to genetically encode for mechanisms that can inhibit the effect of a potential antibiotic for the sake of survival (Martínez, 2008). The importance to detect the susceptibility of a bacterial isolate to certain antibiotics brought the value of the Antibiotic Susceptibility Test (AST). Currently, the AST is being carried out using phenotypic methods in the microbiology labs, for instance, the disc diffusion method or MALDI-TOF MS (Idelevich et al., 2018). The disc diffusion phenotypic method for AST holds the same disadvantages as conventional culture to detect the causative microorganism in sepsis or any other infectious disease. For example, the disc diffusion method takes a lot of time to determine the resistance pattern and there is a risk of contamination of the used media. The results are highly dependent on the equipment used and the conditions in the microbiology lab, such as PH level in the plates, moisture, and failure to test bacteria that did not grow in satisfactory amount (Khan et al., 2019).

Escherichia coli

Escherichia coli (E. coli) is a type of bacteria that can cause sepsis or other human infections (Manges et al., 2019). It is a gram-negative straight rod, facultative anaerobic bacteria that originates from the family *Enterobacteriaceae* (Tenaillon et al., 2010). *E. coli* is a heterogeneous group of strains and mostly they are unharmful to the human host, it has been living in the intestines of a healthy human (and other warm-blooded animals) for decades in a good health and mutual benefits. Nevertheless, some of the *E. coli* strains acquired specific virulence attributes that can cause severe infections with lethal outcomes (Johnson & Russo, 2005; Russo & Johnson, 2003). It is interesting how some strains of *E. coli* transformed from commensalism to pathogenicity and caused both intraintestinal and extraintestinal diseases to millions of people yearly (Manges et al., 2019; Tenaillon et al., 2010). Extraintestinal pathogenic *E. coli* (ExPEC) strains are responsible for many diseases globally both in healthcare-associated infections and community-acquired infections. Their versatility, complex phylogenetic structure, and harboring a wide range of virulence genes make them cause non-intestinal infections such as urinary tract infections, bloodstream infections, and prostate infections (Manges et al., 2019; Sarowska et al., 2019). The

intestines are thought to be the primary source of the ExPEC strains from where they can emerge to other locations on their human host (or other warm-blooded animals) and cause non-intestinal infections (Johnson & Russo, 2005; Tenaillon et al., 2010). Also, ExPEC can transmit to humans from zoonotic origin through, for example, contaminated food or water (Ombarak et al., 2016). The ExPEC strains were associated with developing new and troubling antibiotic resistance genes (Kumarasamy et al., 2010; Liu et al., 2016). Furthermore, it was found in a study that the prevalence of fluoroquinolones resistance in *E. coli* was related to a higher mortality rate in adult sepsis patients (Goldstein et al., 2019). The extended-spectrum beta-lactamase-producing (ESBL) *E. coli* bacteria is known to overrepresent multi-drug resistance genes and can be resistant to a wide range of broad-spectrum antibiotics, which may lead to treatment failure (Ojer-Usoz et al., 2017), and in turn result in more severe outcomes (Namikawa et al., 2017). According to the European Committee on Antimicrobial Susceptibility Testing report (EUCAST, 2020), the 30 countries of the European Union/European Economic Area reported 163005 of *E. coli* clinical isolates. More than half of these isolates (57.1%) were resistant to at least one of the antibiotic groups included in the surveillance in 2019, for example, carbapenems, third-generation cephalosporines, aminopenicillin, aminoglycosides and fluoroquinolones. Although there was a noticeable decrease in the resistances against aminopenicillins, fluoroquinolones, and aminoglycosides among *E. coli* clinical isolates in the reported data, there was also an increase in the third generation cephalosporine resistance and a slight increase in carbapenems resistance. Although those statistics show an improvement in overcoming the *E. coli* resistance to certain antibiotics in comparison to the period 2002-2018, still the road is long to improve the susceptibility of *E. coli* to several antibiotics.

It was revealed, using Multi Locus Sequencing Type (MLST), that *E. coli* strains are divided unequally into several phylogenetic groups A, B1, B2, C, D, E, F, and clad I (Sarowska et al., 2019; Dale & Woodford, 2015). Virulence factors (VF) genes that can cause extraintestinal diseases were mostly harbored by strains located in B2 and D phylogenetic groups. In addition to group E which is related to group D and group F which is related to group B2 (Sarowska et al., 2019). Regardless of the overrepresentation of VF genes in ExPEC strains that are causing infections, there are only a few traits to differentiate between commensal *E. coli* and pathogenic ExPEC (Alqasim et al., 2014). Furthermore, until now there is no classification of the severity of ExPEC strains according to VFs and their relation to MLST typing (Manges et al., 2019).

The mechanisms behind the transmission and the selection of resistant clones of ExPEC are still not well-understood and need further research (Riley, 2014). One of the well-known differences between the commensal *E. coli* and ExPEC is the emergence of pandemic lineages such as *E. coli* sequence type 131 (ST131) and its sub-lineages that are known to be associated with high drug resistance (Kallonen et al., 2017; Price et al., 2013; Stoesser et al., 2016). Other studies used WGS to explore the genomic basis of other non-ST131, for example, VFs and the presence of resistance genes were investigated in ST73 (Alhashash et al., 2016; Kallonen et al., 2017), and others investigated ST95 *E. coli* isolates from bloodstream infections (Stephens et al., 2015). More studies are still needed to investigate the epidemiological and biological causes of ExPEC pandemic lineages and to discover the success of more STs lineages. Data about antimicrobial resistance and VFs can lead to a proper selection of treatment, understanding resistance regimens, and identifying areas of interventions that can monitor the emergence of antimicrobial resistance and the incidence of infection lineages (World Health Organization *[WHO]*, 2014).

Next-generation sequencing (NGS)

Next-generation sequencing (NGS) or massively parallel sequencing are terms used to describe technologies that are being used to sequence DNA and led to a revolution in genomic research (Behjati & Tarpey, 2013). The NGS technologies reduced the cost of sequencing drastically, besides being high-throughput technologies. Additionally, the ability of multiplexing samples, which is adding different bar codes to different samples and then to be pooled and sequenced together, also contributed to the reduction of cost (Mardis, 2017). In microbiology, the main usefulness of NGS is to replace the traditional methods of identifying the microorganism, such as morphology, staining properties, and metabolic characteristics with a genomic description. The genomes of a microorganism contain valuable information about drug sensitivity, antibiotic resistance, and virulence (Behjati & Tarpey, 2013; Besser et al., 2018). The utilization of culture-independent molecular methods, such as NGS revolutionized research of microbial communities. The NGS technology proved its usefulness as an unbiased analysis tool to identify organisms in bloodstream infections, where the classic laboratory approaches might fail (Brenner et al., 2018; Decker et al., 2017; Grumaz et al., 2016).

Illumina sequencing technology

Illumina short reads producing platforms sequence DNA by the reversible termination sequencing by synthesis (SBS) method (Alekseyev et al., 2018; van Dijk et al., 2018). A summary of the SBS process by Illumina is demonstrated in Figure 1. Illumina sequencing platforms offer several advantages that make them preferably used in several biomedical research. It is a high throughput technology that can produce millions of reads. Besides, the accuracy that can reach 99.7%, or error rate per base is 1/1000 (Goodwin et al., 2016). Also, the platforms are capable of paired-end sequencing. The advantage of paired-end sequencing allows the users to sequence the DNA from both ends, which in turn generate more data, and high-quality reads with better accuracy (Ambardar et al., 2016; Reuter et al., 2015). Additionally, paired-end sequencing facilitates the identification of genome rearrangement, such as, insertion, deletion, and inversion. Yet, Illumina platforms have some limitations. The platforms produce relatively short reads (~50-250 bp) to keep the quality as high as possible because it was found that reads longer than 500bp can have a higher error rate (Tan et al., 2019). Moreover, the reads may contain artifacts such as adapter sequences, duplicated reads, or low-quality reads (Goodwin et al., 2016; Kircher et al., 2011). The quality of reads decreases with the increase in cycle numbers; Due to the degradation of chemicals used by the sequencer machine (Kircher et al., 2011; van Dijk et al., 2018). The platform produces a huge number of short reads data, which make the needed bioinformatics analysis is the bottleneck of the Illumina sequencing process. Still, the valuable information that can be achieved by Illumina sequencing, at such affordable cost and accuracy is undeniable (Goodwin et al., 2016; Mardis, 2017).

The high-quality reads produced by Illumina platforms can only be interpreted and understood with the help of bioinformatics tools for NGS analysis (Yohe & Thyagarajan, 2017). The sequencer generates large raw files that are referred to as fastq files. Those files can reach hundreds of megabytes that need to be checked for quality, preprocessing for the removal of undesired reads, and assembly to generate the fasta files. The assembly can either be *de novo* or mapping to a reference genome. The assembly tools output is a set of contigs (contiguous sequences), each contig represents a part of the sequenced template. The longer the contigs the better they can be annotated as *de novo* or mapped to a reference genome (Heydari et al., 2017). This series of bioinformatics analysis of NGS data is called a pipeline, and most pipelines require expert

knowledge in applying bioinformatics to identify significant information from the reads (DePristo et al., 2011). Several tools are freely available to analyze short reads, some of these tools are summarized in Appendix 1, Table 1. Recently cloud-based bioinformatics platforms to analyze raw NGS data are now available. These cloud-based platforms can provide reliable results from only uploading the raw fastq files in literally less than an hour. The 1928Diagnostics is a cloud-based platform that is designed to analyze the WGS of several types of microorganisms to detect gene mutations, toxin genes, and antimicrobial resistance genes from raw reads (1928Diagnostics, 2020). The 1928Diagnostics cloud-based platform offers different pipelines to analyze different species of microorganisms. In addition, the platform supports paired-end short reads of Illumina platforms and other NGS long reads platforms. Furthermore, the platform provides epidemiological marker typing required for MLST and core genome MLST (cgMLST). All these advantages from the 1928Diagnostics platform do not need experience in coding either scripting bioinformatics skills, which allow NGS data to be more utilized in clinical settings and research purposes. In this thesis an in-house bioinformatics pipeline that was developed by the systems biology center team at the University of Skövde was used to analyze WGS of *E. coli* clinical isolates (I Shemirani et al., 2019; Saxenborn et al., 2019), and the obtained results were compared with the results generated by the 1928Diagnostics *E. coli* pipeline.

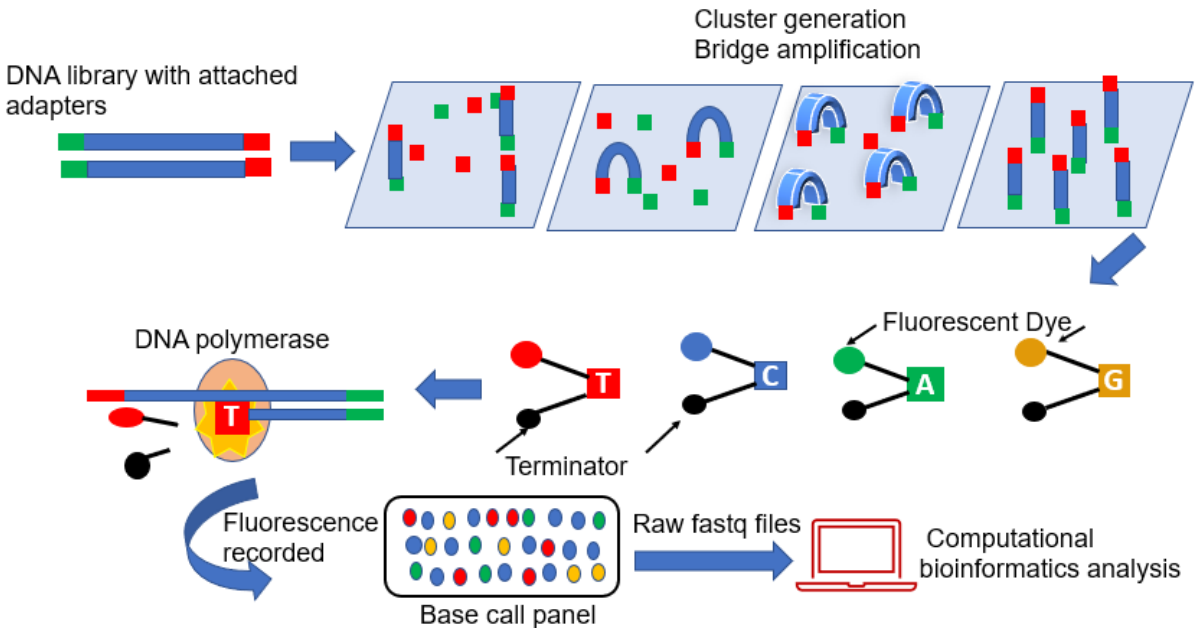


Figure 1. Shows the process of SBS by an Illumina sequencer. Illumina requires the extraction of DNA before sequencing. DNA then is fragmented into smaller sizes, and specific adapters are attached to the fragmented template (and bar codes if applicable). This process is called library preparation. The prepared library is attached to a solid surface (flow cell) using the attached adapters. This is followed by bridge amplification using polymerase chain reaction before the actual sequencing. The purpose of bridge amplification is template enrichment to increase its signal to be detected by the sequencer (cluster generation). The next step is the actual sequencing. The platform utilizes the reversible termination technology. This method is based on the detection of emitted fluorescence after the incorporation of the complementary fluorescently labeled reversible terminator nucleotide, then cleavage of the reversible fluorescent dye by DNA polymerase. Before the cleavage of the reversible fluorescent dye from the reversible terminator nucleotide, the signals are imaged and are stored in a base-calling panel. The sequencer then generates the raw fastq files by recalling the stored signals and give them quality scores. The quality scores represent how confident the sequencer recalled a specific base. The sequencer will try to remove the adapters sequences from the raw fastq files, but usually this process is not fully successful, and part of the adapters sequences will remain in the generated reads. The following step is the computational bioinformatics analysis.

Aim

The primary aim of the project was to study the biodiversity of *E. coli* genomes that were collected from clinical isolates from adult patients with suspected community-onset sepsis or septic shock. The dataset used in this thesis was WGS of isolates extracted from patients with SOFA scores 2-3. The isolates were previously identified as *E. coli* by using classical laboratory methods during the "Sepsis study Skaraborg" that was conducted from September 8, 2011, to June 7, 2012, Skaraborg, Sweden (Ljungström et al., 2017; Ljungström et al., 2019). The Secondary aims for the project was to study biodiversity and pathogenicity characteristics of the genome such as: (i) Identify antibiotic resistance genes, (ii) Identify virulence and toxin genes, (iii) Identify phylogenetic typing/genotyping traits, (vi) classification of ExPEC by virulence in sepsis patients (SOFA score 2-3) or by MLST if possible, then evaluation the utility of the results in clinical settings. In this thesis, short reads of paired-end NGS data of the WGS of *E. coli* isolates were analyzed using two bioinformatics pipelines, the first is the in-house developed pipeline, and the second is the 1928Diagnostics *E. coli* pipeline. In comparison to other studies that were conducted to identify the causative microorganisms in sepsis patients, this study provides a detailed analysis of the biodiversity of ExPEC, its pathogenicity characteristics, and traits, regardless of the isolates resistance profiles. The study may provide an evidence of the usefulness of utilizing WGS-based analysis as an unbiased analysis tool to identify the characteristics and pathogenic traits from clinical isolates of bacteria. The obtained information can be greatly useful for surveillances and tracking the emergence of antibiotic resistances.

Materials and Methods

Collection of isolates

The dataset that was used in this thesis was WGS data of *E. coli* clinical isolates. The isolates were collected during the prospective observational "Sepsis study Skaraborg" that was conducted from September 8, 2011, to June 7, 2012, in Skaraborg hospital in cooperation with Unilabs, and the systems biology center at Skövde University in Västra Götland region in southwestern, Sweden (Ljungström et al., 2017; Ljungström et al., 2019). The Skaraborg hospital is a 640-bed secondary care hospital that serves around 256700 inhabitants of which 206900 are adults (December 31, 2011). During the study period, all adult patients (≥18 years) who were admitted from the emergency department and were clinically evaluated for developing community-onset sepsis, or septic shock and administered intravenous antibiotics within 48 hours after admission were asked to be included in the study. In total 1800 pathogenic bacterial isolates were collected from different sample locations before empirical antibiotic administration. These isolates were cryopreserved at the time of recovery by transferring colonial material to MicrobankTM vials (Pro-Lab Diagnostics, Ontario, Canada) and stored at -80°C. The isolates were collected from four different locations, blood, upper respiratory tract, urine, and general cultures for wounds (Table 1). Some isolates were collected from different sample locations or at different time points from the same patient. In total 18 patients had two or more isolates collected from different sample locations. Fifteen patients had two isolates collected from blood and urine at the same time point. One patient had two isolates collected both from urine at different time points. One patient had three isolates, one of them collected from blood, while the other two collected from urine at different time points. One patient had three isolates collected from blood, urine, and upper respiratory tract at the same time point. Due to the change in color and/or appearance at the culture plates, four samples were analyzed independently although they were obtained from only two cultures (EC1904VIT and EC1904GRA – EC490 and EC490nr2).

Species identification (phenotypic methods)

Species identification of isolates was performed in the clinical microbiology laboratory Unilabs at Skaraborg Hospital, Skövde, on a MicroFlex LT mass spectrometer (Bruker Daltonics, Germany) with BioTyper software v2.0 using default parameter settings as part of the routine clinical practice. For correct species identification, spectral scores above 2.0 were used as a cut-off. At the time of the study, the Bruker microorganism database MBT Compass Library DB-4110 (Bruker Daltonics, Germany) released in April 2011 was used. Based on culture and followed by MALDI-TOF MS (DB-4110) for species identification, 422 isolates were identified to be *E. coli*, and some of the isolates were classified as ESBL *E. coli*.

Antibiotic susceptibility testing

Antibiotic susceptibilities were determined *in vitro* in the clinical microbiology lab, Unilabs, Skövde by accredited laboratory methods using the disc diffusion method on Mueller-Hinton media according to EUCAST guidelines (www.eucast.org). As the phenotypic AST was performed as part of the routine clinical practice, the sample type mainly determined which antibiotics that was to be tested for each bacterial isolate.

DNA extraction

Out of the 422 collected *E. coli* isolates, 420 isolates were recovered after freezing and could be re-cultured. DNA of *E. coli* isolates were extracted from pure cultures using the MagNA Pure 96 DNA and Viral NA Small Volume kit (200 µl sample) on a MagNA Pure 96 instrument according to the Pathogen Universal 200 protocol (Roche Diagnostics, Switzerland). Extracted DNA concentrations were measured by using Qubit 3.0, and purities were determined by NanoDrop spectrophotometer (Thermo Fisher Scientific, USA), then samples were kept in -20°C.

Library preparation, sequencing and NGS data generation

The extracted genomic DNA samples were sent for paired-end sequencing to SciLifeLab in Solna, Sweden. NexteraXT libraries and quality control were performed according to the manufacturer´s protocol (Illumina, San Diego, CA). The samples were sequenced using the Illumina Hiseq 2500 platform using the high-throughput protocol for bacterial genomes. The generated raw Fastq files were uploaded at the Skövde University as zipped files (.gz) to be analyzed. The raw data contained information about sequencer identifier, sequences (reads), and quality scores. The data was divided into three groups according to the SOFA score of the patients from which the isolates were originated. In this thesis, the analysis focused on isolates collected from patients with SOFA scores 2 and 3 who were diagnosed with confirmed bacterial sepsis according to the sepsis-3 definition. In total 150 isolates WGS were included in this thesis.

Table1. The demographic information for the patients from whom the *E. coli* isolates were collected.

SOFA Score	**SOFA score 2**	**SOFA score 3**
Isolates Count (%)	77 (51%)	73 (49%)
Gender of patients (%)	Female 54 (70%) Male 23 (30%)	Female 46 (63%) Male 27(37%)
Mean age in years (sd*)	Female 75 (±14.7) Male 76 (±14.6)	Female 80 (±11.1) Male 77 (±9.3)
Sample location (counts)	Urine 57 Blood 17 Wound 3	Urine 58 Blood 14 Upper respiratory tract 1

*sd= Standard deviation

Merging of raw data files

The paired-end reads of the *E. coli* isolates were divided into several files for each read type during sequencing, for instance, the forward read for a certain isolate was divided into two files, as well as the reverse read. To have one sequence file for each read (one sequence file for forward reads and one sequence file of revers reads) for each isolate, several files that belong to the same read type were merged using the 'zcat' command in Linux operating system (Ubuntu 5.8.0-45generic), before proceeding to the next step (Ward, 2021). The used code for merging files is presented in Appendix 2, Figure 1. This was the first step before employing the in-house developed bioinformatics pipeline and the 1928Diagnostics *E. coli* pipeline to analyze the raw fastq files. A summary of the workflow by the two bioinformatics pipelines is presented in Figure 2.

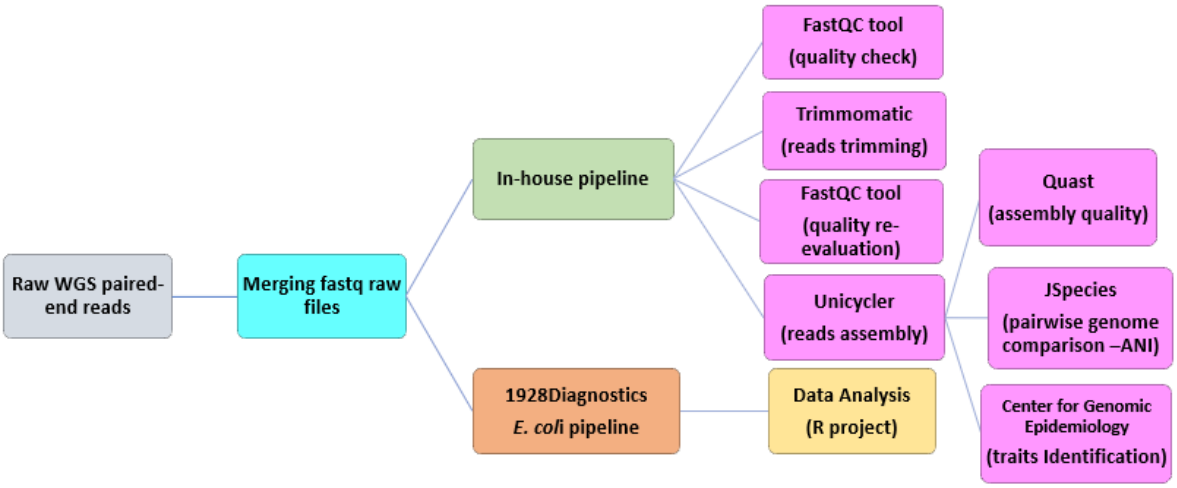


Figure 2. Shows a summary of the workflow used for the WGS-based analysis. Two bioinformatics pipelines were used to analyze the raw WGS paired-end reads. The pink boxes represent the in-house pipeline, and the orange box represents the 1928Diagnostics *E. coli* pipeline. The obtained results were statistically analyzed by R software to identify their significance.

Quality evaluation

The quality of the raw Fastq files was evaluated using the freely available FastQC tool (version 0.11.9) for Windows (Babraham bioinformatics, 2020; Cock et al., 2010). The modules that were used to check the quality of the raw Fastq files were per base sequence quality, per sequence quality score, per base sequence content, per sequence GC content, and adapter content (Appendix 2, figures 2-6). Assessing the raw reads guided the trimming and quality filtering parameters. After trimming, the FastQC tool was used again to re-evaluate the trimmed files (Appendix 2, figures 2-6).

Pre-processing (trimming)

For data trimming of technical sequences (adapter sequences) and quality filtering of reads, the freely available tool Trimmomatic (version 0.39) for windows was used (Bolger et al., 2014). The parameters that were used to trim adapters and quality filtering were: (i) ILLUMINACLIP function was used to trim the Nextera PE adapters with the following parameter: not allowing mismatches, the palindromeClipThreshold accuracy was set to 30 and the simpleClipThreshod was set to allow alignment of at least ten bases of the adapter sequence against a read. (ii) The SLIDINGWINDOW option was used to filter out low-quality bases. The parameters were set to calculate the mean quality score of each four bases and clip if the mean Phred quality score was below 20. (iii) The option MINLIN was used to drop any read had length below 30 bases (Appendix 2, figure 7). Trimmomatic generates four files for each isolate. Two files for the survived trimmed paired reads (forward and reverse reads), and two files for the survived unpaired reads, which were survived without their mates (unpaired forward reads and unpaired reverse reads). Only paired reads files were used as inputs for the assembly tool.

Assembly and evaluation of quality

The trimmed fastq files were assembled as *de novo* using the Unicycler assembly tool (version 0.4.8) for Linux (Ubuntu 5.8.0-45generic; Wick et al., 2017). The Unicycler assembly tool is a freely available tool to resolve bacterial genomes from short, long, or hybrid reads (Wick et al., 2017). The unicycler assembler uses St. Petersburg genome assembler (SPAdes version 3.13.0) (Bankevich et al., 2012) to construct De Bruijn graph. Unicycler was used with default settings allowing read error correction by SPAdes (Appendix 2, figure 8). The fasta files were successfully

generated and the quality of the assembled contigs was evaluated using the Quality Assessment Tool for Genome Assembly (QUAST; version 5.0.2) (Gurevich et al., 2013; Wick et al., 2017). QUAST is a freely available tool to evaluate the quality of the assembly. It was used with default settings and with the reference genome *E. coli* UTI89 (NCBI accession number NC_007946.1).

Genotypic analysis (Center for Genomic Epidemiology)

For genotypic analysis of the fasta files, freely available tools in the Center for Genomic Epidemiology website (CGE, 2021) were used to identify the genetic traits of the isolates. To identify antibiotics resistance genes, ResFinder (version 4.1) was used with default settings. The default settings were a minimum of 90% identity threshold and a minimum gene length of 60%. ResFinder 4.1 was used to identify acquired antibiotic resistance genes and mutations in chromosomal genes that can cause antibiotic resistance using the PointFinder option in ResFinder tool (Bortolaia et al., 2020; Zankari et al., 2017). To identify acquired virulence genes, VirulenceFinder (version 2.0) was used with default settings. The default settings were a minimum of 90% identity threshold and a minimum gene length of 60% (Joensen et al., 2014; Tetzschner et al., 2020). For MLST typing, the tool MLST (version 2.0) was used with the Achtman schema (Larsen et al., 2012). The typing is based on seven housekeeping genes (*adk, fumC, gyrB, icd, mdh, recA, and purA*) that are coded in the chromosome of *E. coli* (Wirth et al., 2006). CSIphylogeny tool was used to create a clustering tree of all the clinical isolates compared to reference genome *E. coli* UTI89 (Kaas et al., 2014).

Species identification (genotypic method)

Species identification was performed on the fasta files using JSpeciesWS freely available web services (JSpeciesWS, 2021). Pairwise genome comparisons were performed with reference genome *E. coli* UTI89. The tool measures the probability of a group of genomes to be belonging to the same species (or in comparison to a reference genome) using the average nucleotide identity (ANI) method. The ANI method that was applied on the studied isolates was based on Blast+ (ANIb). The cutoff value of similarity is 95%. Genome drafts with similarity ≥ 95% were considered to be belonging to the same species and < 95% similarity were considered a different species (Richter et al., 2009; Richter et al., 2016).

1928Diagnostics genotypic analysis

The raw Fastq files (untrimmed) were uploaded to the cloud-based platform 1928Diagnostics and were analyzed by a specifically developed pipeline to analyze *E. coli* paired-end reads (1928Diagnostics, 2020). The *E. coli* pipeline offers species identification, quality measurements of the reads, phylogenetic group, identification of antibiotic resistance genes, identification of virulence/toxin genes, MLST typing, and *fimH* typing. Most of the results were automatically integrated into one excel sheet which was downloaded from the platform webpage. In addition to the previous tools, Single Nucleotide Polymorphism (SNP) analysis was performed with reference genome *E. coli* UTI89.

Statistical analysis

Statistical analysis was performed using R (version 4.04; R Foundation for Statistical Computing, Austria). Descriptive statistical analysis was used to compare the genotypic results obtained from the in-house pipeline, and the 1928Diagnostics *E. coli* pipeline, with the phenotypic results from the clinical laboratory. The presence of antibiotics resistance genes and virulence genes were presented as mean counts with their confidence interval of 95% (CI 95%). Chi-square test was used for comparisons between groups, SOFA score 2 and 3 to evaluate the significant difference

between the two groups for the frequency of MLST type, presence of resistance and virulence genes. For the comparisons of the counts of resistance genes detected by 1928Diagnostics and ResFinder, paired t-test or Wilcoxon signed-rank test was used. For virulence genes results, One-way ANOVA or Kruskal Wallis test was used determine the significant difference in means of virulence genes distribution in three STs, ST73, ST95, and ST131. Normality and equal variance were checked by Shapiro-Wilk test and Levene test, respectively. P values < 0.05 were considered significant (Ross, 2014).

Ethical approval

This thesis is a continuation of the research that started with the Skaraborg sepsis study which was approved by The Regional Ethical Review Board in Gothenburg (376-11) (Ljungström et al., 2017). In this thesis, the genomes of *E. coli* clinical isolates were investigated, and bacterial isolates do not require informed consent. Still, all included patients in the "Skaraborg sepsis study" have provided written informed consents (Ljungström et al., 2017). No personal information such as names or personal security numbers were published in this thesis.

Results

1928Diagnostics platform

The 1928Diagnostics platform *E. coli* pipeline guided the analysis during this thesis. After the 150 raw fastq files (untrimmed) were uploaded platform, the *E. coli* pipeline excluded 14 isolates from the analysis. Three isolates were excluded because their files were highly contaminated with other species, and eleven isolates were excluded due to low sequence depth, below 30X (Appendix 3, Table 1). The excluded isolates from 1928Diagnostics pipeline were excluded from further analysis by the in-house pipeline, and those were not analyzed after trimming. Further analysis was performed with these 136 isolates. The obtained results from the *E. coli* pipeline facilitated narrowing down the number of reference genome candidates. The reference genome *E. coli* UTI89 (NCBI accession number NC_007946.1) was selected and used as a reference genome when needed in the analysis steps. The selection of the reference was based on its similarity to the majority of the isolates in regards of: phylogroup, genome fraction similarity, MLST, and being an ExPEC. The obtained results from the platform included: antibiotic resistance genes, MLST, and phylogroup; In addition to the results from the SNP analysis. VFs and toxin genes could not be detected by the 1928Diagnostics platform, because currently, the platform can detect only virulence/toxin genes for intestinal *E. coli* strains but not ExPEC.

Quality evaluation, trimming, and assembly

Using the FastQC tool the quality of the raw reads was evaluated. A common quality paradigm was noticed and according to the results from the FastQC tool the trimming parameters were determined, and all isolates were trimmed using Trimmomatic (code is shown in Appendix 2, figure 7). In Appendix 2 (figures 2-6), the output from the FastQC tool for the forward file of one isolate before and after trimming is presented. In the summary output of this file before trimming (Appendix 2, figure2), there were four warnings, per tile sequence quality, per base sequence content, per sequence GC content and sequence duplication levels. The adapter content was the only module that the file failed. After trimming, the quality improved and all adapter contents were removed (see Appendix2, figures 2- 6).

Assembly of the trimmed fastq files was performed using Unicycler assembler (code is shown in Appendix 2, figure 8). The fasta files for all isolates were successfully generated and the quality of the assembly was checked by Quast with the reference genome *E. coli* UTI89. The parameters that were considered to evaluate the quality of assembly were: Total length of the assembly, counts of contigs, and genome fraction (%), GC content (%), N50, L50, and NG50 (Appendix 2, Table 1). The mean values of total length, genome fraction (%), and GC content (%) with 95% CI is shown below in Table 2. The obtained results were satisfactory, and the analysis continued.

Table 2. The mean values with their 95% CI of Total length, GC content and genome fraction of all the studied *E. coli* isolates.

Parameter	Mean (95% CI)
Total length (bp*)	4960803 bp (4932575 bp-4989030 bp)
GC content (%)	50.7% (50.5% - 50.8%)
Genome fraction (%)	84.7% (83.7% - 85.8%)

*bp= Base pair

Species identification, MLST, and phylogroup

The 136 isolates studied in this thesis were previously identified by MALDI-TOF MS (DB-4110) as *E. coli.* Genotypic species identification was performed using ANIb through JSpeciesWS web services with the reference genome *E. coli* UTI89. All the 136 isolates were identified as *E. coli*, with a mean similarity to the reference genome of 97.8% (95%CI 97.6%- 98%). The paired isolates (EC490- EC490nr2, EC1904VIT-EC1904GRA) all were identified as *E. coli* by ANIb. Each pair showed the exact similarity to the reference genome. These results match species identification results from 1928Diagnostics pipeline. Phylogroups were predicted by the 1928Diagnostics pipeline, and the dominant phylogroup was B2 64% (n=86), followed by B1 and A 10% each (each n= 13), D 7% (n=9), F 5% (n=7), C 3% (n=4), and two isolates with unknown phylogroup. The isolates (EC490- EC490nr2) belonged to the exact phylogroup B1. Similarly, the isolates (EC1904VIT-EC1904GRA) both belonged to phylogroup A.

MLST typing using Achtman schema was already included in the *E. coli* 1928Diagnostics pipeline and was determined for all isolates. In total, 50 different STs were identified. The most common STs were ST73 (17%), ST95 (9%), ST131 (6%) and novel STs (3%). The cgMLST tree was generated (Appendix 3, figure 1). The tree shows that isolates with similar STs cluster together. All isolates that belonged to ST73, ST95 and ST131 fall under the B2 phylogroup. The MLST tool in CGE was used with the Achtman schema as well, and the results matched 1928Diagnostics MLST results. Though, the MLST tool from CGE could predict the nearest STs for the novel STs. A summary of the distributions of STs among isolates is presented in Appendix 3, Table 2. The two isolates (EC490- EC490nr2) showed the exact MLST (ST224), and the isolates (EC1904VIT-EC1904GRA) showed the same MLST as well (ST48). These results confirm that the change in color or appearance by culture was not related to species, phylogroup, or MLST. There was no evidence to consider the paired isolates (EC490- EC490nr2, EC1904VIT-EC1904GRA) as four independent samples instead of two. For the further analysis, those four isolates were considered as two, as the previous results proved that each pair was the same clinical isolate. The distribution of the most common MLST among the isolates that were collected from patients with SOFA scores 2 and 3 was observed (figure 3). To evaluate the significance of the difference in frequency of distribution of the MLST between the SOFA scores, a Chi-square test was performed. No significance in the frequency of MLST distribution between the two SOFA scores was observed (all p-values > 0.05). In other words, the severity of sepsis could not be related to the MLST of the clinical isolates.

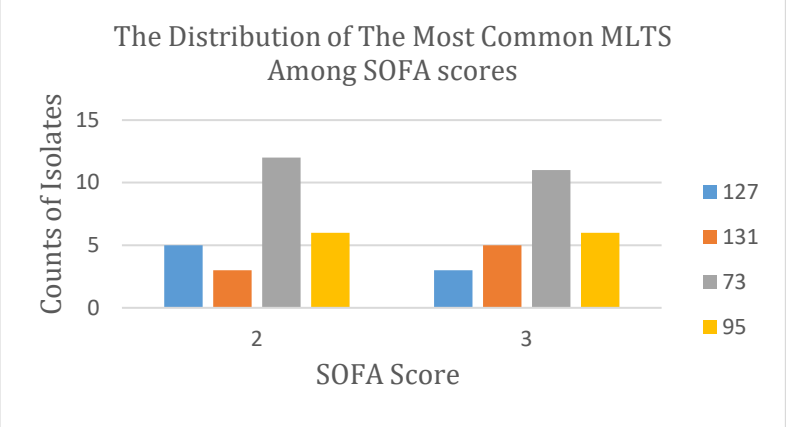


Figure 3. The distribution of the most common MLST among SOFA scores (SOFA score 2 & 3). The Y-axis represents the counts of isolates, and the X-axis represents the SOFA scores. On the left part of the plot SOFA

score 2 is shown and on the right part SOFA score 3 is shown. The bars represent the MLST and the explanatory legend for MLST is shown on the far right of the plot.

Antibiotic resistance genes

The presence of antibiotic resistance genes was determined by 1928Diagnostics *E. coli* pipeline and by ResFinder to identify acquired and point mutation chromosomal resistance genes. Comparisons between both genotypic methods with the phenotypic results were conducted. Both genotypic methods could detect antibiotic resistance genes in the clinical isolates. The genotypic methods showed a high agreement of the predicted number of the resistant isolates. Though, ResFinder predicted more isolates to be resistant (n=6) than 1928Diagnostics. The 1928Diagnostics pipeline predicted that ~28% (n=38) of the isolates were resistant to at least one class of the investigated antibiotic classes. ResFinder predicted that ~33% (n=44) were resistant to at least one class of the investigated antibiotic classes. The antibiotic classes that were investigated by the genotypic methods were: Aminoglycosides, Broad spectrum beta-lactams, Extended-spectrum beta-lactams (ESBL), Chloramphenicol, Colistin, Macrolides, Lincosamide, and Streptogramin (MLS), Quinolones, Sulfonamides, and Tetracyclines (the genes and a comparison between 1928Diagnostics and ResFinder findings are presented in Appendix 4, Table 1). A comparison between the predicted resistant isolates between 1928Diagnostics and ResFinder is presented in Table 3. Besides specific antibiotic resistance genes, the 1928Diagnostics and ResFinder reported that 100% of the isolates (n=134) encoded for multi-drug efflux pumps. The discovered multi-drug efflux pumps genes were unique to each tool (Appendix 4, Table 1).

Table 3. A comparison between the predicted genotypic antibiotic resistance results from 1928Diagnostics and ResFinder for the *E. coli* isolates (n=134).

Antibiotic Class	RR	SS	RS	SR	Discordant across methods
Aminoglycosides	25	108	0	1	1
Broad Beta-lactams	23	109	1	1	2
ESBL	2	131	1	0	1
Chloramphenicol	4	128	0	2	2
Colistin	1	131	0	2	2
MLS	3	131	0	0	0
Quinolones	11	113	0	10	10
Sulfonamides	24	109	0	1	1
Tetracycline	14	120	0	0	0
Trimethoprim	14	120	0	0	0

RR= The count of isolates that showed resistance in both 1928Diagnostics and ResFinder, SS=The count of isolates that showed susceptibility in both 1928Diagnostics and ResFinder, RS=The count of isolates that showed predicted resistance by 1928Diagnostics but were predicted to be susceptible by ResFinder, SR=The count of isolates that showed predicted susceptibility by 1928Diagnostics but were predicted to be resistant by ResFinder.

The highest counts of detected resistance genes by both tools were for Aminoglycosides, followed by resistance genes for Sulfonamides, and then resistance genes for Broad spectrum Beta-lactams. ResFinder predicted significantly higher counts of resistance genes for Quinolones than 1928Diagnostics (p-value < 0.001). Out of all the resistant isolates, 52% coded for multi-drug resistance (resistance genes for three or more antibiotic classes). A summary of the counts for the

detected resistance genes for all the investigated antibiotic classes, by 1928Diagnostics and ResFinder is presented in Table 4. In addition, a detailed description of the resistance patterns in all the resistant isolates is presented in Appendix 4, Table 2. According to ResFinder, it was observed that 100% of the isolates belonged to ST131 (n=8) showed resistance to at least one of the investigated antibiotic classes, while according to the findings from 1928Diagnostics pipeline, 50% of ST131 (n=4) isolates showed a resistance to at least one antibiotic class.

Table 4. The counts of resistance genes detected by 1928Diagnostics and ResFinder for each antibiotic class, from the clinical isolates (n= 134).

Antibiotics class	1928Diagnostics	ResFinder	P-value*
Aminoglycosides	44	51	0.05
Broad Beta-lactams	23	24	0.32
Chloramphenicol	5	6	0.32
Colistin	1	3	0.16
ESBL	3	2	0.32
MLS	3	3	-
Quinolones	12	43	<0.001
Sulfonamides	28	29	0.32
Tetracyclines	14	14	-
Trimethoprim	13	13	-
Total	146	188	-

*The p-values are derived from paired t-test/ or Wilcoxon sign-ranked test, the choice of the test was based on the result of the normality test. The comparison was performed for each antibiotic class by the two methods.

The *in vitro* phenotypic AST predicted that only 5% of the tested isolates were resistant to at least one class of the investigated antibiotics. The *in vitro* phenotypic results were limited on several occasions, for example, colistin, MLS, and Tetracycline were not tested phenotypically at all, so the comparison between the genotypic results and phenotypic results for these classes was not possible. In addition, Aminoglycosides and Chloramphenicol were among the least tested classes. The most tested classes were Broad-spectrum beta-lactams and Quinolones, followed by Sulfonamides and Trimethoprim. Most of the isolates that were tested *in vitro* to be resistant by phenotypic AST, were predicted to be resistant by the genotypic methods, except for a few occasions (n=9). A comparison between the genotypic methods, 1928Diagnostics, and ResFinder, for predicting antibiotic resistance and AST phenotypic method are presented in Appendix 3, Table 3, and Table 4, respectively.

Virulence genes

Virulence genes were detected by the VirulenceFinder tool from CGE web services. The tool detected 85 genes. The count of the detected genes reached 2672 genes in all the *E. coli* isolates (n=134) with a mean count of 31 genes (95% CI 24-39) per isolate. All the isolates coded for a combination of several genes from several gene families. For instance, most of the isolates coded for a combination of adhesins, cell outer membrane, lipopolysaccharide capsule, tellurium iron resistance, iron reuptake and increased serum survival. The count of the most frequent genes in all isolates are summarized below in Table 8. All the detected genes were related to ExPEC species and none of the isolates coded for Shiga toxins.

Table 5. A summary of the counts for the most frequent virulence genes that were detected in the *E. coli* clinical isolates (n=134).

Virulence genes (n=2672)	Counts of genes (%)
Iron reuptake (*air, chuA, iroN, iucC, iutA, sitA*)	385 (14%)
Adhesins fimbriae genes (*A, F, P and S fimbriae*)	313 (12%)
Capsular antigen (*kpsE, kpsM_K15, kpsMII, kpsMII_K1, kpsMII_K23, kpsMII_K5, kpsMII_K52, kpsMIII_K96*)	199 (7%)
Tellurium ion resistance (*terC*)	134 (5%)
Glutamate Decarboxylase (*gad*)	130 (5%)
Outer membrane protein (*ompT*)	111 (4%)
Increased serum survival (*iss*)	106 (4%)

Virulence genes were further investigated in threes STs, that were most frequent in the studied dataset, ST73 (n=23), ST95 (n=12), and ST 131 (n=8). ST73 isolates harbored 670 genes with a mean of 29 genes per isolate (95%CI 27-31). ST95 isolates harbored 304 genes with a mean of 25 genes per isolate (95%CI 24-27). ST131 isolates harbored 162 genes with a mean of 20 genes per isolate (95%CI 17-23). One-way ANOVA reported a significant difference in means of virulence genes counts between the three STs (p-value <0.05). Tukey's test was used as a post-hoc test and all reported p-values were significant (p-value <0.05) which confirms the difference in means. In addition, it was observed that although isolates belonged to ST73 and ST95 showed a more susceptible antibiotic profile than isolates that belonged to ST131, those harbored higher means of virulence genes per isolate. The detected genes in the three STs correlated to the general observation of the common genes summarized in Table 8 and explained in detail in Appendix 5, Table 1. An independent t-test was used to determine if there was a significant difference in the mean virulence gene counts of isolates belonging to SOFA scores 2 and 3. The reported p-value was insignificant (p-value >0.05), and the mean of the gene counts in the two SOFA groups is equal.

SNP analysis (1928Diagnostics) and CSIphylogeny (CGE)

SNP analysis was performed on the 1928Diagnostics platform in comparison to the reference genome *E. coli* UTI89. It was observed that isolates that were collected from the same patient but from different sample locations (ex. blood and urine) at the same time point clustered together, while isolates that were collected from the same patient at different time points, did not cluster together. This may suggest that isolates collected at different time points are not the same *E. coli* isolate type. In addition, it was observed that isolates that belonged to the same STs clustered together. The CSIphylogeny tool from CGE was used to create a clustering tree in comparison to the reference genome *E. coli* UTI89. This tree showed similar results to the SNP analysis from 1928Diagnostics. The obtained trees are presented in Appendix 6, figures 1 and 2.

Discussion

In this thesis, WGS-based analysis was used to describe the characteristics and traits of *E. coli* clinical isolates that were collected from confirmed sepsis patients. The analysis was performed using two bioinformatics pipelines, 1928Diagnostics *E. coli* pipeline and the in-house developed pipeline. The analysis started by uploading 150 raw fastq files to the 1928Diagnostics platform. The *E. coli* pipeline excluded 14 isolates from the analysis, three isolates were excluded because those were highly contaminated with other species (Appendix 3, Table 1), and eleven isolates were excluded due to a low sequence depth (below 30X). The excluded isolates from the 1928Diagnostics platform were excluded from the in-house pipeline analysis as well. Contamination in WGS data is already known to be negatively affecting and misleading the findings from WGS-based analysis (Ballenghien et al., 2017; Cornet et al., 2018). Contamination can be introduced at any step during data generation. For example, contamination may occur during DNA isolation, library preparation, or even from the sequencer itself (Salter et al., 2014). In addition, it was reported that contamination in WGS data is common even when sequenced from pure culture and can lead to erroneous findings (Goig et al., 2020). Therefore, the exclusion of the contaminated isolates was crucial for the accuracy of the results. Although the eleven isolates with low sequence depth were identified as *E. coli* by 1928Diagnostics pipeline, they were excluded due to the low theoretical depth (below 30X). It is highly recommended to only analyze data with theoretical depth of at least 30X and above for meaningful results (Doyle et al., 2020). Species identification was also a part of the in-house pipeline, and it was performed on the 136 isolates by calculating the pairwise ANIb through JSpeciesWS web services. The ANI method is considered one of the most accurate methods for species identification. The DNA-DNA hybridization was considered earlier to be the gold standard for species identification of prokaryotes. Nonetheless, in the era where the complex genomic information is generously available, the DNA-DNA hybridization approach seemed outdated, and an alternative standard was needed. According to previous research, ANI is the best alternative, because it uses pairwise genome comparison to identify similarity of nucleotide identity between a group of genomes (Ciufo et al., 2018; Richter et al., 2009).

It is believed that *E. coli* that fall under phylogroups A and B1 are either commensal or can cause intestinal infections, while ExPEC are expected to fall under B2 or D phylogroup (Köhler & Dobrindt, 2011; Sarowska et al., 2019). The phylogroup analysis results from 1928Diagnostics revealed that 64% of the isolates belonged to B2 as expected for isolates collected from sepsis patients (caused by extraintestinal infection). Surprisingly, A and B1 phylogroup represented ~ 10% each. In other words, A and B1 phylogroup represented one fifth of the whole data. These results can correlate to the suggestion that there is a thin line between commensalism and being an ExPEC (Leimbach et al., 2013; Tetzschner et al., 2020). One of the reasons behind a commensal or intestinal *E. coli* to cause ExPEC infections is the ability of *E. coli* to horizontally exchange mobile genetic elements with related bacteria or even with other species (Sarowska et al., 2019). In a German study (McKinnon et al., 2018), two isolates that belonged to B1 phylogroup and ST85 were collected from blood and urine from a patient with urosepsis. The isolates were found to carry CoIV plasmid that coded for multi-drug resistance. Interestingly, before this study, ST58 phylogenetic group B1 *E. coli* was previously isolated from several avian and mammalian hosts and was not known to be a threat to humans nor animals. Currently, the most used approach for *E. coli* genotyping is typing by MLST. In the present study the Achtman schema was used. The most common detected STs were ST73 (17%), ST95 (9%), and ST131 (6%). Most of the reported infections with *E. coli* from Sweden (blood stream and other infections) were focusing only on

ST131 (Brolund et al., 2014; Fröding et al., 2020; Ny et al., 2016). This is highly understandable, as ST131 strains are well-known for harboring troubling antibiotic resistance genes globally (Manges et al., 2019). This led to under reporting of other successful lineages such as ST73 and ST95. Their success is not primarily related to antibiotic resistance as in the case of ST131 (Alhashash et al., 2016; Kallonen et al., 2017). It was observed in the present study that isolates belonged to the most common MLST ST73 and ST95 had more susceptible antibiotic profiles than isolates that belonged to ST131. These results are in concordance with other European studies that reported predicted genotypic antibiotic resistances. Studies from Denmark, Germany, France, Spain, Ireland ,and the United Kingdom reported ST73 and ST95 beside ST131 as being among the most common identified STs from blood stream infections and urinary tract infections caused by *E. coli;* Additionally, they reported that isolates belonged to ST73 and ST95 had a more susceptible antibiotic profile in comparison to isolates belonged to ST131 (Day et al., 2016; Flament-Simon et al., 2020; Hertz et al., 2016; Kallonen et al., 2017; Toval et al., 2014). The success of other lineages other than ST131 need to be further studied in Sweden, regardless of their antibiotic resistance/susceptibility profile. Exploring the relationship between the genotypic resistance predictions and MLST, a high diversity in the STs of the resistant isolates was noted. Although isolates belonged to ST131 showed the highest resistance profile, they harbored the lowest mean of VFs genes in comparison to the other most common STs, ST73 and ST95. In a Canadian study (Manges et al., 2015) ST73 and ST95 were identified to be related to harboring a large number of VFs from human infections. These findings suggest that harboring a high count of VFs genes may not be directly related to the ability of an isolate to develop a resistance.

The 136 clinical *E. coli* isolates were assessed for the presence or absence of antibiotic resistance genes by the 1928Diagnostics *E. coli* platform and ResFinder tool from CGE. Genotypic predictions of antibiotic resistance in *E. coli* have proved to be as accurate as phenotypic AST on several occasions (Moran et al., 2017; Neuert et al., 2018; Tyson et al., 2015). Yet, predicting resistance in open genomes such as *E. coli* can be challenging (Balloux et al., 2018). To achieve the clinical utility from the genotypic resistance predictions, it is important to utilize tools and resistance genes databases that provide resistance predictions for single antibiotic rather than predictions for the whole class (Mahfouz et al., 2020). Currently, ResFinder (version 4.1) outputs a phenotypic table that can predict the phenotypic resistance for individual antibiotics based on the detected genes (Bortolaia et al., 2020), while 1928Diagnostics platform outputs resistance genes for the broad antibiotic classes and does not state individual antibiotics that the genes may confer resistance against. Predictions of single antibiotic resistance can explain a lot of the discordant results observed between phenotypic resistance and genotypic resistance predictions (Appendix 4, Tables 3 and 4). The Comprehensive Antibiotic Resistance Database (CARD) was used to assist identifying the resistance genes' families, origins, and common coding genetic elements, for instance, plasmids or chromosomes (Alcock et al., 2020). The two tools detected the exact resistance genes for Aminoglycosides. Those genes belong to the *O-phosphotransferases (APHs)* and *O-nucleotidyltransferases (ANTs)* Aminoglycosides modifying enzymes families (AME). Three genes belong to the family *APHs* (*APH (3')-Ia, APH (3'')-Ib and APH(6)-Id*), and one gene belong to the *ANTs* family (*ANT(3'')-Ia*) (Ramirez & Tolmasky, 2010). The genes existed in the isolates either individually or in combination with each other. Interestingly, most of the AMEs resistance genes have other origins than *E. coli*, which confirms the spread of this type of resistance through mobile genetic elements between different species (Zárate et al., 2018). Not every AME gene can deactivate every aminoglycoside antibiotic. In this thesis, the only tested Aminoglycoside phenotypically was Tobramycin, which is known to inhibit *APH (3')* enzymes. In other words,

bacteria that codes for *APH (3')-la* and *APH (6)-ld* might not be resistant to Tobramycin; In addition, genes *APH (3'')-lb,* and *ANT (3'')-la* are not known to have activity against Tobramycin (Zárate et al., 2018). These assumptions were in concordance with the phenotypic AST and the predictions from ResFinder. In contrast, in a recent Polish study it was proved that the presence of the gene *APH (3'')-lb* can cause resistance to Tobramycin (Ojdana et al., 2018). Due to the limited phenotypic data in this thesis, the effect of *APH (3'')-lb* on Tobramycin therefore cannot be determined.

The broad -spectrum beta lactam antibiotics class is another large class of antibiotics. All broad-spectrum penicillins, cephalosporines, and carbapenems are counted as broad-spectrum beta lactam antibiotics. Here again comes the utmost importance of identifying which resistance genes can cause resistance to which antibiotic or group of antibiotics. The 1928Diagnostics pipeline reported the genes *TEM-1, TEM-30* and *TEM40*, while ResFinder reported only the *TEM-1* gene. The *TEM-1* gene is coded on a plasmid that originates from *E. coli* and can cause resistance to ampicillin, amoxicillin and early cephalosporins, but cannot cause resistance to amoxicillin-clavulanic acid nor third generation cephalosporins starting from cefotaxime (Davies & Davies, 2010; Salverda et al., 2010). The genes *TEM-30* and *TEM-40* in addition to the previously mentioned antibiotics, they can cause resistance to amoxicillin-clavulanic acid (Stapleton et al., 1995; Tsang et al., 2021). These results have a high concordance with the AST phenotypic results (Appendix 4, Tables 3 and 4). The ESBL is a sub-group of broad-spectrum beta lactams that covers antibiotics with activity against *Pseudomonas aeruginosa*, such as ceftazidime and piperacillin-tazobactam. In the present data, both pipelines detected the gene *CTX-M-15* which is well known of causing high resistance to ceftazidime and cefotaxime. The isolates that carry the plasmids that code for *CTX-M-15* will still be susceptible to piperacillin-tazobactam (Poirel et al., 2002). This genotypic prediction of ESBL resistance, perfectly matches the phenotypic AST. The 1928Diagnostics pipeline detected another gene for the ESBL class in one isolate, the gene *OXA-1*. This gene was not reported by ResFinder. The presence of the *OXA-1* gene is somehow rare to be found in *E. coli*. Nevertheless, it is believed to cause resistance to carbapenems, especially if it was found in a trio combination with *CTX-M* resistance gene for ESBL and *ACC (6')-lb-cr* resistance gene for Aminoglycosides (Evans & Amyes, 2014; Livermore et al., 2019). In this dataset, the *OXA-1* gene was found in a different genes combination. The gene was in combination with another Aminoglycoside resistance gene *(ANT(3')-la)*, Sulfonamides resistance gene (*sul1*) and Tetracycline efflux pump (*tet(B)*). In addition, none of the *CTX-Ms* genes were present. Furthermore, carbapenems were not tested for this isolate, so the available information from the phenotypic AST could not be related to the findings from previous studies. Other studies suggested that the OXA-1 resistance gene is associated with resistance to amoxicillin clavulanic acid (Stapleton et al., 1995; Tsang et al., 2021).

Most of the previously discussed resistance genes code on conjugative plasmids and other mobile genetic elements which spread resistances through horizontal gene exchange. There is another type of resistance was observed on the studied dataset, the chromosomal gene mutations. Gene mutations in the *gyrA* and *parC, the* replacement of amino acids at specific locations of these genes can confer resistance to Quinolones. In *E. coli*, the most reported mutations in clinical isolates that can cause resistance are the substitution of S to L at the codon 83 of *gyrA* and *parC* at codon 56 where A is replaced by T (Azargun et al., 2019; Ruiz, 2003). Other mutations of *gyrA* at codons 67, 81, 82, 84, 87 and l06 were reported (Drlica & Zhao, 1997). A single mutation at these positions can confer high resistance to Nalidixic acid but not to fluorinated quinolones (Ciprofloxacin). This information is in concordance with the phenotypic AST and the genotypic results from

1928Diagnostics where the biggest mismatch between 1928Diagnostics with ResFinder was observed. ResFinder reported all the chromosomal point mutations in *gyrA* and *parC* that can confer resistant to all Quinolones not only the fluorinated Quinolones, while 1928Diagnostics reported chromosomal resistance to ciprofloxacin only. The Nalidixic acid was not tested phenotypically at all, so the resistance effect of other mutations cannot be determined for this antibiotic.

The last type of resistance that was observed in the present study, was the resistance caused by antibiotic efflux pumps. The two pipelines both detected genes coding for multi-drug efflux pumps. 1928Diagnostics reported three genes, of which two genes belong to the major facilitator superfamily (*emrD* and *mdtM*), and one gene belongs the to resistance nodulation-cell division family (*acrF*). ResFinder detected one gene (*mdfA*) that was present in all isolates with different lengths and identities, which belongs to the major facilitator superfamily. The direct effect of these multi-drug efflux pumps genes could not be noticed in the AST phenotypic data. As mentioned previously in the results section only 5% of the isolates showed any resistance *in vitro*, while the rest were susceptible even though they were coding for multi-drug efflux pumps. This suggests that their function was more related to the survival and adaptation to the surrounding environment. Gram negative bacteria can control the permeability of their outer membrane by decreasing the permeability of antibiotics and delay their penetration inside the cell by using efflux pumps to pump them out of the bacterial cell. In addition, efflux pumps may pump out toxic materials and extra nutrition as well (Hernando-Amado et al., 2016; Viveiros et al., 2007). Another type of efflux pump genes that were detected in the present dataset are the genes coding for drug-specific efflux pumps, the *tet(A)* and *tet(B)* genes. The genes code to pump out Tetracyclines, causing an acquired resistance to this antibiotic class (Roberts, 2005). These genes are usually code on conjugative mobile genetic elements such as plasmids and transposons (Poole, 2007; Roberts, 2005), and this explains how these genes immigrated to *E. coli* and other gram-negative bacteria facilitating the spread of Tetracyclines resistance worldwide. In a study that collected *E. coli* clinical isolates from worldwide regions, the resistance and susceptibility to Tetracyclines in presence of *tet* genes was examined (Tuckman et al., 2007). The study revealed that the presence of the *tet (B)* gene could pump out Minocycline with high efficiency, while the presence of *tet(A)* gene pumped out Tetracycline. The study also proved that the efficacy of Tigecycline is not affected by both genes, and *E. coli* isolates remained susceptible even if coding for *tet(A)* and *tet(B)* genes. The Tetracycline class was not tested phenotypically at all for the studied *E. coli* isolates, therefore the comparison with the phenotypic AST was not possible.

It is important to keep in consideration that WGS data provide information about the presence of the genes, but not their expression. This can lead to identifying false resistance, because the presence of the gene does not mean that the isolate will be resistant immediately (Ferreira et al., 2020). The presence of the gene may indicate that the isolate is still developing the resistance or even maybe the gene is just silent. In nine cases the phenotypic AST predicted resistance (eight cases for broad spectrum beta lactams and one case for trimethoprim) while the two pipelines predicted those cases to be susceptible (Appendix 4, Tables 3 and 4). One of the reasons behind this discrepancy might be the loss of the plasmid that code for these resistances during the sub-culture of the isolates before sequencing (Neuert et al., 2018). What confirms this assumption, as discussed above, is that most of resistance genes found in the present study are coded on mobile genetic elements, mostly on conjugative plasmids. This is another down side of the genotypic predictions of antibiotic resistance. Besides the possibility of predicting a false resistance, there is

also the possibility to predict a false susceptibility. The genotypic predictions of antibiotic resistance must be handled with caution when they are utilized in clinical settings.

All the detected virulence genes in the present study were documented to be related to ExPEC in the Virulence Factor of Pathogenic Bacteria database (VFDB, 2021). The evidence that the detected virulence genes are related to ExPEC is usually derived from studies that compared virulence factors from ExPEC and isolates from intestinal *E. coli* (Johnson & Russo, 2005). All the isolates harbored adhesins genes, and some isolates even harbored a combination of several adhesins genes. Besides harboring genes coding for iron reuptake, toxins, outer membrane, lipopolysaccharide, and polysaccharide capsules. The encoded VFs are essential for the isolates to escape the urinary tract, wounds and upper respiratory tract to the blood stream and cause blood stream infections. Moreover, those are essential for the horizontal genes exchange and for the cell survival in the surroundings (Sarowska et al., 2019). These findings are in concordance with a study that was conducted in Stockholm (Fröding et al., 2020), where they have reported a similar list of VF genes for beta lactams resistant *E. coli* clinical isolates that were collected from blood stream infections, around the same time the isolates in the current study were collected. The relationship between the amount and type of harbored VFs genes could not be related to the severity of sepsis (SOFA score). More studies are still needed to reveal the relationship between VFs, severity of sepsis (or other infections) and resistance to antibiotics.

Conclusion

To sum up, the present study could provide evidence for the usefulness of utilizing WGS-based analysis to identify traits and characteristics of *E. coli* clinical isolates. The obtained information can be highly utilized in research fields and for surveillance purposes. More studies are needed to explore successful lineages other than ST131, and to understand the relationship between virulence, genotype, and the severity of the disease. According to what was discussed above the WGS-based analysis can predict antibiotic resistance; Nevertheless, this is highly dependent on understanding the detected genes functions and the ability to confer resistance to which antibiotics. In clinical settings identifying resistance pattern, and accurate selection of the suitable antibiotic is crucial. Until all the discrepancies related to the genotypic antibiotic resistance predictions are solved, WGS-based resistance predictions cannot fully replace the phenotypic AST. The WGS-based analysis of *E. coli* clinical isolates can reveal valuable information about the biodiversity in their genomes which are useful for surveillance purposes and tracking the emergence of antibiotic resistances.

Ethical aspects, gender perspectives, and impact on the society

Ethics is a set of moral ideals and a theory that determines what is good for humans, the environment, and society. Ethics in science require regulations that uphold humanity's core values, especially human rights, and dignity (Andorno, 2009). Scientific interest should not overweigh the welfare of the involved subjects, respect them, and explain to them how their samples will be treated. The research team should respect the subjects' right of written informed consents, which must guarantee confidentiality and privacy of the participants (Andorno, 2009). The ethical committees have an important rule of ensuring the morality of scientific research work (Häyry & Takala, 2005). Respecting all the previous considerations, the "Sepsis Skaraborg study" was approved by The Regional Ethical Review Board in Gothenburg (376-11) (Ljungström et al., 2017). In this thesis, the genomes of *E. coli* clinical isolates were investigated, and bacterial isolates do not require informed consent. All included patients in the "Skaraborg sepsis study" provided written informed consents, written and oral information about the planed research project were provided (Ljungström et al., 2017). No personal information such as names or personal security numbers was published in this thesis, and it is not known for this project. The present study provided valuable information about the genotypic characteristics of *E. coli* isolates. The collected isolates were collected from females and males patients and there is no gender was intentionally ignored or over represented (Table 1). The obtained data can be highly useful for surveillances, tracking antibiotic resistance and guide therapeutic strategies.

Future perspectives

NGS has truly revolutionized the biomedical research (Ambardar et al., 2016). All identified findings above were only from analyzing WGS data through bioinformatics computational tools, without the need for lengthy and costly wet lab procedures. The field is still developing and in the near future it might replace clinical microbiological conventional methods. Yet, there are several limitations for the utilization of NGS data analysis in clinical settings. Although most of the used bioinformatics are freely available, the method is still costly for many places in the world. Though, cost is not the only obstacles in face of utilizing NGS in clinical settings. Although SBS is well-known of its accuracy, the time needed for library preparation and the bioinformatics analysis (Goodwin et al., 2016; Kircher et al., 2011), It is too long to be clinically relevant. Some of the bioinformatics steps can be omitted for the sake of application in clinical settings, such assembly which was the bottleneck of the present study. Currently, there are free tools such *K-mer* tools from CGE that can handle raw data and identify STs, resistance genes and virulence genes without the need for assembled contigs. Furthermore, commercial platforms such as 1928Diagnostics are available and can handle raw data and provide a summary of the results in a short time. With the help of freely available antibiotic resistance genes databases such as CARD, the information obtained from the 1928Diagnostics *E. coli* pipeline can be useful for research purposes.

E. coli is one of the pathogens that is responsible for millions of infections yearly worldwide, it can cause infections inside and outside the digestive system. Still, there is not enough studies on susceptible *E. coli* that can cause a variety of serious diseases such as sepsis as presented in this study. Most of the available studies focus on investigating resistant *E. coli*, especially ST131. In Sweden, more studies on *E. coli* are needed regardless of their antibiotic resistance profile to

explore other successful lineages such as ST73 and ST95 (Manges et al., 2019). It will be interesting to compare the results of antibiotic resistance and virulence genes with other studies that investigated other *E. coli* clinical isolates collected from patients with other SOFA scores.

Acknowledgments

I would start by acknowledging my main thesis supervisor professor Helena Enroth Adjunct professor at the University of Skövde. Her support, patience and trust motivated me to do my best and she was the main reason behind the success of the thesis. As she always provided plenty of interesting ideas and allowed me to have my own opinion. This is all besides her continuous recommendations to improve the workflow of the project. Thank you Helena for everything! I am forever grateful for you.

I would also like to express my gratefulness to the professors in the systems biology team at the University of Skövde for their support and continuous recommendations to improve interpreting and understanding the obtained results through all stages of the project. Without their support the success of this thesis would not have been possible: Magnus Fagerlind, Anna-Karin Pernestig, Diana Tilevik, Andreas Tilevik, Kristoffer Nilsson Grimstad, John Baxter and Zelmina Lubovac. I would additionally like to thank Dimitrios Arnellos from the 1928Diagnostics platform for his patience to answer all my questions, for facilitating the use of the platform, and for helping to understand the obtained results. I would further thank my colleagues Atousa Madadipouya and Geivi Lember for their help and encouragement during the thesis work.

I want to thank my parents and my siblings, especially my sister Deena, for their emotional support through all my study period.

Thank you all! I am always grateful for your help and support.

References

1928Diagnostics. Retrieved 2020 December 15 from https://www.1928diagnostics.com/

Alcock, B. P., Raphenya, A. R., Lau, T. T., Tsang, K. K., Bouchard, M., Edalatmand, A., ... & McArthur, A. G. (2020). CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic acids research*, *48*(D1), D517-D525.

Alekseyev, Y. O., Fazeli, R., Yang, S., Basran, R., Maher, T., Miller, N. S., & Remick, D. (2018). A next-generation sequencing primer—how does it work and what can it do?. *Academic pathology*, *5*, 2374289518766521.

Alhashash, F., Wang, X., Paszkiewicz, K., Diggle, M., Zong, Z., & McNally, A. (2016). Increase in bacteraemia cases in the East Midlands region of the UK due to MDR Escherichia coli ST73: high levels of genomic and plasmid diversity in causative isolates. *Journal of Antimicrobial Chemotherapy*, *71*(2), 339-343.

Alqasim, A., Scheutz, F., Zong, Z., & McNally, A. (2014). Comparative genome analysis identifies few traits unique to the Escherichia coli ST131 H30Rx clade and extensive mosaicism at the capsule locus. *BMC genomics*, *15*(1), 830.

Ambardar, S., Gupta, R., Trakroo, D., Lal, R., & Vakhlu, J. (2016). High throughput sequencing: an overview of sequencing chemistry. *Indian journal of microbiology*, *56*(4), 394-404.

Andorno, R. (2009). Human dignity and human rights as a common ground for a global bioethics. *Journal of Medicine and Philosophy*, *34*(3), 223-240.

Angus, D. C., Linde-Zwirble, W. T., Lidicker, J., Clermont, G., Carcillo, J., & Pinsky, M. R. (2001). Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. *Critical care medicine*, *29*(7), 1303-1310.

Azargun, R., Barhaghi, M. H. S., Kafil, H. S., Oskouee, M. A., Sadeghi, V., Memar, M. Y., & Ghotaslou, R. (2019). Frequency of DNA gyrase and topoisomerase IV mutations and plasmid-mediated quinolone resistance genes among Escherichia coli and Klebsiella pneumoniae isolated from urinary tract infections in Azerbaijan, Iran. *Journal of global antimicrobial resistance*, *17*, 39-43.

Babraham bioinformatics. (2020). Home page. Retrieved 2020 November 16 from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Ballenghien, M., Faivre, N., & Galtier, N. (2017). Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. *BMC biology*, *15*(1), 1-16.

Balloux, F., Brynildsrud, O. B., Van Dorp, L., Shaw, L. P., Chen, H., Harris, K. A., ... & Eldholm, V. (2018). From theory to practice: translating whole-genome sequencing (WGS) into the clinic. *Trends in microbiology*, *26*(12), 1035-1048.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., ... & Pyshkin, A. V. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, *19*(5), 455-477.

Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing?. *Archives of Disease in Childhood-Education and Practice*, *98*(6), 236-238.

Besser, J., Carleton, H. A., Gerner-Smidt, P., Lindsey, R. L., & Trees, E. (2018). Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clinical microbiology and infection*, *24*(4), 335-341.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120.

Bone, R. C., Balk, R. A., Cerra, F. B., Dellinger, R. P., Fein, A. M., Knaus, W. A., ... & Sibbald, W. J. (1992). Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest*, *101*(6), 1644-1655.

Bortolaia, V., Kaas, R. S., Ruppe, E., Roberts, M. C., Schwarz, S., Cattoir, V., ... & Aarestrup, F. M. (2020). ResFinder 4.0 for predictions of phenotypes from genotypes. *Journal of Antimicrobial Chemotherapy*, *75*(12), 3491-3500.

Brenner, T., Decker, S. O., Grumaz, S., Stevens, P., Bruckner, T., Schmoch, T., ... & Weigand, M. A. (2018). Next-generation sequencing diagnostics of bacteremia in sepsis (Next GeneSiS-Trial): study protocol of a prospective, observational, noninterventional, multicenter, clinical trial. *Medicine*, *97*(6).

Brolund, A., Edquist, P. J., Mäkitalo, B., Olsson-Liljequist, B., Söderblom, T., Wisell, K. T., & Giske, C. G. (2014). Epidemiology of extended-spectrum β-lactamase-producing Escherichia coli in Sweden 2007–2011. *Clinical Microbiology and Infection*, *20*(6), O344-O352.

Center for Genomic Epidemiology. (CGE). Retrieved 2021 February 15 from http://www.genomicepidemiology.org/

Ciufo, S., Kannan, S., Sharma, S., Badretdin, A., Clark, K., Turner, S., ... & DiCuccio, M. (2018). Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *International journal of systematic and evolutionary microbiology*, *68*(7), 2386.

Coburn, B., Morris, A. M., Tomlinson, G., & Detsky, A. S. (2012). Does this adult patient with suspected bacteremia require blood cultures?. *Jama*, *308*(5), 502-511.

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, *38*(6), 1767-1771.

Cornet, L., Meunier, L., Van Vlierberghe, M., Léonard, R. R., Durieu, B., Lara, Y., ... & Baurain, D. (2018). Consensus assessment of the contamination level of publicly available cyanobacterial genomes. *PloS one*, *13*(7), e0200323.

Dale, A. P., & Woodford, N. (2015). Extra-intestinal pathogenic Escherichia coli (ExPEC): disease, carriage and clones. *Journal of Infection*, *71*(6), 615-626.

Davies, J., & Davies, D. (2010). Origins and evolution of antibiotic resistance. *Microbiology and molecular biology reviews*, *74*(3), 417-433.

Day, M. J., Doumith, M., Abernethy, J., Hope, R., Reynolds, R., Wain, J., ... & Woodford, N. (2016). Population structure of Escherichia coli causing bacteraemia in the UK and Ireland between 2001 and 2010. *Journal of Antimicrobial Chemotherapy*, *71*(8), 2139-2142.

Decker, S. O., Sigl, A., Grumaz, C., Stevens, P., Vainshtein, Y., Zimmermann, S., ... & Brenner, T. (2017). Immune-response patterns and next generation sequencing diagnostics for the detection of mycoses in patients with septic shock—results of a combined clinical and experimental investigation. *International journal of molecular sciences*, *18*(8), 1796.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., ... & McKenna, A. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, *43*(5), 491.

Doyle, R. M., O'Sullivan, D. M., Aller, S. D., Bruchmann, S., Clark, T., Pelegrin, A. C., ... & Harris, K. A. (2020). Discordant bioinformatic predictions of antimicrobial resistance from whole-genome sequencing data of bacterial isolates: an inter-laboratory study. *Microbial genomics*, *6*(2).

Drlica, K., & Zhao, X. (1997). DNA gyrase, topoisomerase IV, and the 4-quinolones. *Microbiology and molecular biology reviews*, *61*(3), 377-392.

Evans, B. A., & Amyes, S. G. (2014). OXA β-lactamases. *Clinical microbiology reviews*, *27*(2), 241-263.

Fan, S. L., Miller, N. S., Lee, J., & Remick, D. G. (2016). Diagnosing sepsis–The role of laboratory medicine. *Clinica chimica acta*, *460*, 203-210.

Ferreira, I., Beisken, S., Lueftinger, L., Weinmaier, T., Klein, M., Bacher, J., ... & Posch, A. E. (2020). Species identification and antibiotic resistance prediction by analysis of whole-genome sequence data by use of ARESdb: An analysis of isolates from the unyvero lower respiratory tract infection trial. *Journal of clinical microbiology*, *58*(7).

Ferrer, R., Martin-Loeches, I., Phillips, G., Osborn, T. M., Townsend, S., Dellinger, R. P., ... & Levy, M. M. (2014). Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program. *Critical care medicine*, *42*(8), 1749-1755.

Flament-Simon, S. C., Nicolas-Chanoine, M. H., García, V., Duprilot, M., Mayer, N., Alonso, M. P., ... & Blanco, J. (2020). Clonal structure, virulence factor-encoding genes and antibiotic resistance of Escherichia coli, causing urinary tract infections and other extraintestinal infections in humans in Spain and France during 2016. *Antibiotics*, *9*(4), 161.

Frieri, M., Kumar, K., & Boutin, A. (2017). Antibiotic resistance. *Journal of infection and public health*, *10*(4), 369-378.

Fröding, I., Hasan, B., Sylvin, I., Coorens, M., Nauclér, P., & Giske, C. G. (2020). Extended-Spectrum-β-Lactamase-and Plasmid AmpC-Producing Escherichia coli Causing Community-Onset Bloodstream Infection: Association of Bacterial Clones and Virulence Genes with Septic Shock, Source of Infection, and Recurrence. *Antimicrobial Agents and Chemotherapy*, *64*(8).

Goig, G. A., Blanco, S., Garcia-Basteiro, A. L., & Comas, I. (2020). Contaminant DNA in bacterial sequencing experiments is a major source of false genetic variability. *BMC biology*, *18*(1), 1-15.

Goldstein, E., MacFadden, D. R., Karaca, Z., Steiner, C. A., Viboud, C., & Lipsitch, M. (2019). Antimicrobial resistance prevalence, rates of hospitalization with septicemia and rates of mortality with sepsis in adults in different US states. *International journal of antimicrobial agents*, *54*(1), 23-34.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333.

Grumaz, S., Stevens, P., Grumaz, C., Decker, S. O., Weigand, M. A., Hofer, S., ... & Sohn, K. (2016). Next-generation sequencing diagnostics of bacteremia in septic patients. *Genome medicine*, *8*(1), 1-13.

Guo, S. Y., Zhou, Y., Hu, Q. F., Yao, J., & Wang, H. (2015). Procalcitonin is a marker of gram-negative bacteremia in patients with sepsis. *The American Journal of the Medical Sciences*, *349*(6), 499-504.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072-1075.

Häyry, M., & Takala, T. (2005). Human dignity, bioethics, and human rights. *Developing World Bioethics*, *5*(3), 225-233.

Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*, *56*(2), 61-77.

Hernando-Amado, S., Blanco, P., Alcalde-Rico, M., Corona, F., Reales-Calderón, J. A., Sánchez, M. B., & Martínez, J. L. (2016). Multidrug efflux pumps as main players in intrinsic and acquired resistance to antimicrobials. *Drug Resistance Updates*, *28*, 13-27.

Hertz, F. B., Nielsen, J. B., Schønning, K., Littauer, P., Knudsen, J. D., Løbner-Olesen, A., & Frimodt-Møller, N. (2016). Population structure of drug-susceptible,-resistant and ESBL-producing Escherichia coli from community-acquired urinary tract infections. *BMC microbiology*, *16*(1), 1-6.

Heydari, M., Miclotte, G., Demeester, P., Van de Peer, Y., & Fostier, J. (2017). Evaluation of the impact of Illumina error correction tools on de novo genome assembly. *BMC bioinformatics*, *18*(1), 374.

How, C. K., Hou, S. K., Shih, H. C., Huang, M. S., Chiou, S. H., Lee, C. H., & Juan, C. C. (2015). Expression profile of MicroRNAs in gram-negative bacterial sepsis. *Shock*, *43*(2), 121-127.

I Shemirani, M., Tilevik, A., Tilevik, D., Pernestig, A. K., & Enroth, H. (2019). Comparison of Whole Genome Sequencing Pipelines for Analysis of Staphylococcus aureus Isolates from Sepsis Patients. In *14th Annual Workshop in Systems Biology, University of Skövde, Sweden*.

Idelevich, E. A., Sparbier, K., Kostrzewa, M., & Becker, K. (2018). Rapid detection of antibiotic resistance by MALDI-TOF mass spectrometry using a novel direct-on-target microdroplet growth assay. *Clinical Microbiology and Infection*, *24*(7), 738-743.

Joensen, K. G., Scheutz, F., Lund, O., Hasman, H., Kaas, R. S., Nielsen, E. M., & Aarestrup, F. M. (2014). Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. *Journal of clinical microbiology*, *52*(5), 1501-1510.

Johnson, J. R., & Russo, T. A. (2005). Molecular epidemiology of extraintestinal pathogenic (uropathogenic) Escherichia coli. *International journal of medical microbiology*, *295*(6-7), 383-404.

JSpeciesWS. Retrieved 2021 April 20 from http://jspecies.ribohost.com/jspeciesws/

Kaas, R. S., Leekitcharoenphon, P., Aarestrup, F. M., & Lund, O. (2014). Solving the problem of comparing whole bacterial genomes across different sequencing platforms. PloS one, 9(8), e104984.

Kallonen, T., Brodrick, H. J., Harris, S. R., Corander, J., Brown, N. M., Martin, V., ... & Parkhill, J. (2017). Systematic longitudinal survey of invasive Escherichia coli in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome research*, *27*(8), 1437-1449.

Kawahara-Matsumizu, M., Yamagishi, Y., & Mikamo, H. (2018). Misidentification of Neisseria cinerea as Neisseria meningitidis by matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS). *Japanese journal of infectious diseases*, *71*(1), 85-87.

Khan, Z. A., Siddiqui, M. F., & Park, S. (2019). Current and emerging methods of antibiotic susceptibility testing. *Diagnostics*, *9*(2), 49.

Kircher, M., Heyn, P., & Kelso, J. (2011). Addressing challenges in the production and analysis of illumina sequencing data. *BMC genomics*, *12*(1), 382.

Köhler, C. D., & Dobrindt, U. (2011). What defines extraintestinal pathogenic Escherichia coli?. *International Journal of Medical Microbiology*, *301*(8), 642-647.

Kumarasamy, K. K., Toleman, M. A., Walsh, T. R., Bagaria, J., Butt, F., Balakrishnan, R., ... & Krishnan, P. (2010). Emergence of a new antibiotic resistance mechanism in India, Pakistan, and the UK: a molecular, biological, and epidemiological study. *The Lancet infectious diseases*, *10*(9), 597-602.

Larsen, M. V., Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R. L., ... & Lund, O. (2012). Multilocus sequence typing of total-genome-sequenced bacteria. *Journal of clinical microbiology*, *50*(4), 1355-1361.

Leimbach, A., Hacker, J., & Dobrindt, U. (2013). E. coli as an all-rounder: the thin line between commensalism and pathogenicity. *Between pathogenicity and commensalism*, 3-32.

Levy, M. M., Fink, M. P., Marshall, J. C., Abraham, E., Angus, D., Cook, D., ... & Ramsay, G. (2003). 2001 sccm/esicm/accp/ats/sis international sepsis definitions conference. *Intensive care medicine*, *29*(4), 530-538.

Liu, Y. Y., Wang, Y., Walsh, T. R., Yi, L. X., Zhang, R., Spencer, J., ... & Yu, L. F. (2016). Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. *The Lancet infectious diseases*, *16*(2), 161-168.

Livermore, D. M., Day, M., Cleary, P., Hopkins, K. L., Toleman, M. A., Wareham, D. W., ... & Woodford, N. (2019). OXA-1 β-lactamase and non-susceptibility to penicillin/β-lactamase inhibitor combinations among ESBL-producing Escherichia coli. *Journal of Antimicrobial Chemotherapy*, *74*(2), 326-333.

Ljungström, L., Andersson, R., & Jacobsson, G. (2019). Incidences of community onset severe sepsis, Sepsis-3 sepsis, and bacteremia in Sweden–A prospective population-based study. *PloS one*, *14*(12), e0225700.

Ljungström, L., Pernestig, A. K., Jacobsson, G., Andersson, R., Usener, B., & Tilevik, D. (2017). Diagnostic accuracy of procalcitonin, neutrophil-lymphocyte count ratio, C-reactive protein, and lactate in patients with suspected bacterial sepsis. *PloS one*, *12*(7).

Mahfouz, N., Ferreira, I., Beisken, S., von Haeseler, A., & Posch, A. E. (2020). Large-scale assessment of antimicrobial resistance marker databases for genetic phenotype prediction: a systematic review. *Journal of Antimicrobial Chemotherapy*, *75*(11), 3099-3108.

Manges, A. R., Geum, H. M., Guo, A., Edens, T. J., Fibke, C. D., & Pitout, J. D. (2019). Global extraintestinal pathogenic Escherichia coli (ExPEC) lineages. *Clinical microbiology reviews*, *32*(3).

Manges, A. R., Harel, J., Masson, L., Edens, T. J., Portt, A., Reid-Smith, R. J., ... & Boerlin, P. (2015). Multilocus sequence typing and virulence gene profiles associated with Escherichia coli from human and animal sources. *Foodborne pathogens and disease*, *12*(4), 302-310.

Mardis, E. R. (2017). DNA sequencing technologies: 2006–2016. *Nature protocols*, *12*(2), 213.

Mariansdatter, S. E., Eiset, A. H., Søgaard, K. K., & Christiansen, C. F. (2016). Differences in reported sepsis incidence according to study design: a literature review. *BMC medical research methodology*, *16*(1), 137.

Martínez, J. L. (2008). Antibiotics and antibiotic resistance genes in natural environments. *Science*, *321*(5887), 365-367.

McKinnon, J., Chowdhury, P. R., & Djordjevic, S. P. (2018). Genomic analysis of multidrug-resistant Escherichia coli ST58 causing urosepsis. *International journal of antimicrobial agents*, *52*(3), 430-435.

Mellhammar, L., Wullt, S., Lindberg, Å., Lanbeck, P., Christensson, B., & Linder, A. (2016, October). Sepsis incidence: a population-based study. In *Open forum infectious diseases* (Vol. 3, No. 4). Oxford University Press.

Moran, R. A., Anantham, S., Holt, K. E., & Hall, R. M. (2017). Prediction of antibiotic resistance from antibiotic resistance genes detected in antibiotic-resistant commensal Escherichia coli using PCR or WGS. *Journal of Antimicrobial Chemotherapy*, *72*(3), 700-704.

Namikawa, H., Yamada, K., Fujimoto, H., Oinuma, K. I., Tochino, Y., Takemoto, Y., ... & Kakeya, H. (2017). Clinical characteristics of bacteremia caused by extended-spectrum beta-lactamase-producing Escherichia coli at a tertiary hospital. *Internal Medicine*, *56*(14), 1807-1815.

Neuert, S., Nair, S., Day, M. R., Doumith, M., Ashton, P. M., Mellor, K. C., ... & Dallman, T. J. (2018). Prediction of phenotypic antimicrobial resistance profiles from whole genome sequences of non-typhoidal Salmonella enterica. *Frontiers in microbiology*, *9*, 592.

Ny, S., Löfmark, S., Börjesson, S., Englund, S., Ringman, M., Bergström, J., ... & Byfors, S. (2016). Community carriage of ESBL-producing Escherichia coli is associated with strains of low pathogenicity: a Swedish nationwide study. *Journal of Antimicrobial Chemotherapy*, dkw419.

Ojdana, D., Sieńko, A., Sacha, P., Majewski, P., Wieczorek, P., Wieczorek, A., & Tryniszewska, E. (2018). Genetic basis of enzymatic resistance of E. coli to aminoglycosides. *Advances in medical sciences*, *63*(1), 9-13.

Ojer-Usoz, E., González, D., & Vitas, A. I. (2017). Clonal diversity of ESBL-producing Escherichia coli isolated from environmental, human and food samples. *International journal of environmental research and public health*, *14*(7), 676.

Ombarak, R. A., Hinenoya, A., Awasthi, S. P., Iguchi, A., Shima, A., Elbagory, A. R. M., & Yamasaki, S. (2016). Prevalence and pathogenic potential of Escherichia coli isolates from raw milk and raw milk cheese in Egypt. *International Journal of Food Microbiology*, *221*, 69-76.

Phua, J., Ngerng, W. J., See, K. C., Tay, C. K., Kiong, T., Lim, H. F., ... & Capistrano, R. (2013). Characteristics and outcomes of culture-negative versus culture-positive severe sepsis. *Critical care*, *17*(5), R202.

Poirel, L., Gniadkowski, M., & Nordmann, P. (2002). Biochemical analysis of the ceftazidime-hydrolysing extended-spectrum β-lactamase CTX-M-15 and of its structurally related β-lactamase CTX-M-3. *Journal of Antimicrobial Chemotherapy*, *50*(6), 1031-1034.

Poole, K. (2007). Efflux pumps as antimicrobial resistance mechanisms. *Annals of medicine*, *39*(3), 162-176.

Poonawala, H., Conner, T. M., & Peaper, D. R. (2018). The brief case: misidentification of Brucella melitensis as Ochrobactrum anthropi by matrix-assisted laser desorption ionization–time of flight mass spectrometry (MALDI-TOF MS).

Price, L. B., Johnson, J. R., Aziz, M., Clabots, C., Johnston, B., Tchesnokova, V., ... & Andersen, P. S. (2013). The epidemic of extended-spectrum-β-lactamase-producing Escherichia coli ST131 is driven by a single highly pathogenic subclone, H30-Rx. *MBio*, *4*(6).

Ramirez, M. S., & Tolmasky, M. E. (2010). Aminoglycoside modifying enzymes. *Drug resistance updates*, *13*(6), 151-171.

Reuter, J. A., Spacek, D. V., & Snyder, M. P. (2015). High-throughput sequencing technologies. *Molecular cell*, *58*(4), 586-597.

Rhodes, A., Evans, L. E., Alhazzani, W., Levy, M. M., Antonelli, M., Ferrer, R., ... & Rochwerg, B. (2017). Surviving sepsis campaign: international guidelines for management of sepsis and septic shock: 2016. *Intensive care medicine*, *43*(3), 304-377.

Richter, M., & Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences*, *106*(45), 19126-19131.

Richter, M., Rosselló-Móra, R., Oliver Glöckner, F., & Peplies, J. (2016). JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics*, *32*(6), 929-931.

Riley, L. W. (2014). Pandemic lineages of extraintestinal pathogenic Escherichia coli. *Clinical Microbiology and Infection*, *20*(5), 380-390.

Roberts, M. C. (2005). Update on acquired tetracycline resistance genes. *FEMS microbiology letters*, *245*(2), 195-203.

Ross, S. M. (2014). Introduction to probability and statistics for engineers and scientists. Academic Press.

Ruiz, J. (2003). Mechanisms of resistance to quinolones: target alterations, decreased accumulation and DNA gyrase protection. *Journal of Antimicrobial Chemotherapy*, *51*(5), 1109-1117.

Russo, T. A., & Johnson, J. R. (2003). Medical and economic impact of extraintestinal infections due to Escherichia coli: focus on an increasingly important endemic problem. *Microbes and infection*, *5*(5), 449-456.

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., ... & Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC biology*, *12*(1), 1-12.

Salverda, M. L., De Visser, J. A. G., & Barlow, M. (2010). Natural evolution of TEM-1 β-lactamase: experimental reconstruction and clinical relevance. *FEMS microbiology reviews*, *34*(6), 1015-1036.

Sarowska, J., Futoma-Koloch, B., Jama-Kmiecik, A., Frej-Madrzak, M., Ksiazczyk, M., Bugla-Ploskonska, G., & Choroszy-Krol, I. (2019). Virulence factors, prevalence and potential transmission of extraintestinal pathogenic Escherichia coli isolated from different sources: recent reports. *Gut pathogens*, *11*(1), 10.

Saxenborn, P. (2018). Sepsis: Genotypic analysis of clinical Klebsiella spp. using next-generation sequencing. (Master's thesis, Skövde university, Skövde). Retrieved April 5, 2021 from https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1225335&dswid=-8889

Saxenborn, P., Tilevik, A., Fagerlind, M., Lubovac-Pilav, Z., Pernestig, A. K., Tilevik, D., & Enroth, H. (2019). Using next-generation sequencing to study biodiversity in Klebsiella spp. isolated from patients with suspected sepsis. In *European Congress of Clinical Microbiology and Infectious Diseases*.

Seymour, C. W., Gesten, F., Prescott, H. C., Friedrich, M. E., Iwashyna, T. J., Phillips, G. S., ... & Levy, M. M. (2017). Time to treatment and mortality during mandated emergency care for sepsis. *New England Journal of Medicine*, *376*(23), 2235-2244.

Shankar-Hari, M., Phillips, G. S., Levy, M. L., Seymour, C. W., Liu, V. X., Deutschman, C. S., ... & Singer, M. (2016). Developing a new definition and assessing new clinical criteria for septic shock: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *Jama*, *315*(8), 775-787.

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., ... & Hotchkiss, R. S. (2016). The third international consensus definitions for sepsis and septic shock (Sepsis-3). *Jama*, *315*(8), 801-810.

Stapleton, P., Wu, P. J., King, A., Shannon, K., French, G., & Phillips, I. (1995). Incidence and mechanisms of resistance to the combination of amoxicillin and clavulanic acid in Escherichia coli. *Antimicrobial agents and chemotherapy*, *39*(11), 2478.

Stephens, C. M., Skerker, J. M., Sekhon, M. S., Arkin, A. P., & Riley, L. W. (2015). Complete genome sequences of four Escherichia coli ST95 isolates from bloodstream infections. *Genome announcements*, *3*(6).

Stoesser, N., Sheppard, A. E., Pankhurst, L., De Maio, N., Moore, C. E., Sebra, R., ... & Kos, V. (2016). Evolutionary history of the global emergence of the Escherichia coli epidemic clone ST131. *MBio*, *7*(2).

Tan, G., Opitz, L., Schlapbach, R., & Rehrauer, H. (2019). Long fragments achieve lower base quality in Illumina paired-end sequencing. *Scientific reports*, *9*(1), 1-7.

Tenaillon, O., Skurnik, D., Picard, B., & Denamur, E. (2010). The population genetics of commensal Escherichia coli. *Nature Reviews Microbiology*, *8*(3), 207-217.

Tetzschner, A. M. M., Johnson, J. R., Johnston, B. D., Lund, O., & Scheutz, F. (2020). In silico genotyping of Escherichia coli isolates for extraintestinal virulence genes by use of whole-genome sequencing data. *Journal of clinical microbiology*, *58*(10).

The European Committee on Antimicrobial Susceptability Teasting [EUCAST] . (2020). Surveillance antimicrobial resistance Europe report 2019. Retrieved 2020 December 31 from https://www.ecdc.europa.eu/sites/default/files/documents/surveillance-antimicrobial-resistance-Europe-2019.pdf

Torsvik, M., Gustad, L. T., Mehl, A., Bangstad, I. L., Vinje, L. J., Damås, J. K., & Solligård, E. (2016). Early identification of sepsis in hospital inpatients by ward nurses increases 30-day survival. *Critical care*, *20*(1), 244.

Toval, F., Köhler, C. D., Vogel, U., Wagenlehner, F., Mellmann, A., Fruth, A., ... & Dobrindt, U. (2014). Characterization of Escherichia coli isolates from hospital inpatients or outpatients with urinary tract infection. *Journal of clinical microbiology*, *52*(2), 407-418.

Tsang, K. K., Maguire, F., Zubyk, H. L., Chou, S., Edalatmand, A., Wright, G. D., ... & McArthur, A. G. (2021). Identifying novel β-lactamase substrate activity through in silico prediction of antimicrobial resistance. *Microbial Genomics*, 000500.

Tuckman, M., Petersen, P. J., Howe, A. Y., Orlowski, M., Mullen, S., Chan, K., ... & Jones, C. H. (2007). Occurrence of tetracycline resistance genes among Escherichia coli isolates from the phase 3 clinical trials for tigecycline. *Antimicrobial agents and chemotherapy*, *51*(9), 3205-3211.

Tyson, G. H., McDermott, P. F., Li, C., Chen, Y., Tadesse, D. A., Mukherjee, S., ... & Zhao, S. (2015). WGS accurately predicts antimicrobial resistance in Escherichia coli. *Journal of Antimicrobial Chemotherapy*, *70*(10), 2763-2769.

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The third revolution in sequencing technology. *Trends in Genetics*, *34*(9), 666-681.

Vincent, J. L., Brealey, D., Libert, N., Abidi, N. E., O'Dwyer, M., Zacharowski, K., ... & Picard-Maureau, M. (2015). Rapid diagnosis of infection in the critically ill, a multicenter study of molecular detection in bloodstream infections, pneumonia, and sterile site infections. *Critical care medicine*, *43*(11), 2283.

Vincent, J. L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., ... & Thijs, L. G. (1996). The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure.

Vincent, J. L., Opal, S. M., Marshall, J. C., & Tracey, K. J. (2013). Sepsis definitions: time for change. *Lancet (London, England)*, *381*(9868), 774.

Virulence Factor Database. (VFDB). Retrieved 2021 May 5 from http://www.mgc.ac.cn/VFs/main.htm

Viveiros, M., Dupont, M., Rodrigues, L., Couto, I., Davin-Regli, A., Martins, M., ... & Amaral, L. (2007). Antibiotic stress, genetic response and altered permeability of E. coli. *PloS one*, *2*(4), e365.

Ward, B. (2021). How Linux works: What every superuser should know. no starch press.

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS computational biology*, *13*(6), e1005595.

Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L. H., ... & Achtman, M. (2006). Sex and virulence in Escherichia coli: an evolutionary perspective. *Molecular microbiology*, *60*(5), 1136-1151.

World Health Organization. (2014). *Antimicrobial resistance: global report on surveillance*. World Health Organization. Retrieved April 5, 2021 from https://www.who.int/drugresistance/documents/AMR_report_Web_slide_set.pdf?ua=1

Yohe, S., & Thyagarajan, B. (2017). Review of clinical next-generation sequencing. *Archives of pathology & laboratory medicine*, *141*(11), 1544-1557.

Zankari, E., Allesøe, R., Joensen, K. G., Cavaco, L. M., Lund, O., & Aarestrup, F. M. (2017). PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. *Journal of Antimicrobial Chemotherapy*, *72*(10), 2764-2768.

Zárate, S. G., De la Cruz Claure, M. L., Benito-Arenas, R., Revuelta, J., Santana, A. G., & Bastida, A. (2018). Overcoming aminoglycoside enzymatic resistance: design of novel antibiotics and inhibitors. *Molecules*, *23*(2), 284.

Zhang, D., Micek, S. T., & Kollef, M. H. (2015). Time to appropriate antibiotic therapy is an independent determinant of postinfection ICU and hospital lengths of stay in patients with sepsis. *Critical care medicine*, *43*(10), 2133-2140.

Appendix 1: Freely available tools for NGS analysis

Table 1. Some of the freely available tools to analyze NGS short reads data

Tool's Name	Application	Link
FastQC	Quality check and assessment	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Trimmomatic	Adaptor trimming and quality filtering	http://www.usadellab.org/cms/?page=trimmomatic
CutAdapt	Adaptor trimming and quality filtering	https://github.com/marcelm/cutadapt
BBDuk plugin	Adaptor trimming and quality filtering	https://www.geneious.com/plugins/bbduk/
SPAdes	Assembly for bacterial genomes (*de novo*)	http://bioinf.spbau.ru/spades
Velvet	Assembly (*de novo* and reference-based)	https://www.applied-maths.com/velvet-1104-windows-executables-zip (for windows) https://www.ebi.ac.uk/~zerbino/velvet/ (for Unix)
Unicycler	Assembly for bacterial genomes (*de novo*)	https://github.com/rrwick/Unicycler
QUAST	Quality assessment for genomes assembly	http://quast.sourceforge.net/
k-mer ID	Choice of a reference genome	https://github.com/phe-bioinformatics/kmerid
Mauve	- Contigs comparison against a reference genome - Contigs rearrangements	http://darlinglab.org/mauve/download.html
Galaxy server	The server contains several tools to analyze NGS data, from quality assessment to assembly	https://usegalaxy.org/
Center for genomic epidemiology	The web-based server contains several tools to detect resistance genes, virulence genes, MLST typing, species identification and more tools to identify microorganisms' traits	http://www.genomicepidemiology.org/
The Comprehensive Antibiotic Resistance Database (CARD)	Antibiotic resistance genes database	https://card.mcmaster.ca/

Appendix 2: Material and Methods

Merging Fastq files

Fastq files were downloaded to the University of Skövde as zipped files. Forward reads and reverse reads for each isolate were divided into several files. The bash script below shows the code that was used to merge the forward files and the reverse files into single separate files for each isolate (figure 1).

```
zcat file1.fastq.gz file2.fastq.gz | gzip > merged_file.fastq.gz
```

Figure 1. Shows the code that was used to merge several forward and reverse files that belong to the same isolate into single separate files.

Quality evaluation of the reads: FastQC

In the figures below are examples of the analysis modules produced by FastQC tool for the forward file for the isolate EC27 before and after trimming with Trimmomatic (Figures 2-6). The summary of the analysis of the file before trimming is presented in Figure 2. For this file, there were four warnings, per tile sequence quality, per base sequence content, per sequence GC content and sequence duplication levels. There was only one module that the file failed, the adapter content module. For the trimmed EC27 forward file, the quality has obviously improved. A summary of the quality after trimming is presented in Figure 3.

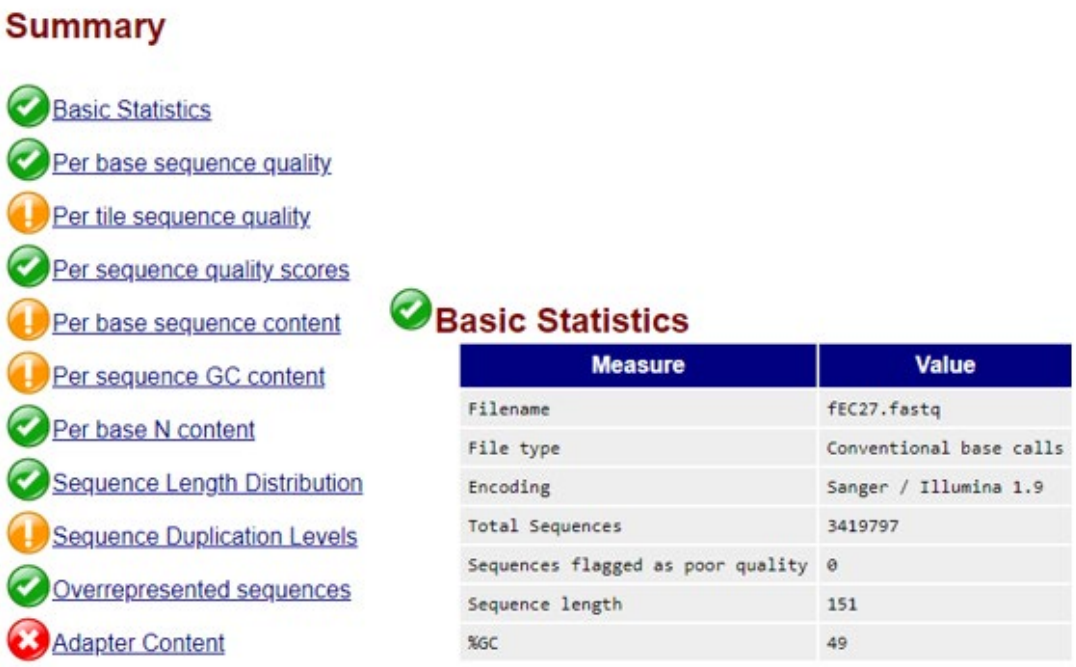


Figure 2. Shows the summary of the analysis modules and the basic statistics of the forward file for the isolate EC27 before trimming. The green tick means that the file has passed the module. The yellow tick means that there is a warning, and the module needs to be checked to identify the warning reasons. The red tick means that the file failed the module.

Figure 3. Shows the summary of the analysis modules and the basic statistics of the forward file for the isolate EC27 after trimming. The green tick means that the file has passed the module. The yellow tick means that there is a warning, and the module needs to be checked to identify the warning reasons. The quality of forward file was improved after trimming. The per tile sequence quality was turned from warning (yellow tick) into passed (green tick). The adapter content module warning sign (red tick) was turned into passed (green tick) after adapter removal by trimming.

The Per base quality sequence plots (Figure 4) show an overview of the quality scores of all bases in all the positions in the EC27 forward fastq files. On the left side of the figure is the plot for the untrimmed file and on the right side of the figure is the plot for the trimmed file. In the untrimmed file for forward EC27, the quality of reads was already high with an overall mean quality score of 30 until it was declined towards the latest cycles of sequencing. This expected when sequencing with most Illumina sequencer, because of chemical degradation by the end of sequencing (Kircher et al., 2011). After trimming with Trimmomatic, all the low-quality bases were removed, and all reads have fallen in the high-quality area of the graph (with the green background) with mean quality scores between 32- 38.

Before Trimming

After Trimming

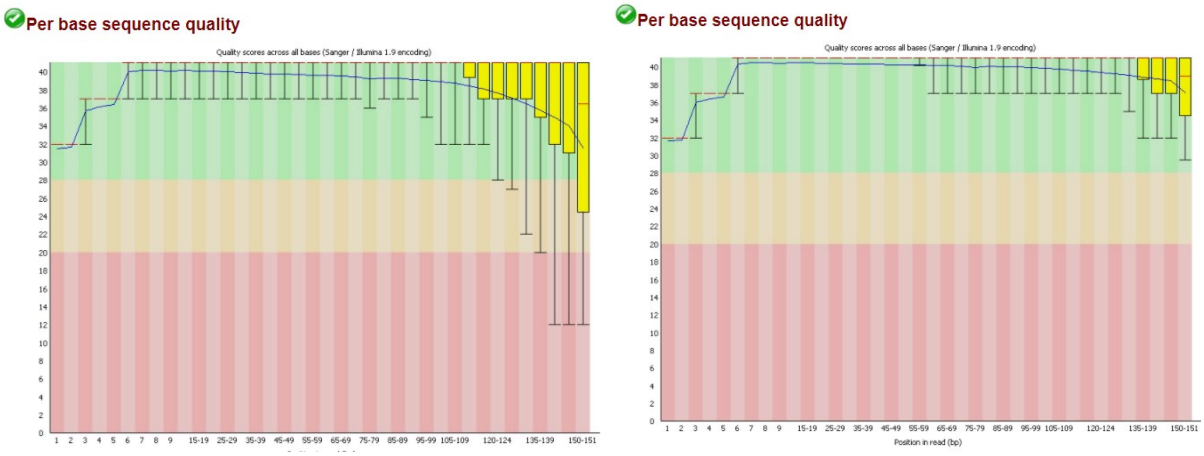
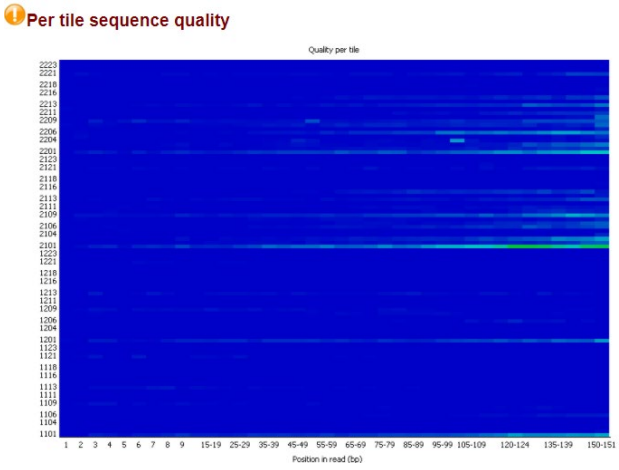


Figure 4. Shows the per-base sequence quality for EC27 forward file before and after trimming. The Y-axis represents Illumina quality scores, and the X-axis represents the position in the red in bp. The background of the plot divides the Y-axis into three sections, the green section is for the reads with the highest quality scores, the orange section is for reads with acceptable quality scores and the red section is for reads with low-quality scores. The yellow boxes represent the interquartile range, a 25% and 75% of reads' quality scores at a specific position. The black whiskers represent the 10% and 90% points of quality scores in a specific position. The red line represents the median value of the reads' quality score at a specific position. The blue lines represent the mean quality score values in a specific position.

The per tile sequence quality plots for EC27 forward file before and after trimming are presented below (Figure 5). ON the left side of the figure is the plot for the untrimmed file, where it is obvious that there were some bases at specific positions on the flow cell that had a lower quality score (indicated by the light blue and yellow colors). The plot on the right side of the graph shows the plot for the trimmed file, most of the bases with lower quality were removed and the plot appeared in a cooler blue color.

Before Trimming



After Trimming

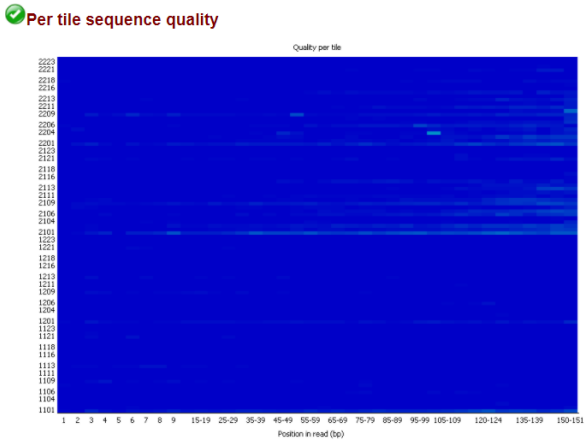


Figure 5. The plot shows the quality of the bases on the flow cell tile. This plot appears only for the Illumina library. The Y-axis represents the position in the flow cell tile and the X-axis represents the position on the read. The plot is on a cold and hot scale color, where cooler colors indicate a better-quality score than hotter colors.

The FastQC tool comes with a built-in database of the most used adapters in NGS sequencing. The module adapter content shows if there are any adapter sequences included in the inspected file. The plots for adapter content for EC27 forward file before and after trimming are presented below in Figure 6.

Before Trimming After Trimming

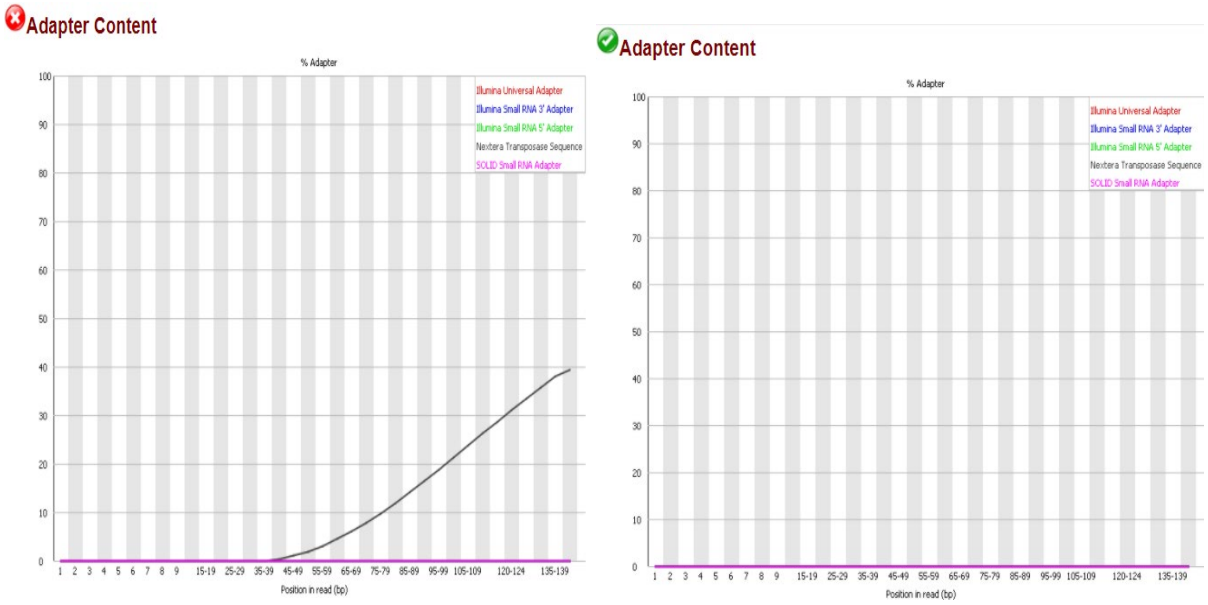


Figure 6. The adapter content plots for the EC27 forward file before and after trimming. The Y-axis represents the percentage of the adapter content and the X-axis represent the position on the read. Nextera Transposase sequence was detected, but after trimming almost all adapter content was removed.

Fastq files trimming: Trimmomatic

```
#!/bin/bash
path_Trim='/home/linux4/tools/Trimmomatic/trimmomatic-0.39.jar' #path to trimmomatic
path_merged='/media/linux4/Elements/Merged_fastq_files' #path to the parent folder
mkdir out_Trim #the output file
all_forw=($(find $path_merged/EC*/*_1.fastq.gz)) #path to all forwad files
all_rev=($(find $path_merged/EC*/*_2.fastq.gz)) #path to all reverse files
for file in ${!all_forw[@]}
do
 echo $file
 names1=$(basename ${all_forw[file]} _1.fastq.gz)
 main=$(echo ${all_forw[file]})
 mate=$(echo ${all_rev[file]})
 java -jar "$path_Trim" PE $main $mate -baseout out_Trim/${names1}.fastq.gz ILLUMINACLIP:NexteraPE-PE.fa:0:30:10 SLIDINGWINDOW:4:20 MINLEN:30
done
rm out_Trim/*U.fastq.gz #delet all unpaired files
```

Figure 7. Bash script that was used to trim all merged forward and reverse files using Trimmomatic tool (version 0.39).

Assembly: Unicycler

```
#Unicycler code
unicycler -1 <path to forward file> -2 <path to reverse file> -o <path to output folder> --spades_path <path to spades v3.13.0>
```

Figure 8. Bash script contains the code used for assembly using unicycler with default settings.

Quast Results

Table 1. Quast output values for Total length, Contigs counts, L50, N50, NG50, GC% and Genome fraction

Strain	Total Length	Contigs counts	L50(*)	N50(**)	NG50(***)	GC Content (%)	Genome fraction (%)
EC17	5099958	111	9	224451	249544	50.4%	90%
EC27	4863058	100	10	151712	149990	50.6%	75%
EC29	5051342	88	6	293910	293910	50.6%	96%
EC38	4891996	70	8	222400	222400	50.7%	78%
EC67	4546033	91	10	127672	108000	50.8%	72%
EC69	5170528	95	6	343381	343381	50.6%	90%
EC72	5184959	112	7	266549	266549	50.3%	91%
EC85	5225655	142	10	182640	186639	50.7%	79%
EC96	4854331	53	4	621440	621440	50.5%	87%
EC135	4830003	60	4	547124	547124	50.7%	89%
EC144	5144310	148	9	170866	170866	50.7%	77%
EC148	5211831	164	10	190833	190833	50.5%	85%
EC167	4975589	65	6	279859	279859	50.5%	91%
EC170	5232638	119	8	227272	287080	50.4%	92%
EC171	5031145	152	15	108634	108634	50.7%	76%
EC181	5071722	64	6	310609	310609	50.5%	93%
EC202	5060538	108	8	203751	203751	50.6%	91%
EC214	5193416	140	9	161234	190667	50.7%	85%
EC236	5085770	230	24	58079	58079	50.6%	79%
EC241	4811407	42	3	720151	720151	50.6%	87%
EC243	5074720	225	25	57707	57707	50.6%	79%
EC244	5151982	146	7	213800	213800	50.4%	91%
EC246	5004898	85	10	200043	200043	50.5%	77%
EC249	5171095	154	8	202436	213800	50.4%	91%
EC250	4990517	133	12	140640	140343	50.7%	77%
EC252	4785581	96	9	156816	155879	50.6%	77%
EC258	5008370	100	8	226922	226922	50.5%	89%
EC259	4998678	101	8	158800	158800	50.7%	75%
EC294	5053852	94	10	190808	190808	50.7%	84%
EC313	4655760	63	7	262451	247907	50.7%	87%
EC314	5051009	119	10	159784	159784	50.5%	90%
EC315	4920628	102	8	165177	160309	50.5%	88%
EC337	4853154	74	8	193934	172005	50.5%	90%
EC342	5044765	125	15	110341	110341	50.6%	78%
EC351	5065180	278	33	50102	50102	50.6%	78%
EC358	5065180	278	33	50102	50102	50.6%	78%
EC386	5085770	138	10	179975	179975	50.4%	91%
EC473	4997091	70	6	262451	262451	50.4%	91%
EC481	4866630	98	10	160003	134367	50.6%	90%

EC490	4847158	101	11	180707	180707	50.8%	76%
EC490nr2	4871504	98	11	181616	181616	50.7%	76%
EC538	4951289	92	10	202340	202340	50.5%	90%
EC554	4872559	70	6	272563	246939	50.4%	85%
EC575	4977407	203	12	132674	131282	50.4%	86%
EC583	4945592	206	16	90911	87699	50.6%	77%
EC598	4801017	85	8	196348	188588	50.7%	76%
EC610	5037531	107	6	221102	221102	50.5%	91%
EC614	4923480	109	10	186312	186312	50.5%	89%
EC622	4848989	103	10	143747	137390	50.7%	84%
EC639	5098757	151	10	162558	162558	50.6%	94%
EC699	5012480	120	8	189458	189458	50.5%	85%
EC724	5085397	155	14	106575	106575	50.7%	80%
EC725	5017332	117	9	199564	199564	50.3%	91%
EC726	4609162	99	11	154443	131928	50.9%	76%
EC733	5108138	121	7	206807	242032	50.3%	92%
EC737	4714302	92	9	167627	155153	50.8%	77%
EC740	5059276	159	12	106575	103487	50.9%	79%
EC797	5074113	106	8	203660	203660	50.5%	92%
EC800	5054527	121	9	189129	189129	50.5%	92%
EC810	4634996	74	10	138353	130533	50.7%	76%
EC813	4841534	160	13	113160	109554	50.8%	76%
EC815	4535010	49	8	184329	166821	50.8%	83%
EC822	4546787	80	12	130533	125944	50.8%	75%
EC848	4939016	120	8	216019	216019	50.4%	92%
EC878	4851670	78	9	175770	175770	50.7%	77%
EC883	4885762	79	8	175770	167720	50.7%	77%
EC895	5012409	140	10	138555	128078	50.4%	91%
EC897	5017872	129	10	130494	121086	50.7%	78%
EC913	4872341	125	11	138554	126956	50.6%	89%
EC925	4920569	95	10	160161	159635	50.7%	84%
EC929	4982870	121	8	202337	202337	50.4%	91%
EC937	5063945	116	11	184850	184850	50.5%	90%
EC949	4960626	114	8	212521	212521	50.4%	89%
EC985	5006871	141	16	107198	107198	60%	76%
EC989	5157365	190	16	111965	111965	50.2%	90%
EC1001	5125967	115	11	170497	170497	50.5%	93%
EC1003	4643102	61	8	209646	196529	50.8%	75%
EC1007	5089813	115	11	163123	163123	50.5%	93%
EC1014	5133420	163	11	168393	168393	50.4%	89%
EC1021	5054720	106	8	219550	219550	50.4%	90%
EC1078	5027462	179	13	132879	132879	50.7%	87%
EC1079	5052120	188	13	131789	131789	50.6%	88%
EC1098	4613623	55	7	245472	201805	50.7%	84%
EC1101	4804184	95	10	184850	153929	50.6%	89%
EC1112	4687675	174	25	66095	61782	50.8%	76%

EC1151	5008332	119	10	185608	179457	50.5%	91%
EC1168	4714807	76	13	126094	122954	50.6%	76%
EC1181	4855449	107	8	203754	202282	50.6%	89%
EC1223	5000947	73	5	284198	270532	50.4%	91%
EC1228	4995685	74	6	267779	267779	50.4%	91%
EC1233	5084162	156	9	184844	184844	50.4%	89%
EC1260	4841507	93	6	307055	202436	50.6%	88%
EC1272	5225258	111	7	216993	216993	50.5%	87%
EC1278	5044847	112	10	139514	139514	50.6%	91%
EC1309	4979002	63	5	383631	383631	50.6%	91%
EC1315	5349045	143	11	172098	172098	50.6%	88%
EC1324	5285970	147	12	172098	180971	50.6%	87%
EC1359	5149142	149	9	199650	199650	50.4%	89%
EC1410	5224892	188	10	168615	172935	50.5%	77%
EC1412	4978427	138	10	147535	147535	50.6%	78%
EC1445	4706096	67	8	195136	194605	50.7%	90%
EC1450	4997784	102	6	204971	204971	50.4%	90%
EC1459	4991344	100	6	270238	270238	50.4%	91%
EC1482	4719207	37	4	566758	566758	50.7%	86%
EC1498	5117277	51	4	528587	528587	50.6%	85%
EC1524	4914150	93	8	192133	168506	50.6%	88%
EC1555	5035943	241	26	58079	58079	50.6%	78%
EC1567	5122660	135	8	260927	268774	50.5%	92%
EC1571	4697933	79	8	208585	177641	50.8%	75%
EC1592	4982630	46	4	409697	358620	50.6%	85%
EC1611	4504874	90	11	123855	121165	50.8%	75%
EC1622	5025007	78	5	345284	345284	50.5%	91%
EC1630	4994825	73	5	345284	345284	50.5%	90%
EC1663	5139634	60	5	358089	358089	50.5%	93%
EC1676	5007176	108	8	286958	286958	50.5%	89%
EC1678	4818556	73	6	249842	240235	50.7%	82%
EC1717	4625942	52	7	273453	273232	50.7%	77%
EC1726	4730651	54	7	273453	273232	50.6%	78%
EC1747	4888038	117	10	160180	160180	50.7%	76%
EC1761	5005777	113	6	204504	204504	50.4%	91%
EC1769	5083212	82	12	126205	126205	50.7%	79%
EC1780	4852450	61	9	197186	197186	50.7%	78%
EC1789	5099631	85	6	318708	318708	50.5%	90%
EC1824	4883566	74	5	326888	326888	50.6%	92%
EC1888	5064693	74	4	559239	559239	50.6%	92%
EC1904GRA	4904332	130	15	105902	100138	50.7%	77%
EC1904VIT	4903101	125	16	109906	105902	50.6%	77%
EC1914	4901460	132	16	105902	100137	50.6%	77%
EC1915	5004728	58	5	357409	357409	50.7%	89%
EC1931	5039716	102	6	287103	287103	50.5%	91%
EC1945	4912842	124	9	171517	171517	50.4%	89%

EC1970	4694933	124	14	106319	93170	50.5%	76%
EC1973	4888340	105	11	131285	125493	50.6%	78%
EC1987	4911914	88	10	183063	183063	50.6%	77%
EC2000	5047275	120	6	287030	287030	50.5%	89%
EC2019	4961172	89	7	199876	199876	50.6%	79%

*L50: The minimum number of contigs that produce half (50%) of the bases of the assembly. **N50: The contig length such that using longer or equal length contigs produces half (50%) of the bases of the assembly. ***NG50: The contig length such that using longer or equal length contigs produces half (50%) of the bases of the reference genome provided (*E. coli* UTI89).

Appendix 3: Species identification

Table 1. The three isolates that were detected by the 1928Diagnostics platform to be contaminated with other species than *E.coli.*

Isolate ID	Species
EC1087	*Burkholderia dolosa* ~34% *Escherichia coli* ~66%
EC1930	*Escherichia coli* ~67% *Klebsiella pneumoniae* ~30%
EC2013	*Escherichia coli* ~34% *Salmonella enterica* ~57%

Table 2. The distribution of MLST from 1928Diagnostics *E.coli* pipeline and MLST tool in CGE using the Achtman schema

ST	Number of isolates	Percentage (%)
ST73	23	16.9%
ST95	12	8.8%
ST131 & ST127	8	5.9%
ST80	6	4.4%
Novel ST*	4	2.9%

The remaining STs	5-1 each ST	61%

*MLST tool from CGE has predicted that two of the novel ST are ST73 with a 99.9% identity of the allele *mdh,* and ST646 with 99.9% identity of the allele *adk.*

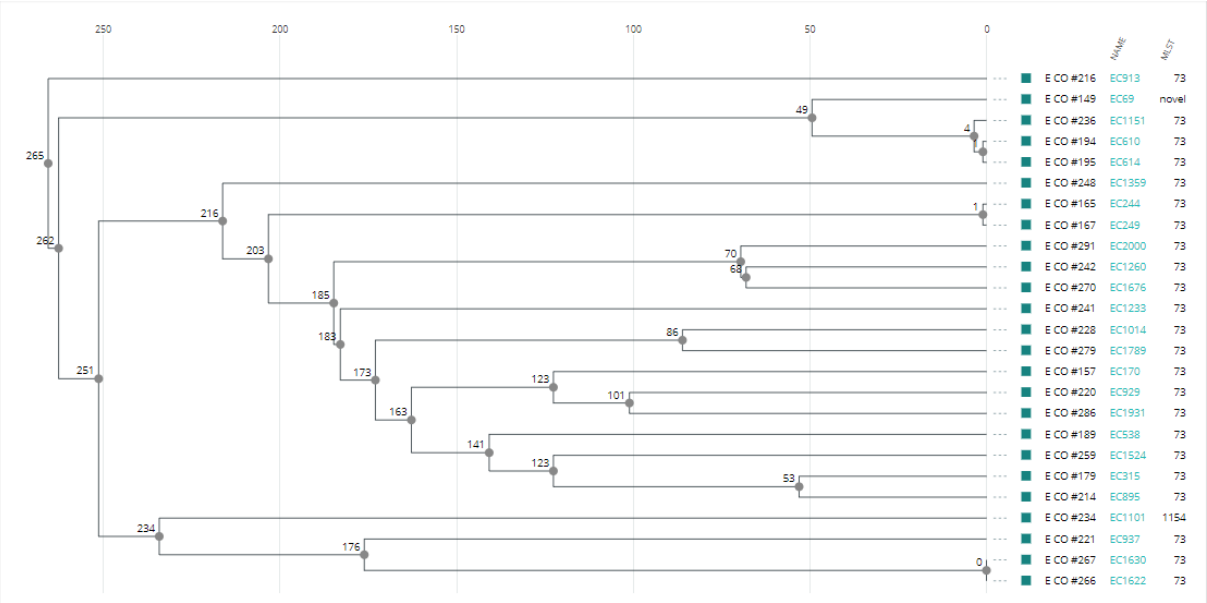


Figure 1. Shows a part of cgMLST cluster tree for 23 isolates belong to ST73, one Novel ST and one ST1154. As shown in tree the novel MLST for the isolate EC69 has clustered with the ST73, that indicated a similarity between the novel ST and ST73. MLST tool from CGE has predicted that the novel ST for EC69 is ST73 with a 99.9 identity of the allele *mdh.*

Appendix 4: Antibiotic resistance predictions

Table 1. The most frequently detected antibiotic resistance genes in all isolates (n=134) by 1928Diagnostics *E.coli* pipeline and ResFinder

Description	**1928Diagnostics (Isolates count)**	**ResFinder (Isolates count)**
Multi-drug efflux pumps		
Major facilitator super family	emrD mdtM	mdfA
Resistance nodulation-cell division family	acrF	
Aminoglycosides		
O-nucleotidyltransferases (ANTs)	ANT(3'')-la	aadA1 aadA2 aadA5
O-phosphotransferases (APHs)	APH(3')-la APH(3'')-lb APH(6)-ld	APH(3')-la APH(3'')-lb APH(6)-ld
Broad-spectrum beta lactams		
TEM-Beta lactamase family	TEM-1 TEM-30 TEM-40	TEM-1A TEM-1B TEM-1C TEM-1D
***Quinolones**		
DNA mutations	gyrA(S83L) parC(A56T)	gyrA(S83L) parC(A56T) parE(I529L)
Sulfonamides		
Sulfonamides resistant sul	Sul1 Sul2 Sul3	Sul1 Sul2 Sul3
Tetracycline		
Major facilitator super family	Tet(A) Tet(B)	Tet(A) Tet(B)
Trimethoprim		
Trimethoprim resistant dihydrofolate reductase dfr	dfrA1 dfrA5	dfrA1 dfrA5

	dfrA8	dfrA8
	dfrA12	dfrA12
	dfrA14	dfrA14
	dfrA17	dfrA17

Table 2. The actual counts of resistance genes for each resistant isolate generated from 1928Diagnostics platform and ResFinder.

Isolate	MLST	Aminoglycosides		Beta-lactam		Chloram-phenicol		Quinolone		Sulfonamide		Tetracycline		Trimethoprim	
		RF	1928	RF	1928	RF	1928	RF	1928	RF	1928	RF	1928	RF	1928
EC67	744	4	4	1	1	1	1	4	2	2	2	1	1	1	1
EC72	novel	2	2	1	1	0	0	0	0	1	1	0	0	0	0
EC85	80	2	1	1	1	1	1	2	1	1	1	1	1	0	0
EC148	131	0	0	1	1	0	0	2	1	0	0	0	0	0	0
EC170	73	1	1	0	0	0	0	0	0	1	1	0	0	0	0
EC171	58	2	2	1	1	0	0	1	1	1	1	0	0	1	1
EC214	131	2	2	1	1	0	0	4	1	2	2	0	0	1	1
EC250	10	3	3	1	1	1	1	0	0	2	2	1	1	1	1
EC294	131	0	0	0	0	0	0	1	0	0	0	0	0	0	0
EC314	95	0	0	1	1	0	0	0	0	0	0	0	0	0	0
EC315	73	0	0	0	0	0	0	1	0	0	0	0	0	0	0
EC351	357	2	0	1	0	0	0	1	0	1	0	0	0	0	0
EC358	59	2	1	1	1	0	0	1	0	1	1	0	0	0	0
EC473	127	0	0	0	0	0	0	1	0	0	0	0	0	0	0
EC490	224	2	1	0	0	1	1	4	1	1	1	1	1	1	1
EC575	12	1	1	1	1	1	0	0	0	1	1	0	0	1	1
EC622	131	0	0	0	0	0	0	1	0	0	0	1	1	0	0
EC724	69	2	2	1	1	0	0	0	0	1	1	1	1	0	0
EC740	69	2	2	1	1	0	0	0	0	1	1	1	1	0	0
EC878	410	1	1	0	0	0	0	0	0	0	0	0	0	1	1
EC883	410	1	1	0	0	0	0	0	0	0	0	0	0	1	1
EC897	88	2	2	1	1	0	0	0	0	1	1	1	1	0	0
EC913	73	1	2	0	0	0	0	0	0	0	0	0	0	1	1
EC925	131	0	0	1	1	0	0	5	1	0	0	0	0	0	0

EC1021	998	1	2	1	1	0	0	1	1	0	0	0	0	1	1
EC1078	144	2	1	1	1	0	0	0	0	1	1	0	0	0	0
EC1079	144	2	1	1	1	0	0	0	0	1	1	0	0	0	0
EC1098	104	0	0	1	1	0	0	0	0	0	0	0	0	0	0
EC1112	10	0	0	0	0	0	0	1	1	0	0	0	0	0	0
EC1233	73	3	2	0	0	0	0	0	0	2	2	0	0	1	1
EC1272	131	0	0	0	0	0	0	1	0	0	0	0	0	0	0
EC1315	404	0	0	1	1	0	0	0	0	1	1	0	0	0	0
EC1324	404	0	0	1	1	0	0	0	0	1	1	0	0	0	0
EC1410	12	2	2	1	1	0	0	0	0	1	1	0	0	1	1
EC1412	56	2	2	0	0	0	0	0	0	1	1	1	1	0	0
EC1498	131	0	0	0	0	0	0	1	0	0	0	0	0	0	0
EC1524	73	1	1	0	0	0	0	0	0	1	1	1	1	0	0
EC1555	59	0	0	0	0	0	0	1	0	1	1	0	0	0	0
EC1592	131	0	0	0	0	0	0	1	0	0	0	0	0	0	0
EC1678	117	0	0	0	0	0	0	0	0	0	0	1	1	0	0
EC1747	1125	2	2	0	0	0	0	0	0	1	1	1	1	0	0
EC1769	393	0	0	0	0	0	0	5	1	0	0	1	1	0	0
EC1973	549	4	3	1	1	1	1	0	0	1	1	1	1	1	1
EC1987	156	0	0	1	1	0	0	4	1	0	0	0	0	0	0

RF=ResFinder

Table 3. A comparison between the phenotypic AST results and 1928Diagnostics antibiotic resistance predictions for the most phenotypically tested antibiotics.

Antibiotic Class	RR	SS	RS	SR	NA/S	NA/R
Aminoglycosides*	0	41	0	8	67	17
Broad Beta-lactams	19	87	8	4	15	1
ESBL**	2	42	0	0	87	1
Chloramphenicol	1	20	0	0	110	3
Quinolones	6	108	0	5	15	0
Sulfonamides	3	92	0	14	18	7
Trimethoprim	12	103	1	2	16	0

AST= Antibiotic susceptibility testing, RR= The count of isolates that have shown resistance in both AST and 1928Diagnostics, SS=The count of isolates that have shown susceptibility in both AST and 1928Diagnostics, RS=The count of isolates that have shown resistance by AST but have been predicted to be susceptible by 1928Diagnostics, SR=The count of isolates the have shown susceptibility by AST but have been predicted to resistant by 1928Diagnostics, NA/S=The count of isolates with no phenotypic data but have been predicted to be susceptible by 1928Diagnostics, NA/R=The count of isolates with no phenotypic data, but have been predicted to be resistant by 1928Diagnostics.*One isolate was excluded from Table 4 because it has shown intermediate susceptibility to aminoglycoside; However, 1928Diagnostics has predicted it to be susceptible.** Two isolates were excluded from Table 4 because it has shown intermediate susceptibility to ESBL; However, 1928Diagnostics has predicted them to be susceptible.

Table 4. A comparison between the phenotypic AST results and ResFinder antibiotic resistance predictions for the most phenotypically tested antibiotics.

Antibiotic Class	RR	SS	RS	SR	NA/S	NA/R
Aminoglycosides*	0	40	0	9	67	17
Broad Beta-lactams	19	86	8	5	16	0
ESBL**	2	42	0	0	88	0
Chloramphenicol	1	20	0	0	108	5
Quinolones	6	99	0	14	14	1
Sulfonamides	3	91	0	15	18	7
Trimethoprim	12	103	1	2	16	0

AST= Antibiotic susceptibility testing, RR= The count of isolates that have shown resistance in both AST and ResFinder, SS=The count of isolates that have shown susceptibility in both AST and ResFinder, RS=The count of isolates that have shown resistance by AST but have been predicted to be susceptible by ResFinder, SR=The count of isolates the have shown susceptibility by AST but have been predicted to resistant by ResFinder, NA/S=The count of isolates with no phenotypic data but have been predicted to be susceptible by ResFinder, NA/R=The count of isolates with no phenotypic data, but have been predicted to be resistant by ResFinder.*One isolate was excluded from Table 4 because it has shown intermediate susceptibility to aminoglycoside; However, ResFinder has predicted it to be susceptible.** Two isolates were excluded from Table 4 because it has shown intermediate susceptibility to ESBL; However, ResFinder has predicted them to be susceptible.

Appendix 5: Virulence genes

Table 1. The counts of the most frequently detected virulence genes in the investigated *E.coli* isolates (n=134)

Description	Virulence gene	Counts of isolates
Adhesins		
A fimbriae	*afaA*	9
	afaB	7
	afaC	9
	afaD	14
	afaE	3
	afaE8	3
F1C fimbriae	*focC*	17
	focCsfaE	25
	focG	23
	focI	22
P fimbriae	*papA_F11*	13
	papA_F12	6
	papA_F13	5
	papA_F14	6
	papA_F43	27
	papA_F48	8
	papA_F7-2	2
	papA_F9	4
	papA_feiA_F8	4
	papA_fsiA_F16	7
	papC	61
S fimbriae	*sfaD*	25
	sfaE	5
	sfaS	8
Immune evasion		
Capsular antigen	*kpsE*	102
	kpsM_K15	1
	kpsMII	17
	kpsMII_K1	31
	kpsMII_K23	9
	kpsMII_K5	35
	kpsMII_K52	3
	kpsMIII_K96	1
Increased serum survival		

Toll-like receptor inhibitor	*iss*	106
Outer membrane Protein	*tcpC*	35
	OmpT	111
Invasion		
Cell invasion to host	ibeA	25
Iron uptake	*air*	12
	chuA	104
	iroN	63
	iucC	55
	iutA	56
	sitA	95
Toxins		
Protease	*pic*	44
Cell lysis	hlyF	20
Miscellaneous		
Glutamate Decarboxylase	gad	130
Tellurium ion resistance	terC	134
Putative fimbrial-like adhesin	yfcV	90
The universal stress proteins	usp	89
Siderophore yersiniabactin receptor	fyuA	100

Appendix 6: SNP and CSIphylogeny trees

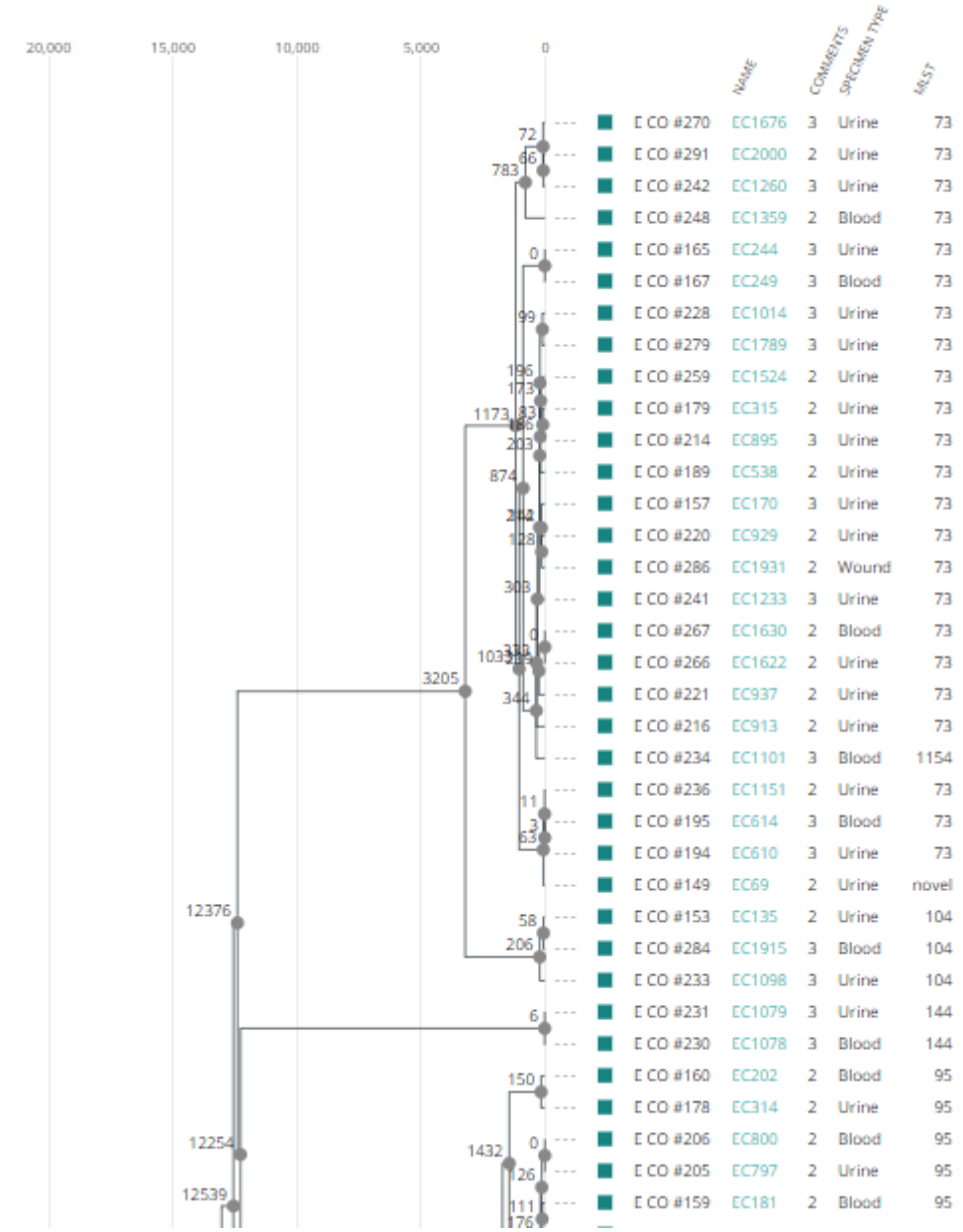


Figure 1. The clustering tree resulted from the SNP analysis on the 1928Diagnostics platform in comparison with the reference genome *E.coli* UTI89. The results have predicted which isolates were obtained from the same patient but were collected from different sample locations at the same time point. However, isolates that were obtained from the same patient at different time point did not cluster together.

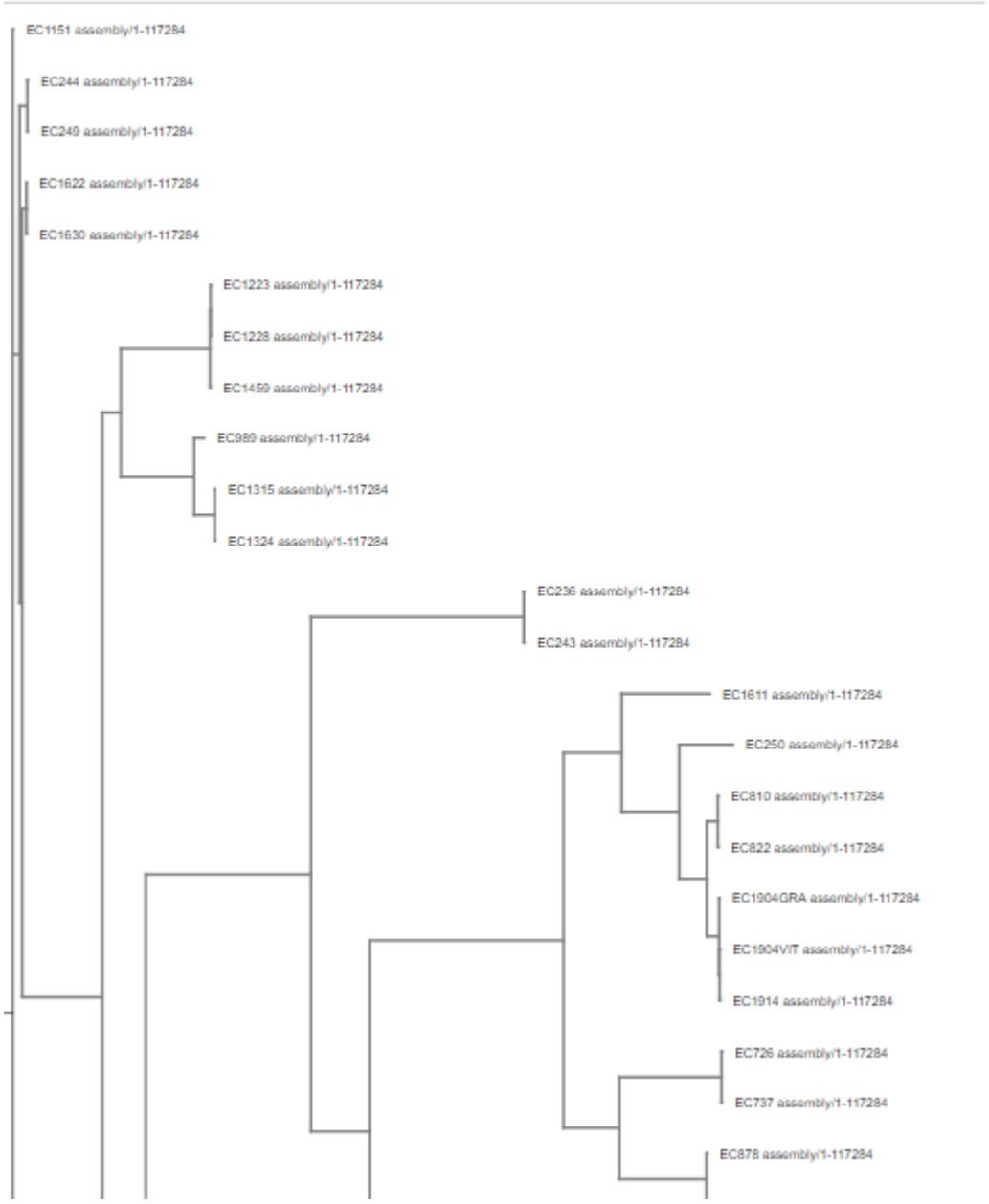


Figure 2. The clustering tree resulted from the CSIphylogeny analysis on the CGE web services in comparison with the reference genome *E.coli* UTI89. The results have predicted which isolates were obtained from the same patient but were collected from different sample locations at the same time point. However, isolates that were obtained from the same patient at different time point did not cluster together.