Degree project

UNIVERSITY OF SKÖVDE

COMPARATIVE ANALYSIS OF ILLUMINA AND HYBRID WHOLE-GENOME SEQUENCING FOR GENOMIC CHARACTERIZATION OF CLINICAL *ESCHERICHIA COLI* ISOLATES ASSOCIATED WITH SEPSIS

Master Degree Project (120 credits) in Systems Biology

Second Cycle 45 credits

Spring term (2026)

Student: Atif Waheed

Supervisor: Sanja Jurcevic, University of Skövde

Examiner: Matthew Herring, University of Skövde

Abstract

Sepsis is a life-threatening condition caused by a dysregulated host response to infection and remains a major global health concern. *Escherichia coli (E. coli)* is a common cause of community-onset sepsis and is increasingly associated with antimicrobial resistance. Conventional diagnostics, such as blood culture, are limited by long turnaround times and low sensitivity. Whole-genome sequencing (WGS) enables comprehensive characterization of clinical isolates, including virulence factors, antimicrobial resistance genes, and strain lineages. This study aimed to update and evaluate an existing Illumina short-read bioinformatics pipeline for clinical *E. coli* isolates and to assess hybrid genome assembly combining Illumina short reads with Oxford Nanopore long reads.

Nine isolates were analyzed using both the in-house and updated Illumina pipelines. The updated pipeline produced significantly higher N50 values, with a mean increase of 43,990 bp and $p = 0.0037$, while species identification and most downstream genomic results remained consistent between the workflows. Multilocus sequence typing (MLST) assigned sequence types to eight of nine isolates in the updated pipeline. The updated workflow also eliminated at least 118 counted manual operations and enabled batch processing of all nine isolates. Hybrid assembly increased N50 for eight of the nine isolates, with a statistically significant overall difference compared with Illumina-only assembly at $p = 0.014$. MLST assignments remained unchanged between the updated Illumina-only and hybrid assemblies.

Overall, the updated Illumina pipeline improved assembly contiguity and workflow automation while preserving species identification and shared MLST assignments. Hybrid assembly further improved contiguity for eight of the nine isolates without changing MLST assignments. Its effects on downstream gene detection require further evaluation. These results support the use of updated, locally executable workflows and hybrid sequencing for genomic characterization of clinical *E. coli* isolates.

Popular scientific summary

Sepsis is a life-threatening condition that occurs when the body's response to an infection becomes uncontrolled and starts to damage its own organs. Instead of only fighting the infection, the immune system triggers widespread inflammation, which can lead to organ failure and, in severe cases, death. Despite modern medicine, sepsis remains one of the leading causes of death worldwide. One of the main challenges is that it develops quickly and is difficult to diagnose.

E. coli is one of the major causes of sepsis. This bacterium normally lives harmlessly in the human gut, but some strains can cause serious infections if they enter the bloodstream or other parts of the body. The harmful strains can also carry genes that make them resistant to antibiotics, meaning that standard treatment may not always work. Traditionally, doctors diagnose infections by growing bacteria from blood or urine. This method is slow and not very sensitive. In a disease like sepsis, every hour is critical, and delays in treatment can significantly increase the risk of death. In addition, some bacteria may not grow well in the lab, which can make diagnosis and treatment even more difficult.

To address these problems, scientists are developing genome sequencing methods that read all of a bacterium's genetic material. This can quickly reveal what type of bacteria is present, whether it carries antibiotic-resistance genes, and how closely related it is to other known strains.

There are two main ways to sequence DNA. One reads DNA in short pieces very accurately but struggles to put the full picture back together. The other reads much longer pieces of DNA, which makes it easier to reconstruct the whole genome, but with more errors along the way. In this project, both methods were combined to take advantage of their complementary strengths: accurate short reads and long reads that can improve the contiguity of genome assemblies. This may make it easier to identify genes associated with antibiotic resistance and virulence.

The wider social benefit of this work lies in improving how we detect and understand infectious diseases. In the future, improved genomic characterization could contribute to faster identification of bacterial pathogens and antimicrobial resistance determinants, potentially supporting more targeted antibiotic treatment. It can also help limit the unnecessary use of broad-spectrum antibiotics, which is one of the main drivers of antibiotic resistance. Antibiotic resistance is a growing global problem. When bacteria evolve to survive antibiotic treatment, infections become harder and sometimes impossible to treat. By improving genomic tools and analysis pipelines, research like this contributes to better surveillance of resistant bacteria and helps public health systems track and control outbreaks more effectively.

In the long term, advances in genome sequencing and bioinformatics may lead to faster diagnostic tools that can be used directly in hospitals. Overall, this work demonstrates how modern sequencing technologies and computational analysis can be combined to improve our understanding of dangerous bacterial infections and support better healthcare decisions in the future.

Table of Contents

Abstract 2

Popular scientific summary 3

Abbreviations 1

Introduction 2

- Sepsis 2
- Sepsis and Bacteria (*E. coli*) 2
- Diagnosis of Sepsis 3
- Next-generation sequencing 3
- Illumina short-read sequencing 4
- Oxford Nanopore long-read sequencing 4
- Hybrid genome assembly 5
- Bioinformatics Pipelines and Genome Assembly 5

Aim 6

Materials and Methods 7

- Bacterial Isolates 7
- DNA extraction and Quality control 7
- Fragment analysis 7
- Library preparation and Nanopore sequencing 8
- Bioinformatic analysis 8
  - In-house Illumina pipeline 8
  - Updated Illumina pipeline 9
  - Comparison of the in-house and updated pipeline 9
  - Pipeline automation assessment 9
  - Nanopore assembly 9
  - Hybrid assembly 9
  - Statistical analysis 10

Results 10

- DNA extraction and quality control 10
- Fragment analysis 10
- Library preparation and Nanopore sequencing 10
- Bioinformatic pipeline 10
  - In-house Illumina pipeline 10

Updated Illumina pipeline 11
Comparison of the in-house and updated Illumina pipelines 11
Workflow automation assessment 13
Nanopore assembly 13
Hybrid assembly 14
Comparison of updated Illumina-only and hybrid assemblies 14
Statistical analysis 14
Discussion 15
DNA extraction and quality control 15
Fragment analysis 15
Library preparation and Nanopore sequencing 15
In-house pipeline 16
Updated Illumina pipeline and comparison with the in-house pipeline 16
Nanopore assembly 18
Hybrid assembly and comparison with Illumina-only assemblies 18
Limitations 19
Conclusion 19
Ethical aspects, gender perspectives, and impact on society 19
Future perspectives 20
Acknowledgments 21
References 21
Appendix A – Sample identification 29
Appendix B – Quality control for DNA extraction and fragment analysis 30
Appendix C – In-house pipeline: commands and parameters 31
Download raw reads from SRA 31
FastQC before Trim 31
Trimmomatic trimming 31
SPAdes assembly 32
QUAST reference-free assembly QC 32
CGE ResFinder 33
CGE VirulenceFinder 33
CGE PlasmidFinder 33
CGE MLST 33

Appendix D – Updated pipeline: commands and parameters 34
Recorded versions: 34
Full updated pipeline script 34
Appendix E– In-house Illumina pipeline results 38
Appendix F– Updated pipeline 41
Appendix G– Comparison of the in-house and updated Illumina pipelines 44
Appendix H – Nanopore-only assembly. 45
Appendix I – Comparison of updated Illumina-only and hybrid assemblies 46
Appendix J- Hybrid assembly results 47
Appendix K– Statistical analysis 49

Abbreviations

AMR	Antimicrobial resistance
ANI	Average nucleotide identity
ANIb	Average nucleotide identity based on BLAST
CGE	Center for Genomic Epidemiology
E. coli	*Escherichia coli*
ExPEC	Extraintestinal pathogenic *E. coli*
GC	Guanine-cytosine
HMW	High-molecular-weight
LB	Lysogeny broth
MLST	Multilocus sequence typing
NCBI	National Center for Biotechnology Information
NGS	Next-generation sequencing
OD	Optical density
OD_{600}	Optical density at 600 nm
ONT	Oxford Nanopore Technologies
PFGE	Pulsed-field gel electrophoresis
SIRS	Systemic inflammatory response syndrome
SOFA	Sequential Organ Failure Assessment
SRA	Sequence Read Archive
ST	Sequence type
TBE	Tris-borate-EDTA
VFDB	Virulence Factor Database
WGS	Whole-genome sequencing

Introduction

Sepsis

The understanding of sepsis has evolved substantially over time. The term originates from the ancient Greek word sepsis, meaning "putrefaction," reflecting early beliefs that the condition resulted from decaying tissue (Baron et al., 2006). The discovery of the association between infectious diseases and bacteria through the work of Louis Pasteur contributed to the development of modern understanding of infection (Jambholkar et al., 2024), while antibiotics later enabled effective treatment of many bacterial infections (Hutchings et al., 2019).

The definition of sepsis has been revised to reflect its complex pathophysiology. Earlier definitions were based on systemic inflammatory response syndrome (SIRS), whereas the 2016 definition emphasized organ dysfunction resulting from a dysregulated host response to infection (Cecconi et al., 2018). Sepsis is therefore a life-threatening condition in which the host response to infection causes tissue and organ damage (Rello et al., 2017; Singer et al., 2016). The most severe form, septic shock, involves profound circulatory and metabolic abnormalities and is associated with a mortality risk exceeding 40% (Dellinger et al., 2013; Liu et al., 2022).

Sepsis remains challenging to diagnose because its clinical symptoms, including fever or hypothermia, tachycardia, and rapid breathing, are often nonspecific (Vincent, 2016). Diagnosis typically combines laboratory findings with clinical scoring systems such as the Sequential Organ Failure Assessment (SOFA) (Chun et al., 2015; Qiu et al., 2023). The SOFA score evaluates respiratory, coagulation, liver, cardiovascular, central nervous, and renal function, with higher scores indicating greater organ dysfunction and mortality risk (Lambden et al., 2019; Kashyap et al., 2021; De Grooth et al., 2017). Although microbiological culture remains the gold standard for pathogen identification, results may require 24–72 hours, delaying treatment decisions, particularly in the presence of antimicrobial resistance (AMR) (Tabak et al., 2018).

Sepsis remains a major global health burden. In 2017, approximately 11 million deaths were attributed to sepsis, representing nearly 20% of global deaths (World Health Organization, 2020). In the United States, reported cases increased from 1.8 million in 2016 to 2.5 million in 2021 (Owens et al., 2024), while in Sweden, more than 4% of hospitalizations involved sepsis and 20% of affected patients died within three months (Mellhammar et al., 2023). The increasing burden has been associated with factors including COVID-19, population aging, immunosuppressive treatments, organ transplantation, invasive procedures, and multidrug-resistant microorganisms (Rhee & Klompas, 2020; Liang, 2016).

Sepsis and Bacteria (*E. coli*)

Sepsis can be caused by bacteria, viruses, fungi, or parasites, with bacteria being the most common. *E. coli* is consistently among the leading bacterial causes of community-onset sepsis. It is a Gram-negative, rod-shaped bacterium that normally inhabits the gastrointestinal tract of humans and animals (Bonten et al., 2021; Vila et al., 2016). While most *E. coli* strains are harmless commensals, extraintestinal pathogenic *E. coli* (ExPEC) possess virulence factors that enable infection beyond the gut, causing urinary tract infections, bacteremia, and severe systemic disease. Increasing antimicrobial exposure has contributed to the emergence of resistance in these pathogenic lineages (Dale & Woodford, 2015).

E. coli is also a major cause of infection worldwide and can readily acquire antimicrobial resistance genes. These genes are frequently carried on plasmids, which facilitate their transfer between bacteria and rapid dissemination (Pitout, 2012). Improved genomic characterization, including more reliable resolution of plasmid-associated regions, is therefore important for understanding the epidemiology and clinical risks associated with *E. coli* sepsis.

Gene acquisition in *E. coli* occurs through horizontal gene transfer involving mobile genetic elements such as plasmids, bacteriophages, insertion sequences, and transposons (Javadi et al., 2017). Pathogenic strains may carry virulence gene clusters known as pathogenicity islands, located on plasmids or chromosomes and generally absent from commensal strains (Dobrindt et al., 2004). A comparison of 17 commensal and pathogenic *E. coli* genomes identified approximately 2,200 shared genes and a collective pangenome containing more than 13,000 genes, demonstrating substantial genetic diversity within the species (Rasko et al., 2008). These virulence factors support immune evasion, tissue invasion, dissemination, and replication within the host, contributing to disease development (Kaper et al., 2004).

Diagnosis of Sepsis

Culture-based methods involve collecting samples from suspected infection sites, such as cerebrospinal fluid, urine, and respiratory secretions, and culturing them on specialized media (Motzkus & Luckmann, 2016). Microbial growth can then be identified and tested for antimicrobial susceptibility. However, cultures are time-consuming, and some organisms may not grow under standard conditions, limiting diagnostic speed and accuracy (Marcello et al., 2016).

Blood culture remains the gold standard for diagnosis of sepsis. Blood is collected before antimicrobial treatment and inoculated into aerobic and anaerobic culture bottles, as prior antibiotic exposure can reduce pathogen recovery (Rand et al., 2019). The bottles are monitored for microbial growth, after which positive samples are identified using methods such as Gram staining, agar culture, or molecular testing, followed by antimicrobial susceptibility testing (Chun et al., 2015).

Despite being the standard method, blood culture has several limitations. Only approximately 30–50% of patients with suspected sepsis have positive blood cultures, and results commonly require 24–72 hours (Hung et al., 2020). Contamination can also cause false-positive results, while fastidious bacteria, anaerobes, and fungi may grow slowly or fail to grow under standard conditions. Detection is further influenced by the volume of blood collected, particularly in infections with low pathogen concentrations (Marcello et al., 2016; Dargère et al., 2018).

These limitations have increased interest in next-generation sequencing (NGS) as a faster and more comprehensive approach to pathogen characterization (Mancini et al., 2010). NGS can provide genomic information on antimicrobial resistance genes, virulence factors, and strain-level characteristics, supporting targeted treatment and clinical decision-making (Agudelo-Pérez et al., 2023).

Next-generation sequencing

In the 1970s, Frederick Sanger developed a method for DNA sequencing, now known as Sanger sequencing. This first-generation technique was highly accurate but limited by the length of DNA fragments it could read, typically less than one thousand base pairs (Eren et al., 2022). Second-generation sequencing platforms such as Illumina enabled massively parallel sequencing of large

numbers of short DNA fragments. These technologies are collectively known as next-generation sequencing (NGS). This approach was also highly accurate and eliminated the need for gel electrophoresis or cloning, as required in Sanger sequencing. However, it had its own limitations: read lengths were very short (100–300 base pairs), homopolymer regions and repetitive sequences could cause errors, and assembling a complete genome remained challenging (Kumar et al., 2024).

Third-generation sequencing overcomes the limitations of short reads by producing long sequences that can span repetitive regions. Oxford Nanopore Technologies (ONT), for example, uses nanopore-based sensing to generate reads ranging from tens to hundreds of kilobases, depending on DNA quality. This allows researchers to assemble bacterial chromosomes and plasmids into single contigs, providing a complete view of genome structure. However, this technology has higher error rates compared to second-generation sequencing (Schadt et al., 2010).

WGS uses NGS technologies to read the entire genome of an organism. WGS captures all genetic material, including both coding and non-coding regions (Ng & Kirkness, 2010). This approach helps researchers study genetic variation, identify mutations, and obtain information such as virulence factors, antibiotic resistance genes, and serotypes (Gilchrist et al., 2015). The cost of sequencing a human genome has dropped rapidly in recent years, making the technology far more accessible (Marshall et al., 2017).

Illumina short-read sequencing

Illumina sequencing is widely used for bacterial WGS because it generates large numbers of highly accurate short reads. This accuracy makes it suitable for nucleotide-level analyses and reliable detection of genomic features. However, short reads may not span repetitive sequences or mobile genetic elements, resulting in fragmented assemblies with multiple contigs rather than complete chromosomes or plasmids (Chen et al., 2025).

For *E. coli*, this limitation can complicate the reconstruction of plasmids and other accessory genomic regions. Although Illumina sequencing provides highly accurate sequence data, short-read assemblies alone may therefore be insufficient to resolve the complete structure of complex bacterial genomes.

Oxford Nanopore long-read sequencing

ONT sequencing provides longer reads that can span repetitive regions and improve the reconstruction of bacterial chromosomes and plasmids. Its main advantage is therefore improved assembly contiguity rather than simply increased sequence output. However, long-read sequencing can have lower base-level accuracy than Illumina sequencing and generally requires high-molecular-weight DNA of sufficient quality. Sequencing yield may also vary between samples, particularly during multiplexed sequencing.

The complementary characteristics of short- and long-read sequencing provide a rationale for hybrid genome assembly. Long reads provide structural continuity, while accurate Illumina reads can improve base-level accuracy (Mink et al., 2025).

Hybrid genome assembly

Hybrid genome assembly combines short- and long-read data to exploit their complementary strengths. Long reads can span repetitive sequences and resolve complex genomic structure, while accurate Illumina reads can support error correction and improve base-level accuracy. Consequently, hybrid assembly can produce more contiguous bacterial genomes than short-read data alone (Wick et al., 2017). Improved assembly contiguity is particularly relevant for *E. coli*, where fragmented assemblies may leave plasmids unresolved or separate genes from their genomic context. More contiguous assemblies therefore may facilitate the characterization of antimicrobial resistance determinants, virulence-associated genes, plasmid replicons, and strain characteristics (Khezri et al., 2021).

Hybrid assembly performance depends on factors such as sequencing depth, read quality, DNA integrity, library preparation, and the relative contribution of short- and long-read data. Evaluating multiple clinical isolates is therefore important for determining whether hybrid assembly consistently improves genome assembly (De Maio et al., 2019).

Bioinformatics Pipelines and Genome Assembly

WGS generates large amounts of data that require structured bioinformatic processing. A bacterial WGS workflow commonly includes quality control, read filtering, genome assembly, assembly assessment, species identification, and downstream characterization of antimicrobial resistance, virulence, plasmids, and sequence types (Satam et al., 2023).

The University of Skövde previously developed an in-house Illumina short-read pipeline for bacterial genome analysis (Saxenborn et al., 2021). The workflow included quality control, trimming, genome assembly, species identification using average nucleotide identity (ANI), plasmid detection with PlasmidFinder, antimicrobial resistance analysis using ResFinder and the 1928 Diagnostics platform, and multilocus sequence typing (MLST). Together, these analyses enabled comprehensive characterization of bacterial genomes. An overview of the previously established in-house Illumina short-read pipeline developed at the University of Skövde is presented in Figure 1, as described by Saxenborn et al. (2021).

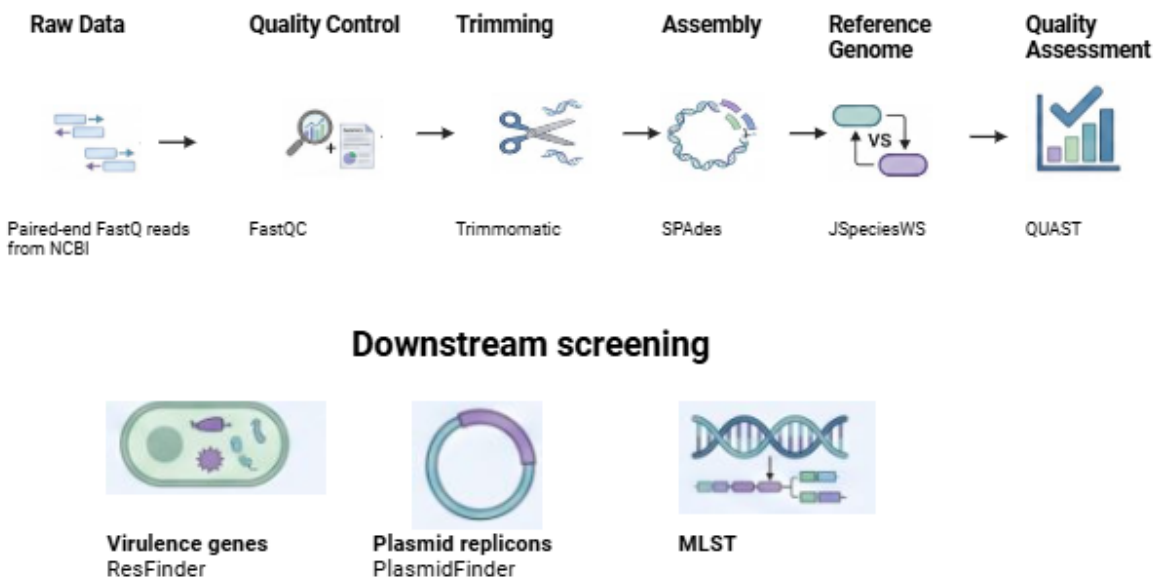

Figure 1. Overview of the previously established in-house Illumina short-read pipeline. Paired-end FASTQ reads retrieved from the National Center for Biotechnology Information (NCBI) undergo quality assessment using FastQC, followed by adapter and low-quality read trimming with Trimmomatic. The processed reads are assembled de novo using SPAdes. Species identification is performed using JSpeciesWS, followed by downstream analysis using ResFinder, PlasmidFinder, and MLST.

Since the development of this workflow, newer tools and software versions have become available. Updating the workflow may improve efficiency and reproducibility, but changes in analytical strategies can affect genomic results. Therefore, updated workflows should be evaluated against previous results to ensure consistency. For *E. coli* sepsis, a reproducible and updated WGS workflow could improve pathogen characterization and support more timely treatment decisions.

Aim

The aim of this thesis was to update and extend an existing in-house Illumina short-read bioinformatics pipeline, developed at the University of Skövde by Saxenborn et al. (2021), for the analysis of WGS data from clinical *E. coli* isolates. Selected manual and web-based steps were replaced with locally executable command-line tools to increase workflow automation. The original and updated workflows were applied to the same sequencing datasets and compared with respect to genome assembly quality, species identification, antimicrobial resistance, virulence genes, plasmid replicons, MLST results, and the level of manual intervention required during analysis.

A second aim was to generate hybrid genome assemblies by combining Illumina short-read and Oxford Nanopore long-read sequencing data from the same clinical *E. coli* isolates.

Finally, the study aimed to compare the contiguity of the hybrid assemblies with that of the updated Illumina-only assemblies, primarily using N50 and L50 values. Downstream genomic comparisons were included as exploratory analyses.

Materials and Methods

Throughout this report, each sample is referred to by its SRR accession number. However, isolate numbers are used in sections describing DNA extraction and fragment analysis. The corresponding original isolate number, strain number, and Nanopore barcode for each sample are provided in Appendix A, Table A1.

Bacterial Isolates

Clinical isolates of *E. coli* used in this thesis originated from a prospective study on community-onset severe sepsis and septic shock conducted at Skaraborg Hospital between 2011 and 2012. The hospital is a secondary care facility located in western Sweden. Patients older than 18 years of age who were admitted to the emergency department with suspected bacterial infection and initiated on intravenous antibiotic treatment within 48 hours were evaluated for severe sepsis or septic shock according to Swedish clinical criteria. Bacterial isolates recovered during routine clinical microbiological diagnostics were included in the study. Ethical approval was obtained from the Regional Ethical Review Board of Gothenburg (376-11) (Saxenborn et al., 2021).

Approximately 1,800 bacterial isolates from various clinical sample types were collected during the study period. The isolates were cryopreserved in Microbank vials and stored at −80 °C until further use. Species identification was performed at the clinical microbiology laboratory Unilabs using routine culture-based diagnostic methods followed by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry analysis (Ljungström et al., 2019).

Ten bacterial isolates were selected and cultured overnight in lysogeny broth (LB). The cultures were then diluted at a 1:50 ratio in fresh LB broth and incubated further until the optical density at 600 nm (OD_{600}) reached 0.9–1.0.

DNA extraction and Quality control

DNA extraction was performed using the Wizard® High-Molecular-Weight (HMW) DNA Extraction Kit (Promega). To increase the concentration of the extracted DNA, the starting culture volume was increased from 1 mL to 5 mL. To compensate for this increased starting material, the centrifugation time at each step was doubled.

The purity of the extracted DNA was assessed using a DS-11 Series spectrophotometer (DeNovix). DNA concentration was measured using the Qubit 4 Fluorometer (Thermo Fisher Scientific) with the Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific).

Fragment analysis

DNA fragment analysis was performed using the CHEF-DR III System (Bio-Rad) for pulsed-field gel electrophoresis (PFGE). A 1% agarose gel was prepared. The CHEF DNA Size Standard, 48.5–1,000 kb, Lambda Ladder (Bio-Rad) was loaded into the well, which was then sealed with molten agarose. After the gel solidified, it was placed into the PFGE system containing cold 0.5× Tris-Borate-EDTA (TBE) buffer.

All samples were prepared by mixing with 6× loading dye (New England Biolabs) and subsequently loaded into the wells. PFGE was performed using the CHEF-DR® III system (Bio-Rad) according to the manufacturer's instructions (CHEF-DR® III Pulsed Field Electrophoresis Systems Instruction Manual and Applications Guide, Bio-Rad Laboratories, 2007). The electrophoresis conditions were set to 6 V/cm with a 120° included angle, a switch time range of

50–90 seconds, and a total run time of 22 hours. After completion of the electrophoresis run, the gel was removed and placed on a rocking platform for staining in a solution containing 30 µL GelRed and 300 mL of 1× TBE buffer for 30 minutes. The gel was subsequently visualized and imaged using the GelDoc Go Imaging System (Bio-Rad).

Library preparation and Nanopore sequencing

Following fragment analysis, library preparation was performed on the genomic DNA to prepare it for nanopore sequencing. Steps including DNA repair and end preparation, native barcode ligation, adapter ligation, clean-up, and priming and loading of the MinION/GridION Flow Cell (DNA) were carried out according to the protocol of the Ligation Sequencing gDNA – Native Barcoding Kit 96 V14 (SQK-NBD114.96) from Oxford Nanopore Technologies. Sequencing was performed using the MinION sequencing platform (Oxford Nanopore Technologies). Prior to sequencing, the flow cell status and pore availability were assessed using the MinKNOW quality control software to ensure that the flow cell met the required quality standards for sequencing.

Bioinformatic analysis

All bioinformatics analyses were performed on a Linux-based system using tool-specific Conda environments. Unless otherwise stated, all software tools were run using their default parameters. Raw Illumina sequencing reads for nine isolates were obtained from the NCBI Sequence Read Archive (SRA). The SRA accession numbers for these samples are listed in Appendix A, Table A1. All datasets were downloaded in paired-end FASTQ format.

In-house Illumina pipeline

Raw sequencing read quality was assessed using FastQC v0.11.8, including per-base sequence quality, guanine–cytosine (GC) content, and adapter contamination (He et al., 2020). FastQC reports were summarized using MultiQC v1.33. Reads were processed using Trimmomatic v0.39 in paired-end mode, with only paired reads retained for assembly. De novo assembly was performed using SPAdes v3.11.1, and assembly quality was evaluated using QUAST v5.0.2 based on total assembly length, number of contigs, largest contig, GC content, N50, and L50.

Three E. coli reference genomes were considered: IAI39 (NC_011750.1), K-12 (CP009685.1), and UTI89 (NC_007946.1). UTI89 was selected following a preliminary QUAST comparison across five isolates because it produced fewer reported misassemblies and a greater mean largest alignment than the other references. ANI was calculated using JSpeciesWS with the BLAST-based average nucleotide identity (ANIb) method. UTI89 was also used for reference-based QUAST assessment and as the common reference for both pipelines.

Downstream analyses were performed using tools from the Center for Genomic Epidemiology (CGE). Acquired antimicrobial resistance genes were identified using ResFinder v4.7.2, and resistance-associated point mutations were identified using PointFinder. Virulence genes and plasmid replicons were identified using VirulenceFinder v3.2.1 and PlasmidFinder v3.0.3. MLST was performed using MLSTFinder v1.0.1 with the *E. coli* Achtman scheme. MLST was performed to determine the genetic lineage of each isolate by analysing the allelic profiles of conserved housekeeping genes and assigning sequence types (STs) based on established MLST schemes.

Detailed commands and parameters for the in-house pipeline are provided in Appendix C.

Updated Illumina pipeline

Raw paired-end reads were processed using fastp v1.3.7 for adapter removal and quality filtering. Filtered reads were assembled using Unicycler v0.5.1 in short-read mode. Assembly quality was evaluated using QUAST v5.0.2 using the same metrics applied to the in-house assemblies.

ANI was calculated using FastANI, with *E. coli* UTI89 (NC_007946.1) as the reference genome. FastANI calculates average nucleotide identity between genome assemblies and a specified reference genome (Amoutzias et al., 2022). The same reference genome was used for both pipelines to allow direct comparison of species-level ANI results.

Antimicrobial resistance genes, plasmid replicons, and virulence-associated genes were screened locally using ABRicate v1.0.1 with the ResFinder, PlasmidFinder, Virulence Factor Database (VFDB), and ecoli_vf databases, all updated on 21 November 2025. ABRicate screening used minimum thresholds of 80% nucleotide identity and 80% reference-sequence coverage. MLST was performed using mlst v2.23.0 with the ecoli_achtman_4 scheme.

Detailed commands, software versions, parameters, and database information for the updated pipeline are provided in Appendix D.

Comparison of the in-house and updated pipeline

The two pipelines were applied to the same nine raw Illumina datasets. Assembly metrics obtained with QUAST were compared between pipelines. ANI values obtained using JSpeciesWS ANIb and FastANI were recorded using UTI89 as the common reference genome. Downstream outputs were compared for antimicrobial-resistance genes, virulence genes, plasmid replicons, and MLST sequence types.

Pipeline automation assessment

Pipeline automation was assessed by recording the number of web-based analysis steps, individual sample submissions, manual result downloads, manual file-management operations, and script executions required to process the nine isolates. In-house pipeline time was measured using a digital stopwatch from the start of Trimmomatic processing to completion of all analyses. Updated pipeline time was measured using GNU time (/usr/bin/time) in verbose mode, covering the complete script from fastp preprocessing through MLST, with timing output saved to a log file.

Nanopore assembly

Nanopore reads were obtained from the fastq_pass directory following sequencing and demultiplexing. FASTQ files belonging to the same barcode were concatenated into a single FASTQ file for each sample. Genome assembly was performed using Flye v2.9, and the resulting assemblies were evaluated using QUAST v5.0.2.

Hybrid assembly

Hybrid assemblies were generated using Unicycler v0.5.1 by combining the filtered Illumina paired-end reads with the corresponding Nanopore long reads for each isolate. Assembly quality and contiguity were assessed using QUAST v5.0.2. The hybrid assemblies were subsequently analyzed for antimicrobial-resistance genes, virulence-associated genes, plasmid replicons, and MLST using the same downstream tools applied to the updated Illumina pipeline.

Statistical analysis

Statistical analyses were performed to compare N50, total assembly length, L50, and number of contigs between the in-house and updated pipelines using paired data from nine samples (n = 9). For each assembly metric, the normality of the paired differences was assessed using the Shapiro–Wilk test. When the differences did not significantly deviate from normality, a paired t-test was used; otherwise, the Wilcoxon signed-rank test was applied. Statistical significance was defined as $p < 0.05$.

Statistical analysis compared N50 values between Illumina-only and hybrid assemblies using a paired design (n = 9). The normality of the paired differences was assessed using the Shapiro-Wilk test. As the differences were not normally distributed, the Wilcoxon signed-rank test was used to compare the two assembly approaches. Statistical significance was set at $p < 0.05$.

Results

DNA extraction and quality control

DNA was successfully extracted from all ten bacterial isolates and evaluated for both concentration and purity prior to downstream applications. DNA concentrations ranged from 10 to 341 ng/µL, with A260/A230 ratios of 1.90–2.17 and A260/A280 ratios of 1.98–2.14. Appendix B, Table B1 presents the DNA concentration and purity measurements for each isolate. Isolate 19 had the lowest DNA concentration, at 10 ng/µL, and was excluded from subsequent sequencing and bioinformatics analyses. The remaining nine isolates proceeded to library preparation and sequencing.

Fragment analysis

PFGE analysis showed high-molecular-weight DNA across the bacterial isolates, although band intensity and definition varied between samples, as shown in Appendix B, Figure B1. Most samples had estimated DNA fragment sizes of 100–200 kb, whereas isolates 20 and 21 showed fragments of approximately 300 kb.

Library preparation and Nanopore sequencing

DNA concentration decreased during library preparation from 46.0 ng/µL after native barcode ligation to 28.2 ng/µL after adapter ligation and clean-up. The flow cell contained 1,538 active pores before sequencing. Nanopore sequencing statistics for each isolate are presented in Appendix A, Table A2.

Bioinformatic pipeline

In-house Illumina pipeline

Following Trimmomatic processing, all 18 read files passed the FastQC Adapter Content module, compared with none before trimming. All files also passed the Overrepresented Sequences and Per Base Sequence Content modules after processing. Complete FastQC results are presented in Appendix E, Table E1.

De novo assemblies were generated for all nine isolates. N50 values ranged from 1,865 bp in SRR11179850 to 217,944 bp in SRR11179853. SRR11179850 also had the highest number of contigs, at 2,681, while SRR11179862 had the largest individual contig, at 468,756 bp. Total

assembly lengths ranged from 3.98 to 5.16 Mbp, and GC content ranged from 50.28% to 51.15%. Full assembly metrics are provided in Appendix E, Table E2.

ANIb values against the UTI89 reference genome ranged from 96.37% to 98.80%, supporting identification of all nine isolates as *E. coli*. Individual ANIb values are presented in Appendix E, Table E3. Reference-based QUAST analysis showed genome fractions ranging from 61.19% in SRR11179849 to 91.58% in SRR11179842, with 39–137 misassemblies per isolate, as reported in Appendix E, Table E4.

VirulenceFinder detected 38–59 virulence-associated genes per isolate, with the lowest count in SRR11179851 and the highest in SRR11179848. These results are presented in Appendix E, Table E5. ResFinder detected 0–17 acquired AMR genes per isolate, with nonzero counts in six isolates and the highest count in SRR11179849. PointFinder identified 0–5 resistance-associated mutations per isolate. Complete AMR and mutation counts are provided in Appendix E, Table E6. PlasmidFinder detected 3–7 plasmid replicons per isolate, as shown in Appendix E, Table E7.

MLST assigned sequence types to seven of the nine isolates, identifying ST10, ST80, ST127, ST131, and ST2448. SRR11179849 and SRR11179850 remained unassigned because complete allele profiles were not obtained. Full allele profiles and sequence-type assignments are presented in Appendix E, Table E8.

Updated Illumina pipeline

Following fastp processing, all 18 read files passed the FastQC Adapter Content module, compared with none before processing. Complete FastQC results are presented in Appendix F, Table F1.

Genome assemblies were generated for all nine isolates using Unicycler. The number of contigs ranged from 50 in SRR11179853 to 802 in SRR11179849. N50 values ranged from 7,916 to 307,935 bp, total assembly lengths ranged from 3.75 to 5.14 Mbp, and GC content ranged from 50.30% to 51.02%. Full assembly metrics are provided in Appendix F, Table F2.

FastANI values against UTI89 ranged from 96.43% in SRR11179849 to 98.95% in SRR11179854, supporting identification of all nine isolates as *E. coli*. Individual ANI values are presented in Appendix F, Table F3. Reference-based QUAST analysis showed genome fractions of 58.52–91.52% and 48–141 misassemblies per isolate, as reported in Appendix F, Table F4.

ABRicate detected 1–15 AMR gene hits per isolate using the ResFinder database and 0–5 plasmid replicons using the PlasmidFinder database. SRR11179850 had the highest count in both analyses. Complete results are provided in Appendix F, Tables F5 and F6. Virulence-associated gene hits ranged from 21 to 95 using VFDB and from 130 to 248 using the *E. coli*-specific ecoli_vf database, as shown in Appendix F, Table F7.

MLST assigned sequence types to eight of the nine isolates. SRR11179849 remained unassigned because a complete allele profile was not obtained. Full allele profiles and sequence-type assignments are presented in Appendix F, Table F8.

Comparison of the in-house and updated Illumina pipelines

The updated pipeline produced higher N50 values, lower L50 values, and fewer contigs for all nine isolates compared with the in-house pipeline. The mean increase in N50 was 43,990 bp, and the difference was statistically significant (paired t-test, $p = 0.0037$). Species identification was consistent between the workflows. ANIb values from JSpeciesWS ranged from 96.37% to 98.80%, while FastANI values ranged from 96.43% to 98.95% against the same UTI89 reference genome. Both analyses supported identification of all nine isolates as *E. coli*.

AMR gene counts were identical between the workflows for eight of the nine isolates after excluding mdfA from the ABRicate results. SRR11179849 was the exception, with 17 genes detected by web-based ResFinder and six detected by ABRicate after excluding mdfA. Table 1 presents the counts before and after this exclusion.

Table 1. Comparison of antimicrobial resistance gene counts detected using the web-based ResFinder pipeline and the local ABRicate pipeline for nine isolates.

Sample	Original web-based AMR count	Updated ABRicate (including *mdfA*)	ABRicate (excluding *mdfA*)
SRR11179842	0	1	0
SRR11179848	7	8	7
SRR11179849	17	7	6
SRR11179850	14	15	14
SRR11179851	6	7	6
SRR11179853	0	1	0
SRR11179854	0	1	0
SRR11179861	8	9	8
SRR11179862	1	2	1

Virulence-associated gene counts differed between the workflows and databases. ABRicate with VFDB returned higher counts than CGE's VirulenceFinder for seven isolates and lower counts for two. The ecoli_vf database returned higher counts than either approach for all nine isolates, as shown in Table 2.

Table 2. Comparison of virulence-associated gene counts detected using CGE VirulenceFinder and ABRicate with the VFDB and *E. coli*-specific ecoli_vf databases.

Sample	Original CGE VirulenceFinder count	Updated ABRicate VFDB count	Updated ABRicate *E. coli* VF count
SRR11179842	56	87	222
SRR11179848	59	77	189
SRR11179849	43	38	130
SRR11179850	43	55	165
SRR11179851	38	46	184
SRR11179853	51	73	193
SRR11179854	50	90	244
SRR11179861	47	21	130
SRR11179862	54	95	248

Plasmid replicon detection also differed between the workflows. The updated ABRicate workflow detected fewer plasmid replicons in all nine isolates, with differences ranging from one to five replicons per isolate. The complete comparison is provided in Appendix G, Table G1.

The in-house pipeline assigned sequence types to seven isolates, whereas the updated pipeline assigned sequence types to eight. All seven isolates typed by both workflows received identical assignments. SRR11179850 was assigned ST10 only by the updated pipeline, while SRR11179849 remained unassigned in both. Table 3 presents the sequence-type comparison.

Table 3. Comparison of ST assignments between the in-house and updated workflows.

Sample	In-house workflow ST	Updated workflow ST
SRR11179842	80	80
SRR11179848	131	131
SRR11179849	NA	NA
SRR11179850	NA	10
SRR11179851	2448	2448
SRR11179853	131	131
SRR11179854	127	127
SRR11179861	10	10
SRR11179862	127	127

Workflow automation assessment

The in-house workflow required 45 individual web submissions, 37 manual result downloads, and at least 36 file-organization or renaming operations, totaling at least 118 counted manual operations. The submission count included nine submissions to each of five web-based tools. Downloads comprised 36 individual downstream result files and one combined JSpeciesWS results table. The updated workflow processed all nine isolates using a batch script without requiring these manual submission, download, or result-organization steps.

Recorded elapsed analysis time decreased from 11 h 25 min 13 s for the in-house workflow to 7 h 56 min 14 s for the updated workflow, corresponding to a reduction of 3 h 28 min 59 s. The complete workflow automation comparison is presented in Table 4.

Table 4. Comparison of workflow automation metrics for the in-house and updated Illumina workflows.

Automation metric	In-house workflow	Updated workflow
Samples analyzed	9	9
Web-based analysis tools	5	0
Individual downstream web submissions	45	0
Manual result downloads	37	0
Manual result organization/renaming	≥36	0
Minimum counted manual operations	≥118	0
Batch processing of all nine samples	Limited	Yes
Analysis parameters stored in executable scripts	Partial	Yes
Recorded elapsed analysis time	11 h 25 min 13 s	7 h 56 min 14 s

Nanopore assembly

Nanopore-only assemblies were obtained for eight of the nine isolates. SRR11179861 produced the most contiguous assembly, comprising two contigs with an N50 of 5,088,210 bp. SRR11179851 and SRR11179849 produced three and four contigs, respectively, with N50 values of approximately 4.67 Mbp. SRR11179850 produced the shortest assembly, with a total

length of 677,102 bp, 48 contigs, and an N50 of 17,869 bp. No Nanopore-only assembly was obtained for SRR11179848. Complete assembly metrics are presented in Appendix H, Table H1.

Hybrid assembly

Assemblies were generated for all nine isolates using combined Illumina and Nanopore reads. N50 values ranged from 81,225 bp in SRR11179850 to 4,668,311 bp in SRR11179851, with corresponding L50 values of 18 and 1. SRR11179853 and SRR11179862 had N50 values of 2,018,938 bp and 1,628,182 bp, respectively, both with an L50 of 2. Complete assembly metrics are presented in Appendix J, Table J1.

MLST assigned sequence types to eight isolates, while SRR11179849 remained unassigned because a complete allele profile was not obtained. Full allele profiles are provided in Appendix J, Table J2. Virulence-associated gene hits ranged from 26 to 100 using VFDB and from 130 to 276 using ecoli_vf. AMR gene counts ranged from 1 to 17 per isolate. These results are presented in Appendix J, Tables J3 and J4, respectively.

Comparison of updated Illumina-only and hybrid assemblies

Hybrid assemblies had higher N50 values and lower L50 values than the updated Illumina-only assemblies for eight of the nine isolates. Both metrics remained unchanged for SRR11179848. The largest increase in N50 occurred in SRR11179851, from 198,106 bp to 4,668,311 bp. N50 values were significantly higher in the hybrid assemblies overall (Wilcoxon signed-rank test, $V = 36$, $p = 0.014$). Table 5 presents the N50 and L50 comparisons.

MLST assignments were unchanged between the two approaches. Eight isolates received identical sequence types, while SRR11179849 remained unassigned in both, as shown in Appendix I, Table I1. Virulence-associated gene counts obtained using VFDB were higher in all nine hybrid assemblies, with increases of 2–12 hits per isolate. AMR gene counts were identical for eight isolates, while the count for SRR11179850 increased from 15 to 17. These comparisons are presented in Appendix I, Tables I2 and I3, respectively.

Table 5. Comparison of N50 and L50 values between updated Illumina-only and hybrid genome assemblies.

Sample	Illumina N50 (bp)	Hybrid N50 (bp)	Illumina L50	Hybrid L50
SRR11179842	241,830	1,104,353	7	2
SRR11179848	190,808	190,808	9	9
SRR11179849	7,916	823,610	114	3
SRR11179850	14,908	81,225	82	18
SRR11179851	198,106	4,668,311	7	1
SRR11179853	307,935	2,018,938	5	2
SRR11179854	35,936	131,918	30	9
SRR11179861	133,852	242,038	11	7
SRR11179862	262,451	1,628,182	6	2

Statistical analysis

The Shapiro–Wilk test showed that the paired differences in N50 values between the in-house and updated pipelines did not significantly deviate from normality ($W = 0.902$, $p = 0.265$). Therefore, a paired t-test was used. The updated pipeline produced significantly higher N50 values than the in-house pipeline ($t(8) = 4.04$, $p = 0.0037$), with a mean increase of 43,990 bp

(95% confidence interval: 18,873–69,108 bp). Statistical results for the other assembly metrics, including L50, total assembly length, and number of contigs, are provided in Appendix K, Table K1.

For the comparison between Illumina-only and hybrid assemblies, the paired differences in N50 values were not normally distributed (Shapiro–Wilk: W = 0.751, p = 0.006). Therefore, the Wilcoxon signed-rank test was used. A statistically significant difference in N50 was observed between the two assembly approaches (V = 36, p = 0.014), with higher N50 values in the hybrid assemblies.

Discussion

DNA extraction and quality control

The DNA extracted using the Promega DNA extraction kit showed high purity, with $A_{260/280}$ ratios ranging between 1.98 and 2.14 for all samples as shown in Appendix B, Table B1. The $A_{260/280}$ ratio is an indicator of protein and RNA contamination (Wilfinger et al., 1997), which can affect DNA quantification and nanopore library preparation and at high levels, interfere with sequencing (Kyei-Tuffuor et al., 2026). The $A_{260/230}$ ratio is another indicator of DNA quality, reflecting possible contaminations such as phenol, EDTA, and chaotropic salts (Imran, 2025). For nanopore sequencing, an optimal $A_{260/230}$ ratio typically falls within the range of 2.0–2.2. Most samples showed A260/A230 ratios close to or within this range, although several values were slightly below 2.0. Maintaining appropriate $A_{260/230}$ values is important for downstream applications (Kruasuwan et al., 2024). To increase DNA yield, the starting culture volume was increased from 1 mL to 5 mL and the centrifugation time was extended. This provided sufficient DNA for nine of the ten isolates to continue to library preparation, while isolate 19 was excluded because of its low DNA concentration. Sufficient DNA quantity is important for Nanopore library preparation and sequencing performance (Heavens et al., 2021).

Fragment analysis

PFGE was used instead of conventional gel electrophoresis because the extraction kit is designed to produce high-molecular-weight DNA fragments in the range of 50 kb to 500 kb, which cannot be reliably resolved using standard agarose gels (Volarić et al., 2021). Most samples contained fragments of approximately 100–200 kb, indicating that DNA integrity was generally preserved and suitable for downstream Nanopore sequencing.

Library preparation and Nanopore sequencing

DNA concentration decreased from 46.0 to 28.2 ng/µL between native barcode ligation and adapter ligation with clean-up. DNA loss during purification may have contributed to this decrease (Howeler et al., 2003). Nevertheless, the prepared library supported sequencing. Previous work has demonstrated that usable Nanopore data can be generated with reduced DNA input, although sequencing performance may be affected (Žemaitis et al., 2025).

The flow cell contained 1,538 active pores before sequencing, exceeding the approximately 800 pores generally considered sufficient for sequencing (Pugh, 2023; Chaves et al., 2025). The run was stopped after approximately 20 hours. Although extending the run could potentially increase sequencing yield and coverage (Afonso & Afonso, 2023), longer sequencing times

would also reduce the potential time advantage of Nanopore sequencing in applications where rapid results are important.

In-house pipeline

Trimmomatic processing improved the FastQC Adapter Content and Overrepresented Sequences results, supporting its effectiveness in removing technical sequences from the reads (Bolger et al., 2014). However, improved read-quality indicators do not necessarily guarantee contiguous assemblies. Trimming involves a balance between removing unreliable sequence and retaining sufficient information for assembly, as excessive trimming can reduce the usable data available (Del Fabbro et al., 2013).

SPAdes assembly quality varied substantially between isolates. SRR11179850 and SRR11179849 produced particularly fragmented assemblies, providing a useful baseline for evaluating the updated workflow. Differences in sequencing depth and read characteristics may contribute to variation in assembly performance (Lander & Waterman, 1988). Short reads also have limited ability to resolve repetitive regions, which can interrupt genome reconstruction (Kircher & Kelso, 2010). These factors were not evaluated individually here, so the causes of fragmentation cannot be established from the assembly metrics alone.

Despite variation in contiguity, total assembly lengths ranged from 3.98 to 5.16 Mbp and GC content remained between 50.28% and 51.15%, broadly consistent with expected *E. coli* genome characteristics (Muto & Osawa, 1987). ANIb results supported the identification of all nine isolates as *E. coli* despite differences in assembly contiguity. This distinction is important because species identification and genome reconstruction assess different aspects of the data: an assembly may support species identification while remaining fragmented. The observed ANI values were consistent with commonly used species boundaries (Jain et al., 2018). Reference-based genome fractions varied from 61.19% to 91.58%, suggesting that lower reference coverage may reflect both assembly fragmentation and genuine genomic differences between isolates and UTI89 rather than assembly failure alone (Goris et al., 2007).

Web-based virulence, antimicrobial-resistance and plasmid profiles also varied among the isolates. Such variation is expected in extraintestinal pathogenic *E. coli*, where virulence and resistance determinants may occur on chromosomes, plasmids and other mobile genetic elements (Dobrindt et al., 2004). However, the presence of virulence genes indicates pathogenic potential rather than confirmed pathogenicity (Wassenaar & Gaastra, 2001). MLST assigned sequence types to seven of nine isolates. SRR11179849 and SRR11179850 also had relatively fragmented assemblies, suggesting that incomplete reconstruction of MLST loci may have contributed to the failed assignments, although this study cannot establish fragmentation as the direct cause. Together, these results provided the baseline against which the updated pipeline was assessed.

Updated Illumina pipeline and comparison with the in-house pipeline

Both Trimmomatic and fastp could be run using scripts. Although fastp provided automatic adapter detection and integrated reporting (S. Chen, 2025), the main improvement in automation came from replacing manual web-based analyses with locally executable tools and processing all nine isolates in a batch.

A clearer improvement was observed at the assembly stage. Unicycler produced significantly higher N50 values than the original SPAdes workflow, with a mean increase of 43,990 bp and p =

0.0037. Unicycler was originally developed to improve bacterial genome assembly by using an assembly framework designed specifically for bacterial genomes, which may explain its stronger performance in this dataset (Wick et al., 2017). However, higher N50 alone does not guarantee a more accurate assembly, and assembly quality should therefore be considered together with other metrics and downstream biological results.

Improved contiguity did not indicate improvement in every measure of assembly quality. Reference-based QUAST reported more misassemblies in the updated assemblies for all nine isolates. This finding limits the conclusion that the updated assemblies were universally better: higher N50 values describe longer contiguous sequences but do not establish their structural accuracy.

Species identification remained consistent between the two pipelines. JSpeciesWS produced ANI values of 96.37–98.80%, while FastANI produced values of 96.43–98.95% against the same UTI89 reference genome. Using the same reference provided a common basis for comparison, although the ANI calculation methods differed. Both approaches therefore produced highly similar identification results. However, FastANI allowed the analysis to be performed locally as part of a command-line workflow rather than through manual web submission.

The web-based AMR analysis and ABRicate produced matching AMR gene counts for eight isolates after excluding mdfA from the ABRicate results. Differences in gene counts were mainly related to the detection of *mdfA*, a multidrug efflux-pump gene commonly found in *E. coli* (Edgar & Bibi, 1997). The inconsistency for SRR11179849 remains unresolved. Comparing individual hits, database versions, and detection thresholds would help determine its cause.

Virulence and plasmid results showed larger differences between the two pipelines. ABRicate with the *E. coli* specific ecoli_vf database detected more virulence-associated genes than either VFDB or the web-based CGE VirulenceFinder analysis. These differences may reflect variation in database content, gene definitions, and reporting criteria rather than differences in software performance alone. Similarly, the updated workflow detected fewer plasmid replicons than the CGE PlasmidFinder analysis, which could also have reflected differences between database versions. Higher gene counts do not necessarily indicate more accurate detection, and plasmid-replicon counts do not represent the number of complete plasmids (Carattoli et al., 2014).

MLST assignments were identical for the seven isolates successfully typed by both pipelines. The additional assignment of SRR11179850 to ST10 suggests that the updated workflow recovered sufficient sequence information to resolve its allelic profile. However, SRR11179849 remained unassigned despite improved contiguity, demonstrating that improved assembly metrics did not resolve every downstream analysis.

The clearest improvement in practical workflow performance was the reduction in manual intervention. The in-house workflow required five web-based tools, 45 individual web submissions, 37 manual result downloads, and at least 36 file-organization steps, corresponding to at least 118 counted manual operations. The updated workflow eliminated these web-based operations and allowed all nine isolates to be processed in batches using executable scripts. Recorded elapsed time also decreased from 11 h 25 min 13 s to 7 h 56 min 14 s. Although execution time can be affected by hardware and system load, the reduction in manual handling and the ability to batch-process samples provide direct evidence that the updated workflow achieved greater automation while preserving species identification and the ST assignments obtained by both workflows.

Nanopore assembly

Nanopore-only assembly produced variable results across the isolates. Several assemblies contained only a few contigs, whereas others remained highly fragmented, as shown in Appendix H, Table H1. These findings demonstrate that long reads supported contiguous assemblies for some isolates but did not provide consistent results across the dataset. This variability may be related to the multiplex sequencing strategy, where uneven barcode distribution can lead to unequal read depth across samples (Thornval et al., 2026). DNA extraction and sample preparation are important considerations for long-read sequencing (Maghini et al., 2020). However, their individual effects were not tested here, so they cannot be identified as the causes of the observed assembly differences. Sample SRR11179848 did not produce a usable assembly, despite its barcode showing some passed reads in MinKNOW. Further examination of read length, quality, and coverage would be needed to explain why this sample did not produce an assembly.

Hybrid assembly and comparison with Illumina-only assemblies

Hybrid assembly combines the high base-level accuracy of short reads with the ability of long reads to span repetitive regions, helping to resolve complex genomic structures, including plasmids (Henson et al., 2012). In this study, Illumina and Nanopore reads were combined to generate hybrid assemblies, fulfilling the second aim. Comparing these assemblies with the updated Illumina-only assemblies addressed the third aim by evaluating changes in contiguity. N50 increased and L50 decreased in eight of the nine isolates, with a statistically significant increase in N50 overall ($p = 0.014$). These results demonstrate improved assembly contiguity for most isolates and are consistent with the benefits of hybrid bacterial genome assembly described by Wick et al. (2017).

The largest improvement occurred in SRR11179851, where N50 increased from 198,106 bp to 4,668,311 bp. However, the benefit varied among isolates. SRR11179848 showed no change in N50 or L50, while SRR11179850 remained relatively fragmented despite improvement. Previous comparisons have also shown that hybrid assembly performance varies with the sequencing data and assembly approach used (De Maio et al., 2019). The present study did not separately test the effects of read depth, read length, or DNA quality, so the reasons for these differences remain uncertain.

Higher N50 and lower L50 values indicate improved contiguity but do not establish that an assembly is complete or structurally correct. Consequently, the results support improved contiguity in most isolates rather than a general claim of complete genome reconstruction (Wang & Wang, 2022). Further assessment of assembly accuracy and completeness would be needed to establish whether the longer contigs correctly represent bacterial chromosomes and plasmids.

Improved hybrid assembly contiguity did not change MLST assignments, with SRR11179849 remaining unassigned. AMR gene counts were unchanged for eight isolates but increased from 15 to 17 in SRR11179850. VFDB screening detected 2–12 additional hits per hybrid assembly. These additional hits require individual examination before concluding that hybrid assembly improved gene detection. Plasmid reconstruction was not assessed in detail, so the downstream findings remain exploratory.

Limitations

The study included nine isolates, limiting the generalizability of the findings. Several pipeline components changed together, so their individual contributions could not be determined. Differences in screening methods, databases, and thresholds also limited comparisons of downstream gene counts. Although reference-based QUAST identified potential assembly discrepancies, structural accuracy was not independently validated. Improved N50 and L50 values therefore demonstrate greater contiguity rather than confirmed completeness or accuracy. Timing measurements were not repeated under controlled conditions, and clinical diagnostic performance was not evaluated.

Conclusion

The study addressed its main objectives by updating the existing Illumina workflow and evaluating hybrid genome assembly for clinical *E. coli* isolates. The updated Illumina pipeline improved the existing in-house workflow in both assembly contiguity and workflow automation. N50 increased by a mean of 43,990 bp and was significantly higher in the updated assemblies, with $p = 0.0037$. The updated workflow also replaced five web-based analysis tools, eliminated at least 118 counted manual operations, and enabled batch processing of all nine isolates. Species identification remained consistent, while AMR, virulence-associated gene, and plasmid-replicon counts showed varying agreement. MLST assignments were identical for isolates successfully typed by both approaches, while SRR11179850 was additionally assigned ST10 by the updated pipeline and SRR11179849 remained unassigned in both.

The second objective was achieved by generating hybrid assemblies that combined short-read Illumina data with long-read Oxford Nanopore data. Addressing the final objective, comparison with the updated Illumina-only assemblies showed a statistically significant overall increase in N50, indicating improved assembly contiguity. These findings support hybrid assembly as an approach to improving contiguity in this dataset, which may facilitate downstream characterization of virulence-associated and AMR determinants. However, the effects of improved contiguity on downstream genomic analyses were not fully established.

Overall, the updated workflow reduced manual intervention, while hybrid assembly improved contiguity for most isolates. These findings support further evaluation of the workflow for genomic characterization of clinical isolates. However, this study did not evaluate clinical diagnostic performance, analytical sensitivity, and clinical validity would need further assessment before routine clinical implementation.

Ethical aspects, gender perspectives, and impact on society

This study was conducted in accordance with ethical approval granted by the Ethical Review Board at the University of Gothenburg (approval number 376/2011). The bacterial isolates used in this project originate from a previously approved clinical study by Ljungström et al. (2019), which investigated sepsis patients in Sweden. The use of these samples falls within the scope of the original ethical authorization, and individual informed consent was not required for secondary analysis of bacterial isolates that could not be linked to individual patients.

All work involving WGS data was performed in compliance with established ethical guidelines for handling clinical microbial material. Particular attention was given to data governance, including secure storage, controlled access, and appropriate data retention policies. To protect

patient privacy, all datasets were fully de-identified prior to analysis, with removal of any indirect or direct identifiers. Sequencing data were stored in secure environments and were not transferred to unsecured external or cloud-based platforms. The data were used exclusively for the intended research purpose of improving genomic analysis workflows for clinically relevant pathogens.

Given that the study involves bacterial isolates rather than direct human or animal experimentation, the ethical considerations primarily relate to data protection, responsible use of clinical-derived material, and the prevention of misuse of genomic information. No attempts were made to re-identify individual patients, thereby minimizing risks related to confidentiality breaches.

From a societal perspective, this work has potential implications for improving the genomic characterization of sepsis-causing *E. coli* isolates. Enhanced genome assembly and analysis pipelines may contribute to more accurate detection of antimicrobial resistance determinants and virulence factors, thereby supporting improved diagnostic accuracy and informing treatment strategies. Furthermore, benchmarking short-read and hybrid assembly workflows may assist clinical laboratories and public health institutions in adopting more effective sequencing strategies for pathogen surveillance and outbreak investigation. In the long term, such improvements may contribute to a better understanding of antimicrobial resistance and support the development of more targeted therapeutic interventions.

As this study is based on microbial genomic data and does not involve direct human participation or experimental animal models, sex and gender perspectives are not directly applicable to the research design or analysis

Future perspectives

This thesis demonstrated the advantages of combining Illumina short-read and nanopore long-read sequencing for the characterization of clinically relevant *E. coli* isolates. Several areas could benefit from further development and improvement in future studies.

Other studies, such as Chakrawarti et al. (2024), have already used hybrid assembly to assign MLST sequence types and identify antimicrobial resistance genes in *Staphylococcus aureus* isolated from dairy farms. Their study also showed that the hybrid approach outperformed Illumina sequencing alone. This suggests that the future potential of this approach could extend far beyond a single disease.

Another future direction would be to investigate the ability of hybrid assemblies to fully resolve plasmids carrying antimicrobial resistance genes. In a study conducted by George et al. (2017), Nanopore sequencing was shown to be valuable for reconstructing plasmid structures and tracking horizontal gene transfer between bacteria. In this study, hybrid assemblies generated more contiguous genomes, suggesting that plasmid reconstruction may also be improved using this approach. Future studies could compare plasmid structures between isolates to investigate how resistance genes are transmitted and maintained. Similarly, Beh et al. (2025) highlight ongoing methodological challenges in resolving AMR-carrying plasmids from whole-genome sequencing data, underscoring that further tool development remains necessary even as hybrid approaches improve overall assembly contiguity.

The workflow developed in this thesis was evaluated using a limited number of isolates. Future studies should expand the number of clinical isolates analyzed in order to evaluate the robustness of the workflow across a wider range of *E. coli* STs, virulence profiles, and antimicrobial resistance patterns. High-risk clones such as ST131 have been frequently associated with multidrug resistance in large-scale genomic surveillance studies. Therefore, including a larger and more diverse collection of isolates could improve our understanding of how resistant and virulent strains spread within healthcare systems and society.

Acknowledgments

I would like to begin by acknowledging my thesis supervisor, Sanja Jurcevic. She guided me, motivated me, and helped me understand the project. I could not have successfully completed the project without her. I am very grateful for the support I received during both the practical and writing parts of the project. She carefully reviewed my work and provided valuable comments that helped me improve it. I would also like to thank Anna-Karin Pernestig and Magnus Fagerlind for their help in the lab and with the data analysis.

On a personal note, I would like to thank my partner, Patrik Andersson, for his time, care, and motivation throughout this journey. He supported me and gave me strength even when I was not feeling my best. I would also like to thank my sisters for their support and words of encouragement when I needed them most, and my brother for checking up on me so often.

References

Achudhan, A. B., Kannan, P., Gupta, A., & Saleena, L. M. (2024). A review of web-based metagenomics platforms for analysing next-generation sequence data. *Biochemical Genetics*, 62(2), 621–632. https://doi.org/10.1007/s10528-023-10467-w

Afonso, C. L., & Afonso, A. M. (2023). Next-Generation sequencing for the detection of microbial agents in avian clinical samples. *Veterinary Sciences*, 10(12), 690. https://doi.org/10.3390/vetsci10120690

Agudelo-Pérez, S., Fernández-Sarmiento, J., León, D. R., & Peláez, R. G. (2023). Metagenomics by next-generation sequencing (mNGS) in the etiological characterization of neonatal and pediatric sepsis: A systematic review. *Frontiers in Pediatrics*, 11, 1011723. https://doi.org/10.3389/fped.2023.1011723

Amoutzias, G. D., Nikolaidis, M., & Hesketh, A. (2022). The notable achievements and the prospects of bacterial pathogen genomics. *Microorganisms*, 10(5), 1040. https://doi.org/10.3390/microorganisms10051040

Baron, R. M., Baron, M. J., & Perrella, M. A. (2006). Pathobiology of sepsis: Are we still asking the same questions? *American Journal of Respiratory Cell and Molecular Biology*, 34(2), 129–134. https://doi.org/10.1165/rcmb.f308

Beh, J. Q., Wick, R. R., Howden, B. P., Connor, C. H., & Webb, J. R. (2025). Challenges and considerations for whole-genome-based antimicrobial resistance plasmid investigations. *Antimicrobial Agents and Chemotherapy*, 69(12), e0109725. https://doi.org/10.1128/aac.01097-25

Bio-Rad Laboratories. (2007). *CHEF-DR® III pulsed field electrophoresis systems instruction manual and applications guide* (Catalog Nos. 170-3690 through 170-3703).

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bonten, M., Johnson, J. R., Van Den Biggelaar, A. H. J., Georgalis, L., Geurtsen, J., De Palacios, P. I., Gravenstein, S., Verstraeten, T., Hermans, P., & Poolman, J. T. (2021). Epidemiology of Escherichia coli Bacteremia: A Systematic Literature Review. *Clinical Infectious Diseases*, 72(7), 1211–1219. https://doi.org/10.1093/cid/ciaa210

Carattoli, A., Zankari, E., García-Fernández, A., Larsen, M. V., Lund, O., Villa, L., Aarestrup, F. M., & Hasman, H. (2014). In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrobial Agents and Chemotherapy, 58*(7), 3895–3903. https://doi.org/10.1128/AAC.02412-14

Cecconi, M., Evans, L., Levy, M., & Rhodes, A. (2018). Sepsis and septic shock. *The Lancet*, 392(10141), 75–87. https://doi.org/10.1016/s0140-6736(18)30696-2

Chakrawarti, A., Eckstrom, K., Laaguiby, P., & Barlow, J. W. (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. *Access Microbiology, 6*(3), 000766.v3. https://doi.org/10.1099/acmi.0.000766.v3

Chaves, M., Hashish, A., Goraichuk, I. V., Casserta, L. C., Mears, M. C., Gadu, E., Bakre, A., Morris, E. R. A., Shelkamy, M. M. S., Nadendla, S., Perez, D. R., & El-Gazzar, M. (2025). Nanopore sequencing in veterinary medicine: from concepts to clinical applications. *Frontiers in Cellular and Infection Microbiology*, 15, 1701570. https://doi.org/10.3389/fcimb.2025.1701570

Chen, S. (2025). fastp 1.0: An ultra-fast all-round tool for FASTQ data quality control and preprocessing. *iMeta*, 4(5), e70078. https://doi.org/10.1002/imt2.70078

Chen, X., Lu, X., Sh, X., Yu, S., & Zhao, J. (2025). Long-Read sequencing outperforms Short-Read sequencing in detecting most structural variations. *Serican Journal of Medicine*, 2(2). https://doi.org/10.17161/sjm.v2i2.23671

Chun, K., Syndergaard, C., Damas, C., Trubey, R., Mukindaraj, A., Qian, S., Jin, X., Breslow, S., & Niemz, A. (2015). Sepsis Pathogen Identification. *Slas Technology*, 20(5), 539–561. https://doi.org/10.1177/2211068214567345

Dale, A. P., & Woodford, N. (2015). Extra-intestinal pathogenic Escherichia coli (ExPEC): Disease, carriage and clones. *Journal of Infection*, 71(6), 615–626. https://doi.org/10.1016/j.jinf.2015.09.009

Dargère, S., Cormier, H., & Verdon, R. (2018). Contaminants in blood cultures: importance, implications, interpretation and prevention. *Clinical Microbiology and Infection*, 24(9), 964–969. https://doi.org/10.1016/j.cmi.2018.03.030

de Grooth, H.-J., Geenen, I. L., Girbes, A. R., Vincent, J.-L., Parienti, J.-J., & Oudemans-van Straaten, H. M. (2017). SOFA and mortality endpoints in randomized controlled trials: A systematic review

and meta-regression analysis. *Critical Care, 21*(1), 38. https://doi.org/10.1186/s13054-017-1609-1

De Maio, N., Shaw, L. P., Hubbard, A., George, S., Sanderson, N. D., Swann, J., Wick, R., AbuOun, M., Stubberfield, E., Hoosdally, S. J., Crook, D. W., Peto, T. E. A., Sheppard, A. E., Bailey, M. J., Read, D. S., Anjum, M. F., Walker, A. S., & Stoesser, N. (2019). Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microbial Genomics*, 5(9). https://doi.org/10.1099/mgen.0.000294

Del Fabbro, C., Scalabrin, S., Morgante, M., & Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS ONE*, 8(12), e85024. https://doi.org/10.1371/journal.pone.0085024

Dellinger, R. P., Levy, M. M., Rhodes, A., Annane, D., Gerlach, H., Opal, S. M., Sevransky, J. E., Sprung, C. L., Douglas, I. S., Jaeschke, R., Osborn, T. M., Nunnally, M. E., Townsend, S. R., Reinhart, K., Kleinpell, R. M., Angus, D. C., Deutschman, C. S., Machado, F. R., Rubenfeld, G. D., . . . Moreno, R. (2013). Surviving Sepsis campaign. *Critical Care Medicine*, 41(2), 580–637. https://doi.org/10.1097/ccm.0b013e31827e83af

Dobrindt, U., Hochhut, B., Hentschel, U., & Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nature Reviews Microbiology*, 2(5), 414–424. https://doi.org/10.1038/nrmicro884

Edgar, R., & Bibi, E. (1997). MdfA, an Escherichia coli multidrug resistance protein with an extraordinarily broad spectrum of drug recognition. *Journal of Bacteriology*, 179(7), 2274–2280. https://doi.org/10.1128/jb.179.7.2274-2280.1997

Eren, K., Taktakoğlu, N., & Pirim, I. (2022). DNA sequencing methods: From past to present. *The Eurasian Journal of Medicine, 54*(Suppl. 1), S47–S56. https://doi.org/10.5152/eurasianjmed.2022.22280

George, S., Pankhurst, L., Hubbard, A., Votintseva, A., Stoesser, N., Sheppard, A. E., Mathers, A., Norris, R., Navickaite, I., Eaton, C., Iqbal, Z., Crook, D. W., & Phan, H. T. T. (2017). Resolving plasmid structures in Enterobacteriaceae using the MinION nanopore sequencer: assessment of MinION and MinION/Illumina hybrid data assembly approaches. *Microbial Genomics*, 3(8), e000118. https://doi.org/10.1099/mgen.0.000118

Gilchrist, C. A., Turner, S. D., Riley, M. F., Petri, W. A., & Hewlett, E. L. (2015). Whole-Genome sequencing in outbreak analysis. *Clinical Microbiology Reviews*, 28(3), 541–563. https://doi.org/10.1128/cmr.00075-13

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, 57(Pt 1), 81–91. https://doi.org/10.1099/ijs.0.64483-0

He, X., Chen, S., Li, R., Han, X., He, Z., Yuan, D., Zhang, S., Duan, X., & Niu, B. (2020). Comprehensive fundamental somatic variant calling and quality management strategies for human cancer genomes. *Briefings in Bioinformatics*, 22(3). https://doi.org/10.1093/bib/bbaa083

Heavens, D., Chooneea, D., Giolai, M., Cuber, P., Aanstad, P., Martin, S., Alston, M., Misra, R., Clark, M. D., & Leggett, R. M. (2021). How low can you go? Driving down the DNA input requirements

for nanopore sequencing. *bioRxiv* (Cold Spring Harbor Laboratory). https://doi.org/10.1101/2021.10.15.464554

Henson, J., Tischler, G., & Ning, Z. (2012). Next-Generation sequencing and large genome assemblies. *Pharmacogenomics*, 13(8), 901–915. https://doi.org/10.2217/pgs.12.72

Howeler, M., Ghiorse, W. C., & Walker, L. P. (2003). A quantitative analysis of DNA extraction and purification from compost. *Journal of Microbiological Methods*, 54(1), 37–45. https://doi.org/10.1016/s0167-7012(03)00006-x

Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Human Immunology*, 82(11), 801–811. https://doi.org/10.1016/j.humimm.2021.02.012

Hung, S., Lan, H., Han, S., Wu, C., & Chen, K. (2020). Current evidence and limitation of biomarkers for detecting sepsis and systemic infection. *Biomedicines*, 8(11), 494. https://doi.org/10.3390/biomedicines8110494

Hutchings, M. I., Truman, A. W., & Wilkinson, B. (2019). Antibiotics: past, present and future. *Current Opinion in Microbiology*, 51, 72–80. https://doi.org/10.1016/j.mib.2019.10.008

Imran, I. (2025). Enhancing the Quality of DNA Retrieved from Long-Term Human Blood Preservation Suitable for Genetic Analysis. *Russian Journal of Genetics*, 61(4), 485–490. https://doi.org/10.1134/s1022795424701886

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), 5114. https://doi.org/10.1038/s41467-018-07641-9

Jambholkar, P. C., Choudhari, S. G., & Sharma, M. (2024). Louis Pasteur: A legacy unmasked. *Cureus*, 16(8), e68080. https://doi.org/10.7759/cureus.68080

Javadi, M., Bouzari, S., & Oloomi, M. (2017). Horizontal Gene Transfer and the Diversity of *Escherichia coli. InTech* eBooks. https://doi.org/10.5772/intechopen.68307

Kaper, J. B., Nataro, J. P., & Mobley, H. L. T. (2004). Pathogenic Escherichia coli. *Nature Reviews Microbiology*, 2(2), 123–140. https://doi.org/10.1038/nrmicro818

Kashyap, R., Sherani, K. M., Dutt, T., Gnanapandithan, K., Sagar, M., Vallabhajosyula, S., Vakil, A. P., & Surani, S. (2021). Current Utility of Sequential Organ Failure Assessment Score: A Literature review and future directions. *The Open Respiratory Medicine Journal*, 15(1), 1–6. https://doi.org/10.2174/1874306402115010001

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), 2560. https://doi.org/10.3390/microorganisms9122560

Kircher, M., & Kelso, J. (2010). High-throughput DNA sequencing – concepts and limitations. *BioEssays*, 32(6), 524–536. https://doi.org/10.1002/bies.200900181

Kumar, K. R., Cowley, M. J., & Davis, R. L. (2024). Next-Generation Sequencing and Emerging Technologies*. *Seminars in Thrombosis and Hemostasis*, 50(07), 1026–1038. https://doi.org/10.1055/s-0044-1786397

Kruasuwan, W., Sawatwong, P., Jenjaroenpun, P., Wankaew, N., Arigul, T., Yongkiettrakul, S., Lunha, K., Sudjai, A., Siludjai, D., Skaggs, B., & Wongsurawat, T. (2024). Comparative evaluation of commercial DNA isolation approaches for nanopore-only bacterial genome assembly and plasmid recovery. *Scientific Reports*, 14(1), 27672. https://doi.org/10.1038/s41598-024-78066-2

Kyei-Tuffuor, L., Agordzo, S. K., Asante, A. K., Boateng, D. S., Adjei, W. N. S., Frimpong, V. N. B., Frimpong, R. N., Agyemang-Yeboah, F., Kobialka, R. M., Wahed, A. a. E., Truyen, U., Amoako, Y. A., Phillips, R. O., & Frimpong, M. (2026). Comparative performance of portable DNA extraction protocols and bioinformatics workflows for rapid detection of pathogens and antimicrobial resistance using Oxford Nanopore sequencing. *bioRxiv* (Cold Spring Harbor Laboratory). https://doi.org/10.64898/2026.02.23.707422

Lambden, S., Laterre, P. F., Levy, M. M., & Francois, B. (2019). The SOFA score—development, utility and challenges of accurate assessment in clinical trials. *Critical Care*, 23(1), 374. https://doi.org/10.1186/s13054-019-2663-7

Lander, E. S., & Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics, 2*(3), 231–239. https://doi.org/10.1016/0888-7543(88)90007-9

Liang, S. Y. (2016). Sepsis and other infectious disease emergencies in the elderly. *Emergency Medicine Clinics of North America*, 34(3), 501–522. https://doi.org/10.1016/j.emc.2016.04.005

Liu, B., Zheng, D., Zhou, S., Chen, L., & Yang, J. (2022). VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Research*, 50(D1), D912–D917. https://doi.org/10.1093/nar/gkab1107

Liu, Y., Yao, Y., Yu, M., Gao, Y., Qi, A., Jiang, T., Chen, Z., Shou, S., & Chai, Y. (2022). Frequency and mortality of sepsis and septic shock in China: a systematic review and meta-analysis. *BMC Infectious Diseases*, 22(1), 564. https://doi.org/10.1186/s12879-022-07543-8

Ljungström, L., Andersson, R., & Jacobsson, G. (2019). Incidences of community onset severe sepsis, Sepsis-3 sepsis, and bacteremia in Sweden – A prospective population-based study. *PLoS ONE, 14*(12), e0225700. https://doi.org/10.1371/journal.pone.0225700

Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics Proteomics & Bioinformatics*, 14(5), 265–279. https://doi.org/10.1016/j.gpb.2016.05.004

Maghini, D. G., Moss, E. L., Vance, S. E., & Bhatt, A. S. (2020). Improved high-molecular-weight DNA extraction, nanopore sequencing and metagenomic assembly from the human gut microbiome. *Nature Protocols*, 16(1), 458–471. https://doi.org/10.1038/s41596-020-00424-x

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. *Clinical Microbiology Reviews*, 23(1), 235–251. https://doi.org/10.1128/cmr.00043-09

Marcello, M. G., Tumolo, M. R., De Donno, A., Verri, T., Serio, F., Bagordo, F., & Zizza, A. (2016). In vitro diagnosis of sepsis: a review. *Pathology and Laboratory Medicine International*, 1. https://doi.org/10.2147/plmi.s49800

Marshall, D. A., MacDonald, K. V., Robinson, J. O., Barcellos, L. F., Gianfrancesco, M., Helm, M., McGuire, A., Green, R. C., Douglas, M. P., Goldman, M. A., & Phillips, K. A. (2017). The Price of Whole-Genome Sequencing May be Decreasing, But Who Will be Sequenced? *Personalized Medicine*, 14(3), 203–211. https://doi.org/10.2217/pme-2016-0075

Mellhammar, L., Wollter, E., Dahlberg, J., Donovan, B., Olséen, C., Wiking, P. O., Rose, N., Schwarzkopf, D., Friedrich, M., Fleischmann-Struzek, C., Reinhart, K., & Linder, A. (2023). Estimating sepsis incidence using administrative data and clinical medical record review. *JAMA Network Open*, 6(8), e2331168. https://doi.org/10.1001/jamanetworkopen.2023.31168

Mink, S., Attenberger, C., Busch, Y., Kiefer, J., Peter, W., Cadamuro, J., Steiert, T. A., Franke, A., & Gassner, C. (2025). Merging High-Throughput, Amplicon-Based second and third generation sequencing data: an integrative and modular data analysis framework for haplotype prediction and output evaluation. *International Journal of Molecular Sciences*, 26(7), 3443. https://doi.org/10.3390/ijms26073443

Motzkus, C. A., & Luckmann, R. (2016). Does infection site matter? A Systematic review of infection site mortality in sepsis. *Journal of Intensive Care Medicine*, 32(8), 473–479. https://doi.org/10.1177/0885066615627778

Muto, A., & Osawa, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. *Proceedings of the National Academy of Sciences*, 84(1), 166–169. https://doi.org/10.1073/pnas.84.1.166

Ng, P. C., & Kirkness, E. F. (2010). Whole genome sequencing. *Methods in Molecular Biology*, 628, 215–226. https://doi.org/10.1007/978-1-60327-367-1_12

Ni, Y., Li, J., Tan, Y., Shen, G., & Liu, C. (2025). Advance in the assembly of the plant mitochondrial genomes using high-throughput DNA sequencing data of total cellular DNAs. *Plant Biotechnology Journal*, 23(11), 4944–4965. https://doi.org/10.1111/pbi.70249

Nicolas-Chanoine, M., Bertrand, X., & Madec, J. (2014). Escherichia coli ST131, an Intriguing Clonal Group. *Clinical Microbiology Reviews*, 27(3), 543–574. https://doi.org/10.1128/cmr.00125-13

Opal, S. M. (2009). The evolution of the understanding of sepsis, infection, and the host response: A brief history. *Critical Care Clinics*, 25(4), 637–663. https://doi.org/10.1016/j.ccc.2009.08.007

Owens, P. L., Miller, M. A., Barrett, M. L., & Hensche, M. (2024). Overview of outcomes for inpatient stays involving sepsis, 2016–2021 (HCUP Statistical Brief #306). *Agency for Healthcare Research and Quality*. https://hcup-us.ahrq.gov/reports/statbriefs/sb306-overview-sepsis-2016-2021.pdf

Papp, M., & Solymosi, N. (2022). Review and comparison of antimicrobial resistance gene databases. *Antibiotics*, 11(3), 339. https://doi.org/10.3390/antibiotics11030339

Pitout, J. D. D. (2012). Extraintestinal Pathogenic Escherichia coli: A Combination of Virulence with Antibiotic Resistance. *Frontiers in Microbiology*, 3, 9. https://doi.org/10.3389/fmicb.2012.00009

Pugh, J. (2023). The current state of nanopore sequencing. *Methods in Molecular Biology*, 2632, 3–14. https://doi.org/10.1007/978-1-0716-2996-3_1

Qiu, X., Lei, Y., & Zhou, R. (2023). SIRS, SOFA, qSOFA, and NEWS in the diagnosis of sepsis and prediction of adverse outcomes: a systematic review and meta-analysis. *Expert Review of Anti-infective Therapy*, 21(8), 891–900. https://doi.org/10.1080/14787210.2023.2237192

Rand, K. H., Beal, S. G., Rivera, K., Allen, B., Payton, T., & Lipori, G. P. (2019). Hourly effect of pretreatment with IV antibiotics on blood culture positivity rate in emergency department patients. *Open Forum Infectious Diseases*, 6(5), ofz179. https://doi.org/10.1093/ofid/ofz179

Rasko, D. A., Rosovitz, M. J., Myers, G. S. A., Mongodin, E. F., Fricke, W. F., Gajer, P., Crabtree, J., Sebaihia, M., Thomson, N. R., Chaudhuri, R., Henderson, I. R., Sperandio, V., & Ravel, J. (2008). The Pangenome Structure of Escherichia coli: Comparative Genomic Analysis of E. Coli Commensal and Pathogenic Isolates. *Journal of Bacteriology*, 190(20), 6881–6893. https://doi.org/10.1128/jb.00619-08

Rello, J., Valenzuela-Sánchez, F., Ruiz-Rodriguez, M., & Moyano, S. (2017). Sepsis: A Review of Advances in Management. *Advances in Therapy*, 34(11), 2393–2411. https://doi.org/10.1007/s12325-017-0622-8

Rhee, C., & Klompas, M. (2020). Sepsis trends: increasing incidence and decreasing mortality, or changing denominator? *Journal of Thoracic Disease*, 12(S1), S89–S100. https://doi.org/10.21037/jtd.2019.12.51

Rice, E. S., & Green, R. E. (2018). New approaches for genome assembly and scaffolding. *Annual Review of Animal Biosciences*, 7(1), 17–40. https://doi.org/10.1146/annurev-animal-020518-115344

Rodriguez-R, L. M., Rosselló-Móra, R., & Konstantinidis, K. T. (2020). Identification and Classification of Prokaryotes using whole-genome sequences. *In CABI eBooks* (pp. 217–230). https://doi.org/10.1079/9781789244984.0217

Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R. P., Banday, S., Mishra, A. K., Das, G., & Malonia, S. K. (2023). Next-Generation Sequencing Technology: Current trends and advancements. *Biology*, 12(7), 997. https://doi.org/10.3390/biology12070997

Saxenborn, P., Baxter, J., Tilevik, A., Fagerlind, M., Dyrkell, F., Pernestig, A., 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, 640408. https://doi.org/10.3389/fmicb.2021.640408

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, 19(R2), R227–R240. https://doi.org/10.1093/hmg/ddq416

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J., Coopersmith, C. M., Hotchkiss, R. S., Levy, M. M., Marshall, J. C., Martin, G. S., Opal, S. M., Rubenfeld, G. D., Van Der Poll, T., Vincent, J., & Angus, D. C. (2016). The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*, 315(8), 801. https://doi.org/10.1001/jama.2016.0287

Tabak, Y. P., Vankeepuram, L., Ye, G., Jeffers, K., Gupta, V., & Murray, P. R. (2018). Blood culture turnaround time in U.S. acute care hospitals and implications for laboratory process optimization. *Journal of Clinical Microbiology*, 56(12). https://doi.org/10.1128/jcm.00500-18

Thornval, N. R., Lacy-Roberts, N., Rebelo, A. R., Nilsson, P., Mourão, J., Gibson, C., Hasman, H., & Hendriksen, R. S. (2026). Optimization of nanopore sequencing for surveillance of antimicrobial

resistance in low-resource settings. *Frontiers in Public Health*, 14, 1755877. https://doi.org/10.3389/fpubh.2026.1755877

Vila, J., Sáez-López, E., Johnson, J. R., Römling, U., Dobrindt, U., Cantón, R., Giske, C. G., Naas, T., Carattoli, A., Martínez-Medina, M., Bosch, J., Retamar, P., Rodríguez-Baño, J., Baquero, F., & Soto, S. M. (2016). Escherichia coli: an old friend with new tidings. *FEMS Microbiology Reviews*, 40(4), 437–463. https://doi.org/10.1093/femsre/fuw005

Vincent, J. (2016). The clinical challenge of sepsis identification and Monitoring. *PLoS Medicine*, 13(5), e1002022. https://doi.org/10.1371/journal.pmed.1002022

Volarić, M., Veseljak, D., Mravinac, B., Meštrović, N., & Despot-Slade, E. (2021). Isolation of High Molecular Weight DNA from the Model Beetle Tribolium for Nanopore Sequencing. *Genes,* 12(8), 1114. https://doi.org/10.3390/genes12081114

Wang, P., & Wang, F. (2022). A proposed metric set for evaluation of genome assembly quality. *Trends in Genetics*, 39(3), 175–186. https://doi.org/10.1016/j.tig.2022.10.005

Wassenaar, T. M., & Gaastra, W. (2001). Bacterial virulence: can we draw the line? *FEMS Microbiology Letters*, 201(1), 1–7. https://doi.org/10.1111/j.1574-6968.2001.tb10724.x

World Health Organization. (2020). *Global report on the epidemiology and burden of sepsis: Current evidence, identifying gaps and future directions*. https://www.who.int/publications/i/item/9789240010789

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Computational Biology*, 13(6), e1005595. https://doi.org/10.1371/journal.pcbi.1005595

Wilfinger, W. W., Mackey, K., & Chomczynski, P. (1997). Effect of pH and ionic strength on the spectrophotometric assessment of nucleic acid purity. *BioTechniques*, 22(3), 474–481. https://doi.org/10.2144/97223st01

Wingett, S. W., & Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research*, 7, 1338. https://doi.org/10.12688/f1000research.15931.2

Yang, S., Lu, C., Yao, C., & Hung, C. (2019). To Trim or Not to Trim: Effects of Read Trimming on the De Novo Genome Assembly of a Widespread East Asian Passerine, the Rufous-Capped Babbler (Cyanoderma ruficeps Blyth). *Genes*, 10(10), 737. https://doi.org/10.3390/genes10100737

Žemaitis, L., Palepšienė, R., Juzėnas, S., Alzbutas, G., Burgi, P., Heinis, T., Charmet, J., Suter, S. A., Jost, M., Raišutis, R., Simmel, F., & Galminas, I. (2025). High-performance protocol for ultra-short DNA sequencing using Oxford Nanopore Technology (ONT). *PLoS ONE*, 20(4), e0318040. https://doi.org/10.1371/journal.pone.0318040

Appendix A – Sample identification

Table A1. Correspondence between isolate numbers, SRR accessions, strain numbers, and Nanopore barcodes.

Isolate	SRR accession	Strain number	Barcode
1	SRR11179862	473	1
2	SRR11179861	884	2
3	SRR11179850	1068	3
4	SRR11179848	1422	4
10	SRR11179842	733	5
17	SRR11179854	1164	6
18	SRR11179853	914	7
19	*	*	*
20	SRR11179851	1819	8
21	SRR11179849	1804	9

*Isolate 19 was excluded from sequencing due to insufficient DNA concentration.

Table A2. Nanopore sequencing yield and read statistics for each isolate. It is possible to see the total number of bases sequenced, total number of reads, and their pass percentages.

Isolate	Total bases (Mb)	Total reads (k)	Pass reads (%)
1	257.08	29.10	96.40
2	235.43	28.51	95.80
3	333.67	37.50	95.20
4	85.35	18.72	93.70
10	186.35	28.45	94.10
17	145.87	21.75	93.40
18	323.74	39.64	94.50
19	*	*	*
20	223.56	34.10	96.70
21	105.50	19.95	95.90

*Isolate 19 was excluded from sequencing due to insufficient DNA concentration.

Appendix B – Quality control for DNA extraction and fragment analysis

Table B1. DNA concentration and purity measurements for the ten bacterial isolates.

Isolate	$A_{260/230}$	$A_{260/280}$	Conc. ng/µL
1	1.90	1.99	127
2	1.93	2.05	50.2
3	2.03	2.11	90.8
4	2.01	2.14	341
10	2.17	2.03	264
17	1.97	2.07	75
18	2.05	2.03	111
19	1.93	1.98	10
20	2.16	2.05	38
21	2.13	2.04	108

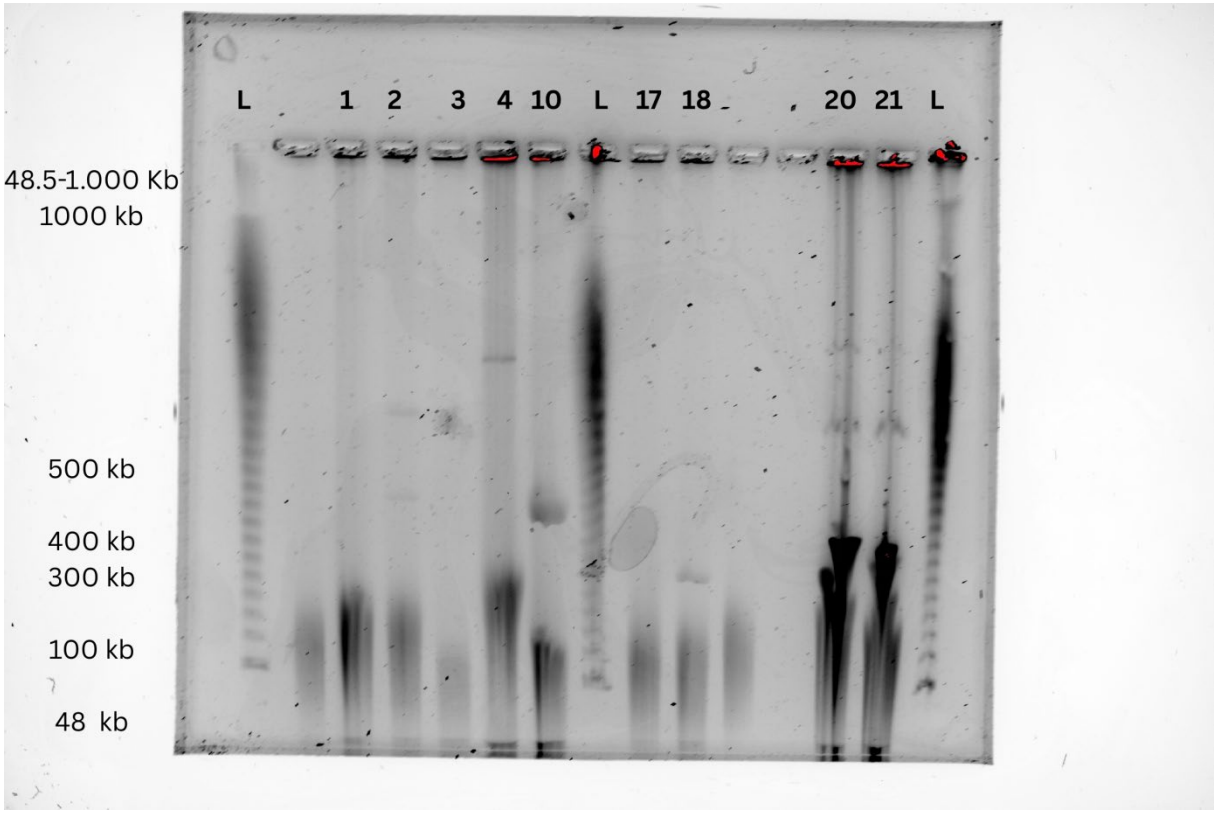

Figure B1. PFGE analysis of high-molecular-weight DNA from selected bacterial isolates. Lanes marked L contains the ladder CHEF DNA Size Standard (48.5–1,000 kb). Sample lanes correspond to isolates 1, 2, 3, 4, 10, 17, 18, 20, and 21. Most isolates showed high-molecular-weight DNA migrating predominantly between approximately 100 and 200 kb, whereas isolates 20 and 21 showed DNA extending to approximately 300 kb. Band intensity and definition varied between samples.

Appendix C – In-house pipeline: commands and parameters

Download raw reads from SRA

```
cd ~/atif_pipeline_comparison

while read SRR
do
  echo "Downloading $SRR"

  prefetch "$SRR" \
    --output-directory sra_download

  fasterq-dump "sra_download/$SRR" \
    --split-files \
    --threads 8 \
    --outdir raw_reads \
    --temp fasterq_tmp

done < srr_accessions.txt
```

FastQC before Trim

```
cd ~/atif_pipeline_comparison


fastqc \
  -t 8 \
  -o original_pipeline/01_fastqc_raw \
  raw_reads/*.fastq \
  2>&1 | tee logs/original_fastqc_raw.log
```

Trimmomatic trimming

```
while read SRR
do
  echo "Running Trimmomatic for $SRR"


  java -jar /usr/share/java/trimmomatic-0.39.jar PE \
    -threads 8 \
    -phred33 \
    raw_reads/${SRR}_1.fastq \
    raw_reads/${SRR}_2.fastq \
    original_pipeline/02_trimmomatic/${SRR}_1_paired.fastq \
```

```
    original_pipeline/02_trimmomatic/${SRR}_1_unpaired.fastq \
    original_pipeline/02_trimmomatic/${SRR}_2_paired.fastq \
    original_pipeline/02_trimmomatic/${SRR}_2_unpaired.fastq \
    ILLUMINACLIP:/usr/share/trimmomatic/NexteraPE-PE.fa:0:30:10 \
    SLIDINGWINDOW:4:20 \
    HEADCROP:12 \
    MINLEN:30 \
    2>&1 | tee logs/${SRR}_trimmomatic.log

done < srr_accessions.txt
```

SPAdes assembly

```
cd ~/atif_pipeline_comparison

while read SRR
do
  echo "Running SPAdes for $SRR"

  spades.py \
    -1 original_pipeline/02_trimmomatic/${SRR}_1_paired.fastq \
    -2 original_pipeline/02_trimmomatic/${SRR}_2_paired.fastq \
    -o original_pipeline/04_spades/${SRR} \
    -t 8 \
    2>&1 | tee logs/${SRR}_spades.log

done < srr_accessions.txt
```

QUAST reference-free assembly QC

```
cd ~/atif_pipeline_comparison

quast.py \
  original_pipeline/04_spades/*/contigs.fasta \
  -o original_pipeline/05_quast/reference_free \
```

-t 8 \

2>&1 | tee logs/original_quast_reference_free.log

CGE ResFinder

Parameter

Tool: ResFinder 4.7.2

Species: *Escherichia coli*

Input: assembled contigs / FASTA

Acquired AMR genes: ON

Point mutations: ON

Minimum identity: 80%

Minimum coverage: 60%

Disinfectant genes: OFF

Unknown mutations: OFF

CGE VirulenceFinder

Tool: VirulenceFinder 3.2.1

Input: assembled contigs / FASTA

Species/database: *Escherichia coli*

Minimum identity: 90%

Minimum coverage: 60%

CGE PlasmidFinder

Tool: PlasmidFinder 3.0.3

Database: Enterobacteriaceae

Input: assembled contigs / FASTA

Minimum identity: 95%

Minimum coverage: 60%

CGE MLST

Tool: MLSTFinder 1.0.1

Scheme/profile: *Escherichia* / Achtman

Input: assembled contigs / FASTA

Appendix D – Updated pipeline: commands and parameters

Recorded versions:
fastp 1.3.7

Unicycler 0.5.1

SPAdes 4.3.0

FastANI (v1.34)

QUAST 5.0.2

ABRicate 1.0.1 Database updated 21 November 2025

mlst 2.23.0

Full updated pipeline script

```
cd ~/atif_pipeline_comparison

cat > scripts/run_updated_pipeline.sh <<'BASH'
#!/usr/bin/env bash
set -euo pipefail

THREADS=8

PROJECT="$HOME/atif_pipeline_comparison"
RAW="$PROJECT/raw_reads"
OUT="$PROJECT/updated_pipeline"
LOGS="$PROJECT/logs"

REF_UTI89="$PROJECT/references/candidate_ecoli_refs/UTI89_NC_007946.1.fasta"

mkdir -p "$OUT"/{01_fastp,02_unicycler,03_quast,04_fastani,05_abricate,06_mlst}
mkdir -p "$LOGS"

SAMPLES=(
SRR11179842
SRR11179848
```

```
SRR11179849
SRR11179850
SRR11179851
SRR11179853
SRR11179854
SRR11179861
SRR11179862
)

echo "Updated/local pipeline started: $(date)"
echo "Threads: $THREADS"
echo

for SRR in "${SAMPLES[@]}"
do
  echo "================================"
  echo "Sample: $SRR"
  echo "Start: $(date)"
  echo "================================"

  R1="$RAW/${SRR}_1.fastq"
  R2="$RAW/${SRR}_2.fastq"

  FP_R1="$OUT/01_fastp/${SRR}_R1.fastq.gz"
  FP_R2="$OUT/01_fastp/${SRR}_R2.fastq.gz"

  ASSEMBLY_DIR="$OUT/02_unicycler/${SRR}"
  ASSEMBLY="$ASSEMBLY_DIR/assembly.fasta"

  echo "[1/6] fastp trimming/QC: $SRR"
  conda run -n improved_assembly fastp \
```

```
    --in1 "$R1" \
    --in2 "$R2" \
    --out1 "$FP_R1" \
    --out2 "$FP_R2" \
    --detect_adapter_for_pe \
    --qualified_quality_phred 20 \
    --length_required 30 \
    --thread "$THREADS" \
    --html "$OUT/01_fastp/${SRR}_fastp.html" \
    --json "$OUT/01_fastp/${SRR}_fastp.json" \
    > "$LOGS/${SRR}_fastp.log" 2>&1

echo "[2/6] Unicycler assembly: $SRR"
mkdir -p "$ASSEMBLY_DIR"
conda run -n improved_assembly unicycler \
    -1 "$FP_R1" \
    -2 "$FP_R2" \
    -o "$ASSEMBLY_DIR" \
    -t "$THREADS" \
    > "$LOGS/${SRR}_unicycler.log" 2>&1

if [ ! -s "$ASSEMBLY" ]; then
    echo "ERROR: assembly not found or empty for $SRR: $ASSEMBLY"
    exit 1
fi

echo "[3/6] QUAST: $SRR"
conda run -n skovde_quast5 quast.py \
    "$ASSEMBLY" \
    -o "$OUT/03_quast/${SRR}" \
    -t "$THREADS" \
```

```
    > "$LOGS/${SRR}_quast.log" 2>&1

  echo "[4/6] FastANI against UTI89: $SRR"
  conda run -n skovde_fastani fastANI \
    -q "$ASSEMBLY" \
    -r "$REF_UTI89" \
    -o "$OUT/04_fastani/${SRR}_vs_UTI89.tsv" \
    > "$LOGS/${SRR}_fastani.log" 2>&1 || true

  echo "[5/6] ABRicate databases: $SRR"
  for DB in resfinder plasmidfinder vfdb ecoli_vf
  do
    conda run -n cge abricate \
      --db "$DB" \
      "$ASSEMBLY" \
      > "$OUT/05_abricate/${SRR}_${DB}.tsv"
  done

  echo "[6/6] MLST: $SRR"
  conda run -n cge mlst "$ASSEMBLY" \
    > "$OUT/06_mlst/${SRR}_mlst.tsv" || true

  echo "Finished sample $SRR: $(date)"
  echo
done

echo "Updated/local pipeline finished: $(date)"
BASH

chmod +x scripts/run_updated_pipeline.sh
```

Appendix E– In-house Illumina pipeline results

Table E1. FastQC module status before and after Trimmomatic processing of Illumina reads.

FastQC module	Raw	After Trimmomatic
Adapter Content	18 FAIL	18 PASS
Basic Statistics	18 PASS	18 PASS
Overrepresented sequences	13 PASS, 5 WARN	18 PASS
Per base N content	18 PASS	18 PASS
Per base sequence content	18 WARN	18 PASS
Per base sequence quality	18 PASS	18 PASS
Per sequence GC content	17 WARN, 1 FAIL	16 WARN, 2 FAIL
Per sequence quality scores	18 PASS	18 PASS
Sequence Duplication Levels	6 PASS, 7 WARN, 5 FAIL	6 PASS, 10 WARN, 2 FAIL
Sequence Length Distribution	18 PASS	18 WARN

Table E2. Reference-free QUAST assembly metrics for the in-house Illumina pipeline.

Sample	Contigs	Largest contig (bp)	Total length (bp)	GC%	N50 (bp)	L50
SRR11179842	146	394,231	5,105,418	50.28	160,137	11
SRR11179848	186	375,960	5,163,879	50.77	137,217	12
SRR11179849	1,787	39,831	3,984,975	50.82	3,766	249
SRR11179850	2,681	11,361	4,061,200	51.15	1,865	652
SRR11179851	106	386,746	4,789,179	50.71	182,606	11
SRR11179853	100	432,417	4,924,170	50.58	217,944	9
SRR11179854	1,019	123,558	4,814,030	50.43	10,689	78
SRR11179861	153	373,448	4,725,537	50.71	98,704	16
SRR11179862	128	468,756	5,028,189	50.30	184,902	9

Table E3. JSpeciesWS ANIb values for the nine isolates against the *E. coli* UTI89 reference genome.

Sample	JSpeciesWS ANIb vs UTI89 (%)
SRR11179842	98.80
SRR11179848	98.09
SRR11179849	96.50
SRR11179850	96.40
SRR11179851	96.45
SRR11179853	98.15
SRR11179854	98.76
SRR11179861	96.37
SRR11179862	98.68

Table E4. Reference-based QUAST metrics for in-house Illumina assemblies against *E. coli* UTI89.

Sample	Genome fraction Original %	Misassemblies Original

SRR11179842	91.583	39
SRR11179848	83.770	70
SRR11179849	61.193	57
SRR11179850	62.385	40
SRR11179851	75.620	137
SRR11179853	85.321	71
SRR11179854	87.787	43
SRR11179861	74.250	118
SRR11179862	91.144	50

Table E5. Virulence-associated genes detected using VirulenceFinder in the in-house pipeline.

Sample	Virulence genes
SRR11179842	56
SRR11179848	59
SRR11179849	43
SRR11179850	43
SRR11179851	38
SRR11179853	51
SRR11179854	50
SRR11179861	47
SRR11179862	54

Table E6. Acquired antimicrobial-resistance genes and resistance-associated point mutations detected using ResFinder and PointFinder.

Sample	Unique AMR genes	Point mutations
SRR11179842	0	0
SRR11179848	7	5
SRR11179849	17	3
SRR11179850	14	3
SRR11179851	6	4
SRR11179853	0	1
SRR11179854	0	0
SRR11179861	8	0
SRR11179862	1	1

Table E7. Plasmid replicons detected using PlasmidFinder in the in-house pipeline.

Sample	Plasmid replicons
SRR11179842	3
SRR11179848	7
SRR11179849	4
SRR11179850	6
SRR11179851	6
SRR11179853	5

SRR11179854	3
SRR11179861	4
SRR11179862	3

Table E8. MLST allele profiles and sequence-type assignments for the in-house Illumina assemblies.

Sample	ST	adk	fumC	gyrB	icd	mdh	purA	recA
SRR11179842	80	13	24	19	14	23	1	10
SRR11179848	131	53	40	47	13	36	28	29
SRR11179849	-	-	-	14	1679?	~9	8	14
SRR11179850	-	-	-	-	-	-	-	-
SRR11179851	2448	6	23	14	18	9	8	14
SRR11179853	131	53	40	47	13	36	28	29
SRR11179854	127	13	14	19	36	23	11	10
SRR11179861	10	10	11	4	8	8	8	2
SRR11179862	127	13	14	19	36	23	11	10

- = No allele number was assigned

~ = Uncertain match / low confident

? = Partial match

Appendix F– Updated pipeline

Table F1. FastQC module status before and after fastp processing of Illumina reads.

FastQC module	Raw	After fastp
Adapter Content	18 FAIL	18 PASS
Basic Statistics	18 PASS	18 PASS
Overrepresented sequences	13 PASS, 5 WARN	18 PASS
Per base N content	18 PASS	18 PASS
Per base sequence content	18 WARN	18 WARN
Per base sequence quality	18 PASS	18 PASS
Per sequence GC content	17 WARN, 1 FAIL	16 WARN, 2 FAIL
Per sequence quality scores	18 PASS	18 PASS
Sequence Duplication Levels	6 PASS, 7 WARN, 5 FAIL	6 PASS, 7 WARN, 5 FAIL
Sequence Length Distribution	18 PASS	18 WARN

Table F2. Reference-free QUAST assembly metrics for the updated Illumina pipeline.

Sample	Contigs	Total length (bp)	Largest contig (bp)	GC%	N50 (bp)	L50
SRR11179842	102	5,111,719	464,699	50.30	241,830	7
SRR11179848	102	5,135,598	399,805	50.84	190,808	9
SRR11179849	802	3,751,853	65,890	51.02	7,916	114
SRR11179850	569	4,636,391	126,709	50.98	14,908	82
SRR11179851	64	4,755,315	652,082	50.73	198,106	7
SRR11179853	50	4,875,152	666,998	50.67	307,935	5
SRR11179854	409	4,763,185	251,581	50.45	35,936	30
SRR11179861	97	4,680,526	377,152	50.79	133,852	11
SRR11179862	63	4,995,520	738,322	50.36	262,451	6

Table F3. FastANI values for the updated Illumina assemblies against *E. coli* UTI89.

Sample	FastANI vs UTI89 (%)
SRR11179842	98.8681
SRR11179848	98.3698
SRR11179849	96.4309
SRR11179850	96.8638
SRR11179851	96.6890
SRR11179853	98.5114
SRR11179854	98.9457
SRR11179861	96.7281
SRR11179862	98.9169

Table F4. Reference-based QUAST metrics for updated Illumina assemblies against *E. coli* UTI89.

Sample	N50 (bp)	Genome fraction %	Misassemblies Updated
SRR11179842	241,830	91.519	48

SRR11179848	190,808	83.756	76
SRR11179849	7,916	58.516	68
SRR11179850	14,908	73.014	82
SRR11179851	198,106	75.557	141
SRR11179853	307,935	85.245	72
SRR11179854	35,936	87.311	48
SRR11179861	133,852	73.938	126
SRR11179862	262,451	90.965	53

Table F5. Antimicrobial-resistance gene hits detected using ABRicate with the ResFinder database.

Sample	ResFinder AMR genes
SRR11179842	1
SRR11179848	8
SRR11179849	7
SRR11179850	15
SRR11179851	7
SRR11179853	1
SRR11179854	1
SRR11179861	9
SRR11179862	2

Table F6. Plasmid replicons detected using ABRicate with the PlasmidFinder database.

Sample	PlasmidFinder replicons
SRR11179842	0
SRR11179848	3
SRR11179849	2
SRR11179850	5
SRR11179851	1
SRR11179853	1
SRR11179854	0
SRR11179861	2
SRR11179862	0

Table F7. Virulence-associated gene hits detected using ABRicate with the VFDB and ecoli_vf databases.

Sample	VFDB genes	ecoli_vf genes
SRR11179842	87	222
SRR11179848	77	189
SRR11179849	38	130
SRR11179850	55	165
SRR11179851	46	184
SRR11179853	73	193

SRR11179854	90	244
SRR11179861	21	130
SRR11179862	95	248

Table F8. MLST allele profiles and sequence-type assignments for the updated Illumina assemblies.

Sample	ST	adk	fumC	gyrB	icd	mdh	purA	recA
SRR11179842	80	13	24	19	14	23	1	10
SRR11179848	131	53	40	47	13	36	28	29
SRR11179849	-	6	117?	14	1679?	~9	8	14
SRR11179850	10	10	11	4	8	8	8	2
SRR11179851	2448	6	23	14	18	9	8	14
SRR11179853	131	53	40	47	13	36	28	29
SRR11179854	127	13	14	19	36	23	11	10
SRR11179861	10	10	11	4	8	8	8	2
SRR11179862	127	13	14	19	36	23	11	10

- = No allele number was assigned

~ = Uncertain match

? = Partial match

Appendix G– Comparison of the in-house and updated Illumina pipelines

Table G1. Comparison of plasmid replicon counts between the in-house and updated Illumina pipelines.

Sample	Original CGE PlasmidFinder replicon count	Updated ABRicate PlasmidFinder replicon count
SRR11179842	3	0
SRR11179848	7	3
SRR11179849	4	2
SRR11179850	6	5
SRR11179851	6	1
SRR11179853	5	1
SRR11179854	3	0
SRR11179861	4	2
SRR11179862	3	0

Appendix H – Nanopore-only assembly.

Table H1. QUAST assembly metrics for Nanopore-only genome assemblies.

Sample	Total Length (bp)	Contigs	N50 (bp)	L50	GC (%)
SRR11179862	4,229,917	138	56,292	25	50.44
SRR11179861	5,091,736	2	5,088,210	1	50.40
SRR11179850	677,102	48	17,869	13	51.38
SRR11179848	N/A	N/A	N/A	N/A	N/A
SRR11179842	4,764,466	106	113,286	12	50.38
SRR11179854	2,197,088	106	38,592	19	50.61
SRR11179853	4,898,262	43	294,124	5	50.66
SRR11179851	4,780,344	3	4,674,021	1	50.75
SRR11179849	4,830,648	4	4,675,276	1	50.79

*N/A = Not available.

Appendix I – Comparison of updated Illumina-only and hybrid assemblies

Table I1. Comparison of MLST sequence-type assignments between updated Illumina-only and hybrid assemblies.

Sample	Illumina ST	Hybrid ST
SRR11179842	80	80
SRR11179848	131	131
SRR11179849	N/A	N/A
SRR11179850	10	10
SRR11179851	2448	2448
SRR11179853	131	131
SRR11179854	127	127
SRR11179861	10	10
SRR11179862	127	127

N/A= No ST assigned.

Table I2. Comparison of virulence-associated gene counts between updated Illumina-only and hybrid assemblies.

Sample	Illumina	Hybrid
SRR11179842	87	89
SRR11179848	77	81
SRR11179849	38	50
SRR11179850	55	60
SRR11179851	46	50
SRR11179853	73	77
SRR11179854	90	98
SRR11179861	21	26
SRR11179862	95	100

Table I3. Comparison of antimicrobial-resistance gene counts between updated Illumina-only and hybrid assemblies.

Sample	Illumina AMR genes	Hybrid AMR genes
SRR11179842	1	1
SRR11179848	8	8
SRR11179849	7	7
SRR11179850	15	17
SRR11179851	7	7
SRR11179853	1	1
SRR11179854	1	1
SRR11179861	9	9
SRR11179862	2	2

Appendix J- Hybrid assembly results

Table J1. QUAST assembly metrics for hybrid genome assemblies.

Sample	Total Length (bp)	N50 (bp)	L50	GC (%)
SRR11179842	5,206,387	1,104,353	2	50.30
SRR11179848	5,137,108	190,808	9	50.84
SRR11179849	4,841,720	823,610	3	50.74
SRR11179850	4,802,730	81,225	18	50.91
SRR11179851	4,830,327	4,668,311	1	50.77
SRR11179853	4,916,734	2,018,938	2	50.68
SRR11179854	4,956,401	131,918	9	50.48
SRR11179861	4,779,132	242,038	7	50.74
SRR11179862	5,073,754	1,628,182	2	50.40

Table J2. MLST allele profiles and sequence-type assignments for hybrid genome assemblies.

Sample	ST	adk	fumC	gyrB	icd	mdh	purA	recA
SRR11179842	80	13	24	19	14	23	1	10
SRR11179848	131	53	40	47	13	36	28	29
SRR11179849	-	6	117?	14	1679?	~9	8	14
SRR11179850	10	10	11	4	8	8	8	2
SRR11179851	2448	6	23	14	18	9	8	14
SRR11179853	131	53	40	47	13	36	28	29
SRR11179854	127	13	14	19	36	23	11	10
SRR11179861	10	10	11	4	8	8	8	2
SRR11179862	127	13	14	19	36	23	11	10

- = No allele number was assigned

~ = Uncertain match / low confident

? = Partial match

Table J3. Virulence-associated gene hits detected using ABRicate with the VFDB and ecoli_vf databases.

Sample	VFDB	*Ecoli_vf*
SRR11179842	89	265
SRR11179848	81	201
SRR11179849	50	130
SRR11179850	60	180
SRR11179851	50	198
SRR11179853	77	207
SRR11179854	98	276
SRR11179861	26	142
SRR11179862	100	265

Table J4. AMR gene hits detected in hybrid assemblies using ABRicate with the ResFinder database.

Sample	AMR genes
SRR11179842	1
SRR11179848	8

SRR11179849	7
SRR11179850	17
SRR11179851	7
SRR11179853	1
SRR11179854	1
SRR11179861	9
SRR11179862	2

Appendix K– Statistical analysis

Table K1. Statistical comparison of assembly metrics between the in-house and updated Illumina pipelines.

Metric	Shapiro–Wilk test	Comparison test	p-value	Significant?
Total length	p = 0.0002377	Wilcoxon signed-rank	0.1551	No
L50	p = 1.567×10^{-5}	Wilcoxon signed-rank	0.00885	Yes
Number of contigs	p = 0.0006284	Wilcoxon signed-rank	0.009152	Yes