Degree project

UNIVERSITY OF SKÖVDE

Different pipelines for Genome Assembly of clinical *E. coli* from patients with suspected Sepsis

Master's degree Project (120 credits) in Systems Biology

Second Cycle 45 credits

Spring term 2026

Student: Madalena Monteiro

Supervisor: Anna Karin Pernestig, University of Skövde

Examiner: Matthew Herring, University of Skövde

Abstract

Sepsis is a life-threatening organ dysfunction caused by the host´s dysfunctional response to an infection. Culture-based diagnosis is the standard protocol to identify bacterial pathogens and antibiotic resistance; however, it can take longer, delaying diagnosis. Sequencing techniques enable earlier bacterial identification. Illumina is the most accurate sequencing technique, but short reads reduce genome assembly quality. Oxford Nanopore Technologies (ONT) sequencing technique produces long DNA reads but has lower sequencing accuracy. This work aimed to compare three different pipelines that used different bioinformatic tools and genome assembly techniques. DNA was extracted from *E. coli* isolates from patients with suspected sepsis and sequenced using ONT. Illumina reads were downloaded from NCBI. Reads were trimmed using FastQC and Trimomatic in $Pipeline_{In\text{-}house}$ and Fastp in $Pipeline_{Improved}$ and $Pipeline_{Hybrid}$. Genomes were assembled and analysed with Unicycler and QUAST using only Illumina reads in $Pipeline_{In\text{-}house}$ and $Pipeline_{Improved}$ and using Illumina and ONT reads in $Pipeline_{Hybrid}$. Genome Annotation analysis was performed using the CGE website features in $Pipeline_{In\text{-}house}$ and FastANI, ABRicate, NCBI-amrfinderplus and MLST in $Pipeline_{Improved}$ and $Pipeline_{Hybrid}$. $Pipeline_{Hybrid}$ showed statistically more significant quality in genome assemblies, since it combined the accuracy of Illumina with the less fragmentation of ONT; this led, for example, to the ability to identify the Sequence Type from isolate 3, not possible with $Pipeline_{In\text{-}house}$ and $Pipeline_{Improved}$. The different tools used by $Pipeline_{Improved}$ and $Pipeline_{Hybrid}$ led to a more automatic and less repetitive analysis when compared to $Pipeline_{In\text{-}house}$. Overall, this study showed the potential of $Pipeline_{Hybrid}$ for an enhancement in sepsis diagnosis and patients' treatment; Further studies should be performed to address the accuracy of the different pipelines.

Popular scientific summary

The human organism lives in association with different advantageous microorganisms, which can be called symbiotic microorganisms, that together form the human microbiota. *Escherichia coli* is one of the most common bacterial species in the human microbiota, especially in the gastrointestinal tract, responsible for the production of vitamins K and B12. Some of these bacteria can become dangerous in sterile environments within the human body, such as the blood or urinary tract. Their presence in these new environments can be responsible for multiple diseases, such as diarrhoea, urinary tract infections and sepsis. *E. coli* is considered the most frequently isolated microorganism from adults' bloodstream and is associated with 17% of the total sepsis mortality. Sepsis is a life-threatening organ dysfunction associated with a dysfunctional host response to an infection, with a mortality of 1 in 5 cases, higher than the 1 in 6 deaths associated with a breast cancer diagnosis. For more personalised and effective treatments is important to analyse the organism causing the infection, such as species identification or antibiotic resistance. Nowadays, the gold-standard method for this diagnosis is bacterial cultivation, which can take days and is not possible for every species. The delay caused by the diagnosis waiting time increases the mortality rate.

A faster and more precise alternative for this diagnosis is sequencing the bacterial genome, which works as a manual instruction, where all the information required for bacterial survival is stored. Sequencing it allows translation into a more human-understandable language, where letters represent the DNA molecules that form the DNA fragments that work together for bacterial regulation. When all the sequenced DNA fragments are assembled, it is possible to recreate the original bacterial genome. Several sequencing techniques have been developed over time and grouped into three different generations according to the system used. Second- and Third-generation sequencing techniques are most frequently used, being faster and allowing different genomes to be sequenced at the same time.

Second-generation sequencing techniques are considered to have the highest accuracy rate; however, sequencing is performed in smaller DNA fragments, leading to more complexity and lower quality in genome assembly. Since there are many small fragments, there are more possible ways for them to be assembled together, increasing the chance of wrong assembly. Another disadvantage is the requirement for DNA amplification, which can produce biased sequences. Illumina is the most used technique in this generation for its accuracy, low cost and scalability

Third-generation sequencing techniques do not require DNA amplification, and the resulting sequence fragments are relatively long, thereby improving genome assembly. Nevertheless, they are considered less accurate, producing more sequencing errors, hampering the knowledge of the exact information stored. The development of the MinION device by Oxford Nanopore Technologies reduced sequencing costs and equipment requirements, allowing sequencing outside of the laboratory infrastructure.

To overcome the challenges of these two generations, a hybrid assembly combining Illumina and MinION has been tested for genome reconstruction with higher quality and accuracy. In this hybrid assembly, the long and less accurate DNA reads produced by MinION are used as a template for the short and more accurate reads produced by Illumina. This allows a more trustworthy understanding of bacteria and an improvement in the patient's diagnosis. Thus, allows a more specific and individualised treatment, improving the patient's outcome.

Table of Contents

Abstract 2

Popular scientific summary 3

Abbreviations 1

Introduction 2

Escherichia coli 2

Sepsis disease and diagnosis 2

Whole genome sequencing 3

Bioinformatic analysis of the sequencing data 4

Raw data processing 4

Genome assembly 5

Genome annotation 5

Aim 7

Materials and Methods 8

Bacterial Isolation 8

DNA extraction 8

Short-read Illumina Analysis 9

Quality check, trimming and genome assembly 9

Assembled genome analysis 10

Third-generation sequencing and hybrid assembly 11

Statistical Analysis 12

Results 13

DNA extraction and Gel electrophoresis 13

Quality of Third-Generation Sequencing 13

Quality check, trimming and Genome Assembly 14

Species and Virulence Identification 16

Comparison between $Pipeline_{In\text{-}house}$ and $Pipeline_{Improved}$ 18

Discussion 18

DNA Extractions 18

Quality of Third-Generation Sequencing 19

Quality Check, Trimming and Genome Assembly 20

Species and Virulence Identification 20

Comparison between $Pipeline_{In\text{-}house}$ and $Pipeline_{Improved}$ 22

Conclusion 22
Limitations 23
Ethical aspects, sex and gender perspectives, and impact on society 23
Future perspectives 24
Acknowledgments 24
References 25
Appendix 32
Appendix I: Code for Pipeline$_{\text{In-house}}$ 32
Appendix II: Code for Pipeline$_{\text{Improved}}$ 34
Appendix III: Code for pipeline$_{\text{Hybrid}}$ 37
Appendix IV: DNA extraction and Gel electrophoresis 39
Appendix V: Quality of Third-Generation Sequencing 41
Appendix VI: Trimming 42
Appendix VII: Antibiotic Resistance 48

Abbreviations

AMR	Antimicrobial Resistance
ANI	Average Nucleotide Identity
ARG	Antibiotic Resistance Gene
Bp	Base-pairs
CGE	Centre for Genomic Epidemiology
EUCAST	European Committee on Antimicrobial Susceptibility Testing
ESBL	Extended-Spectrum Beta-Lactamase
FGS	First Generation Sequencing
GC Content	Guanine-Cytosine Content
LB agar	Luria-Bertani agar
MLST	Multi-Locus Sequence Typing
NCBI	National Centre for Biotechnology Information
NGS	Next Generation Sequencing
ONT	Oxford Nanopore Technologies
PCR	Polymerase Chain Reaction
PFGE	Pulse Field Gel Electrophoresis
PSS	Post Sepsis Syndrome
QC	Quality Control
QUAST	Quality Assessment Tool for Genome Assembly
SRR	Sequence Read Run
SSC	Survival Sepsis Campaign
ST	Sequence Type
TGS	Third Generation Sequencing
WGS	Whole-Genome Sequencing

Introduction

Escherichia coli

Escherichia coli (*E. coli*) is part of the normal microflora of the gastrointestinal tract (Jauréguy et al., 2007) but it can be found in other environments, like soil, water, plants, food, etc. In the human gut, *E. coli* is responsible for about 1-5% of the total microbiota and is considered to be very helpful to the host, producing, for example, vitamins K and B12 that are vital for most mammals (Blount, 2015). This bacterium has proven to be very effective in laboratory experiments due to its fast and easy growth and the large amount of knowledge available on genetic information and metabolic activities (Taj et al., 2014). Despite its benefits to humans, *E. coli* can be responsible for human infections, such as diarrhoea, peritonitis, colitis, urinary tract infections or sepsis (Jauréguy et al., 2007). These opportunistic *E. coli* infections are caused by normally beneficial strains when they encounter immunocompromised hosts or a new part of the organism, such as the blood or the urinary tract (Blount, 2015). These are normally sterile parts of the human body, and the appearance of bacteria or other microorganisms indicates a very serious state of disease (Gosiewski et al., 2017). *E. coli* is the most frequent organism isolated in the bloodstream of adult patients (Mora-Rillo et al., 2015). *E. coli* was associated with 17% of all deaths related to sepsis (Tilevik et al., 2022).

Sepsis disease and diagnosis

By definition, sepsis is a life-threatening organ dysfunction caused by the host´s dysfunctional response to the infection (Singer et al., 2016), with mortality rates up to 30% (Fleischmann et al., 2016), where the intensity increases with age (Ljungström et al., 2019). The alterations of the microflora in the gut have proven to influence several diseases, such as sepsis (Zuo et al., 2023). The treatment of Sepsis has been evolving and optimised over the last few years. The last version of the Surviving Sepsis Campaign (SSC) guidelines was developed in 2021 and suggested a variety of tools and programs that can be used to improve the diagnosis of sepsis patients (Srzić et al., 2022). Despite these guidelines, sepsis mortality is increasing due to the lack of personalised treatments for the specific microorganism responsible for the infection (Wang & Liu, 2023).

The current gold standard protocol to diagnose the pathogen causing sepsis is bacterial cultivation in special culture media. This cultivation includes blood cultures and further on to phenotypic antibiotic sensitivity testing, which can be accomplished through the disc diffusion method on Mueller-Hinton medium, following the guidelines from the European Committee on Antimicrobial Susceptibility Testing (EUCAST) (Mancini et al., 2010). However, this method has some disadvantages, such as not all bacteria can be cultivated, and it can take up to two days, which delays the diagnosis process (Gosiewski et al., 2017) and increases the waiting time until the patient's treatment, increasing the mortality rate (Peri et al., 2022). Sepsis treatment relies on early interventions, reducing Post-Sepsis Syndrome (PSS) (Torres et al., 2025). PSS are long-term consequences caused by the deregulation of the organism, such as the loss of bioenergetic capacity due to damage to mitochondria (Gorecki et al., 2025). The development of sequencing technologies enables earlier species identification without the need for blood culture and more detailed analysis of their genomes (Grumaz et al., 2020). Sequencing techniques have already proven to be a useful tool to understand bacterial genomes, including genes that encode proteins with antibiotic resistance, Antibiotic Resistance Genes (ARGs) and the virulence of *E. coli* strains (Tilevik et al., 2022).

Whole genome sequencing

DNA sequencing was introduced by Sanger, who determined DNA sequences, or DNA reads, by primed synthesis with DNA polymerase based on the chain-terminating dideoxynucleotide analogues (Shendure et al., 2017). The commercialisation of this technique led to its broad dissemination throughout the global research community (Pareek et al., 2011). Since then, the study and development of DNA sequencing techniques increased and in 2003, through ABI Prism 3700 with 96 capillaries, the human genome was successfully sequenced for the first time (Grada & Weinbrecht, 2013). Sanger sequencing techniques were used on single genes or small genomes due to the amount of time and money required (McCombie et al., 2019).

Next-generation Sequencing (NGS) overcame the challenges of Sanger sequencing, allowing the generation of around five hundred million bases of raw sequence to billions of bases in a single run (Pareek et al., 2011). There are several NGS platforms; one of the most used is Illumina sequencing (Beggio et al., 2025). Illumina sequencing requires library preparation by adding a specific adapter to the ends of fragmented DNA, amplification using Polymerase Chain Reaction (PCR) and immobilisation (Emiyu & Lelisa, 2022). The update to Illumina sequencing techniques using Genome Analyser or Paired-End modules enabled the use of sequencing adapters with a sample index. This led to the possibility of sequencing multiple samples at the same time in the same lane and then being separated based on their index (Kircher et al., 2011). Illumina sequencing has been proven to be able to sequence *E. coli* DNA reads from isolates collected from sepsis patients (Lember, 2021). Illumina is considered one of the sequencing techniques with the highest accuracy and the best throughput in terms of data produced; however, one of the biggest limitations of Illumina is the size of the resulting genetic fragments. Illumina fragments can be sequenced within a range of 50-600 base pairs (bp) (Weirather et al., 2017). Another limitation of Illumina is the requirement for DNA amplification. Although PCR is a great solution for DNA analysis, it can introduce amplification errors and favour certain DNA sequences over others (Aird et al., 2011). To overcome this issue in NGS, Third Generation Sequencing (TGS) was developed (Pareek et al., 2011).

The TGS enabled sequencing long reads of data (> 1000 bp) without using PCR-based techniques, overcoming the challenges faced by NGS (Weirather et al., 2017). Oxford Nanopore Technologies (ONT) is one of the most used TGS platforms and relies on the conversion of nucleotides' electrical signals as they pass through an α-hemolysin pore attached to a cyclodextrin molecule (Petersen et al., 2019). While the DNA molecule passes through the pore, different characteristics of the ionic current are extracted. The nucleotides block the ionic current for a certain period of time that is characteristic of each base, allowing the determination of the DNA strand (Pareek et al., 2011). The MinION platform was one of the first introduced methods by ONT and has a lower cost and equipment size than Illumina sequencing, which allows it to be used in both laboratory analysis and in field work for a faster diagnosis (Petersen et al., 2019). The biggest drawback faced by nanopore sequencing platforms is the low accuracy and throughput of the sequencing transcripts (Weirather et al., 2017). MinION is considered a useful sequencing platform for the diagnosis of various clinical samples due to its portability, the fact that it can read long reads (>100 kb) and the real-time processing of the generated reads (Grumaz et al., 2020).

Nowadays, the large amount of data obtained through sequencing enables sequencing the organism's whole genome. Whole-genome Sequencing (WGS) increases the accuracy of medical diagnostics, which leads to more personalised medicine (Brlek et al., 2024). One of the main challenges for WGS is analysing the data, which can be a very complex task that demands specific

knowledge and bioinformatics tools, and often requires the development of new processes to handle the large volume of data being produced. (Wyres et al., 2014).

Bioinformatic analysis of the sequencing data

There are several pipelines and bioinformatics pathways for analysing raw data obtained from sequencing techniques. The choice of which to use depends on the data generated, also called reads, and the analysis goal. For WGS, both short and long reads usually demand Quality Control (QC), trimming, genome assembly and genome annotation.

Raw data processing

Processing the raw sequencing data involves QC, trimming and filtering to remove low-quality reads and adapter contamination before downstream analysis. There are several bioinformatic tools available and specialised for each processing step, depending on the sequencing platform, data quality and downstream application. Common raw data processing steps and tools are summarised in Table 1; these tools can be used in both short and long reads (Soto-Serrano et al., 2024;Wan et al., 2025).

Table 1- Summary of commonly used tools for quality checking, trimming and filtering in raw sequencing data

Processing step	Purpose	Example tool	Main feature	Reference
Quality Control (QC)	Sequenced reads QC before the following analysis	FastQC	Generates QC graphical reports, per individual sample	(Brown et al., 2017)
QC report aggregation	Aggregation of the QC report from different samples	MultiQC	Combines the QC reports from different samples in one report	(Ewels et al., 2016)
Trimming and filtering	Deletes low-quality bases and adapter sequences	Trimmomatic	Removes low-quality reads and adapter sequences according to the input parameters	(Saxenborn et al., 2021)
QC and Trimming	Integration of QC, trimming and filtering in one step	Fastp	Assesses QC and automatically sets trimming parameters in one tool	(Zuo et al., 2023)
Base quality assessment	Calculation of the sequencing accuracy	Phred Quality Score (Qscore)	Calculates the base-calling accuracy according to $Q = -10log_{10}(P)$, where P is the probability of an incorrect basecall	(Delahaye & Nicolas, 2021)

Genome assembly

After raw data processing, it is very common to assemble the reads to reconstruct the genome under study. There are many tools capable of genome assembly; SPAdes is an example of such software (Tilevik et al., 2022). Genome assembly using only short Illumina reads often results in fragmented assemblies due to repetitive regions and short fragments (Ruan et al., 2020). To overcome the accuracy challenges of ONT and the short-read limitations of NGS, a hybrid genome assembly can improve the performance and resolution of experiments (Weirather et al., 2017). Using a hybrid genome assembly technique improves contiguity, completeness and accuracy when compared to assemblies performed using only Illumina or ONT reads (Eleiwa et al., 2024). The ONT reads are used as scaffolds by the Illumina reads, which generally have a more accurate base call score (Chen et al., 2020). The use of hybrid assembly tends to improve the assembly quality parameters, such as the N50 value, the length of the read where 50% of the genome assembly is that length or longer, or the number of contigs (Eleiwa et al., 2024).

MaSuRCA and Unicycler are examples of different tools proven to perform hybrid assembly. The main features of SPAdes, MaSuRCA and Unicycler are summarised in Table 2. MaSuRCA's biggest disadvantage is the very high computational requirements, and it takes longer to assemble compared to other platforms, such as SPAdes (Gorman et al., 2023). The assembly made by Unicycler is the most elaborate when comparing these different techniques. According to a 2020 study, neither MaSuRCA nor Unicycler could handle reads with a Q-score<9. However, Unicycler seem to be very useful for genomic analysis, such as antimicrobial resistance (AMR) and virulence potential (Chen et al., 2020).

Table 2- Summary of commonly used tools for genome assembly

Bioinformatic tool	Main feature	Reference
SPAdes Genome Assembler	Open-source software which solves assembly issues, such as sequencing errors, non-uniform coverage, size variation, chimeric reads and bireads	(Bankevich et al., 2012; Sreeremya, 2026)
MaSuRCA	Highly accurate super-reads assembly and solving of repetitive reads	(Zimin et al., 2013)
Unicycler	Uses SPAdes genome assembler and increases the assembly accuracy by using smaller gaps and errors between the reads	(Wick et al., 2016; Wick et al., 2017)

Genome annotation

Genome annotation is the identification of a genome's key features, such as genes and their products (Stein, 2001). Individual genes can be classified, for example, according to their Multi-locus Sequence Type (MLST) or Virulence and Antibiotic Resistance genotypes, to improve epidemiological investigations (Wyres et al., 2014). A gene can be classified as virulent when it facilitates the bacterium to cause damage to the host, and antibiotic-resistant when its product enables the bacterium to survive antibiotic action (Thomas & Wigneshweraraj, 2014). One of the most commonly used tools for genome annotation is the Centre for Genomic Epidemiology (CGE) website (CGE, 2026), which includes several services with an easy-to-comprehend output. The

biggest disadvantage of this tool is that each assembled genome needs to be analysed independently in each service, becoming a very repetitive task (Wyres et al., 2014).

Some command-line tools allow data to run in loops, such as FastANI, ABRicate, and NCBI-amrfinder-plus. FastANI uses the Averaged Nucleotide Identity (ANI) to compare the genetic relatedness between the genome under analysis and the reference genomes input (Jain et al., 2018). ABRicate identifies genes that can be classified as virulent (Chen et al., 2020) and NCBI-amrfinder-plus is an antibiotic resistance identification tool developed by NCBI and uses the NCBI Bacterial Antimicrobial Resistance Reference Gene Database, which provides information of the genes that confer resistance to antibiotic subclasses (Feldgarden et al., 2021).

When talking about antibiotic resistance identification, the biggest drawback of WGS compared to more traditional techniques is that WGS cannot identify new resistance pathways that have not been described or updated into the database used, which indicates a necessity for traditional methods of antibiotic resistance identification (Wyres et al., 2014).

The bioinformatic tools can be combined in pipelines, creating a workflow for the analysis of the sequencing data. An In-house pipeline ($Pipeline_{In\text{-}house}$) published by Shemirani et al. (2023) allows the analysis of Illumina Sequencing data for species identification and for genome annotation (Figure 1). This pipeline started with an input of raw Illumina sequences that are then processed by quality checking using FastQC, trimming using the Trimmomatic tool and quality checking again. The short Illumina reads are then assembled using Unicycler Assembler, and then the assembly passes through a quality check using the QUAST tool. Genome annotation following this pipeline is done using the services on the CGE website. After following this pipeline, it was possible to see an improvement in the diagnosis of severe infections caused by *Staphylococcus aureus*.

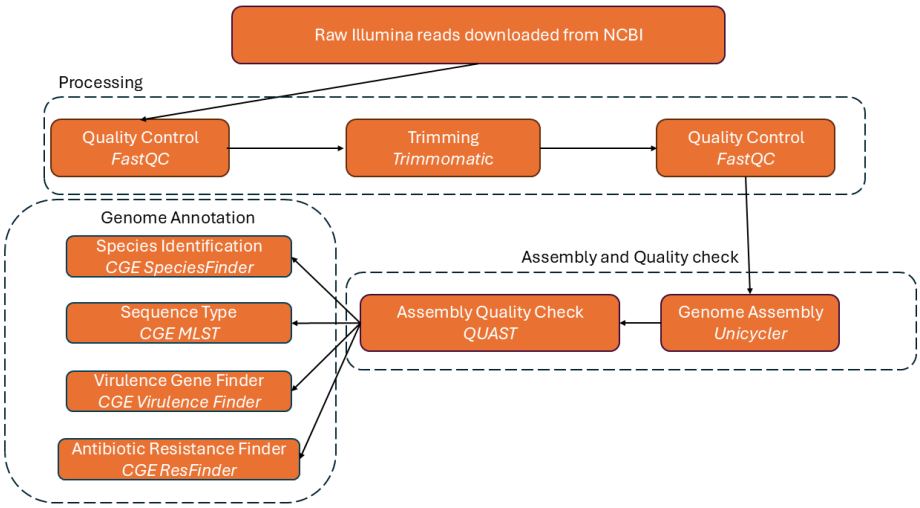

Figure 1- Outline of the in-house pipeline ($Pipeline_{In\text{-}house}$) published by Shemirani et al. (2023). The $Pipeline_{In\text{-}house}$ consists of several manual steps: Input of Raw Illumina reads downloaded in the FastQ format, processing of data, Assembly and Quality check, and genome annotation of the assembled contigs in the FASTA format. The output from genome annotation was stored and converted into tables in a Word document. The bioinformatic tools used in each step of the data analysis are indicated in Italics.

Aim

The present project aimed to compare three different bioinformatic pipelines to analyse raw sequencing data generated from DNA extracted from *E. coli* isolates collected from patients with suspected sepsis. The first pipeline ($Pipeline_{In\text{-}house}$) used only short reads from Illumina Sequencing for genome assembly and characterised the assembled genomes using tools such as FastQC, Trimmomatic, QUAST and the Centre for Genome Epidemiology (CGE) for genome annotation, including the identification of virulence and antimicrobial resistance genes. The second pipeline ($Pipeline_{Improved}$) was created based on $Pipeline_{In\text{-}house}$, where some bioinformatic tools were changed for a more automatic process, to achieve more reproducibility, reliability and to decrease human errors. This pipeline used only command-line tools, such as Fastp, FastANI and ABRicate. The third pipeline ($Pipeline_{Hybrid}$) used a hybrid genome assembly combining short reads from Illumina and long reads produced by Oxford Nanopore Technologies (ONT). The bioinformatic tools used in this pipeline were equal to the ones used by $Pipeline_{Improved}$. The hybrid assembly is expected to improve the understanding of the bacterium's genome and, in the future, improve patients' treatment. The *E. coli* genomes assembled by the three different pipelines were analysed, and several genes were examined, including those associated with virulence and antimicrobial resistance. The results of this study not only compared different genome assembly methods but also provided insights into clinically relevant *E. coli* strains. This comparison could improve sepsis patients' treatment, providing the possibility for a more specific diagnosis.

Materials and Methods

Bacterial Isolation

An observational study conducted by Ljungström et al. (2019) during nine months (September 2011- June 2012) collected bacterial isolates from adult patients (≥ 18 years) admitted to the emergency room with suspected sepsis or septic shock at Skaraborg's Hospital, with 640 beds, in Sweden. In total, approximately 1800 isolates were collected and stored at -80ºC in MicrobankTM vials (Pro-Lab Diagnosis, Ontario, Canada). All the isolates were collected during routine patient care and were approved by the Regional Ethical Review Board of Gothenburg (376-11). Based on the characterisation of all collected samples, 422 bacterial isolates were identified as *E. coli*, some of which were ESBL-resistant, using the disc diffusion method on Mueller-Hinton medium following the guidelines from the European Committee on Antimicrobial Susceptibility Testing (EUCAST) (Ljungström et al., 2019).

In 2022, Tilevik et al. assembled and characterised the genome from 21 of the *E. coli* isolates collected from Ljungström et al. (2019), nine ESBL-resistant and 13 non-ESBL-resistant, using Whole Genome Sequencing Techniques (WGS) and bioinformatic tools. From these 21 bacterial isolates, 10 were randomly assigned and used in this experiment: four ESBL-resistant and six non-ESBL-resistant. They were identified using their isolate number attributed by Tilevik et al. (2022), their number from the BioProject: PRJNA606666 created by Ljungström et al. (2019), and their Sequence Read Run (SRR) number (Table 3).

Table 3- Identification of ten different isolates used in this experiment

Isolate number[1]	BioProject number [2]	SRR Number [3]
1	473 ESBL	SRR11179862
2	884ESBL	SRR11179861
3	1068 ESBL	SRR11179850
4	1422 ESBL	SRR11179848
10	733	SRR11179842
12	1526	SRR11179859
13	152	SRR11179858
14	574	SRR11179857
15	810	SRR11179856
16	822	SRR11179855

[1] Isolate number given by Tilevik et al. (2022)
[2] Number from the BioProject PRJNA606666 given by Ljungström et al. (2019)
[3] Sequence Read Run Number

DNA extraction

All 10 bacterial isolates used in this study and listed in Table 3 were inoculated and incubated at 37ºC for 24 h in freshly prepared LB agar plates, according to the "Making LB agar plates" Protocol (Addgene). From these agar plates, one colony was inoculated into 10 mL of LB broth in a 50 mL Falcon tube and incubated overnight at 37ºC and 220rpm in the Orbital Shaker-Incubator ES-20 (BioSan), according to the "Inoculating a liquid Bacterial Culture" protocol (Addgene).

From the overnight culture, 1mL was used to measure the OD_{600} using the CO7500 colourimeter machine and the User Manual (Biochrom WPA). At an OD_{600} of 0.8, 1mL of the overnight culture was used for DNA extraction using the QIAamp DNA mini-Kit and the QIAamp DNA and Blood mini

kit handbook (QIAGEN GmbH, Hilden, Germany). From the handbook, Appendix D: Protocols for Bacteria was used, following the section "Isolation of genomic DNA from bacterial suspension cultures", which follows the protocol: "DNA purification from tissue" from step 3. The protocol followed the manufacturer's instructions except for step 3, where incubation with proteinase K was 1 hour; steps 8-11, where the centrifugation speed was changed to 14,000 rpm; and step 12 was not performed due to a decrease in DNA concentration. DNA concentration was measured using a Qubit™ Fluorometer and the Qubit™ dsDNA Broad Range Assay kit (Invitrogen, Thermo Fisher Scientific, Waltham, USA) following the manufacturer's instructions; and purity was measured in the DS11 FX Spectrophotometer/ Fluorometer (DeNovix Inc., Wilmington, USA).

Two different electrophoresis techniques were used for exact fragment size detection. Pulse-Field Gel Electrophoresis (PFGE), which identifies longer fragments, was performed using the CHEF-DR III Pulse Field Electrophoresis System (Bio-Rad Laboratories, Hercules, USA) according to the Instruction Manual and Application Guide. DNA fragment size was estimated using the CHEF DNA Size Standard, 48.5- 1000 kb, Lambda Ladder (Bio-Rad Laboratories, Hercules, USA). Standard agarose Gel Electrophoresis, used for smaller fragments, was performed using a 0.6% agarose gel with Gel Loading Dye Purple (6x) (New England Biolabs, Ipswich, USA), and run for 1 h at 100V according to the "Agarose Gel Electrophoresis" protocol (Addgene). Quick-load purple 1kb plus DNA ladder (New England Biolabs, Ipswich, USA) was used for DNA fragment size estimation. The results from both electrophorese were observed using the ChemiDoc™ Go Imaging System with Image Lab Touch Software (Bio-Rad Laboratories, Hercules, USA).

Short-read Illumina Analysis

Using the SRR number (Table 3) for the given isolates, short reads were downloaded from NCBI using the prefetch command in SRA Toolkit v3.2.1. and analysed according to two different pipelines. An existing in-house pipeline ($Pipeline_{In\text{-}house}$) (Figure1) (Lember, 2021; Mahmoud & Enroth, 2021; Shemirani et al., 2023) that uses both command-line interfaces in a Linux environment and tools from the Centre for Genomic Epidemiology website (CGE, 2010); and a new pipeline that used only the Linux environment, created based on $Pipeline_{In\text{-}house}$ and designated as the improved pipeline ($Pipeline_{Improved}$). Both pipelines are designed for genome reconstruction and analysis from the extracted DNA. This allows species identification and analysis of genes encoding virulence proteins and genes encoding proteins that confer antibiotic resistance.

Quality check, trimming and genome assembly

According to $Pipeline_{In\text{-}house}$ (Figure 1), the quality of the reads was checked using FastQC v0.12 (Brown et al., 2017) and the trimming and filtering process was performed through Trimmomatic v0.40 (Saxenborn et al., 2021). The report provided for each read by FastQC, before and after trimming, showed the desired trimming parameters and confirmed whether the trimming process was successful. Trimmomatic v0.40 was used in a loop for all 10 isolates (Table 3) to remove Illumina Nextera Adapter Content, 4-base windows with quality lower than 25 (Phred Quality Score), the initial and final bases with quality lower than 3, and reads shorter than 50 bp. In $Pipeline_{Improved}$ (Figure 2), the read quality check and the trimming process were performed automatically in one tool, Fastp v1.0.1 (Zuo et al., 2023). For both pipelines, assembly was performed using the SPAdes genome assembler v4.2.0 (Bankevich et al., 2012; Sreeremya, 2026) in the Unicycler tool v0.5.1 (Wick et al., 2016; Wick et al., 2017), and the Quality of the assemblies was checked using QUAST v5.3.0 (Gurevich et al., 2013).

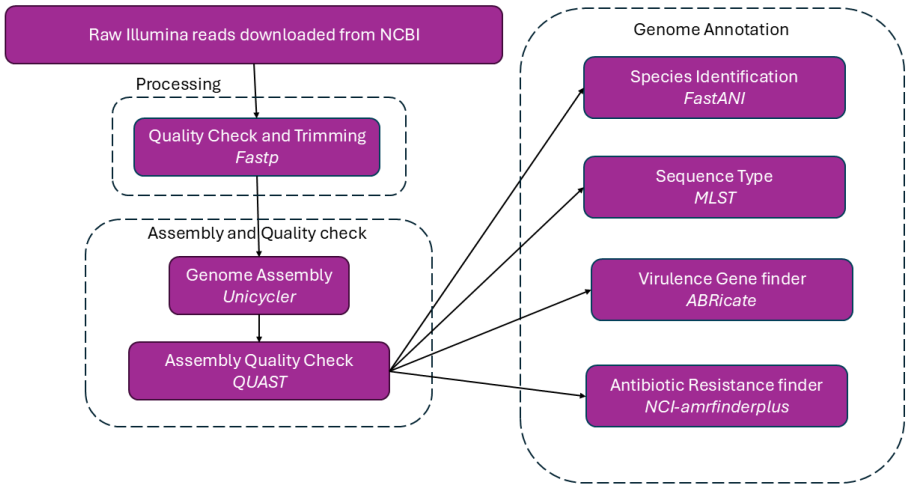

Figure 1-Outline of the improved pipeline ($Pipeline_{Improved}$). The $Pipeline_{Improved}$ consists of several manual steps: input of Raw Illumina reads downloaded in the FASTQ format, Assembly and Quality check, and Automatic command-line steps such as Processing and genome annotation of the assembled contigs in the FASTA format. The output from each genome annotation tool in a text file was stored and converted into tables in a Word document. The bioinformatic tools used in each step of the data analysis are indicated in Italics.

Assembled genome analysis

According to $Pipeline_{In\text{-}house}$ (Figure 1), all genome annotation analyses after assembly were conducted on the Centre for Genomic Epidemiology (CGE) website (Access: https://genomicepidemiology.org/services/), which provides genome analysis services and is currently updated with newly discovered genomic information. These services include VirulenceFinder v2.0.5 (Database last update: 2022-12-02), which identifies virulence genes; Multi-locus sequence typing (MLST) v2.0.9 (Database last update: 2025-09-15), which identifies the sample category based on alleles; and ResFinder v4.7.2, which identifies acquired antibiotic-resistance genes. The different genes were grouped according to the different antibiotic classes that confer resistance. These groups were already defined by the NCBI-amrfinderplus tool used in $Pipeline_{Improved}$. All tools were used with a 90% identity threshold and a minimum length of 60.

In $Pipeline_{Improved}$ (Figure 2), the analysis of the assembled genome was performed on Linux using command-line tools. Complete *E. coli* genome assemblies were downloaded in FASTA format from the NCBI RefSeq database using the ncbi-genome-download tool. The complete genomes were then used for Average Nucleotide Identity (ANI) comparisons to the assembled genomes using FastANI v1.34 (Jain et al., 2018) to identify the closest *E. coli* strains. ABRicate v1.0.1 (Chen et al., 2020) was used to find Virulence genes, and NCBI-amrfinderplus v3.12.8 (Feldgarden et al., 2021) was used to identify the resistance that the isolates showed to the different antibiotic classes, and MLST v2.23.0 with the PubMLST database (Jolley & Maiden, 2010) showed the Sequence Type (ST) of each isolate according to the alleles present in the assembled genome. NCBI-AmrFinderPlus finds the bacterial resistance gene to different antibiotics and separates them into different already predefined resistance subclasses from the NCBI-AmrFinderPlus database, such as Sulfonamide, Quaternary ammonium, Cephalosporin, Quinolone, Phenicol, Gentamicin, β-Lactam, Trimethoprim, Tetracycline and Streptomycin. All the code used to analyse the Illumina

reads following Pipeline$_{\text{In-house}}$ (Figure 1) and Pipeline$_{\text{Improved}}$ (Figure 2) can be found in Appendix I and II, respectively.

After following Pipeline$_{\text{In-house}}$ (Figure 1) and Pipeline$_{\text{Improved}}$ (Figure 2), some information was analysed to compare the automation of the different tools used and to understand the human requirements of each pipeline. Differences in the number of files analysed in read processing, the number of web-based tools, the number of files submitted, the number of parameters set manually, and the total number of manual operations made by these two different pipelines were studied.

Third-generation sequencing and hybrid assembly

From the extracted DNA, which was previously quality-checked and assessed for fragment size, a library was prepared using the MinION Ligation Sequencing gDNA-Native Barcoding kit 6 V14 (SQK-NBD114.96) and the protocol from Oxford Nanopore Technologies. For the concentration checkpoints of the DNA in the protocol, they were measured using a Qubit™ Fluorometer and the Qubit™ dsDNA Broad Range Assay kit (Invitrogen, Thermo Fisher Scientific, Waltham, USA) following the manufacturer's instructions; and purity at the same checkpoints was measured in the DS11 FX Spectrophotometer/ Fluorometer (DeNovix Inc., Wilmington, USA). The library containing nine barcodes attached to the DNA extracted from the different isolates, from barcode number 39 to number 46, was loaded and sequenced for 20 h using an R10.4.1 flow cell (FLO-MIN114), with 1367 available pores, which was then washed using the MinION Flow Cell Wash Kit and protocol (EXP-WASH004 or EXP-WSH004-XL) from Oxford Nanopore Technologies. Sequencing was started and conducted through the MinKnow software v6.2.14, where sequencing tests (n=5) were performed, and a Qscore >9 was chosen together with a high-accuracy model for basecalling. Other parameters for sequencing were set to the default value.

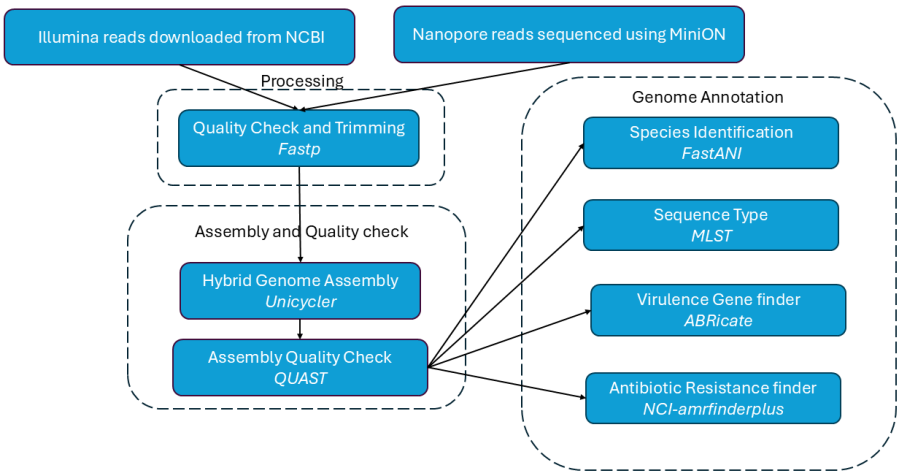

Figure 2- Outline of the Hybrid pipeline (Pipeline$_{\text{Hybrid}}$) analysed using the updated tool from the Pipeline$_{\text{Improved}}$. The Pipeline$_{\text{Hybrid}}$ consists of assembling the genome using Downloaded Illumina reads from NCBI and Nanopore reads produced by MinION. The contigs analysis is done following the same steps of the Pipeline$_{\text{Improved, which}}$ consists of one manual step: input of Raw Illumina reads downloaded in the FastQ format and Automatic command-line steps such as Processing and genome annotation of the assembled contigs in the FASTA format. The output from each genome annotation tool in a text file was stored and converted into tables in a Word document. The bioinformatic tools used in each step of the data analysis are indicated in Italics.

The raw data in the fastq.gz format were then transferred into a computer with a Linux Environment and analysed using $Pipeline_{Hybrid}$ (Figure 3), where the fastq_pass reads were trimmed using fastp v1.0.1. The hybrid assembly between the already trimmed short Illumina reads downloaded using the prefetch command from the SRA-tool and the long ONT reads was done using SPAdes genome assembler v4.2.0 in the Unicycler tool v0.5.1. The assembled genome was then analysed using the improved pipeline used for the Illumina read analysis. All the code used for $Pipeline_{Hybrid}$ (Figure 3) can be found in Appendix III.

Statistical Analysis

Statistical analysis was performed on the data to understand if the output from the different pipelines was significantly different. All the statistical analysis was performed in RStudio software V4.4.2 (Posit Software, Boston, USA), and statistical significance was considered when $p<\alpha$, assuming $\alpha=0.05$.

To address the differences in assembly contiguity, statistical tests were performed on the N50 and the L50 parameters from genome assemblies analysed by QUAST. For a better comparison, only the isolates analysed by the three different pipelines were used (n=7). For each parameter, the normality of the paired differences was assessed using the Shapiro-Wilk test. For normally distributed outputs, where the differences do not significantly deviate from normality, a paired t-test was used; the Wilcoxon signed-rank test was used for samples not normally distributed.

We compared the differences in ARGs found by each pipeline using two statistical measures. We analysed the total number of ARGs each pipeline detected across the different isolates using the Friedman rank-sum test. By using a binary matrix with the presence and absence ARGs in each isolate, the Jaccard similarity was used to compare whether the same ARGs were found by each pipeline. In the Jaccard similarity, 1 was considered as 100% similarity, where the genes found by the pipelines were the same, and 0 as 0% similarity.

Results

DNA extraction and Gel electrophoresis

The DNA concentrations measured using the fluorometer Qubit™ and the purity values from the A260/230 and A260/280 ratios of the DNA extractions are shown in Appendix IV: Table A1. Isolate 4 did not have sufficient concentration for Qubit to measure. Isolates number 3 and 13 had the lowest measured concentrations, less than 20ng/μL, and Isolates number 1 and 12 had the highest concentrations, more than 30ng/μL. In most of the samples, the absorbance ratio A260/280 values were between 1.80 and 2.00. The absorbance ratio A260/230 values varied between 1.71 and 2.18. Most absorbance ratio A260/230 values are between 1.71 and 1.80; however, isolates 1, 2, 13, and 14 have purity values over 2.0.

The different Gel electrophoresis methods measured the sizes of the extracted DNA. On standard agarose gel electrophoresis, all the bands observed are bigger than the biggest band in the ladder (10kbp) (Appendix IV: Figure A1). When running the standard 0.6% agarose gel electrophoresis, no shorter bands migrated and were found after running. On the PFGE, the results showed bands shorter and longer than the ladder (50kbp) (Appendix IV: Figure A2). In some isolates, smaller bands can be observed in the gel that match the ladder bands in size. Although it is not possible to determine the exact size of the extraction, the results from these two gel electrophoresis methods, combined, indicate that all fragments present in the extracted DNA are longer than 10kbp, and some are longer than 50kbp.

Quality of Third-Generation Sequencing

To prepare a library using DNA extracted from all isolates, 400 ng of DNA was needed in 12 μL; for the long fragments, the pool loaded into the MiniON should be equivalent to 20-25 ng/μL. When pooling the barcodes, the concentration of the pool measured with the QubitTM was 1.70 ng/μL; therefore, only the DNA extractions from isolates 1, 2, 3, 10, 12, 15 (Appendix IV: Table A1) were used in the pooling and DNA from isolates 4, 17, 18 and 21 were obtained from Atif Waheed (personal communication, March 2026) (Appendix IV: Table A2) to increase the concentration of the pool to 7 ng/μL.

Sequencing ran for 20 h using 1367 available pores and generated 3.19 Gbases (Gb) and 1.61 million (M) reads. 99.95% of the produced reads were considered to pass with a Qscore over 9, and 2.83 Gb of the generated bases had a Qscore over 9. The pore availability through the sequencing run decreased in the sequencing pores between 6:30 h and 11 h, which corresponds to an increase in the percentage of inactive pores. All these metrics obtained from the sequencing run are presented in Appendix V: Table A1 and Figure 1. In the summary of read quality parameters, the mean Q-score of the reads was 16, and the median Q-score was 17. The average read length was 1953bp, the median Read length was 722bp, and the N50 read length was 5.23kb. From all the barcodes sequenced, Isolate 4 generated the least amount of data: 3.03 Mbases (Mb) and 1.10 thousand (k) reads. The sequenced DNA from Isolate number 2 produced fewer reads than isolates 1 and 15, but more bases, resulting in longer reads. Around 32.3 % of the total reads produced, and 25.7% of the total bases produced, were not classified as any barcode. The parameters from the data obtained can be found in Appendix V: Tables A2-A4.

Quality check, trimming and Genome Assembly

The raw Illumina reads downloaded from NCBI and the raw reads sequenced using nanopore technologies were checked and trimmed through different pipelines (Figures 1-3). In $Pipeline_{In\text{-}house}$ (Figure 1), FastQC showed mostly good parameters; only the duplication levels, GC content, the per-base sequence content and the adapter content needed to be trimmed. The output from FastQC before trimming for isolate 1 is shown in Appendix VI: Figures A1- A5. After trimming, the parameters in the yellow category were kept the same. The per-base sequence content did not change after trimming using Trimmomatic, and the Sequence Duplication levels improved from 20% duplication to less than 5% duplication. The output from FastQC after trimming for isolate 1 is shown in Appendix VI: Figures A6- A9.

All the downloaded Illumina reads used in $Pipeline_{Improved}$ (Figure 2) and $Pipeline_{Hybrid}$ (Figure 3), had the same length before trimming: 151 base pairs (bp). After trimming with Fastp, the mean length decreased, except for reads from isolate number 1 (135bp) and isolate 3 (99bp). More reads from Isolate 1 passed the trimming process, having a Qscore over 9, and in Isolate 4, fewer reads passed. Isolate 3 was found to be the one with the least number of low-quality reads (21.478k), and Isolate 1 had the highest number of low-quality reads (375.064k). All the trimming information for the Illumina reads is presented in Appendix VI: Table A1. Trimming the long reads from the nanopore sequencing in Fastp (Figure 3) increased the mean length of reads from around 3000bp to around 4500bp. Isolate number 4 was the one with the most passed reads and the one with the lowest number of low-quality reads, 642k and 7, respectively. Isolates 1, 12 and 15 were the ones with the highest number of low-quality reads, more than 400 reads. Isolate number 10 was found to be the isolate with the lowest number of passed reads; 14.639k reads were above Q score Qscore. The trimming information of the long reads can be seen in Appendix VI: Table A2.

The genome assembly was performed with Unicycler in all the pipelines: $Pipeline_{In\text{-}house}$ (Figure 1), $Pipeline_{Improved}$ (Figure 2) and $Pipeline_{Hybrid}$ (Figure 3). The quality of the assemblies was measured using the QUAST command-line tool, with different parameters such as the number of contigs with more than 1000bp, largest contig, total length, N50, L50 and GC content. The results from the QUAST report are shown in Tables 4 and 5. Both QUAST reports from the assemblies done only with the Illumina reads ($Pipeline_{In\text{-}house}$ and $Pipeline_{Improved}$) are equal (Table 4). Isolate 3 assembly had the highest number of contigs (1380), the shortest contig (12593 bp), the lowest N50 (2387), the highest L50 (401) and the highest GC content (51.4%). Isolates 1 and 10 had the assemblies with the least number of contigs, the largest contig sizes, the highest N50, the lowest L50 and the lowest GC content. In all the other assemblies, parameter values were considered adequate. In the QUAST report from the Hybrid assemblies ($Pipeline_{Hybrid}$) (Table 5), isolates 2, 10 and 15 were considered to have very good quality assemblies with the lowest number of contigs (6, 4 and 3, respectively) and the largest contigs. In most of the samples, the L50 value is 1, meaning that 1 contig holds more than 50% of the information of the genome. Isolate 4 is the hybrid assembly with the lowest quality, having 69 contigs larger than 1000bp, the largest contig having 2,018,597bp, and an N50 of 285,842bp and an L50 of 7.

The N50 and the L50 values are significantly different between the different pipelines. Hybrid assemblies (Table 5) show higher N50 values (paired t-test, $p=0.005$) and lower L50 values (Wilcoxon signed rank test, $p=0.022$) than when using only Illumina reads from $Pipeline_{In\text{-}house}$ and $Pipeline_{Improved}$ (Table 4).

Table 4- QUAST report from the assembly done by Unicycler using the Illumina reads, according to the $Pipeline_{In\text{-}house}$ and the $Pipeline_{Improved}$

Isolate	Contigs ≥1000 bp	Largest Contig (bp)	Total Length	N50 (bp)	L50	GC (%)
1	55	738,322	4,990,007	262,451	6	50.36
2	80	378,839	4,670,143	120,830	12	50.81
3	1380	12,593	3,028,010	2,387	401	51.4
4	109	382,704	5,070,199	158,700	12	50.82
10	68	662,360	5,079,406	340,508	6	50.3
12	83	406,231	4,607,279	122,240	13	50.74
13	82	399,922	4,564,749	122,240	12	50.74
14	75	594,807	4,908,392	220,643	7	50.63
15	70	381,438	4,623,443	148,502	10	50.68
16	70	417,306	4,535,048	130,532	12	50.77

Table 5- QUAST report from the Hybrid assembly done by Unicycler using the Illumina reads and the Nanopore reads, according to the $Pipeline_{Hybrid}$

Isolate	Contigs ≥1000 bp	Largest Contig (bp)	Total Length	N50 (bp)	L50	GC (%)
1	38	2,775,275	5,225,517	2,775,275	1	50.32
2	6	4,629,823	4,754,184	4,629,823	1	50.83
3	19	2,018,597	5,076,176	1,401,388	2	50.68
4	56	622,191	468,210	285,842	7	50.85
10	4	4,584,411	5,229,461	4,584,411	1	50.3
12	8	2,468,948	4,747,757	2,468,948	1	50.81
15	3	4,634,277	4,715,493	4,634,277	1	50.74

Species and Virulence Identification

Table 6 shows the results from species identification and the Sequence Type (ST) from each Isolate according to the Pipelines used. Isolate numbers 13, 14 and 16 do not show any result for the hybrid assembly because they were not sequenced using Nanopore Sequencing, as mentioned previously. Species identification was performed using SpeciesFinder from the CGE website for the $Pipeline_{In\text{-}house}$ (Figure 1), the FastANI command-line tool for the $Pipeline_{Improved}$ (Figure 2), and the $Pipeline_{Hybrid}$ (Figure 3). In $Pipeline_{In\text{-}house}$ and $Pipeline_{Hybrid}$, species identification shows the genome identifier (ID) of the *E. coli* reference genomes closest to the assembled genome. CGE SpeciesFinder classified isolates 1 and 10 as Shigella sonnei. However, FastANI identified different *E. coli* strains using $Pipeline_{Improved}$ and $Pipeline_{Hybrid}$. Isolates 2, 3, 12, 13, 15 and 16 were classified as the same *E. coli* strain using $Pipeline_{In\text{-}house}$, but the same did not happen using FastANI. Following $Pipeline_{Improved,}$ isolates 2, 12, 13, 15 and 16 were again very similar, being classified with an equal strain ID, which was only attributed to isolate 15 following $Pipeline_{Hybrid.}$ Isolates 4 and 14 were both identified as the strain *Escherichia coli* Nissle 1917 by $Pipeline_{In\text{-}house}$ and as Esc11complete using $Pipeline_{Hybrid}$. Isolate 15 was the only one where FastANI identified the same genome reference when using $Pipeline_{Improved}$ and $Pipeline_{Hybrid}$.

The Sequence Type (ST) produced by allele comparison for each isolate was the same across analyses using the different pipelines (Table 6). In isolate 3, the assembled genome using only the Illumina reads did not allow ST identification in either CGE MLST ($Pipeline_{In\text{-}house}$) or on the MLST command-line tool ($Pipeline_{Improved}$). Using hybrid assembly, this categorisation was possible, and this isolate belongs to category 10 together with Isolate number 2. Although isolates 2, 12, 13, 15 and 16 were all identified as the same strain using CGE SpeciesFinder ($Pipeline_{In\text{-}house}$) and FastANI ($Pipeline_{Improved}$), they were not all classified as the same Sequence Type according to CGE MLST ($Pipeline_{In\text{-}house}$) and the MLST tool ($Pipeline_{Improved}$).

Although the total number of ARGs found in each isolate by the different pipelines (Table 6) was not significantly different (Friedman rank-sum test, $p=0.82$), differences in the specific gene identification by the pipelines were found by the Jaccard similarity (Appendix VII: Table A1). In Appendix VII: Table A2, it is possible to see the different antibiotic subclasses to which these genes confer resistance, according to the NCBI-amrfinderplus database. The genes identified showed resistance to eight different antibiotic classes using the $Pipeline_{In\text{-}house}$ and $Pipeline_{Improved}$ and nine using the $Pipeline_{Hybrid}$. Although not all the genes were identified by all the pipelines used, resistance to the antibiotic subclass was mostly identified through another gene, except in some cases. For example, isolate 1 ESBL resistance was not identified in the $Pipeline_{Hybrid}$. The gene mdf(A) confers resistance to the macrolide subclass of antibiotics, already offered by mph(A). This resistance was only detected in isolate 3 when using Illumina Assembly ($Pipeline_{In\text{-}house}$ and $Pipeline_{Improved}$), but in the $Pipeline_{Hybrid}$, this resistance is shown in all the isolates. The gene lnu(F) brings a new antibiotic subclass into the analysis, the Lincosamides, which were not identified using $Pipeline_{In\text{-}house}$ and $Pipeline_{Improved}$.

Appendix VII: Table A3 shows the presence of ARGs identified by the tools used in the different pipelines. Thus, it is possible to see the differences in the specific gene identification detected by the Jaccard similarity. Isolates 10, 13, 14, 15 and 16 did not show any resistance genes and isolate 1 showed only 1 resistance gene (blaCTX-M-15). Overall, $Pipeline_{Hybrid}$ was able to identify a greater variety of resistance genes, such as mdf(A) and lnu(F), present in all of the samples (Appendix II: Table A3). $Pipeline_{Hybrid}$ is also more specific when categorising the ARGs. For

example, the genes mph(A) and sul2 were divided into two different variants of the same gene, whereas in $Pipeline_{In\text{-}house}$ and $Pipeline_{Improved}$, this division does not happen.

Table 6- Species identification, Sequence Type and Number of Antibiotic Resistance Genes (ARGs) that resulted from the analysis of the data according to the $Pipeline_{In\text{-}house}$, $Pipeline_{Improved}$ and $Pipeline_{Hybrid}$ in each isolate. The Isolate number was previously assigned for identification in Tilevik et al. (2022).

Isolate number	Pipeline	Species identification	Sequence Type	Number of ARGs
1	$Pipeline_{In\text{-}house}$	*Shigella Sonnei*	127	1
	$Pipeline_{Improved}$	*_AI2938v1_cp*[1]	127	1
	$Pipeline_{Hybrid}$	*ASM4050046v1*[1]	127	1
2	$Pipeline_{In\text{-}house}$	*Escherichia coli* str. K-12 substr. MC4100	10	8
	$Pipeline_{Improved}$	51243816-SK1A[1]	10	8
	$Pipeline_{Hybrid}$	38656_G02[1]	10	5
3	$Pipeline_{In\text{-}house}$	*Escherichia coli* str. K-12 substr. MC4100	-	11
	$Pipeline_{Improved}$	F829[1]	-	11
	$Pipeline_{Hybrid}$	*ASM0052030v1*[1]	10	15
4	$Pipeline_{In\text{-}house}$	*Escherichia coli* Nissle 1917	131	8
	$Pipeline_{Improved}$	Esc11complete[1]	131	8
	$Pipeline_{Hybrid}$	*ASM1942858v1*[1]	131	8
10	$Pipeline_{In\text{-}house}$	*Shigella Sonnei*	80	0
	$Pipeline_{Improved}$	*ASM5406547v1*[1]	80	0
	$Pipeline_{Hybrid}$	*ASM5406523v1*[1]	80	1
12	$Pipeline_{In\text{-}house}$	*Escherichia coli* str. K-12 substr. MC4100	744	12
	$Pipeline_{Improved}$	51243816-SK1A[1]	744	12
	$Pipeline_{Hybrid}$	*ASM231071v1*[1]	744	12
13	$Pipeline_{In\text{-}house}$	*Escherichia coli* str. K-12 substr. MC4100	744	0
	$Pipeline_{Improved}$	51243816-SK1A[1]	744	0
	$Pipeline_{Hybrid}$	NA*	NA	NA*
14	$Pipeline_{In\text{-}house}$	*Escherichia coli* Nissle 1917	91	0
	$Pipeline_{Improved}$	Esc11complete[1]	91	0
	$Pipeline_{Hybrid}$	NA*	NA	NA*
15	$Pipeline_{In\text{-}house}$	*Escherichia coli* str. K-12 substr. MC4100	48	0
	$Pipeline_{Improved}$	51243816-SK1A[1]	48	0
	$Pipeline_{Hybrid}$	51243816-SK1A[1]	48	1
16	$Pipeline_{In\text{-}house}$	*Escherichia coli* str. K-12 substr. MC4100	48	0
	$Pipeline_{Improved}$	51243816-SK1A[1]	48	0
	$Pipeline_{Hybrid}$	NA*	NA*	NA*

*NA Not Applicable because the isolate was not sequenced by Nanopore sequencing.

[1] Identifier from the *E. coli* reference genome classified closest to the assembled genome

Comparison between $Pipeline_{In\text{-}house}$ and $Pipeline_{Improved}$

After completely following both $Pipeline_{In\text{-}house}$ (Figure 1) and $Pipeline_{Improved}$ (Figure 2), automation metrics were analysed to determine whether the new tools made the whole analysis less repetitive and more autonomous. The results from this analysis can be found in Table 7. For files analysed in read processing, $Pipeline_{In\text{-}house}$ (Figure 1) generated one different file for each forward and reverse read, before and after trimming, whereas $Pipeline_{Improved}$ (Figure 2) combined the pre- and post-trimming report for each forward and reverse read, reducing the number of files analysed. In the four web-based tools used, each assembly needed to be submitted separately, increasing the number of repetitive steps, whereas using command-line tools ($Pipeline_{Improved}$), there is no need for file submissions. In $Pipeline_{In\text{-}house}$ (Figure 1), both Trimmomatic and the CGE website tools required manually set parameters, increasing the human dependency in the pipeline. Overall, $Pipeline_{In\text{-}house}$ (Figure 1) increased the number of operations and repetitive steps.

Table 7- Differences in the automation metrics following $Pipeline_{In\text{-}house}$ and $Pipeline_{Improved}$.

	$Pipeline_{In\text{-}house}$	**$Pipeline_{Improved}$**
Output analysed in read processing	40	20
Web-based tools analysed	4	0
Web-submitted files	40	0
Manually set parameters	16	0
Manual operations	>100	27
Number of repetitive steps	7	2

Discussion

In this work, two different bioinformatic pipelines were used to assemble WGS reads from *E. coli* isolates collected from patients suspected of sepsis. $Pipeline_{In\text{-}house}$ (Figure 1) used Illumina reads, and $Pipeline_{Hybrid}$ (Figure 3) used Illumina and ONT reads. The use of new tools for the different steps in $Pipeline_{In\text{-}house}$ constructed $Pipeline_{Improved}$ (Figure 2). Illumina reads from the *E. coli* isolates were downloaded in fastq files from NCBI, and ONT reads were generated by MinION using DNA extracted from ten biological isolates. Only seven extractions were sequenced due to low concentration in the sequencing library. The raw reads were processed and assembled following different pipelines, and the assembled genomes were then characterised using different tools.

DNA Extractions

DNA extracted from the biological isolates (Appendix IV: Table A1) had lower concentrations than the 33.3 ng/µl advised for MinION (Oxford Nanopore Technologies, 2025). According to QIAGEN (2010), a 200µl elution leads, on average, to a final concentration of 34.0 ng/µl, and in other studies was not higher than 40ng/µl (Raven et al., 2021). Further studies using the same kit could decrease the final elution volume to achieve higher concentrations; however, this is only advised in the protocol for samples with less than 3 µg, corresponding to 15ng/µl in 200µl (QIAGEN, 2010). In the bloodstream, the concentration of *E. coli* varies with the degree of severity and time. Taxt et al. (2020) was able to extract *E. coli* DNA from spiked blood cultures, with a final concentration of more than 50 ng/µl, higher than in this study.

The MiniION platform has purity requirements for the DNA loaded. The A260/230 ratio identifies chemical contamination and is considered optimal between 2.0 and 2.2 but can be considered

clean above 1.8 (Koetsier et al., 2019). The measurement of pure DNA in the A260/280 ratio results in values between 1.8 and 2.0 (Boesenberg-Smith et al., 2012). DNA extracted from some isolates in the purity ratio A260/230 is lower than 1.8, which can reflect some residual ethanol from the extraction kit (Koetsier et al., 2019). In all of the extractions, the purity ratio A 260/280 is in the 1.8-2.0 range (Appendix IV: Table A1). The extracted DNA is longer than 10kbp (Appendix IV: Figure A1) and, in some isolates, longer than 50kbp (Appendix IV: Figure A2), which aligns with the up to 50 kbp expected size (QIAGEN, 2010) and with the long fragments required (Nanopore, 2023). Some spillover from the ladder can be found in the adjacent lane (Appendix IV: Figure A1). The 50 kbp high-range DNA ladder (GeneON) could have been used, but long DNA fragments were ensured, and Gel electrophoresis is most effective between 100bp and 25kbp (Lee et al., 2012).

Quality of Third-Generation Sequencing

The 33.3 ng/µl required for starting material (Oxford Nanopore Technologies, 2025) was not achieved due to low DNA concentrations (Appendix IV: Table A1), which decreased the pool concentration in Section 4: Native Barcode Ligation (Oxford Nanopore Technologies, 2025) from 20ng/µl to 1.70 ng/µl. When choosing which DNA extractions to keep, discard and substitute, several aspects were taken into consideration. For instance, ESBL classification (Table 3) could be more advantageous in the virulence discussion, and it was also considered which of the isolates were common to those used in this experiment. The combination of the DNA used in this experiment and the one taken from another colleague increased the pool concentration to 7 ng/µl.

In 20 h, the R10.4.1 flow cell (FLO-MIN114) sequenced a total of 3.19 Gb and 1.61 M reads, a massive quantity of data, compared to Taylor et al. (2019), which produced 3.8 Gb and 430 thousand (k) reads in 48 h; this leads to more genetic information available for analysis. Not all pores are active during the 20 h sequencing run; ONT pores have a limit on the DNA molecules they can sequence before becoming inactive (Meyer et al., 2026). A mean Q-score of 16 in the sequencing run indicates 97.5% accuracy, and a median of 17 corresponds to 98% accuracy in half of the reads. The median read length of 722bp suggests that half of the reads were shorter than 722bp, which is considered to be short reads. This reduction in length compared to the DNA input can be a result of fragmentation during library preparation (Udaondo et al., 2021). However, the increase in the mean read length to 1953bp (Appendix II: Table A2) shows that, although in a lower number, the long fragments are much bigger, which can also be seen in the N50 read length of 5.23kb, meaning that half of the bases sequenced belong to fragments longer than 5.23kb (Kim & Kim, 2022). All the overall sequencing run quality metrics analysed can be found in Appendix V, Tables A1 and A2 and Figure A1.

The number of reads and bases assigned to each barcode differed; Isolates 1, 3, and 15 had more than 300Mb attributed; nevertheless, the number of total reads in isolate 3 is less than 100, meaning longer reads. The attributed bases and reads for isolates 4 and 10 were quite low compared with the other barcodes. According to ONT, this difference can be explained by differences in ligation efficiency. The number of unclassified reads and bases in our sequencing run corresponds to an incomplete barcode attribution that can arise from failed or incomplete barcode ligation or an ambiguous barcode signal (Oxford Nanopore Technologies, 2025). All the sequencing run metrics per isolate and the unclassified data information can be found in Appendix V, Tables A3 and A4.

Quality Check, Trimming and Genome Assembly

Quality check and trimming were very different using the different pipelines, mostly following $Pipeline_{In\text{-}house}$ (Figure 1) and $Pipeline_{Improved}$ (Figure 2), due to the differences in the tools used. Since different tools were used, the quality of the trimming process itself cannot be discussed, only in terms of user experience. This discussion can be seen further when addressing the differences between these two pipelines. In $Pipeline_{In\text{-}house}$, after trimming using Trimmomatic, the overall quality of the reads increased (Appendix VI: Figure A6) and was considered sufficient for the rest of the bioinformatic analysis, since the high GC content found after filtering corresponds to the bias introduced during DNA amplification used in Illumina Sequencing (Benjamini & Speed, 2012). Trimming using Fastp in $Pipeline_{Hybrid}$ (Figure 3) increased the average read length (Appendix VI: Table A2), since the fragments trimmed were also longer. When comparing the performance of Fastp in $Pipeline_{Improved}$ and $Pipeline_{Hybrid}$ (Appendix VI: Tables A1 and A2), the number of passed and low-quality reads is higher in $Pipeline_{Improved}$, since ONT produces fewer fragments.

The trimmed DNA reads were then assembled to reconstruct the *E. coli* isolate genome, and the quality of the Unicycler assemblies was checked using QUAST. From the report produced by QUAST, it is possible to compare the quality of the assemblies produced by the three different pipelines using different parameters. The assembly quality from $Pipeline_{In\text{-}house}$ (Figure 1) and $Pipeline_{Improved}$ (Figure 2) was equal, as shown in Table 4. However, $Pipeline_{Hybrid}$ (Figure 3) increased the quality parameters (Table 5), since in QUAST, read quality reflects greater contiguity and fewer fragmented contigs (Wick et al., 2017). This connectivity can be seen by the statistically significant increase in N50 read length (paired t-test, $p=0.005$) and decrease in L50 (Wilcoxon signed rank test, $p=0.022$). The assembly performed using $Pipeline_{Hybrid}$ (Table 5) resulted in a more complete and less fragmented genome, as evidenced by the reduction in the number of contigs and the L50 parameter, and the increase in the total length and the N50 parameter. *E. coli* genome hybrid assemblies were already shown to be less fragmented than Illumina assemblies by Khezri et al. (2021). In Chakrawarti et al. (2024), the quality of the assembly also improved with the hybrid assembly, when compared to assemblies done using only Illumina reads. The assembly with the reads from isolate 3, following $Pipeline_{In\text{-}house}$ and $Pipeline_{Improved}$ (Table 4), was very fragmented. However, the use of $Pipeline_{Hybrid}$ (Table 5) decreased this fragmentation and increased the assembly quality. The poor quality of the assembly that results from the fragments from isolate 4 (Table 5) can be explained by the fewer bases and reads attributed to isolate 4 during sequencing (Appendix V: Table A2), leading to a prevalence of Illumina reads in the Hybrid assembly.

Species and Virulence Identification

Since this study used DNA and reads from isolates of patients with suspected sepsis, an investigation was also performed on genome annotation, such as species identification or ARGs presence. The analysis of the assembled genomes helped to understand whether using $Pipeline_{Hybrid}$ (Figure 3) improves clinical diagnosis. The comparison of the different tools used for analysis following $Pipeline_{In\text{-}house}$ (Figure 1) and $Pipeline_{Improved}$ (Figure 2) can also help in clinical diagnosis by enhancing output information, which in a clinical environment can increase specialised patients' treatment. The results were generated using the CGE website following $Pipeline_{In\text{-}house}$ (Figure 1) and using FastANI, MLST, ABRicate and NCBI-amrfinderplus in $Pipeline_{Hybrid}$ (Figure 3) and $Pipeline_{Improved}$ (Figure 2). MLST divides the species into different Sequence Types (ST), according to their similarity in different genes (Jolley & Maiden, 2014). The

results from FastANI ($Pipeline_{Hybrid}$ and $Pipeline_{Improved}$) and CGE SpeciesFinder ($Pipeline_{In\text{-}house}$), and from the ST categorisation (Table 6), showed some expected similarities with results already found by Tilevik et al. (2022), which used the same isolates as in the present study.

The misidentification in isolates 1 and 10, which were classified as *Shigella Sonnei* (Table 6), is a common mistake due to the similarity between the two species (Ljungström et al., 2019). FastANI only identified the reference genome closer to the genome assembled (Jain et al., 2018); Thus, the misidentification was not a problem in $Pipeline_{Improved}$ (Figure 2) and $Pipeline_{Hybrid}$ (Figure 3), since only *E. coli* genomes were used as reference genomes. Since CGE SpeciesFinder and FastANI use different reference genome databases, an accurate comparison is difficult. However, it is known from Ljungström et al. (2019)that the 10 isolates used in this experiment are *E. coli* isolates; since FastANI uses *E. coli*-only genomes in the reference genomes database, it does not make the species misidentification from CGE SpeciesFinder (Table 6).

Isolates number 2, 12 and 15 were identified as the same *E. coli* strain using both $Pipeline_{In\text{-}house}$ and $Pipeline_{Improved}$; however, only isolate 15 was classified as the same strain following $Pipeline_{Hybrid}$ (Table 6). According to Shaw & Yu (2023), FastANI increases accuracy with the decrease in assembly fragmentation. This can explain the differences in the reference genomes considered closest from $Pipeline_{Improved}$ (Figure 2) and $Pipeline_{Hybrid}$ (Figure 3). Overall, FastANI, following $Pipeline_{Improved}$ and $Pipeline_{Hybrid}$, provided more specific results than CGE SpeciesFinder, following $Pipeline_{In\text{-}house}$. However, in a clinical environment, species identification is necessary and not only strain or reference genome identification. Thus, the reference genome database in the analysis site has to be more complex and store more information.

From the 11000 existing *E. coli* ST (Elankumaran et al., 2022), the isolates used in this experiment belonged to eight different groups: ST127, ST10, ST131, ST80, ST744, ST91 and ST48. All of the ST categories identified for the isolates correspond to the ones previously identified by Tilevik et al. (2022). Overall, $Pipeline_{Hybrid}$ improved MLST identification, enabling isolate 3 identification as ST10 (Table 6), not achieved using Illumina reads ($Pipeline_{In\text{-}house}$ and $Pipeline_{Improved}$). In terms of the number of ARGs (Table 6), the differences across the different pipelines were not statistically significant (Friedman rank-sum test, $p=0.82$); what differed was which genes were detected, according to the Jaccard Similarity (Appendix VII: Table A1). Both $Pipeline_{In\text{-}house}$ and $Pipeline_{Improved}$ identified the same virulence genes using CGE VirulenceFinder/ResFinder and ABRicate/NCBI-amrfinderplus, respectively; however, differences can be found when comparing to $Pipeline_{Hybrid}$ (Appendix VII: Table A1 and A3).

$Pipeline_{Hybrid}$ was not able to identify any gene from the β-lactams antibiotically subclass in isolate 1 (Appendix VII: Table A3) found by $Pipeline_{In\text{-}house}$ and $Pipeline_{Improved}$, which are very commonly associated with ST127 (Table 6), an *E. coli* ST associated with most urinary tract and bloodstream infections (Elankumaran et al., 2022). *E. coli* ST131, identified in isolate 4 (Table 6), was the predominant sequence type found in Sweden by Brolund et al. (2014), and they are frequently found with acquired resistance to β-lactamases and fluoroquinolones (Mackow et al., 2026), also identified by the pipelines used (Appendix VII: Tables A2 and A3). $Pipeline_{Hybrid}$ not only identified genes mdf(A) and lnu(F) but also identified different strains of the same virulent gene, not identified by $Pipeline_{In\text{-}house}$ and $Pipeline_{Improved}$. However, $Pipeline_{Hybrid}$ did not identify some genes in the genomes of isolates 1, 2, 4, and 12, identified by $Pipeline_{In\text{-}house}$ and $Pipeline_{Improved}$. Further studies are needed to test $Pipeline_{Hybrid}$, since the non-identification of the β-lactams antibiotically subclass is a sign of lower accuracy in gene detection. This result was not expected since, in other studies, the knowledge provided by a hybrid assembly improves clinical diagnosis by identifying

genes encoding resistance to different antibiotic classes, normally identifying genes not identified using only Illumina assembly (Khezri et al., 2021). The only possible reason is a problem with the ONT reads used, since in $Pipeline_{Improved}$ the result was more similar to the expected result and the bioinformatic tools used were the same as $Pipeline_{Hybrid}$. The low DNA concentrations (Appendix IV: Table A1) and consequent low library concentration can have an impact, since less genetic information was available, not reaching the starting advised material from ONT (Oxford Nanopore Technologies, 2025). However, this does not match the fact that isolate 1 was one of the isolates with more and longer reads sequenced (Appendix V: Table A3). New phenotypic antibiotic results should be obtained due to the possibility of antibiotic resistance loss with the lack of exposure to decrease fitness cost (Kim et al., 2021).

Comparison between $Pipeline_{In\text{-}house}$ and $Pipeline_{Improved}$

In this study, the new tools used by $Pipeline_{Improved}$ (Figure 2) aimed to enhance the bioinformatic workflow, when compared to $Pipeline_{In\text{-}house}$. Several differences can be found in read processing and genome annotation following $Pipeline_{In\text{-}house}$ and $Pipeline_{Improved}$. $Pipeline_{In\text{-}house}$ requires a read quality check before and after trimming, doubling the number of files analysed in read processing, compared to $Pipeline_{Improved}$ (Table 7). The filtering report provided by Fastp showed reports for the pre- and post-filtered data on the same page, allowing a more direct quality comparison (Chen et al., 2018). The number of manually set parameters also changes, since it is based on the pre-trimming quality that the trimming parameters are set; whereas in $Pipeline_{Improved}$, Fastp sets them automatically. Fastp has already been shown to be more useful due to its simplicity, efficiency, versatility, and reproducibility when compared to Trimmomatic (Chen, 2025).

The use of CGE genome annotation tools also made the analysis with $Pipeline_{In\text{-}house}$ more manual, since in each tool the genome file needed to be manually submitted and the parameters set by the user. Thus, made the pipeline more repetitive and increased the number of manual operations, leading to a more human-dependent pipeline. More automatic pipelines, as $Pipeline_{Improved}$ showed (Table 7), enhance clinical diagnosis by improving reliability and reproducibility and decreasing the number of possible human errors (Alqadi et al., 2024).

Conclusion

In summary, combining Illumina and ONT reads in $Pipeline_{Hybrid}$ improved the quality of the genome assembly. This led to an improvement in the genome annotation analysis: Species identification did not show any mismatches when comparing the assembled genomes with reference genomes from different *E. coli* strains; and MLST categorisation was able to be performed on genomes that $Pipeline_{In\text{-}house}$ and $Pipeline_{Improved}$ could not categorise. Although there were differences in ARG detection across the different pipelines, these results were not considered statistically significant.

The new tools used for post-sequencing bioinformatic analysis by $Pipeline_{Hybrid}$ and $Pipeline_{Improved}$ upgraded the bioinformatic workflow, making the analysis less repetitive and more automatic when compared with $Pipeline_{In\text{-}house}$.

Limitations

In this work, there are some limitations, and here are the major ones: 1) The biggest limitation of this study was the lack of phenotypic antibiotic resistance testing in all the antibiotic subclasses identified by the three different pipelines, which did not allow an understanding of their true accuracy. 2)Another limitation was the low DNA concentrations, which might have affected the amount of genetic information sequenced, influencing the downstream analysis. 3) The different reference genome databases used in species identification inhibit a fair and true comparison between CGE SpeciesFinder and FastANI.

Ethical aspects, sex and gender perspectives, and impact on society

This project impacts sepsis diagnosis in the hospital since hybrid assembly could generate more and more accurate genome information about the bacterium causing the human organism's dysfunction. This will improve specific patient treatment and reduce the mortality rate from Sepsis. All of the data and the isolates used in this experiment were provided by Ljungström et al. (2019), who studied Sepsis in adult patients (>18 years) admitted to the Skaraborg Hospital (between 8th September 2011 and 7th June 2012) who were suspected of suffering from sepsis and treated with intravenous antibiotics within 48 h after entering the hospital. The ethical approval for Ljungström et al. (2019) was approved by the Ethical Review Board at the University of Gothenburg (permit 376/2011). This permit was used for this project since no new samples and data were collected. Ljungström et al. (2019) was an observational study, with no active intervention for the participants. Thus, the samples were collected as events normally occur, not adding any risk to the patients' health. In observational studies, a written informed consent document signed by the patients is not mandatory. However, working with Whole Genome Sequencing can have some ethical challenges that should be considered in non-observational studies. It can be very difficult to have truly informed consent from the patients since they cannot understand the complex mechanisms behind these technologies, as well as their potential. Alongside this, the risk and cost of storage for these large amounts of data and the handling of the actionable and non-actionable results should be taken into consideration (Pinxten & Howard, 2014). Sepsis is caused by microorganisms, which do not differ in sex/gender, so when analysing the microorganism itself, gender is not important to consider. However, according to Merdji et al. (2023) sepsis infections are more frequent and severe in males than females, leading also to an increase in the male mortality rate due to a sepsis infection. This is due to differences in the immune system response caused by gene expression in the XX and XY chromossomes in females and males, respectively. Thus, studies that involve patients should consider these differences.

Future perspectives

To better test the accuracy of the hybrid assembly, a new study should be done including phenotypic antibiotic resistance, as previously done by Tilevik et al. (2022) using all the antibiotics identified by the genome annotation analysis. That would be the most accurate way to understand the viability of this pipeline.

The knowledge that hybrid assembly provides of the genomes assembled has already shown to be very helpful in understanding the symbiotic relationship between hosts and microbes, such as their potential to degrade carbohydrates (Ravi et al., 2022) . More studies using hybrid assembly to assemble the genomes of the human microbiota can amplify the knowledge about human organism function.

Not only in human health but also in plant health, Hybrid assembly has already proven to improve genome analysis. Hybrid assembly improved the diagnosis and the treatment of *Cannabis Sativa* against *Berkeleyomyces rouxiae,* which causes black root rot (Dumigan et al., 2023). Thus, further studies can also improve botany, especially in the medical field, helping in pathogen diagnosis.

Acknowledgments

First, I would like to express my gratitude to my supervisor Anna-Karin Pernesting, professor at the University of Skövde, for the guidance, support and constructive feedback. Her patience, motivation and expertise helped not only this thesis but also my academic and personal growth.

Then, I would like to thank professors Sanja Jurcevic and Magnus Fagerlind for all the support and academic knowledge provided. In this line, I would like to thank Atif, Jonas and Nafees for all the great environment during the laboratory part and for the discussions about the results. I would also like to thank the staff from the University of Skövde, particularly Frida, Ines, Karin and Sara, for all the support in the academic environment and friendly relationship during this time.

Most importantly, I would like to thank my family for all the support provided during my studies abroad. Not only for the emotional support through encouragement, love, and understanding, which helped me achieve this goal on the days when the sun did not come out. But also all the financial support that helped me follow my dreams. Even from afar, they brought peace and happiness to my life. I would also like to thank all of my friends, the new and the old ones. To my friends back in Portugal, I would like to thank all the video calls and messages, all the goodbye tears that turned into huge smiles and hugs in the re-encounter. Specifically, I would like to thank Maria, Ines and Claudia for all the patience and loving words that helped me not give up. For the new friends that I made in Skövde, I would like to thank all the meetups and all the fun that made me feel like I had a home away from home. I would like to say a special thank you to Mojca, Cata, Lena, Taylor and Lorenz for all the support in the bad moments. For being there and making me feel secure in the worst days.

Thank you to everyone who made me part of their lives directly or indirectly during these 2 years. Everyone who helped me grow and changed a little bit of me. Lastly, I thank God for all the guidance and strength to complete this work.

References

Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., & Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, *12*(2). https://doi.org/10.1186/gb-2011-12-2-r18

Alqadi, O. mohamed, Majed Musa Saad Alsuwairi , Dalal Murdhi Alshammari, Abdulrahman Saeed Alshehri , Torki Nasser Aloraifi, Reada Nasser Alsayegh, Alaa Ahmed Abdulaziz Alraddadi, & Anas Mohammed ALiabal. (2024). The Impact of Automation on Clinical Laboratory Efficiency and Error Reduction. *Saudi Journal of Medicine and Public Health*, *1*(2), 927–939. https://doi.org/10.64483/202412249

Beggio, M., Peroni, E., Greco, E., Favretto, G., Degiorgio, D., Rosato, A., & Favarato, M. (2025). Methodological Assessment of High-Throughput Sequencing Platforms: Illumina vs. MGI in Clinical-Grade CFTR Genotyping. *International Journal of Molecular Sciences*, *26*(23). https://doi.org/10.3390/ijms262311701

Benjamini, Y., & Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, *40*(10). https://doi.org/10.1093/nar/gks001

Blount, Z. D. (2015). The unexhausted potential of E. coli. In *eLIFE*. The Natural History of Model Organisms. https://doi.org/10.7554/eLife.05826.001

Boesenberg-Smith, K. A., Pessarakli, M. M., & Wolk, D. M. (2012). Assessment of DNA Yield and Purity: an Overlooked Detail of PCR troubleshooting. *Clinical Microbiology Newsletter*, (34).

Brlek, P., Bulić, L., Bračić, M., Projić, P., Škaro, V., Shah, N., Shah, P., & Primorac, D. (2024). Implementing Whole Genome Sequencing (WGS) in Clinical Practice: Advantages, Challenges, and Future Perspectives. In *Cells* (Vol. 13, Number 6). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/cells13060504

Brolund, A., Edquist, P. J., Mäkitalo, B., Olsson-Liljequist, B., Söderblom, T., Tegmark Wisell, K., & 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. https://doi.org/10.1111/1469-0691.12413

Brown, J., Pirrung, M., & Mccue, L. A. (2017). FQC Dashboard: Integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics*, *33*(19), 3137–3139. https://doi.org/10.1093/bioinformatics/btx373

Chakrawarti, A., Eckstrom, K., Laaguiby, P., & Barlow, J. (2024). Hybrid Illumina-Nanopore assembly improves identification of multilocus sequence types and antimicrobial resistance genes of Staphylococcus aureus isolated from Vermont dairy farms: comparison to Illumina-only and R9.4.1 nanopore-only assemblies. In *Access Microbiology*. https://doi.org/10.1099/acmi.0.000766.v2

Chen, S. (2025). fastp 1.0: An ultra-fast all-round tool for FASTQ data quality control and preprocessing. *IMeta*, *4*(5). https://doi.org/10.1002/imt2.70078

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Chen, Z., Erickson, D. L., & Meng, J. (2020). Benchmarking hybrid assembly approaches for genomic analyses of bacterial pathogens using Illumina and Oxford Nanopore sequencing. *BMC Genomics*, *21*(1). https://doi.org/10.1186/s12864-020-07041-8

Delahaye, C., & Nicolas, J. (2021). Sequencing DNA with nanopores: Troubles and biases. *PLoS ONE*, *16*(10 October). https://doi.org/10.1371/journal.pone.0257521

Dumigan, C. R., Maddock, S., Bray-Stone, D., & Deyholos, M. K. (2023). Hybrid Genome Assembly of Berkeleyomyces rouxiae, an Emerging Cannabis Fungal Pathogen Causing Black Root Rot in an Aeroponic Facility. *Plant Disease*, *107*(9), 2679–2686. https://doi.org/10.1094/PDIS-11-22-2690-RE

Elankumaran, P., Browning, G. F., Marenda, M. S., Reid, C. J., & Djordjevic, S. P. (2022). Close genetic linkage between human and companion animal extraintestinal pathogenic Escherichia coli ST127. *Current Research in Microbial Sciences*, *3*. https://doi.org/10.1016/j.crmicr.2022.100106

Eleiwa, A., Nadal, J., Vilaprinyo, E., Marin-Sanguino, A., Sorribas, A., Basallo, O., Lucido, A., Richart, C., Pena, R. N., Ros-Freixedes, R., Usie, A., & Alves, R. (2024). Hybrid assembly and comparative genomics unveil insights into the evolution and biology of the red-legged partridge. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-70018-0

Emiyu, K., & Lelisa, K. (2022). Review on Illumina Sequecing Technologies. *Austin Journal of Veterinary Science & Animal Husbandry*. www.austinpublishinggroup.com

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Feldgarden, M., Brover, V., Gonzalez-Escalona, N., Frye, J. G., Haendiges, J., Haft, D. H., Hoffmann, M., Pettengill, J. B., Prasad, A. B., Tillman, G. E., Tyson, G. H., & Klimke, W. (2021). AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-91456-0

Fleischmann, C., Scherag, A., Adhikari, N. K. J., Hartog, C. S., Tsaganos, T., Schlattmann, P., Angus, D. C., & Reinhart, K. (2016). Assessment of global incidence and mortality of hospital-treated sepsis current estimates and limitations. *American Journal of Respiratory and Critical Care Medicine*, *193*(3), 259–272. https://doi.org/10.1164/rccm.201504-0781OC

Gorecki, G. P., Bodor, A., Novac, M. B., Costea, D. G., Costea, D. O., Costea, A. C., Grasa, C. N., & Tomescu, D. R. (2025). Rethinking post-sepsis syndrome: linking cellular dysfunction to the clinical picture. In *Critical Care* (Vol. 29, Number 1). BioMed Central Ltd. https://doi.org/10.1186/s13054-025-05491-8

Gorman, Z., Chen, J., de Leon, A. A. P., & Wallis, C. M. (2023). Comparison of assembly platforms for the assembly of the nuclear genome of Trichoderma harzianum strain PAR3. *BMC Genomics*, *24*(1). https://doi.org/10.1186/s12864-023-09544-6

Gosiewski, T., Ludwig-Galezowska, A. H., Huminska, K., Sroka-Oleksiak, A., Radkowski, P., Salamon, D., Wojciechowicz, J., Kus-Slowinska, M., Bulanda, M., & Wolkow, P. P. (2017). Comprehensive detection and identification of bacterial DNA in the blood of patients with sepsis and healthy

volunteers using next-generation sequencing method - the observation of DNAemia. *European Journal of Clinical Microbiology and Infectious Diseases*, *36*(2), 329–336. https://doi.org/10.1007/s10096-016-2805-7

Grada, A., & Weinbrecht, K. (2013). Next-generation sequencing: Methodology and application. *Journal of Investigative Dermatology*, *133*(8). https://doi.org/10.1038/jid.2013.248

Grumaz, C., Hoffmann, A., Vainshtein, Y., Kopp, M., Grumaz, S., Stevens, P., Decker, S. O., Weigand, M. A., Hofer, S., Brenner, T., & Sohn, K. (2020). Rapid Next-Generation Sequencing–Based Diagnostics of Bacteremia in Septic Patients. *Journal of Molecular Diagnostics*, *22*(3), 405–418. https://doi.org/10.1016/j.jmoldx.2019.12.006

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-07641-9

Jauréguy, F., Carbonnelle, E., Bonacorsi, S., Clec'h, C., Casassus, P., Bingen, E., Picard, B., Nassif, X., & Lortholary, O. (2007). Host and bacterial determinants of initial severity and outcome of Escherichia coli sepsis. *Clinical Microbiology and Infection*, *13*(9), 854–862. https://doi.org/10.1111/j.1469-0691.2007.01775.x

Jolley, K. A., & Maiden, M. C. J. (2010). BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*, *11*. https://doi.org/10.1186/1471-2105-11-595

Jolley, K. A., & Maiden, M. C. J. (2014). Using MLST to study bacterial variation: Prospects in the genomic era. *Future Microbiology*, *9*(5), 623–630. https://doi.org/10.2217/fmb.14.24

Khezri, A., Avershina, E., & Ahmad, R. (2021). Hybrid assembly provides improved resolution of plasmids, antimicrobial resistance genes, and virulence factors in escherichia coli and klebsiella pneumoniae clinical isolates. *Microorganisms*, *9*(12). https://doi.org/10.3390/microorganisms9122560

Kim, J., & Kim, C. (2022). A beginner's guide to assembling a draft genome and analyzing structural variants with long-read sequencing technologies. *STAR Protocols*, *3*(3). https://doi.org/10.1016/j.xpro.2022.101506

Kim, M., Park, J., Kang, M., Yang, J., & Park, W. (2021). Gain and loss of antibiotic resistant genes in multidrug resistant bacteria: One Health perspective. In *Journal of Microbiology* (Vol. 59, Number 6, pp. 535–545). The Korean Society for Mocrobiology / The Korean Society of Virology. https://doi.org/10.1007/s12275-021-1085-9

Kircher, M., Heyn, P., & Kelso, J. (2011). Addressing challenges in the production and analysis of illumina sequencing data. In *BMC Genomics* (Vol. 12). https://doi.org/10.1186/1471-2164-12-382

Koetsier, G., Cantor, E., & Biolabs, E. (2019). *A Practical Guide to Analyzing Nucleic Acid Concentration and Purity with Microvolume Spectrophotometers*.

Lee, P. Y., Costumbrado, J., Hsu, C. Y., & Kim, Y. H. (2012). Agarose gel electrophoresis for the separation of DNA fragments. *Journal of Visualized Experiments*, (62). https://doi.org/10.3791/3923

Lember, G. (2021). *SEPSIS-ASSOCIATED ESCHERICHIA COLI WHOLE-GENOME SEQUENCING ANALYSIS USING IN-HOUSE DEVELOPED PIPELINE AND 1928 DIAGNOSTICS TOOL*.

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). https://doi.org/10.1371/journal.pone.0225700

Mackow, N. A., Shao, W., Ge, L., Komarow, L., Jiang, J., Boutzoukas, A., Chen, L., Garcia-Diaz, J., Herc, E. S., Doi, Y., Arias, C. A., Albin, O., Saade, E., Miller, L. G., Jacob, J. T., Satlin, M. J., Krsak, M., Huskins, W. C., Dhar, S., … Doernberg, S. (2026). Escherichia coli ST131 Drives Carbapenem Use for E. coli Bloodstream Infections. *Clinical Infectious Diseases*. https://doi.org/10.1093/cid/ciag160

Mahmoud, N., & Enroth, H. (2021). *NEXT-GENERATION DIAGNOSTICS OF ESCHERICHIA COLI FROM COMMUNITY-ONSET SEPSIS PATIENTS IN SWEDEN: STUDYING THE BIODIVERSITY OF ESCHERICHIA COLI GENOMES*.

Mancini, N., Carletti, S., Ghidoli, N., Cichero, P., Burioni, R., & Clementi, M. (2010). The era of molecular and other non-culture-based methods in diagnosis of sepsis. In *Clinical Microbiology Reviews* (Vol. 23, Number 1, pp. 235–251). https://doi.org/10.1128/CMR.00043-09

McCombie, W. R., McPherson, J. D., & Mardis, E. R. (2019). Next-generation sequencing technologies. *Cold Spring Harbor Perspectives in Medicine*, *9*(11). https://doi.org/10.1101/cshperspect.a036798

Merdji, H., Long, M. T., Ostermann, M., Herridge, M., Myatra, S. N., De Rosa, S., Metaxa, V., Kotfis, K., Robba, C., De Jong, A., Helms, J., & Gebhard, C. E. (2023). Sex and gender differences in intensive care medicine. In *Intensive Care Medicine* (Vol. 49, Number 10, pp. 1155–1167). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s00134-023-07194-6

Meyer, D., Goettsch, W., Spangenberg, J., Stieber, B., Krautwurst, S., Hölzer, M., Brandt, C., Linde, J., Höner zu Siederdissen, C., Srivastava, A., Zarkovic, M., Wollny, D., & Marz, M. (2026). Unlocking the full potential of nanopore sequencing: tips, tricks, and advanced data analysis techniques. *Nucleic Acids Research*, *54*(3). https://doi.org/10.1093/nar/gkag023

Mora-Rillo, M., Fernández-Romero, N., Navarro-San Francisco, C., Díez-Sebastian, J., Romero-Gómez, M. P., Fernández, F. A., Lopez, J. R. A., & Mingorance, J. (2015). Impact of virulence genes on sepsis severity and survival in Escherichia coli bacteremia. *Virulence*, *6*(1), 93–100. https://doi.org/10.4161/21505594.2014.991234

Oxford Nanopore Technologies. (2025). *Ligation sequencing gDNA - Native Barcoding Kit 96 V14 (SQK-NBD114.96)*.

Pareek, C. S., Smoczynski, R., & Tretyn, A. (2011). Sequencing technologies and genome sequencing. In *Journal of Applied Genetics* (Vol. 52, Number 4, pp. 413–435). https://doi.org/10.1007/s13353-011-0057-x

Peri, A. M., Harris, P. N. A., & Paterson, D. L. (2022). Culture-independent detection systems for bloodstream infection. In *Clinical Microbiology and Infection* (Vol. 28, Number 2, pp. 195–201). Elsevier B.V. https://doi.org/10.1016/j.cmi.2021.09.039

Petersen, L. M., Martin, I. W., Moschetti, W. E., Kershaw, C. M., & Tsongalis, G. J. (2019). *Third-Generation Sequencing in the Clinical Laboratory: Exploring the Advantages and Challenges of Nanopore Sequencing*. https://journals.asm.org/journal/jcm

Pinxten, W., & Howard, H. C. (2014). Ethical issues raised by whole genome sequencing. In *Best Practice and Research: Clinical Gastroenterology* (Vol. 28, Number 2, pp. 269–279). Bailliere Tindall Ltd. https://doi.org/10.1016/j.bpg.2014.02.004

QIAGEN. (2010). *QIAamp ® DNA Mini and Blood Mini Handbook Sample & Assay Technologies*.

Raven, K. E., Girgis, S. T., Akram, A., Blane, B., Leek, D., Brown, N., & Peacock, S. J. (2021). A common protocol for the simultaneous processing of multiple clinically relevant bacterial species for whole genome sequencing. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-020-80031-8

Ruan, Z., Wu, J., Chen, H., Draz, M. S., Xu, J., & He, F. (2020). Hybrid genome assembly and annotation of a pandrug-resistant klebsiella pneumoniae strain using nanopore and illumina sequencing. *Infection and Drug Resistance*, *13*, 199–206. https://doi.org/10.2147/IDR.S240404

Saxenborn, P., Baxter, J., Tilevik, A., Fagerlind, M., Dyrkell, F., Pernestig, A. K., Enroth, H., & Tilevik, D. (2021). Genotypic Characterization of Clinical Klebsiella spp. Isolates Collected From Patients With Suspected Community-Onset Sepsis, Sweden. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/fmicb.2021.640408

Shaw, J., & Yu, Y. W. (2023). Fast and robust metagenomic sequence comparison through sparse chaining with skani. *Nature Methods*, *20*(11), 1661–1665. https://doi.org/10.1038/s41592-023-02018-3

Shemirani, M. I., Tilevik, D., Tilevik, A., Jurcevic, S., Arnellos, D., Enroth, H., & Pernestig, A. K. (2023). Benchmarking of two bioinformatic workflows for the analysis of whole-genome sequenced Staphylococcus aureus collected from patients with suspected sepsis. *BMC Infectious Diseases*, *23*(1). https://doi.org/10.1186/s12879-022-07977-0

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., & Waterston, R. H. (2017). DNA sequencing at 40: Past, present and future. *Nature*, *550*(7676). https://doi.org/10.1038/nature24286

Singer, M., Deutschman, C. S., Seymour, C., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J. D., Coopersmith, C. M., Hotchkiss, R. S., Levy, M. M., Marshall, J. C., Martin, G. S., Opal, S. M., Rubenfeld, G. D., Poll, T. Der, Vincent, J. L., & Angus, D. C. (2016). The third international consensus definitions for sepsis and septic shock (sepsis-3). In *JAMA - Journal of the American Medical Association* (Vol. 315, Number 8, pp. 801–810). American Medical Association. https://doi.org/10.1001/jama.2016.0287

Soto-Serrano, A., Li, W., Panah, F. M., Hui, Y., Atienza, P., Fomenkov, A., Roberts, R. J., Deptula, P., & Krych, L. (2024). *Matching Excellence: ONT's Rise to Parity with PacBio in Genome*

Reconstruction of Non-Model Bacterium with High GC Content. https://doi.org/10.1101/2024.02.26.582104

Sreeremya, S. (2026). Bacterial Whole Genome Assembly Using SPAdes-Review. *Journal of Research in Microbiology and Immunology*, *8*(1).

Srzić, I., Adam, V. N., & Pejak, D. T. (2022). SEPSIS DEFINITION: WHAT'S NEW IN THE TREATMENT GUIDELINES. *Acta Clinica Croatica*, *61*, 67–72. https://doi.org/10.20471/acc.2022.61.s1.11

Stein, L. (2001). Genome annotation: from sequence to biology. *Nature Reviews Genetics*, *2*(7), 493–503. https://doi.org/10.1038/35080529

Taj, M. K., Samreen, Z., Ling, J. X., Taj, I., Hassan, T., & Yunlin, W. (2014). Escherichia coli as a model organism. *International Journal of Engineering Research and Science and Technology*, *3*(2), 1–8.

Taxt, A. M., Avershina, E., Frye, S. A., Naseer, U., & Ahmad, R. (2020). Rapid identification of pathogens, antibiotic resistance genes and plasmids in blood cultures by nanopore sequencing. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-64616-x

Taylor, T. L., Volkening, J. D., DeJesus, E., Simmons, M., Dimitrov, K. M., Tillman, G. E., Suarez, D. L., & Afonso, C. L. (2019). Rapid, multiplexed, whole genome and plasmid sequencing of foodborne pathogens using long-read nanopore technology. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-52424-x

Thomas, M. S., & Wigneshweraraj, S. (2014). Regulation of virulence gene expression. In *Virulence* (Vol. 5, Number 8, pp. 832–834). Landes Bioscience. https://doi.org/10.1080/21505594.2014.995573

Tilevik, D., Pernestig, A. K., Fagerlind, M., Tilevik, A., Ljungström, L., Johansson, M., & Enroth, H. (2022). Sequence-based genotyping of extra-intestinal pathogenic Escherichia coli isolates from patients with suspected community-onset sepsis, Sweden. *Microbial Pathogenesis*, *173*. https://doi.org/10.1016/j.micpath.2022.105836

Torres, J. S. S., Tamayo-Giraldo, F. J., Bejarano-Zuleta, A., Nati-Castillo, H. A., Quintero, D. A., Ospina-Mejía, M. J., Salazar-Santoliva, C., Suárez-Sangucho, I., Ortiz-Prado, E., & Izquierdo-Condoy, J. S. (2025). Sepsis and post-sepsis syndrome: a multisystem challenge requiring comprehensive care and management—a review. In *Frontiers in Medicine* (Vol. 12). Frontiers Media SA. https://doi.org/10.3389/fmed.2025.1560737

Udaondo, Z., Sittikankaew, K., Uengwetwanit, T., Wongsurawat, T., Sonthirod, C., Jenjaroenpun, P., Pootakham, W., Karoonuthaisiri, N., & Nookaew, I. (2021). Comparative analysis of pacbio and oxford nanopore sequencing technologies for transcriptomic landscape identification of penaeus monodon. *Life*, *11*(8). https://doi.org/10.3390/life11080862

Wan, Y., Pike, R., Harley, A., Mumin, Z., Potterill, I., Meunier, D., Ganner, M., Getino, M., Coelho, J., Jauneikaite, E., Moganeradj, K., Brown, C. S., Holmes, A. H., Demirjian, A., Hopkins, K. L., & Pichon, B. (2025). Complete genome assemblies and antibiograms of 22 Staphylococcus capitis isolates. *BMC Genomic Data*, *26*(1). https://doi.org/10.1186/s12863-025-01303-8

Wang, W., & Liu, C. F. (2023). Sepsis heterogeneity. In *World Journal of Pediatrics* (Vol. 19, Number 10, pp. 919–927). Zhejiang University School of Medicine Children's Hospital. https://doi.org/10.1007/s12519-023-00689-8

Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., Buck, D., & Au, K. F. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*, *6*, 100. https://doi.org/10.12688/f1000research.10571.1

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). https://doi.org/10.1371/journal.pcbi.1005595

Wyres, K. L., Conway, T. C., Garg, S., Queiroz, C., Reumann, M., Holt, K., & Rusu, L. I. (2014). WGS analysis and interpretation in clinical and public health microbiology laboratories: What are the requirements and how do existing tools compare? *Pathogens*, *3*(2), 437–458. https://doi.org/10.3390/pathogens3020437

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics*, *29*(21), 2669–2677. https://doi.org/10.1093/bioinformatics/btt476

Zuo, Z., Pei, L., Liu, T., Liu, X., Chen, Y., & Hu, Z. (2023). Investigation of Gut Microbiota Disorders in Sepsis and Sepsis Complicated with Acute Gastrointestinal Injury Based on 16S rRNA Genes Illumina Sequencing. *Infection and Drug Resistance*, *16*, 7389–7403. https://doi.org/10.2147/IDR.S440335

Appendix

Appendix I: Code for Pipeline$_{\text{In-house}}$

```
#Download of the Illumina samples

mkdir ecoli_project

cd ecoli_project

prefetch SRR11179862 SRR11179861 SRR11179850 SRR11179848 SRR11179842 SRR11179859 SRR11179858 SRR11179857 SRR11179856 SRR11179855

for srr in SRR11179862 SRR11179861 SRR11179850 SRR11179848 SRR11179842 SRR11179859 SRR11179858 SRR11179857 SRR11179856 SRR11179855

do

  fasterq-dump $srr --split-files --threads 8

  gzip ${srr}_1.fastq

  gzip ${srr}_2.fastq

done

ls *.fastq.gz

#Quality control using FastQC

for dir in SRR*

do

  fastqc \

    ${dir}/${dir}_1.fastq.gz \

    ${dir}/${dir}_2.fastq.gz \

    -o fastqc_raw

Done

#Trimming using Trimmomatric

for sample in SRR11179862 SRR11179861 SRR11179850 SRR11179848 SRR11179842 SRR11179859 SRR11179858 SRR11179857 SRR11179856 SRR11179855

do

trimmomatic PE -threads 4 \

${sample}_1.fastq ${sample}_2.fastq \

 ${sample}_1_paired.fastq ${sample}_1_unpaired.fastq \

 ${sample}_2_paired.fastq ${sample}_2_unpaired.fastq \
```

ILLUMINACLIP:/usr/share/trimmomatic/NexteraPE-PE.fa:2:40:15 \ SLIDINGWINDOW:4:25 \ LEADING:3 TRAILING:3 \ MINLEN:50 done Explain this line by line to me please

Post-trimming Quality Check using FastQC

fastqc \

trimmed_reads/${dir}_1_paired.fastq.gz \

trimmed_reads/${dir}_2_paired.fastq.gz \

-o fastqc_trimmed

done

#Unicycler Assembly

for sample in SRR11179862 SRR11179861 SRR11179850 SRR11179848 SRR11179842 SRR11179859 SRR11179858 SRR11179857 SRR11179856 SRR11179855

do

unicycler \

-1 ${sample}_1_paired.fastq \

-2 ${sample}_2_paired.fastq \

-o assembly_${sample} \

-t 4

Done

#Assembly Quality check with QUAST

quast assembly_*/assembly.fasta -o quast_report

xdg-open quast_SAMPLE/report.html

#Post assembly analysis made in the Centre for Genomic Epidmiologic website

Appendix II: Code for Pipeline$_{Improved}$

```
#Download of the Illumina samples

mkdir ecoli_project

cd ecoli_project

prefetch SRR11179862 SRR11179861 SRR11179850 SRR11179848 SRR11179842 SRR11179859 SRR11179858 SRR11179857 SRR11179856 SRR11179855

for srr in SRR11179862 SRR11179861 SRR11179850 SRR11179848 SRR11179842 SRR11179859 SRR11179858 SRR11179857 SRR11179856 SRR11179855

do

  fasterq-dump $srr --split-files --threads 8

  gzip ${srr}_1.fastq

  gzip ${srr}_2.fastq

done

ls *.fastq.gz

#Trimming and Quality check with Fastp

mkdir fastp_reads

for sample in SRR11179862 SRR11179861 SRR11179850 SRR11179848 SRR11179842 \
        SRR11179859 SRR11179858 SRR11179857 SRR11179856 SRR11179855

do

  fastp \
    -i ${sample}_1_paired.fastq.gz \
    -I ${sample}_2_paired.fastq.gz \
    -o fastp_reads/${sample}_1_clean.fastq.gz \
    -O fastp_reads/${sample}_2_clean.fastq.gz \
    --detect_adapter_for_pe \
    --thread 8 \
    --html fastp_reads/${sample}.html \
    --json fastp_reads/${sample}.json

done

#Unicycler assembly

cd fastp_reads

for sample in SRR11179862 SRR11179861 SRR11179850 SRR11179848 SRR11179842 SRR11179859 SRR11179858 SRR11179857 SRR11179856 SRR11179855
```

```
do
unicycler \
-1 ${sample}_1_paired.fastq \
-2 ${sample}_2_paired.fastq \
-o assembly_${sample} \
-t 4
done

#Assembly quality check with QUAST
quast assembly_*/assembly.fasta
-o quast_report
xdg-open quast_SAMPLE/report.html
#Download E. coli reference genomes
ncbi-genome-download bacteria \
--genera "Escherichia coli" \
--assembly-level complete \
--formats fasta \
--output-folder ecoli_refseq
find /home/fagm/ecoli_refseq -name "*.fna.gz" | wc -l
find /home/fagm/ecoli_refseq -name "*.gz" -exec gunzip {} \;
find /home/fagm/ecoli_refseq -name "*.fna" | wc -l
#FastANI genome comparison
for i in assembly_*/assembly*.fasta
do
 name=$(basename $i .fasta)
 fastANI \
 -q $i \
 --rl ref_list.txt \
 -o ${name}_fastani.txt \
 -t 4
Done
```

```
#MLST categorisation
mlst *.fasta > mlst_results.txt
cat mlst_results.txt
#ABRicate
#!/bin/bash
cd ~/refseq_ecoli/assemblies_for_abricate
# Create folder for results
mkdir -p abricate_results
db=resfinder   # change to another DB if you want
for f in *.fasta; do
   sample=$(basename "$f" .fasta)
   abricate --db $db "$f" > abricate_results/${sample}_${db}.tab
done
# Summarise all results
abricate --summary abricate_results/*.tab > abricate_results/summary_${db}.txt
cd ~/refseq_ecoli/assemblies_for_abricate/abricate_results
less summary_resfinder.txt
#AMRfinderplus
for f in *_amrfinder.txt; do
   sample=$(basename "$f" _amrfinder.txt)
   awk -v s="$sample" 'BEGIN{OFS="\t"} NR>1 {print s,$0}' "$f"
done > amrfinder_table.txt
```

Appendix III: Code for pipeline$_{Hybrid}$

```
#Illumina reads download, quality check and trimming equal to Pipeline_Improved
#Combine long reads from all the barcodes produced by MiniON
for i in barcode*; do
  cat $i/*.fastq.gz > ${i}.fastq.gz
done

#Process of long reads
for bc in barcode*.fastq.gz; do
  sample=$(basename "$bc" .fastq.gz)
  fastp \
   -i "$bc" \
   -o "${sample}_filtered.fastq.gz" \
   -h "${sample}_fastp.html" \
   -j "${sample}_fastp.json"
done

#Unicycler Hybrid Assembly
#!/bin/bash
declare -A samples
samples[39]=SRR11179862
samples[40]=SRR11179861
samples[41]=SRR11179850
samples[42]=SRR11179848
samples[43]=SRR11179842
samples[44]=SRR11179859
samples[45]=SRR11179856
for bc in 39 40 41 42 43 44 45
do
  srr=${samples[$bc]}
  unicycler \
```

```
    -1 /home/fagm/${srr}/${srr}_1_trimmed.fastq.gz \
    -2 /home/fagm/${srr}/${srr}_2_trimmed.fastq.gz \
        -l /home/fagm/Downloads/Madalena_2026/17032026M/20260317_1720_MD-100502_FBF66481_caac88c1/fastq_pass/barcode${bc}_filtered.fastq.gz \
    -o ${srr}_unicycler \
    -t 16 \
    --mode bold
done
``
#QUAST hybrid assembly quality check
quast \
    ${srr}_unicycler/assembly.fasta \
    -o ${srr}_quast \
    -t 16 \
    --min-contig 500
Done
#FastANI, MLST, ABRicate and AMRfinderplus code equal to Pipeline_Improved
```

Appendix IV: DNA extraction and Gel electrophoresis

Table A1- The DNA Concentration and the purity values of the DNA extracted from the isolates were measured using the fluorometer Qubit ™ and the DeNovix DS11 FX Spectrophotometer/ Fluorometer, respectively.

Isolate Number	Concentration (ng/µl)	A260/230 ratio	A260/280 ratio
1	32.1	2.18	1.92
2	28.4	2.00	2.02
3	19.4	1.74	1.85
4	NA*	NA*	NA*
10	21.4	1.74	1.78
12	38.6	1.74	1.87
13	16.7	2.14	1.94
14	20.5	2.02	2.04
15	29.4	1.80	1.88
16	28	1.71	1.87

*Not Applicable

Table A2- The DNA Concentration and the purity values of the DNA extracted by another colleague, measured using the fluorometer Qubit ™ and the DeNovix DS11 FX Spectophotometer/ Fluorometer, respectively.

Isolate Number	**Concentration (ng/µl)**	**A260/230 ratio**	**A260/280 ratio**
4	341	2.0	2.14
17	75	1.97	2.07
18	111	2.05	2.03
21	108	2.13	2.04

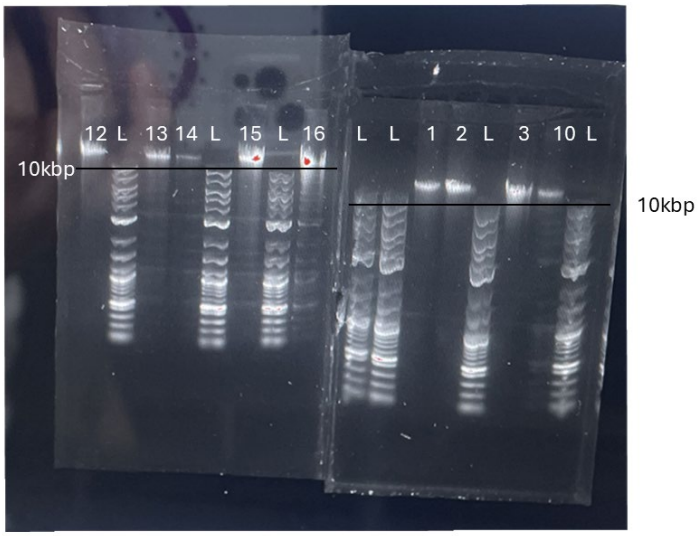

Figure A1- Results from the Standard agarose Gel electrophoresis, where all the isolates ran on a 0.6% agaore gel with the Quick-Load purple 1kb plus DNA ladder from New England Labs. The highest band from the molecular weight is about 10kbp and is marked with a black line. The DNA extracted from each isolate in each well can be identified with the isolate number on top, and the ones where the ladder was loaded were identified with the letter L on top of the well.

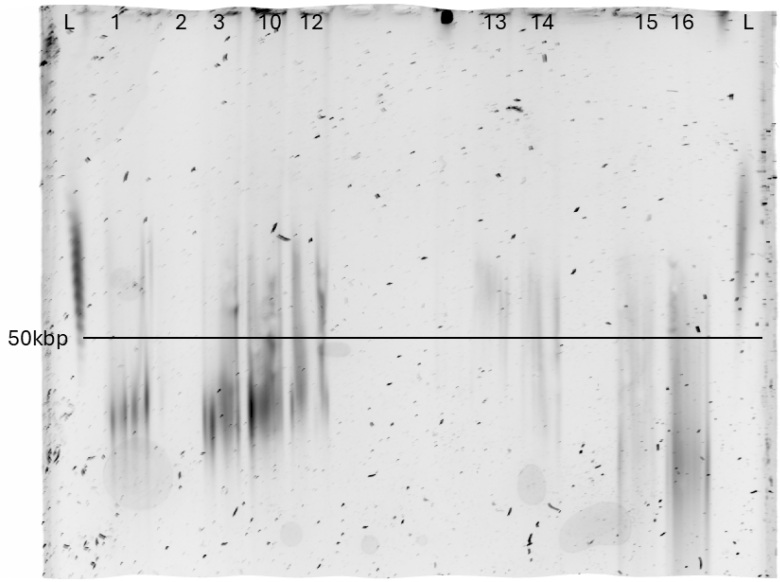

Figure A2- Results from the Pulse Field Gel electrophoresis, where all the isolates ran on a 0.6% agarose gel with the Lambda Ladder #1703635 from Bio-Rad. The lowest band from the molecular weight is about 50kbp and is marked with a black line. The DNA extracted from each isolate in each well can be identified with the isolate number on top, and the wells where the ladder was loaded were identified with the letter L on top of the well.

Appendix V: Quality of Third-Generation Sequencing

Table A1- Run performance data taken from the MinION sequencing final run report

Parameter	Value
Number of raw reads	1.61M[1]
Passed reads	1.38M[1]
Total Bases	3.19Gb[2]
Passed Bases	2.83Gb[2]

[1] Number of reads expressed in Millions (M)

[2] Number of bases in Gigabases (Gb)

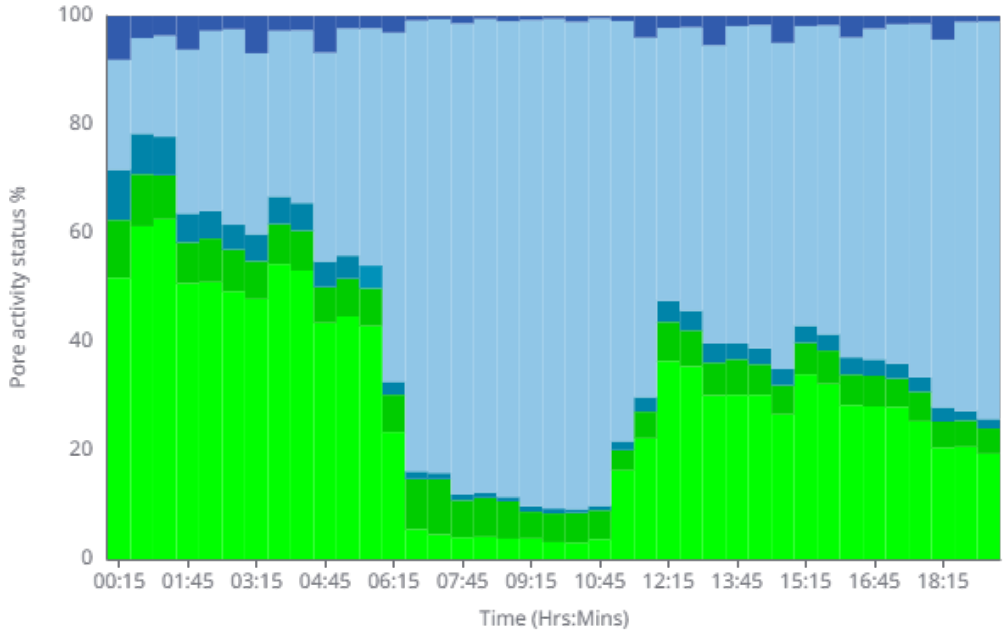

Figure A1- Number of active pores through the 20h sequencing run. The graph shows the percentage of pore activity while the sequencing was being performed. In each timetable is possible to see the percentage of pores that were sequencing (Light green), the percentage of available pores (dark green), the percentage of unavailable pores (mid-tone blue), the percentage of inactive pores (light blue) and the percentage of pores unclassified (dark blue). Source: MinION Sequencing Final Run Report.

Table A2- Summary of the read quality measurements from nanopore sequencing taken from the MinION Sequencing final run report

Parameter	Value
Mean Qscore	16
Median Qscore	17
Mean Read length	1953bp[1]
Median Read length	722bp[1]
N50 read length	5.23kb[2]

1 Read length in base-pairs (bp)

2 read length in kilobases (kb)

Table A3- Information about the sequencing process of each isolate. For each isolate, it is possible to see the total number of bases sequenced and their pass percentage, the total number of reads sequenced, and the percentage of passed reads. Reads and Bases were considered of good quality and passed if the Q score was greater than 9.

Isolate	**Total bases (Mb[1])**	**Passed bases (%)**	**Total reads (k[2])**	**Passed reads (%)**
1	319.77	95.30	104.23	94.10
2	194.68	94.90	62.52	93.30
3	333.67	96.10	97.50	95.20
4	3.03	93.40	1.10	93.50
10	74.53	96.00	24.99	94.70
12	198.95	95.20	73.27	93.80
15	323.73	95.70	102.46	94.50

1 Total number of bases in Megabases (Mb)

2 Total number of reads in thousands (k)

Table A4- Percentage and precise amount of reads and bases that were not classified during sequence.

Unclassified data	**Percentage**	**Precise value**
Reads	32.3%	444.94k[1]
Bases	25.%	727.62Mb[2]

1 Number of unclassified reads in thousands (k)

2 Number of unclassified bases in Megabases (Mb)

Appendix VI: Trimming

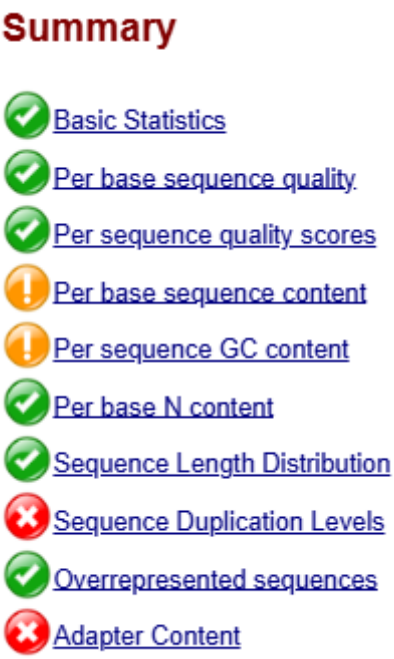

Figure A1- Summary of the quality of the raw reads from isolate 1 in all the parameters checked using FastQC before trimming, following Pipeline$_{\text{In-house}}$. The colour green represents a parameter that is considered to have a good quality, an orange parameter means a moderate quality and a parameter assigned with the colour red is classified as a bad quality.

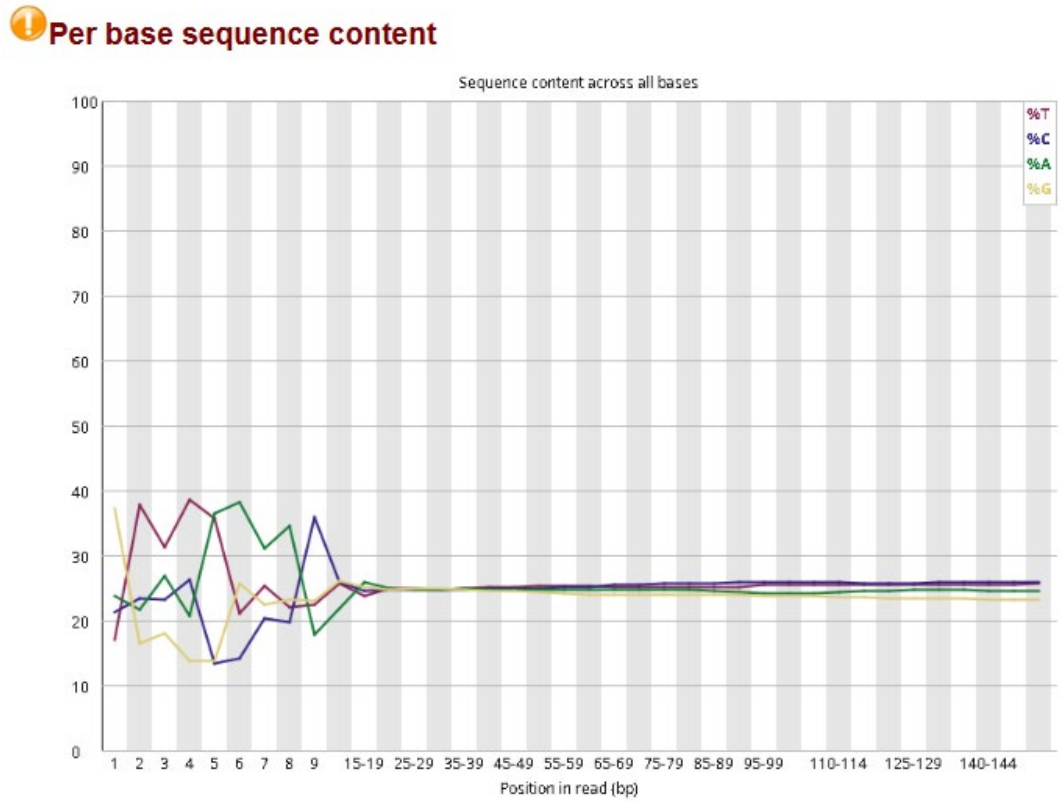

Figure A2- Per-base sequence content Quality from isolate 1 before trimming raw reads, following Pipeline$_{\text{In-house}}$. This parameter is assigned a moderate quality. At the beginning of the read, the different bases have different percentage content, but at the end of the read, the different bases appear with the same frequency

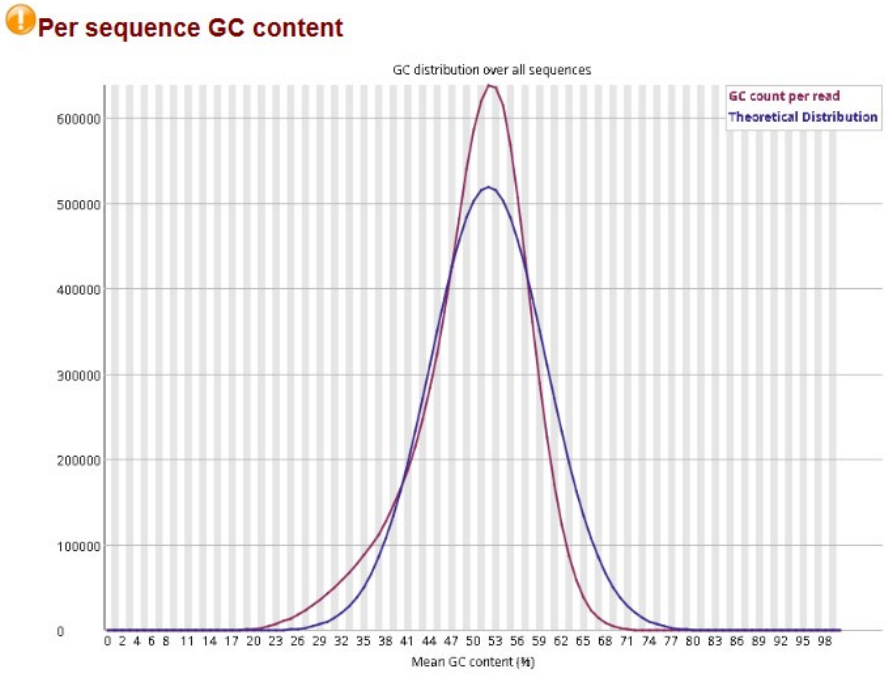

Figure A3- Per-sequence GC content before trimming raw reads from Isolate 1, following Pipeline$_{\text{In-house}}$, with moderate quality. The percentage of GC content is shown to be higher than the theoretical distribution.

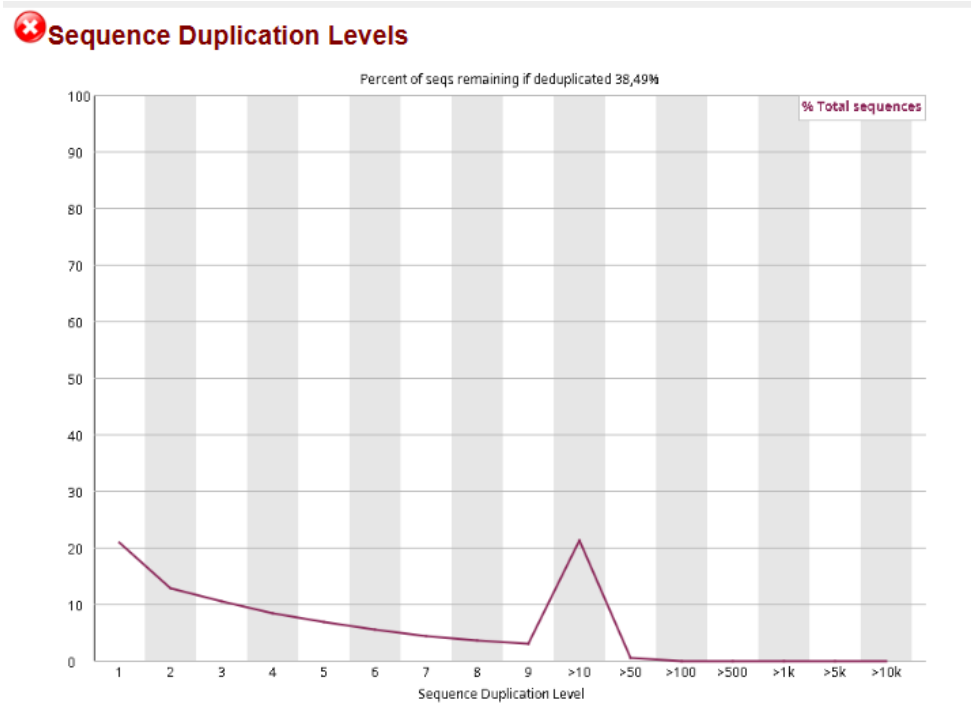

Figure A4- Sequence duplication Levels before trimming raw reads from Isolate 1, following Pipeline$_{\text{In-house}}$, with low quality. Around 20% of the sequences have more than 10 and fewer than 50 duplicates.

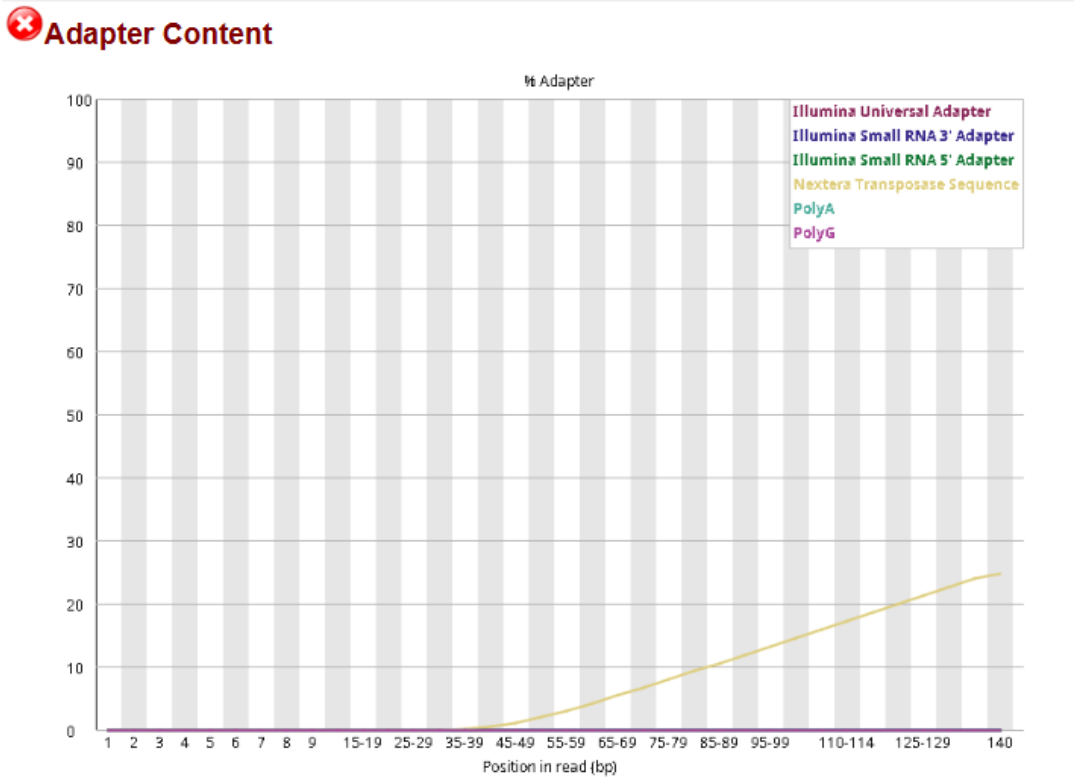

Figure A5- Adapter content before trimming raw reads from Isolate 1, following Pipeline$_{\text{In-house}}$, with low quality. The Nextera Transposase Sequence adapter content increases with the read extent. No other adapter content is found.

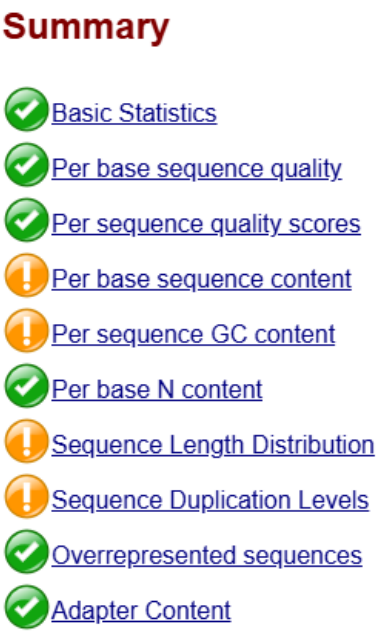

Figure A6- Summary of the quality of the reads from isolate 1 using FastQC after trimming raw reads, following Pipeline$_{\text{In-house}}$. The colour green represents a parameter that is considered to have good quality; an orange parameter means moderate quality, and a parameter assigned the colour red is classified as bad quality.

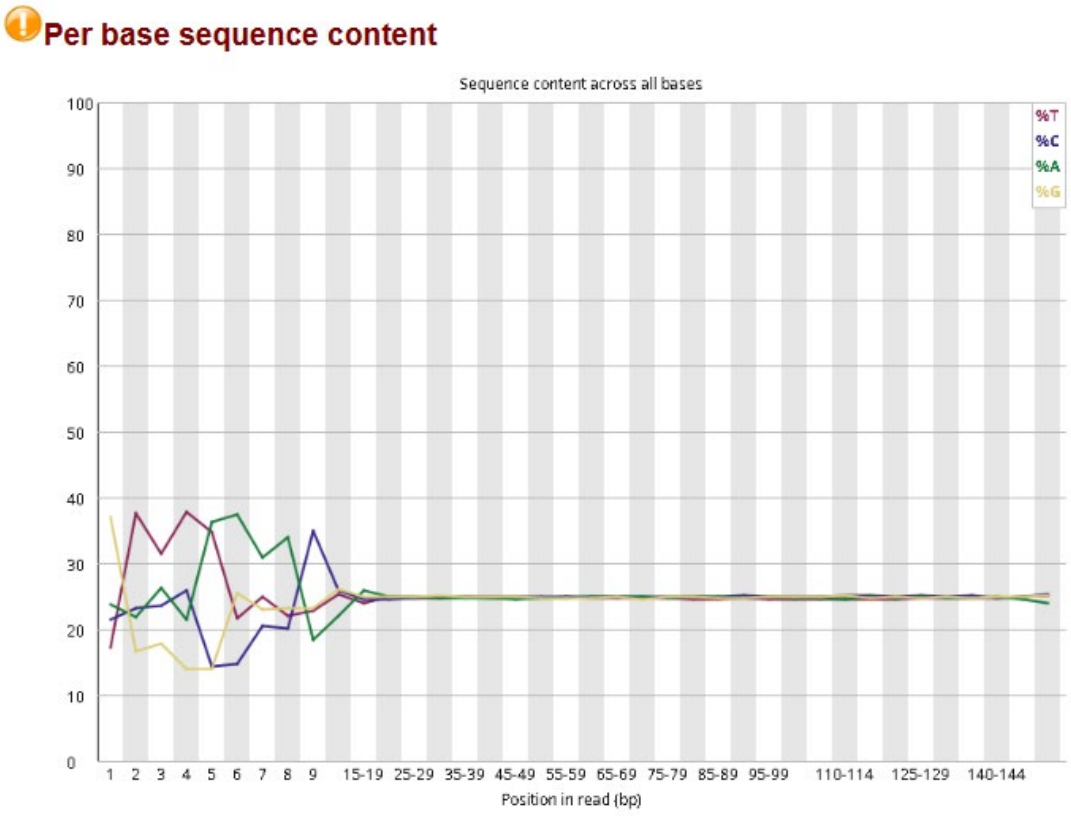

Figure A7- Per-base sequence content Quality from isolate 1 after trimming raw reads, following Pipeline$_{\text{In-house}}$. This parameter is assigned a moderate quality. At the beginning of the read, the different bases have different percentage content, but at the end of the read, the different bases appear with the same frequency

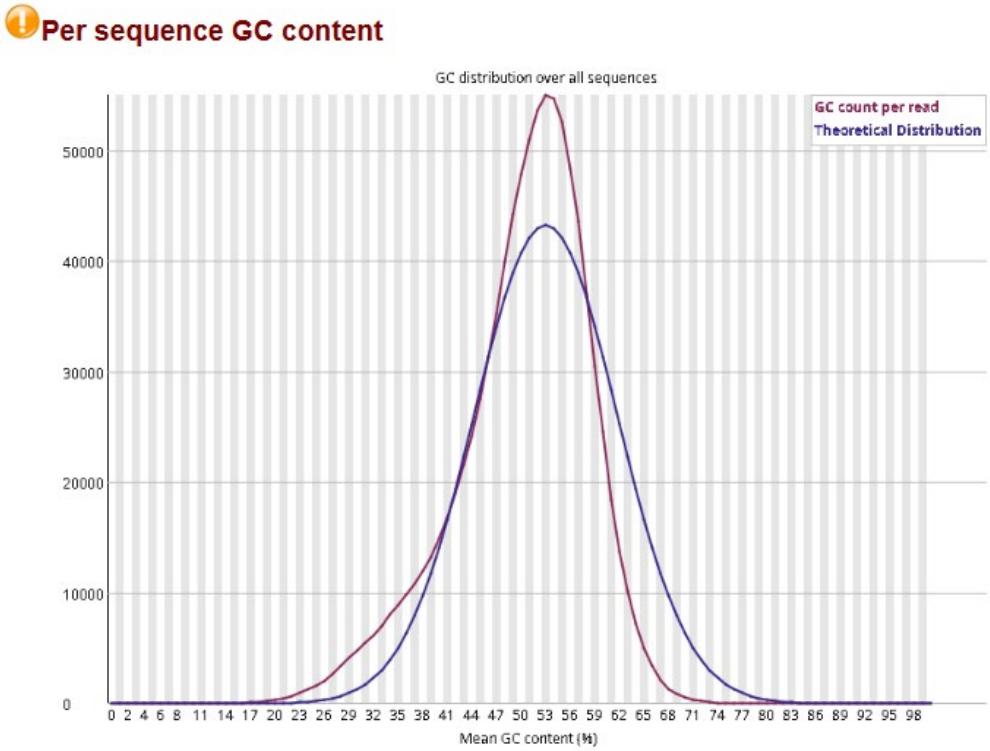

Figure A8- Per-sequence GC content after trimming raw reads, following Pipeline$_{\text{In-house}}$, from Isolate 1 with moderate quality. The percentage of GC content is shown to be higher than the theoretical distribution.

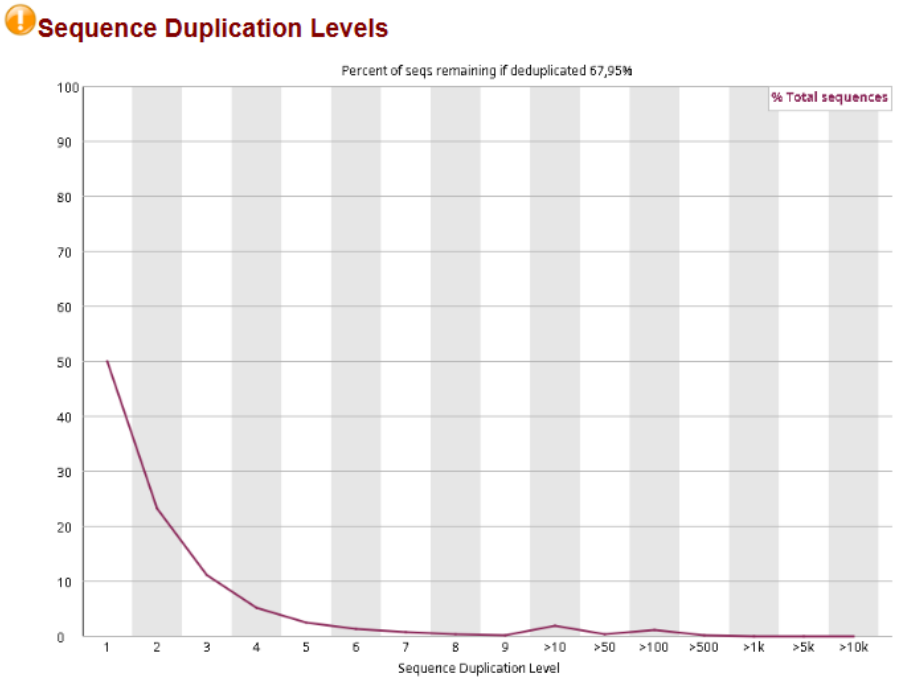

Figure A9- Sequence duplication Levels after trimming raw reads from Isolate 1, following Pipeline$_{\text{In-house}}$, with low quality. Around 5% of the sequences have more than 10and fewer than 50 duplicates.

Table A1- Trimming report for the Illumina reads after trimming using the Fastp tool according to the $Pipeline_{Improved}$.

Isolate number	Mean length before trimming (bp)	Mean length after trimming (bp)	Passed reads*	Low Quality reads*
1	151	135	20.216540M	375.064000k
2	151	141	3.170356M	262.542000k
3	151	99	1.313180M	21.478000k
4	151	142	3.576750k	289.2640000k
10	151	142	9.067418M	410.278000k
12	151	142	4.815846M	340.256000k
13	151	141	2.655552M	233.892000k
14	151	141	3.124386M	286.690000k
15	151	145	4.435636M	301.754000k
16	151	144	3.911786M	323.058000k

*The value can be seen in Millions (M) or thousands (k)

Table A2- Trimming report for the Third-Generation Sequencing reads after trimming using the Fastp tool, according to the $Pipeline_{Hybrid}$

Isolate number	Mean length before trimming	Mean length after trimming	Passed reads*	Low Quality reads*
1	3107bp	4261bp	67.374k	421
2	3165bp	4732bp	35.515k	322
3	3452bp	4648bp	65.624k	389
4	2760bp	4091bp	642k	7
10	3022bp	4564bp	14.639k	138
12	2758bp	4695bp	36.981k	458
15	3201bp	4247bp	69.348k	403

*The value can be seen in Millions (M) or thousands (k)

Appendix VII: Antibiotic Resistance

Table A1- Jaccard similarity calculated between the different pipelines to analyse whether the genes identified are identical. This metric varies between 0 and 1, where 0 means that none of the identified ARGs is the same and 1 means that the same ARGs are identified

Pipelines	**Pipeline$_{\text{In-house}}$**	**Pipeline$_{\text{Improved}}$**	**Pipeline$_{\text{Hybrid}}$**
Pipeline$_{\text{In-house}}$		1.00	0.63
Pipeline$_{\text{Improved}}$	1.00		0.63
Pipeline$_{\text{Hybrid}}$	0.63	0.63	

Table A2- Genes identified that resist the action of the different antibiotic subclasses from the NCBI-amrfinderplus database

Antibiotic subclass	**Genes**
β-lactams	BlaCTX-M-15 (ESBL), BlaCTX-M-1, BlaTEM-1B, BlaDHA-1, BlaCTX-M-55 (ESBL), BlaCTX-M-14 (ESBL)
Aminoglycosides	aadA5, aadA2_1, aph(3'')-Ib, aph(6)-Id, aac(3)-lId
Fluoroquinolones	QnrB4
Macrolides	Mph(A), mdf(A)
Tetracyclines	Tet(B)
Sulfonamides	Sul1, Sul2
Trimethoprim	drfA17
Chloramphenicol	catA1
Lincosamides	lnu(F)
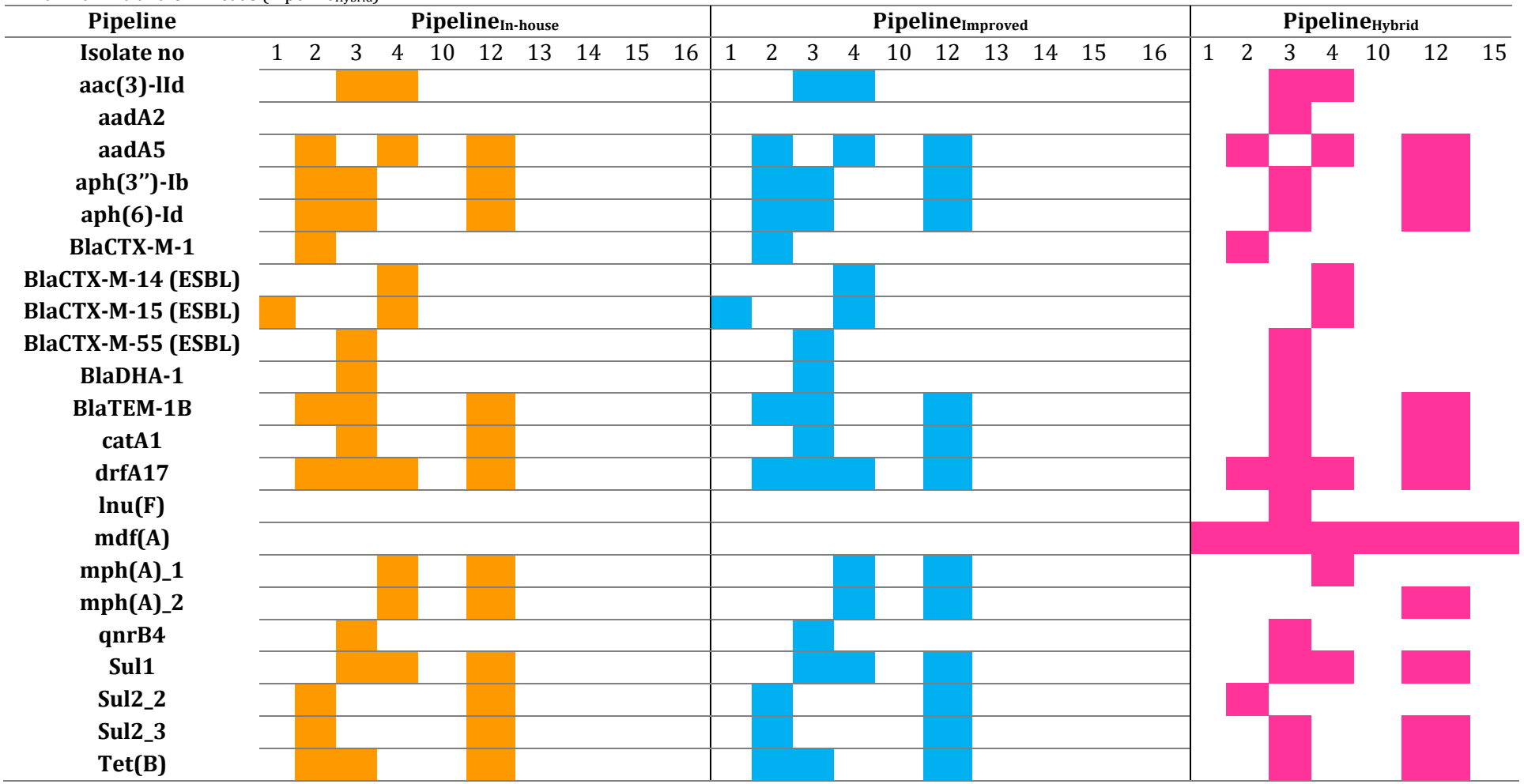
Table A3-Genes detected by VirulenceFinder ($Pipeline_{In-house}$) and by ABRicate and NCBI-amrfinderplus ($Pipeline_{Improved}$ and $Pipeline_{Hybrid}$) in the genome assembled with the DNA extracted from the different isolates. The colour of the pipeline used shows gene detection. A white space means that the gene was not detected. An orange space means the detection of the gene using VirulenceFinder from the genome assembled with Illumina reads ($Pipeline_{in-house}$). A blue space shows the detection of the gene using ABRicate and NCBI-amrfinderplus from the genome assembled with Illumina reads ($Pipeline_{Improved}$). A pink space indicates detection of the gene using ABRicate and NCBI-amrfinderplus from the genome assembled with Illumina and ONT reads ($Pipeline_{Hybrid}$).