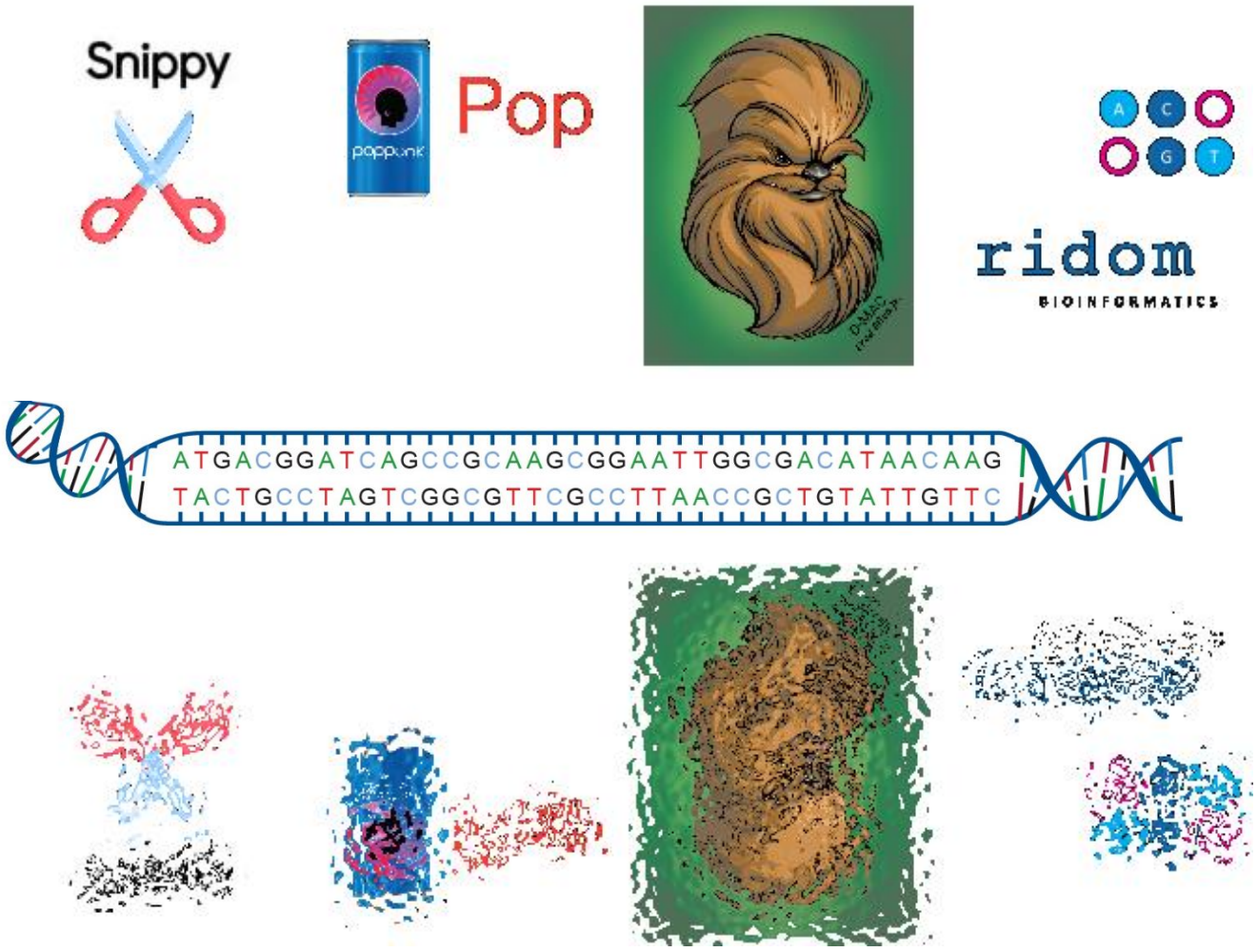
MICROVIDA
MEDISCHE MICROBIOLOGIE & IMMUNOLOGIE
BRABANT & ZEELAND

Whole genome typing methods for differentiating between epidemiologically related and unrelated *E. coli* ST131 isolates



Freek de Kreek[1, 2], Joep J. M. Stohr[2], Jaco J. Verweij[2], Suzan Pas[2], Eefje Schrauwen[1, 3]

1 Academy of Life Sciences & Technology, Avans University of Applied Sciences, Breda, Noord-Brabant, Netherlands
2 Microvida, Tilburg, Noord-Brabant, Netherlands
3 Lectorate Analysistechniques in Life Sciences, Avans University of Applied Sciences, Breda, Noord-Brabant, Netherlands

avans hogeschool

Abstract

Introduction

Current typing methods have difficulty to distinguish epidemiologically related and unrelated E. coli ST131 samples. In this study, whole genome typing methods Ridom SeqSphere, chewBBACA, PopPUNK and Snippy were compared for their ability to distinguish related from unrelated *E. coli* ST131 samples and form clusters based on cutoffs determined in this study.

Methods

Samples collected from the same patient within 3 months were defined as related and unrelated when collected from different institutions. Samples were sequenced with Illumina and typed using the stable *E. coli* wgMLST scheme in Ridom Seqsphere, a specific *E. coli* ST131 wgMLST scheme in Ridom SeqSphere and chewBBACA, PopPUNK, and Snippy. The medians of genetic distances were compared using the Mann-Whitney U test and the correlation between the typing methods was calculated using the Spearman Rank correlation. The cutoff threshold and percentile misclassifications were determined at three cutoffs: the lowest cutoff where all epidemiologically related were genetically classified as related (cutoff 1), the highest cutoff where all epidemiologically unrelated were classified as genetically unrelated (cutoff 2) and the maximum accuracy (cutoff 3). The percentile misclassifications as either related or unrelated were calculated for cutoffs 1 and 2 respectively, and both misclassifications for cutoff 3. The cutoffs were then applied to a real-world dataset and clusters were evaluated using the Simpson's diversity index and Silhouette score.

Results

All typing methods showed a significant genetic distance difference between related and unrelated samples. The percentage of misclassifications at cutoff 1 was between 35.9% - 47.1% for the wgMLST schemes methods, 18.3% for PopPUNK and 3.3% for Snippy. Cutoffs 2 & 3 showed a lower percentage of misclassifications, 20.0% and 16.0% for chewBBACA and 8.0% for the remaining typing methods. Cutoff 1 had a poor performance compared to cutoff 2, which formed well-defined clusters, as shown by a high Simpson's diversity index and Silhouette score greater than 0.85.

Conclusion

Contrary to expectation, the typing methods are capable of distinguishing strictly defined related from unrelated samples. Cutoffs 2 and 3 seem to be best suited for clustering and the typing methods have a better clustering performance than using a stable *E. coli* scheme in Ridom SeqSphere. It is recommended to use Snippy, which had the best performance in misclassifications and the highest Silhouette score at cutoffs 2 & 3. Although the clusters require to be validated based on epidemiological data. For future studies, the adjusted Rand index can be used to evaluate the clusters based on epidemiological data.

Samenvatting

Introductie

De huidige typeringsmethode heeft moeite om epidemiologisch verwante en niet-verwante *E. col*i ST131 te onderscheiden. In deze studie werden de hele genoom typeermethoden Ridom SeqSphere, chewBBACA, PopPUNK en Snippy vergeleken op hun vermogen om verwante van niet-gerelateerde *E. coli* ST131 te onderscheiden en clusters te vormen op basis van cutoffs bepaald in deze studie.

Methoden

Samples die binnen 3 maanden bij dezelfde patiënt werden verzameld, werden gedefinieerd als verwant en niet-verwant wanneer ze bij verschillende instellingen werden verzameld. Samples werden gesequencet met Illumina en getypeerd met behulp van het stabiele *E. coli* wgMLST-schema in Ridom Seqsphere, een specifiek *E. coli* ST131 wgMLST-schema in Ridom SeqSphere en in chewBBACA, PopPUNK en Snippy. De medianen van genetische afstanden werden vergeleken met behulp van de Mann-Whitney U-test en de correlatie tussen de typeringsmethoden werd berekend met behulp van de Spearman's rank correlatie. De cutoffs en het percentage misclassificaties werden bepaald op drie cutoffs: de laagste afkapwaarde waarbij alle epidemiologisch verwante samples genetisch werden geclassificeerd als verwant (cutoff 1), de hoogste afkapwaarde waarbij alle epidemiologisch niet-verwante samples werden geclassificeerd als genetisch niet-verwant (cutoff 2) en bij de maximale nauwkeurigheid (cutoff 3). Het percentage misclassificaties als oftewel verwant of niet-verwant werd berekend voor cutoff 1 en 2, en beide misclassificaties voor cutoff 3. De cutoffs werden vervolgens toegepast op een reële dataset en clusters werden geëvalueerd met behulp van de Simpson's diversity index en Silhouette score.

Resultaten

Alle typeringsmethoden toonden een significant genetisch verschil tussen verwante en niet-verwante samples. Het percentage misclassificaties bij cutoff 1 was 35.9% - 47.1% voor de wgMLST-schema's methoden, 18.3% voor PopPUNK en 3,3% voor Snippy. Cutoff 2 & 3 toonde een lager percentage misclassificaties, 20,0% en 16,0% voor chewBBACA en 8,0% voor de overige typemethoden. Cutoff 1 presteerde slecht in vergelijking met cutoff 2, dat goed gedefinieerde clusters vormde, zoals blijkt uit een hoge Simpson's diversiteitsindex en Silhouette-score van meer dan 0,85.

Conclusie

Tegen de verwachting in zijn de typemethoden in staat strikt gedefinieerde verwante van niet-verwante monsters te onderscheiden. Cutoffs 2 en 3 lijken het meest geschikt voor clustering en de typemethoden hebben betere clusterprestaties dan het gebruik van een stabiel *E. coli*-schema in Ridom SeqSphere. Het wordt aanbevolen om Snippy te gebruiken, die de beste prestatie had in misclassificaties en de hoogste Silhouette-score op cutoff 2 en 3. Hoewel de clusters moeten worden gevalideerd op basis van epidemiologische gegevens. Voor toekomstig onderzoek kan de adjusted Rand-index worden gebruikt om de clusters te evalueren op basis van epidemiologische gegevens.

Table of contents

Introduction

The increasing prevalence of antimicrobial resistance driven by misuse of antibiotics in humans and animals makes treatment of bacterial infections troublesome and in some cases, no treatment is available, which poses a threat to public health. A prominent example of increasing antimicrobial resistance is observed in one of the most frequently occurring bacterial species in humans; *Escherichia coli,* a Gram-negative bacterium commonly found in the human gastrointestinal tract. Over the years, pathogenic *E. coli* strains have acquired multi-drug resistance through the acquisition of mobile genetic elements (MGEs).

For effective infection prevention and control, it is important to monitor multi-drug resistant microorganisms (MDRM). A typing method such as Multi Locus Sequence Typing (MLST) characterizes a bacterial species based on seven housekeeping genes, but this method lacks the distinctive power to determine the genetic relatedness between two samples compared to whole genome typing methods. The principle of whole genome typing methods rests on using Whole Genome Sequencing (WGS) and bioinformatic analysis, enabling comparisons of genomes on single genetic events (Sabat et al, 2013). A commonly used whole genome typing method is gene-by-gene typing based on a whole genome Multi Locus Sequence Typing (wgMLST) scheme, the genetic comparisons are subsequently clustered based on a predetermined cutoff. In a previous study, a wgMLST scheme was created for *E. coli* and a cutoff for differentiating between epidemiological-related (samples obtained from the same patient) and unrelated (samples obtained with no link between patients) was determined for *E. coli* (Kluytmans et al, 2016). However, the clonal group *E. coli* sequence type 131 is the most common extraintestinal pathogenetic *E. coli* worldwide and is primarily associated with urinary tract and bloodstream infections. (Poolman & Wacker, 2016) This clonal group frequently produces extended-spectrum β-lactamases (ESBL) and has relatively little variation in their genome content (Clark et al, 2012). The conserved genome content makes distinguishing related and unrelated unreliable with the current typing method for *E. coli*, requiring increased discriminatory power.

The current typing method used at Microvida is based on a standard *E. coli c*gMLST scheme in Ridom SeqSphere. This method lacks sufficient discriminatory power to distinguish epidemiologically related genomes from unrelated *E. coli* ST131 genomes. This study aims to evaluate which typing tool is best suited for discriminating epidemiologically related samples from epidemiologically unrelated samples of bacterial genomes. The whole genome typing methods to be evaluated in this study are gene-by-gene typing using a wgMLST scheme (Ridom SeqSphere & chewBBACA), Single Nucleotide Polyphormisms (SNP) typing (Snippy) and variable *k*-mer sequence matching (PopPUNK). The study is divided into two parts: first, the capacity of discriminating strictly defined related samples from unrelated samples of the typing methods was evaluated at three pre-defined cutoffs. Secondly, the discriminatory power of a real-world dataset was analyzed for the previously defined cutoffs by evaluating the clusters that were formed based on the cutoffs.

The two whole genome typing methods based on SNPs and variable *k*-mer sequence matching were expected to have the highest discriminatory power. SNP analysis compares single genomic events on shared sequences, providing a higher resolution compared to wgMLST which uses a comparison of entire genes. SNP analysis and wgMLST are limited to either shared genomic positions or genes, but variable *k*-mer sequence matching has an increased resolution due to its capability of comparing the entire genome. Ideally, a typing method should have zero false positives (i.e., classify no unrelated cases as related) when using a cutoff that includes all related samples and clusters should be close-knit and well-separated from each other.

Theoretical background

Whole genome sequencing

The nucleotide sequence of nucleic acid chains like DNA and RNA contains important information on the biochemical and inheritable properties of life. The DNA sequence consists of four nucleotides, cytosine (C), guanine (G), adenine (A) and thymine (T). This string of letters forms the genome, which contains genes that in turn form a template for the production of proteins. To uncover the genetic information, DNA is extracted from the cell and subsequently sequenced using a sequencing platform like Sanger, Illumina, IonTorrent or Oxford Nanopore. WGS sequences the entire genome of an organism by fragmenting the genomic DNA, sequencing these shorter fragments before assembling the overlapping fragments (*de novo* assembly) or by aligning the fragments to a reference genome to form a complete genome. (Heather & Chain, 2016)

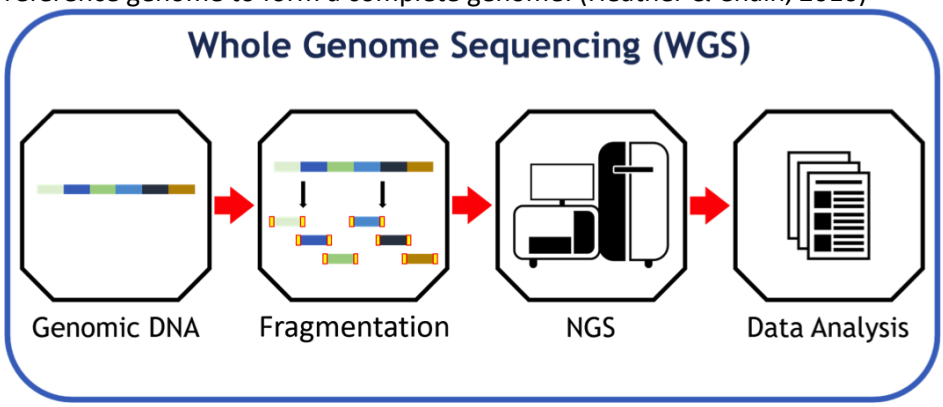


Figure 1. Whole genome sequencing in four steps: collection of genomic DNA, fragmentation of genomic DNA, sequencing of the shorter fragments and finally analysis of sequencing data. (Biobasic, 2021)

Illumina sequencing

Illumina sequencing is a second-generation sequencing technology, it has a high throughput, short reads (50 – 500 base pairs) and a relatively high base calling accuracy. The principle of Illumina sequencing rests on sequencing by synthesis technology. Four types of fluorescently labeled nucleotides are added to a flow cell, the fluorescent labels emit a signal that can be used to 'read' the sequence of a DNA fragment. However, a single strand does not produce enough signal for sequencing, which is why bridge amplification is needed first to obtain more strands (see figure 2).

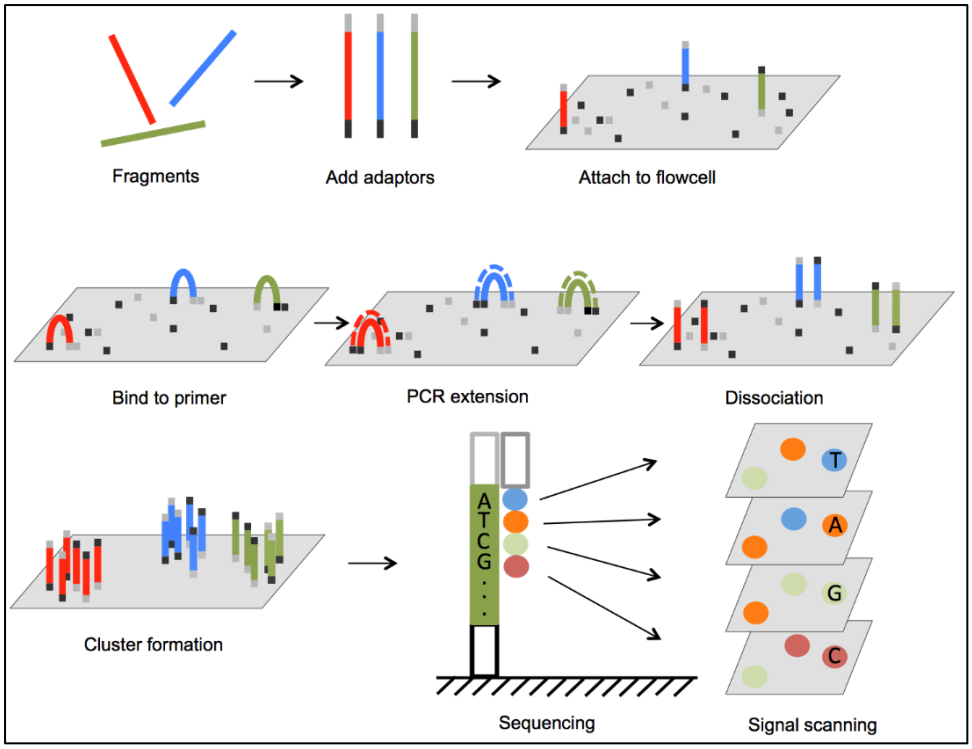


Figure 2. Sequencing by synthesis overview Illumina. Adapters are added to fragments and amplified by bridge amplification on the flow cell. The PCR fragments are then sequenced with signal scanning. (Lu et al, 2016)

Genomic DNA is fragmented into fragments of 200-500 base pairs in length, 5' and 3' adapter is added to the ends of these fragments. The adapters allow the DNA fragments to attach to the surface of the flow cell, bridge amplification is performed to form bundles of copied DNA fragments (clusters). Clusters of the same DNA fragment increase the intensity of the signal to ensure sequencing is possible.

Sequencing primers, fluorescently labeled nucleotides and DNA polymerase are added to sequence the clusters of DNA fragments after amplification. The sequencing primers anneal to the DNA fragments of the cluster and DNA polymerase attaches to elongate the fragments. After the first complementary nucleotide is added to the end of the primer, the base-specific fluorescently labeled nucleotide temporarily blocks DNA polymerase from adding any more nucleotides. The fluorescent signal is excited by a laser, after which the signal is recorded and translated to the corresponding nucleotide. DNA polymerase adds the second complementary nucleotide, the nucleotide is base called and the fluorescent nucleotide is removed. The process repeats itself until all sequences are scanned. (Clark, 2019)

Genome assembly

Assembly is the process of constructing the full genome sequence from raw DNA sequence reads produced by a sequencing platform. There are two main approaches for genome assembly, *de novo* and reference-based assembly. *De novo* assembly does not use a reference genome and constructs the genome sequence only based on the sequenced DNA fragments. The downsides to this method are increased sensitivity to sequencing errors and repetitive regions.

There are several assembly algorithms for *de novo* assembly such as the De Bruijn graph, overlap layout consensus and greedy algorithm. De Bruin graph is the commonly used algorithm for assembly based on short reads, it is a *k*-mers-based approach. *k*-mers are strings of a certain length

(*k*) from a DNA sequence. Reads are divided in *k*-mers and overlap between *k*-mers is used to merge overlapping reads and create a longer sequence, a contig. (Khan et al, 2018) Contigs are subsequently combined to form longer sequences called scaffolds, before finally reconstructing the genome sequence. See figure 3 below for an overview of the *de novo* assembly process.

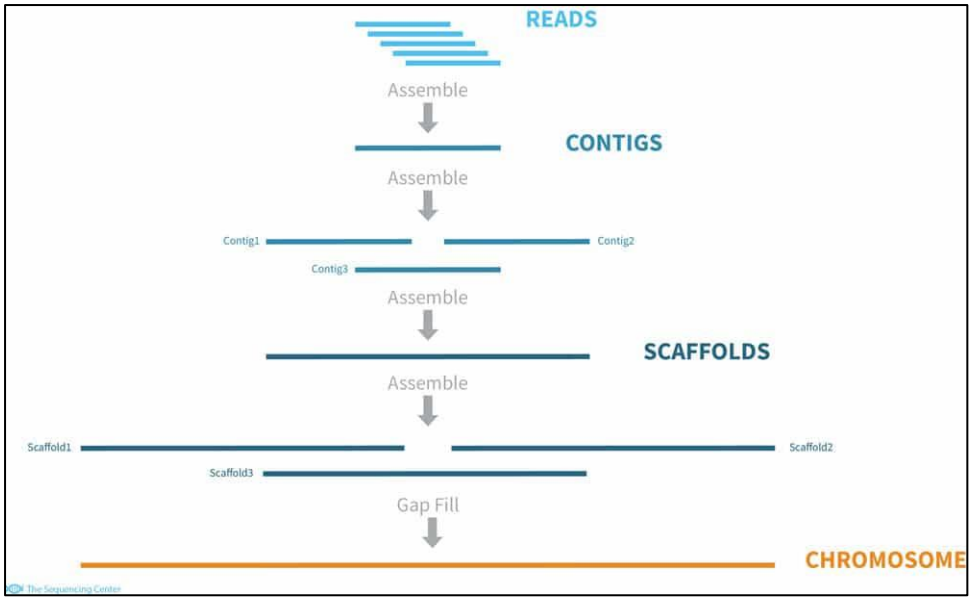


Figure 3, an overview of the de novo assembly process. Overlapping raw sequencing reads are merged into a contig, overlapping contigs are merged into a scaffold. The scaffolds are joined together to form the final draft genome. (The Sequencing Center, 2022)

Reference-based assembly uses a reference genome and maps DNA reads to the reference genome by searching regions of similarity between the reads and the reference genome. A consensus sequence is generated based on comparing the aligned reads, identifying regions of overlap and resolving variations between the reads and the reference genome. This method allows for more efficient assembly and more accuracy in assembling conserved and repetitive regions, but the downside to this method is that divergent gene content compared to the reference assembly is missed and may result in a biased or incomplete assembly when the reference genome is of limited quality and completeness. (Baker, 2012)

Whole genome typing methods

Distinguishing between bacterial isolates plays an important factor in infection control and for outbreaks, to find the source of (person-to-person) transmission in a healthcare setting. Clonally related isolates typically have a significantly higher genotypical and phenotypical likeness than unrelated isolates of the same species. (Belkum, 2007)

A great deal of bacterial typing methods exist, subdivided into phenotyping and genotyping methods. Phenotyping methods are based on biochemical, antigenic or antibiotics susceptibility characteristics of a bacterial species. Genotyping methods are based on the analysis of the genome and plasmids (Tankeshwar, 2021). Whole genome typing methods used in this study are core/whole multilocus sequence typing, core genome SNP-based typing and variable-length *k*-mer comparisons.

Ridom SeqSphere

Ridom SeqSphere (Jünemann et al, 2013) uses whole genome multi-locus sequence typing (wgMLST), a WGS-based gene-by-gene typing method that extends MLST to the genome level (Kluytmans, 2016). Thus, making use of thousands of genes instead of seven reference genes compared to MLST. A wgMLST scheme is usually divided between the core genome and the accessory genome. Genes in the core genome are present in 95% of all strains of a species, genes belong to the accessory genome when the genes are present in less than 95% of all strains of a species. The accuracy and number of alleles in the core and accessory genome are dependent on the variety and number of genomes used to form the wgMLST scheme. (Bruyne et al, 2015)

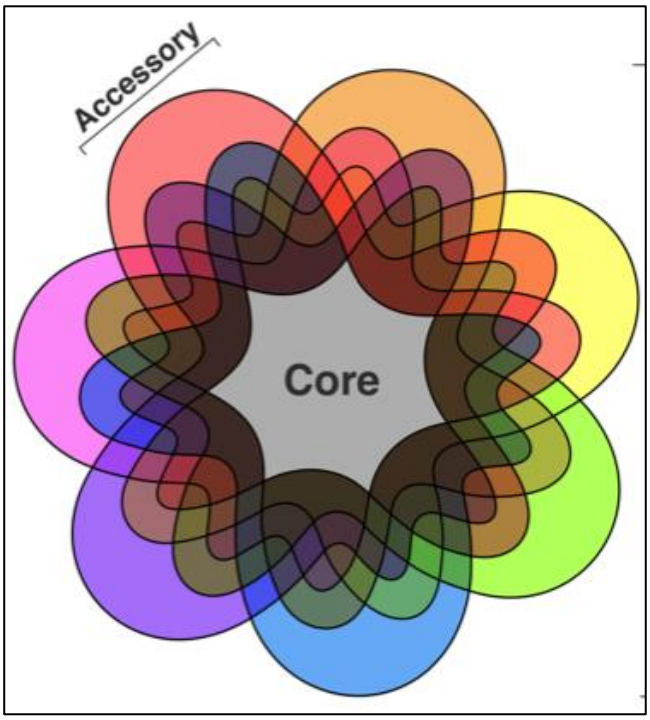


Figure 4. An illustration of the core genome and accessory genome. The core genome presents the genes shared across more than 95% of the members of a species, the accessory genome is made up of genes that are variably shared among members of a species. (McCarthy, 2018)

The principle of wgMLST is based on comparing the assembled DNA sequence to the target alleles in the wgMLST scheme. Differences between the assembled sequence and target alleles in the wgMLST scheme are noted as an allele variant and each allele variant is assigned a numeric allele number. The different allele variants form the allelic profile which is a sequence of these numbers (see figure 5 for an illustration of wgMLST). Differences in genome content can be calculated by comparing the allelic profile of two different genomes of the same species.

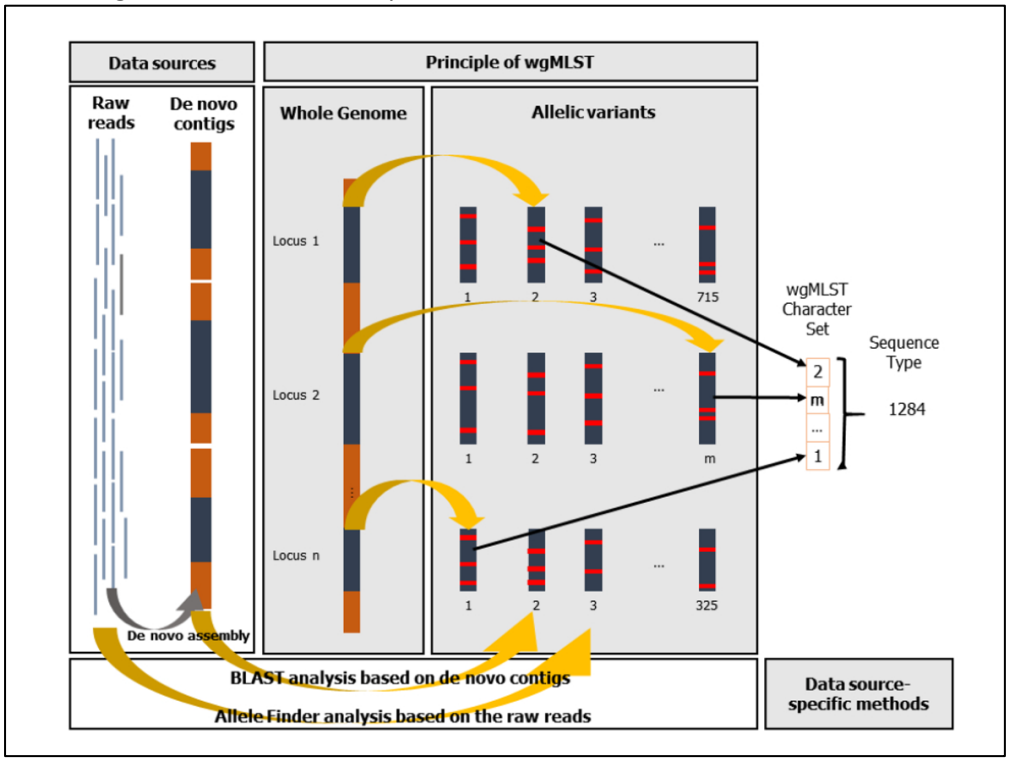


Figure 5. Principle of wgMLST. Raw reads or an assembled genome are referenced to the loci in the wgMLST scheme and typed according to the allelic variants present in the analyzed genome. (Bruyne et al, 2015)

ChewBBACA

chewBBACA (Silva et al, 2018) uses a similar whole genome typing approach to Ridom SeqSphere, but chewBBACA uses coding sequences (CDS) as input instead of the entire unannotated genome. To detect the CDS, genome annotation is needed. Genome annotation is the process of identifying protein-coding sequences and other functional elements from the DNA sequence. CDSs are predicted using an annotation tool such as Bakta (Schwengers et al, 2021) which references the translated DNA sequence to publicly known proteins from protein databases such as RefSeq, SwissProt, AMRFinder and VFDB. (Abril & Castellano, 2019)

chewBBACA first finds CDSs that have 100% nucleotide identity to the alleles in the wgMLST scheme and subsequently searches for alleles with divergent DNA sequences but similar encoded proteins. An advantage of this method over the wgMLST method of Ridom SeqSphere is that this approach allows the identification of alleles that would be considered absent when only comparing DNA sequences. (Silva et al, 2018)

PopPUNK

Population Partitioning Using Nucleotide *k*-mers (PopPUNK) uses variable *k*-mer sequence matching to compare the entire DNA sequence between genomes. PopPUNK calculates the genetic distances between two genomes by splitting the DNA sequences of two genomes into *k*-mers (DNA fragments of a set length) and the similarity between *k*-mers is calculated. Based on the similarity between *k*-mers, PopPUNK classifies the divergence between *k*-mers as core and accessory distances.

The core distance measures the genetic similarity based on the core genome, which consists of genes that are present in all or most members of a bacterial species. It is calculated by comparing the presence or absence of core genes across genomes and quantifying the level of similarity based on the shared core gene content. The accessory distance, on the other hand, focuses on the genetic differences in the accessory genome. The accessory genome consists of genes that are not universally present in all members of a species. It measures the dissimilarity based on variations in the presence or absence of accessory genes between genomes. (Lees et al, 2019) The core and accessory divergences are depicted in figure 6.

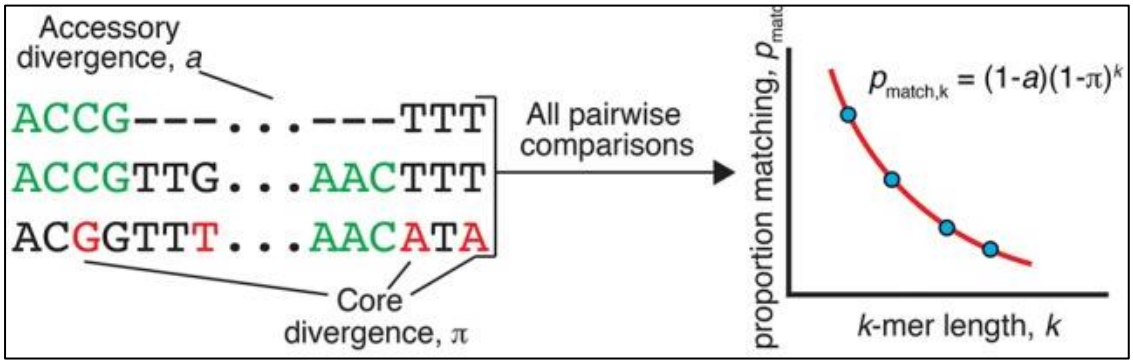


Figure 6. Core and accessory divergence (Lees et al, 2019). Accessory divergence is illustrated as a missing k-mer and core divergence as an SNP on the left. A graph representing the probability a k-mer matches, pmatch, and the changes over the k-mer length.

The similarity of a pair of genomes is expressed in pmatch, the product of the p_{core} and $p_{accessory}$, which is the probability that a *k*-mer matches without any mismatches in the shared core genome sequence and the probability a *k*-mer is not part of the accessory genome unique to one of the compared genomes respectively (Lees et al, 2019).

Snippy

Snippy aligns an input genome to a reference genome and identifies genetic variations in the genome to the reference genome. Genetic variations are SNPs, insertions and deletions in the DNA sequence. A core SNP alignment is performed to compare multiple genomes. All identified SNPs present in all the input genomes are included to generate a multiple sequence alignment as illustrated in figure 7. The number of pairwise differences in SNPs in the multiple sequence alignment is used to determine the genetic divergence between a pair of genomes. (Seemann, 2020)

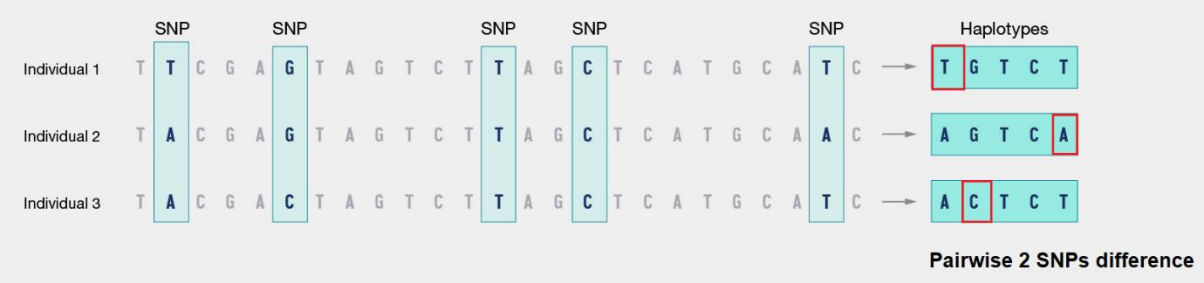


Figure 7. Multiple sequence alignment of core SNP sequences. SNPs at commonly shared genomic positions are identified and the SNP differences are pairwise compared.

Genomic clustering

Genomic clustering is a method to group genomes based on their genetic distances. The genetic variations in this study include SNPs, alleles and sequence similarity. Cluster analysis can be performed at varying degrees, from identifying a subpopulation in a population to outbreak analysis. Clustering can be challenging when the subgroups are unknown and/or there is an indiscrete difference between subgroups.

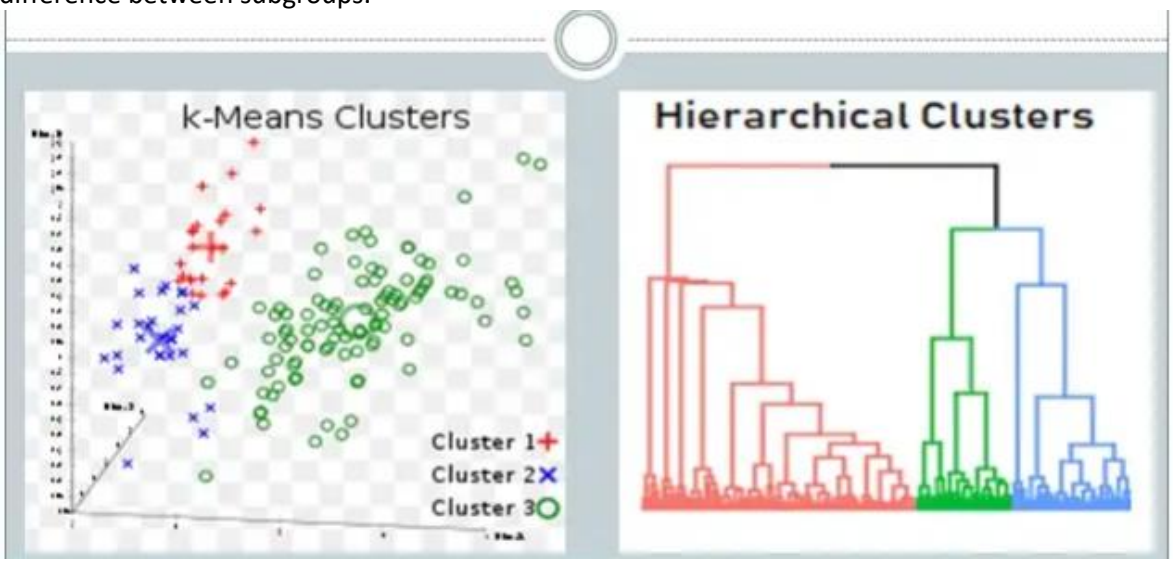


Figure 8. Example of three k-means clusters (left) and hierarchical clusters (right). Three clusters are depicted in the figure, on the left side the clusters have been analyzed using the k-means algorithm and the right side shows the cluster analysis using hierarchical clustering. (Ali, 2021)

Clustering can be performed using different clustering methods, commonly used clustering methods are hierarchical clustering and k-means. The type of clustering method can influence the grouping of clusters as well due to the different approaches. Hierarchical clustering merges clusters step by step, finding the nearest sample or cluster and repeatedly merging them together. K-means assigns samples into a cluster based on the nearest centroid (a mean point of a group of samples). The samples are divided into k clusters. (Kaushik & Mathur, 2014) The clustering method used in this study is single-linkage clustering, a hierarchical clustering method.

Cluster evaluation

Simpson's diversity index

The Simpson's diversity index is a measure to quantify the biological diversity or richness of a population or community. It does so by two measures: the richness, which is the number of species present, and the evenness, which is the abundance of the species present. It provides an estimate of the probability that two randomly selected individuals from the community belong to different species (or groups). The Simpson's index (D) can be calculated with the following formula:
$D = 1 - \Sigma n(n-1)/N(N-1)$,
n is the total number of organisms of a particular species and N is the total number of organisms in all species (Ofwell Woodland & Wildlife Trust, 2000).

Silhouette coefficient

The silhouette coefficient is a measure to quantify if samples form close-knit and well-separated clusters. This means that samples within a cluster should have a lower average distance between samples of their own cluster compared to the average distance between samples from a different cluster. The silhouette ranges from −1 to +1, where a high value indicates close-knit and well-separated clusters and a lower value indicates loose clusters with overlap. Figure 9 shows an illustration of high Silhouette coefficients, the average distance between samples within the same cluster is much smaller than the distance between samples of a different cluster.

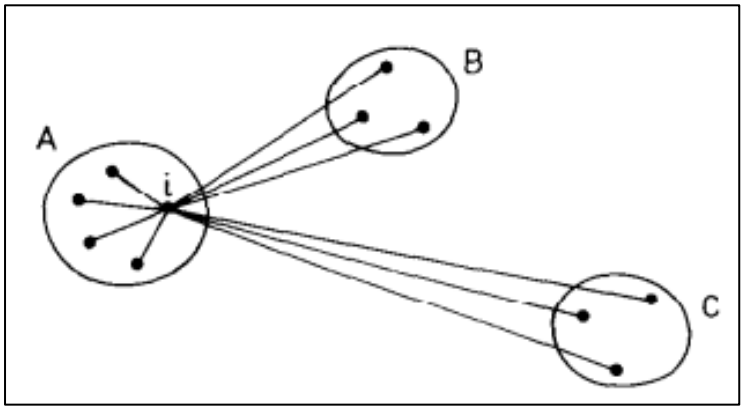


Figure 9. Illustration of the distances to be calculated for the Silhouette coefficient in three clusters: A, B and C. Showing the average distance within a cluster and the average distance between the nearest cluster for a point in cluster A. (Rousseeuw, 1987)

The silhouette coefficient is calculated for each object within a cluster by calculating the average distance between each point within its cluster (a) and the average distance between the nearest cluster (b). The silhouette coefficient can be presented in the formula: $Silhouettecoefficient = (b-a)/\max(a,b)$.

Methods

Abbreviations

Table 1, a list of abbreviations noted in this study.

Abbreviations	
ST	Sequence Type
wgMLST	whole genome Multi Locus Sequence Typing
cgMLST	core genome Multi Locus Sequence Typing
SNP	Single Nucleotide Polyphormism
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
MDRM (BRMO)	Multi-Drug Resistant Microorganisms
AMR	Anti-Microbial Resistance
ESBL	Extended Spectrum Beta-lactamase
BLAST	Basic Local Alignment Tool

Sample/data selection

E. coli ST131 samples were defined as either related or unrelated. Samples were considered related when collected from the same patient and within 3 months, and samples were considered unrelated when collected from a different patient and a different institution. The first dataset will be referred to as the *related and unrelated dataset.* This dataset was constituted only of related and unrelated samples present in the Microvida database, which were collected from 3-mar-2013 to 12-sept-2022 and originated from 8 different institutions (either hospitals or nursery homes) in the region West-Brabant and Zeeland of the Netherlands. Due to a low number of unrelated samples, additional unrelated *E. coli* ST131 genomes were collected from the Genbank database (accessed on 24-nov-2022) with the query "Escherichia coli O25b:H4-ST131" (https://www.ncbi.nlm.nih.gov/data-hub/genome/?taxon=941322) of which only the first sample was selected of each submitter.

The second dataset is referred to as the *BRMO surveillance dataset.* The dataset was constituted of *E. coli* ST131 samples part of multi-drug resistance surveillance from the Microvida database, samples part of the *related and unrelated dataset* were omitted. Samples were collected from 17-jun-2022 to 17-jan-2023 and originated from 6 different institutions (either hospitals or nursery homes) in the regions West-Brabant and Zeeland of the Netherlands. The draft genomes of both datasets were acquired in the same manner.

The studies were observational and laboratory-based, and all samples were anonymous and collected as part of routine clinical care. Thus, the requirement for patient consent was waived.

Sample collection

The *E. coli* ST131 isolates were cultured, and antimicrobial susceptibility testing was performed using either the BD Phoenix (BD, United States) according to the manufacturer's instructions or the Vitek 2 automated system (bioMérieux, France). A 0.5 McFarland suspension in 0.85% physiological saline was prepared and inoculated onto Mueller-Hinton agar plates supplemented with cefotaxime (1 mg/L) and cefoxitin (8 mg/L) on one part of the plate, and ceftazidime (1 mg/L) and cefoxitin (8 mg/L) on the other part of the plate. These plates were incubated at 37°C for 18 to 24 hours. The antimicrobial susceptibility profiles of the *E. coli* ST131 isolates were determined according to the EUCAST guidelines. Results were interpreted as susceptible, intermediate, or resistant.

DNA extraction and Whole Genome Sequencing

A milky white suspension is made by sampling an axenic culture on a sheep blood agar and inoculating it in 800 µl 1x TE The DNA of the bacterial isolates was extracted by using the QIAsymphony DSP Virus/Pathogen Midi Kit (Qiagen, Denmark) and QIAsymphony SP and AS (Qiagen, Denmark). The DNA concentration of the extracted DNA was measured using a Qubit 4 Fluorometer 1x dsDNA HS Assay Kit (ThermoFisher Scientific, United States) and the DNA concentration was normalized.

The DNA was fragmented with the Amplicon Tagment Mix from the Nextera XT DNA Library Prep Kit (Illumina, United States) according to the manufacturer's instructions. Next, the DNA fragments were indexed and amplified with the Nextera XT Index Kit v2 and Nextera PCR Master Mix. They were subsequently purified using AMPure XP beads per the manufacturer's instruction. The DNA fragments were quantified again using a Qubit 4 Fluorometer 1x dsDNA HS Assay Kit (ThermoFisher Scientific, United States) and normalized. Finally, the samples were pooled and loaded onto a v2/v3 MiSeq Cartridge. Before they were subjected to short-read sequencing on an Illumina Miseq platform using Nextera XT 300 (Illumina, United States) base pair paired-end reads.

Assembly and quality control

The short reads were assembled *de novo* in CLC Bio software, CLC genomic workbench (CLC-GW) version 22.0.2 (Qiagen, Denmark). The assembly was passed when the following quality criteria were met: more than 20-fold average coverage, N50 larger than 15,000, and less than 1000 contigs.

MLST

Multilocus sequence typing (MLST) was performed using the Achtman typing scheme in Ridom SeqSphere (Ridom, 2022, client version 8.5.1 and server version 8.3.0).

Bioinformatic analysis

Allele typing

The genetic distances between the draft genomes in both datasets were determined using a stable *E. coli* cgMLST scheme (Kluytmans et al, 2016), and the same cgMLST scheme in addition to the accessory genes was used as the stable *E. coli* wgMLST scheme. The alleles were typed of each genome in Ridom SeqSphere (Ridom, 2022, client version 8.5.1 and server version 8.3.0), and the allele distances between the genomes were extracted by generating a distance matrix in Ridom SeqSphere based on the pairwise allele distances of the corresponding cgMLST or wgMLST scheme used. Pairwise missing values were ignored.

In-between allele distances were likewise determined with a specific *E. coli* ST131 wgMLST scheme. Typing of the genomes and extraction of the distance matrix was similarly executed as the standard *E. coli* cgMLST or wgMLST scheme.

Pairwise allele distances were also determined using chewBBACA, based on the specific *E. coli* ST131 wgMLST scheme created with chewBBACA. Input genomes were annotated with Bakta (Schwengers et al, 2021, version 1.5.1), and the parameters were specified for *genus* as "*Escherichia*" and *species* as "*coli*". The resulting coding sequences were used as input for allele calling using the specific *E. coli* ST131 wgMLST scheme. The *result_alleles.tsv* was uploaded to PHYLoViZ Online (https://online.phyloviz.net/index) as Profile Data. The pairwise distances were computed, and the distance matrix was exported from PHYLoViZ Online.

PopPUNK

The core and accessory distances between the draft genomes were determined using PopPUNK (Lees et al, 2019, version 2.6.0) "--create-db" with default settings. A pairwise comparison was extracted using the *poppunk_extract_distances.py* helper script from PopPUNK.

Snippy

SNPs were identified in each draft genome using Snippy (Seemann, 2015, version 4.6.0) by aligning them to the reference genome. The reference genome was the same hybrid-assembled genome used as a seed genome during the specific wgMLST scheme creation. The default settings were used except for the "--ctgs" parameter because draft genomes (contigs) were used as input. Subsequently, core SNP alignment was performed with snippy-multi and snippy-core. The SNP distance matrix was converted from the core alignment using snp-dists (Seemann, 2021, version 0.8.2) with default settings.

Scheme creation

The specific *E. coli* ST131 scheme was created using the cgMLST target definer at default settings in Ridom SeqSphere (Ridom, 2022, client version 8.5.1 and server version 8.3.0). The wgMLST scheme was based on one hybrid assembly, assembled *de novo* with unicycler (Wick et al, 2017, version 0.4.8) based on Illumina short reads and Oxford Nanopore long reads (Oxford Nanopore Technologies, United Kingdom), annotated with Bakta (Schwengers et al, 2021, version 1.5.1) as the seed genome and 9 short read assemblies *E. coli* ST131 as query genomes. The 9 short read assemblies were defined as unrelated in the *related and unrelated dataset* and picked to include a wide gene variety (pangenome) in the scheme. Mobile genetic elements were included in the scheme, no genes were excluded in the wgMLST scheme.

A specific *E. coli* ST131 scheme was created in chewBBACA (Silva et al, 2018, version 2.8.5) as well using the same set of genomes used for scheme creation in Ridom SeqSphere as the base. The CreateSchema parameter was used with default settings in chewBBACA to create the wgMLST scheme. No genes were excluded in this wgMLST scheme as well.

Statistical analysis

The output of each typing method resulted in a distance matrix of the genetic distances between the dataset except for PopPUNK. PopPUNK produced a pairwise comparison as output and the genetic distance “pmatch” was calculated using the *core* and *accessory* distances: (1 – *accessory*) * (1 - *core*) ^ 29. The distance matrix and a table containing metadata of each sample were used as input for the *Epi_pipeline* Python script (see data availability).

The script executes the following steps: First, the distance matrix was converted to a pairwise comparison and metadata of both samples was added to the pairwise comparison. The metadata was referenced from a table with metadata of each sample holding information such as sample id, patient id, collection date, sample site and sample source (institution, location and department). The pairwise comparisons of related, a pairwise comparison of the same patient within 3 months, and unrelated, a pairwise comparison of one sample from each institution and different patient, were concatenated into a single pairwise comparison. Finally, the distribution of the pairwise genetic distances was visualized in a bar plot.

Using the *Statistics* Python script (see data availability), the threshold for the three different cutoffs, Mann-Whitney U test and Spearman's rank correlation were calculated. The threshold of misclassified as either related or unrelated and max accuracy at different distance thresholds was visualized in a line plot, a boxplot of a median comparison of related & unrelated using the Mann-Whitney U test and a heatmap of the Spearman's rank correlation coefficient between the different typing methods.

Clustering analysis

For each typing method, clustering was performed for the three previously determined cutoffs in line with the single-linkage clustering algorithm. Pairs below the cutoff threshold were selected from the pairwise comparisons and pairs were merged into a single cluster when one sample was shared between pairs. In other words, samples that are in multiple pairs are merged into a single cluster.

The clusters were internally evaluated by calculating the Simpson's diversity index and the Silhouette score was calculated by taking the average Silhouette score of all clusters. Both clustering and cluster evaluation was performed using the *Clustering* Python script (see data availability) for the *BRMO surveillance dataset.*

Results

Overview

This study aimed to differentiate epidemiologically related from epidemiologically unrelated *E. coli* ST131 samples based on their genetic variation using 5 whole genome typing methods: allele typing using stable and specific wgMLST scheme in Ridom SeqSphere, specific wgMLST scheme in chewBBACA, PopPUNK and Snippy. The study can be divided into two parts:

(1) The study first investigated whether the typing methods were capable of distinguishing between strictly related and unrelated. For this purpose, the medians of the genetic distances of epidemiologically related and unrelated samples were compared using the Mann-Whitney U test. The Spearman's rank correlation was used to measure the ranking similarity of genetic distances between the typing methods. The percentage misclassifications as related or unrelated were investigated for three cutoffs: (1) a cutoff at the lowest threshold where all epidemiologically related cases are genetically related, (2) a cutoff at the highest threshold where all epidemiologically unrelated cases are genetically unrelated and (3) a cutoff at the maximum accuracy.

(2) Secondly, these three cutoffs were used later on in the study to determine whether the typing methods could differentiate between epidemiologically related and unrelated *E. coli* ST131 samples on a real-life dataset with less extreme epidemiological links than in the *related and unrelated dataset*. Clusters based on the three cutoffs were evaluated based on their internal clustering with Simpon's diversity index and Silhouette score. The number of samples within a cluster and the cluster sizes were evaluated as well.

Datasets

For each part of the study, a separate dataset was used for analysis. The samples were draft genomes sequenced with Illumina Miseq and assembled *de novo* in CLC genomic workbench or the genome assembly was extracted from GenBank. The two datasets used for the analysis in this study are:

(1) The dataset containing sample pairs of the same patient and within 3 months (related) and different institutions (unrelated) will be referred to as the *related and unrelated dataset.* 50 *E. coli* ST131 samples of 25 related cases were extracted from the Microvida database. 8 *E. coli* ST131 unrelated samples were extracted from the Microvida database as well as an additional 10 unrelated *E. coli* ST131 samples were extracted from the Genbank database. This resulted in 25 related pairwise comparisons and 153 unrelated pairwise comparisons respectively. All samples carried an ESBL-producing gene except for two unrelated *E. coli* ST131 samples.

(2) The dataset of samples part of multi-drug resistance surveillance from the Microvida database will be referred to as the *BRMO surveillance dataset.* 70 ESBL-producing *E. coli* ST131 samples were extracted from the Microvida database. This resulted in 2415 pairwise comparisons.

Scheme creation

At the time of this study, there was no publicly available specific *E. coli* ST131 scheme. For this reason, a specific scheme was created in Ridom SeqSphere using 10 representative *E. coli* ST131 strains from the region of West-Brabant and Zeeland in the Netherlands.

The Ridom SeqSphere specific *E. coli* ST131 wgMLST scheme contained 3602 cgMLST targets and 1083 accessory targets, this totals in 4685 wgMLST targets used for allele typing. The specific *E. coli* ST131 wgMLST scheme created with chewBBACA has 5544 wgMLST targets, which contains notably more targets than the number of wgMLST targets identified using Ridom SeqSphere.

Distinguishing between epidemiologically related and unrelated *E. coli* ST131

The pairwise genetic distances were calculated by each typing method in the *related and unrelated* dataset. The pairwise genetic distances calculated by Ridom SeqSphere and chewBBACA is the difference in alleles, for Snippy it is the difference in SNPs and for PopPUNK it is the probability a *k*-mer will match with 100% identity (this value was inverted for a more intuitive view compared with other typing methods that have a distance which starts from zero). The medians of the genetic distances were compared between epidemiological-related and unrelated samples using the Mann-Whitney U test as shown in figure 10.

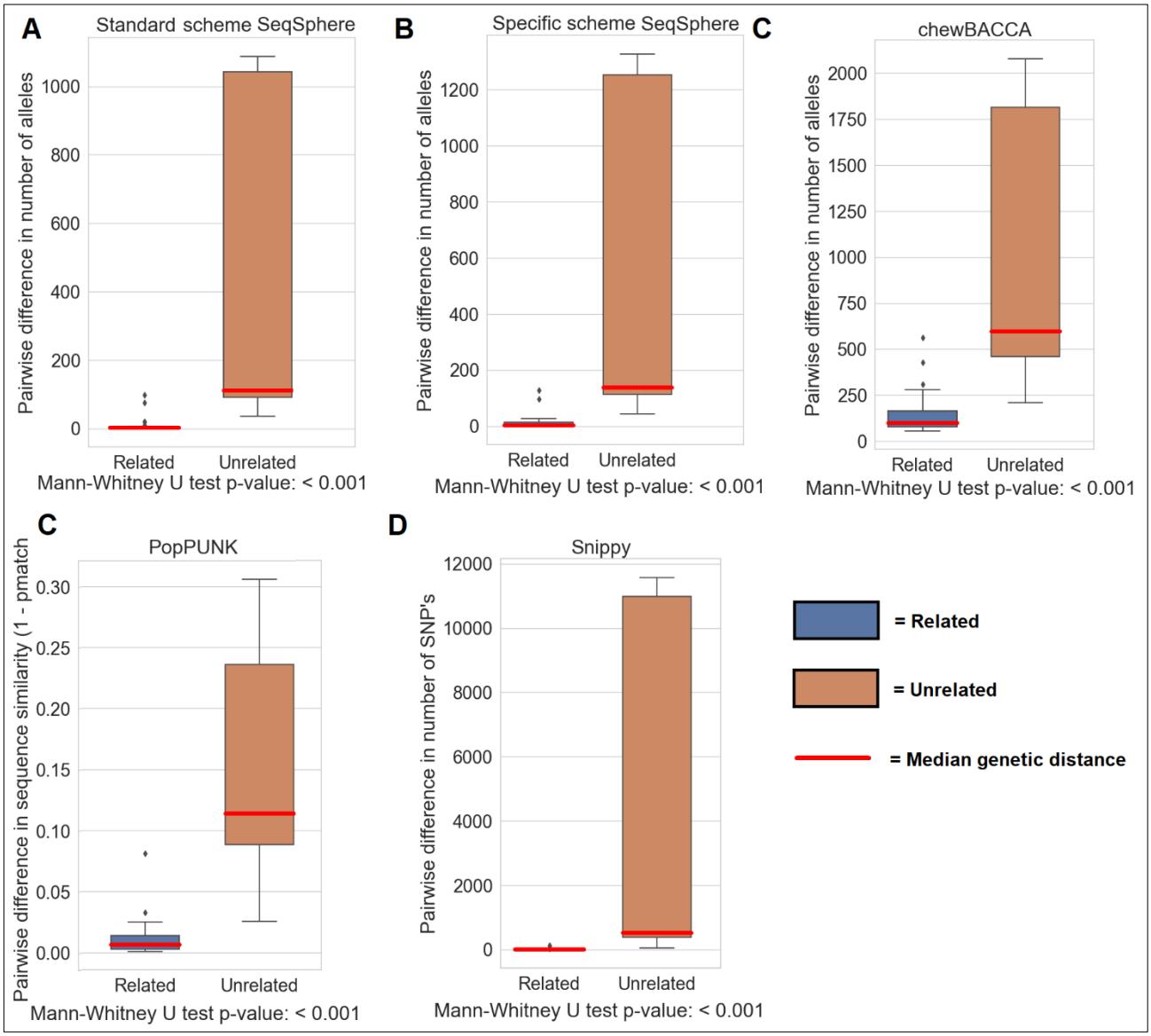


Figure 10. Box plots of the pairwise genetic distances between Related (left) and Unrelated (right) in the related and unrelated dataset as determined by the whole genome typing methods (A - D). The median genetic distances of each group are indicated in red. The spacing between the Related and Unrelated bars in the figure has been cropped.

The p-value as calculated by the Mann-Whitney U test is not greater than 0.001 for all typing methods. This indicates a significant difference in the genetic distance between epidemiologically related and unrelated *E. coli* ST131 samples. A clear distinction can thus be made between epidemiologically related and unrelated *E. coli* ST131 samples.

From the distribution of genetic distances in the boxplots, it is notable that all typing methods except for Snippy show two outliers. The outliers are two related cases and influence the cutoff threshold and misclassifications in the cutoff analysis.

Monotonic relationship between typing methods

To determine the correlation between each typing method, the order of pairwise genetic differences of the *related and unrelated* dataset was compared using Spearman's rank correlation as shown in the heatmap below (figure 11).

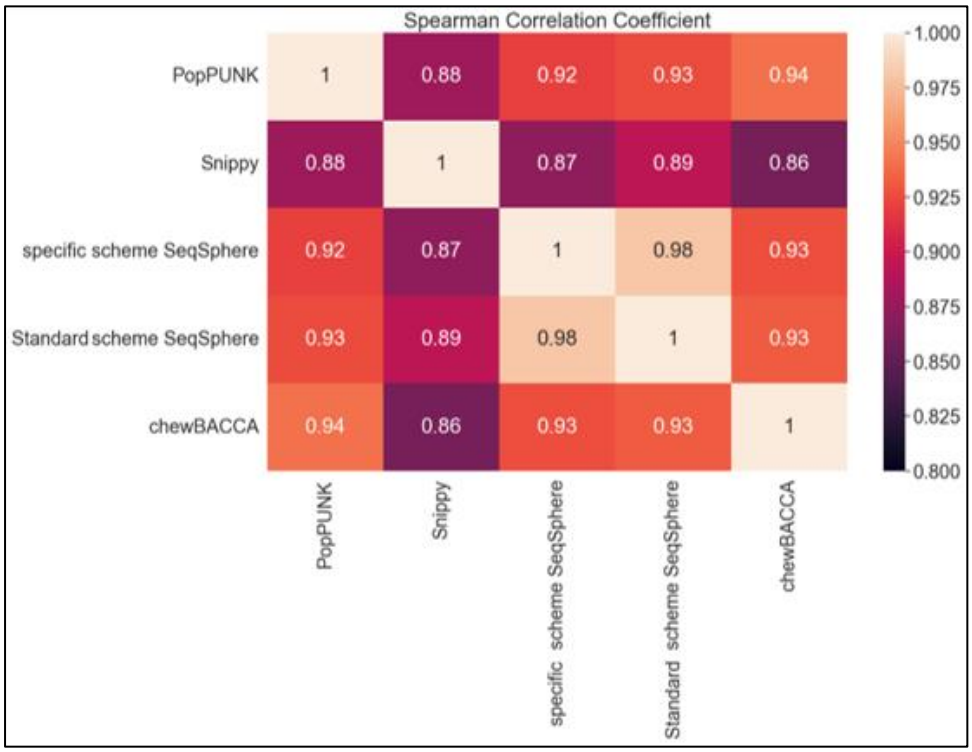


Figure 11, a heatmap of Spearman's rank correlation coefficient between the whole genome typing methods. The Spearman's rank coefficients range from 0.86 - 0.98.

The heatmap in figure 11 shows a high Spearman's rank correlation coefficient ranging from 0.86 to 0.98 across the different whole genome typing methods. This indicates a high correlation between the ranking of pairwise genetic distances. Therefore, the results of the genetic distances can be expected to be relatively similar as shown earlier in the Mann-Whitney U test comparison (see figure 11).

Cutoff determination and misclassifications as either related or unrelated

For clustering based on pairwise genetic distances, it is necessary to determine the cutoff value at which two samples are considered genetically related. In this study, the percentage of misclassifications as either related or unrelated was determined at three cutoffs for the *related and unrelated* dataset:

(1) The cutoff at the lowest threshold where all epidemiologically related cases are genetically related.
(2) The cutoff at the highest threshold where all epidemiologically unrelated cases are genetically unrelated.
(3) The cutoff at the maximum accuracy.

Table 2, the percentage of misclassifications as either related or unrelated for each cutoff per typing method.

Typing method	Cutoff 1[a]	Misclassified as related	Cutoff 2[b]	Misclassified as unrelated	Cutoff 3[c]	Misclassified as unrelated	Misclassified as related
Stable scheme SeqSphere	97	35.9 %	34	8.0 %	34	8.0 %	0.0 %
Specific scheme SeqSphere	128	43.1 %	42	8.0 %	42	8.0 %	0.0 %
chewBBACA	561	47.1 %	208	20.0 %	237	16.0 %	0.7 %
PopPUNK	0.0810	18.3 %	0.0256	8.0 %	0.0339	8.0 %	0.0%
Snippy	120	3.3 %	44	8.0 %	44	8.0 %	0.0%

[a] The cutoff at the lowest threshold where all epidemiologically related cases are genetically related
[b] The cutoff at the highest threshold where all epidemiologically unrelated cases are genetically unrelated
[c] The cutoff at the maximum accuracy

The percentage of misclassifications as related at cutoff 1 was between 35.9% - 47.1% for the wgMLST schemes, 18.3% for PopPUNK and 3.3% for Snippy. When outliers (two related cases) were excluded, the misclassification was less than 1%, but chewBBACA remained unchanged (see table 2 for additional data).

The percentage of misclassifications as unrelated was 8.0% and was the same for cutoffs 2 & 3 for all typing tools except chewBBACA. The percentage of misclassifications as unrelated using chewBBACA was 20.0% for cutoff 1, 16.0% and the percentage misclassified as related was 0.7% for cutoff 3. There is no clear dividing line that separates epidemiologically related from unrelated samples based on these results. Misclassifications are present at each cutoff, which can be attributed to two outliers of related cases as shown in figure 10 (boxplots). Overall, Snippy has the highest discriminatory power at cutoff 1 with 3.3% misclassified as related and 0% misclassifications as unrelated.

Clustering evaluation of *BRMO surveillance dataset* clusters

Clustering was performed on the *BRMO surveillance dataset* based on the single-linkage clustering algorithm at the three previously determined cutoffs. Clusters were evaluated based on the number of clusters, size of the clusters, Simpson's diversity index and Silhouette score.

Table 3. Characteristics of clustering for each typing method and cutoff. Presenting the number of clusters and size of clusters. The Simpson's diversity index and the Silhouette score as a measure of clustering accuracy are shown as well.

Typing method	Cutoff	N clusters	N in cluster	N median in cluster	Range n in cluster	Simpson's diversity index	Silhouette score
PopPUNK	1	3	59	10	[7, 42]	0,616	0,665
PopPUNK	2	7	16	2	[2, 4]	0,995	0,911
PopPUNK	3	6	21	2	[2, 9]	0,981	0,901
Snippy	1	7	66	6	[2, 33]	0,741	0,988
Snippy	2	7	28	2	[2, 14]	0,958	0,996
Snippy	3	7	28	2	[2, 14]	0,958	0,996
Specific wgMLST Seqsphere	1	6	66	5.5	[2, 44]	0,588	0,921
Specific wgMLST Seqsphere	2	7	16	2	[2, 4]	0,995	0,971
Specific wgMLST Seqsphere	3	7	16	2	[2, 4]	0,995	0,971
Stable wgMLST Seqsphere	1	6	66	5	[2, 45]	0,572	0,926
Stable wgMLST Seqsphere	2	8	24	2,5	[2, 6]	0,987	0,961
Stable wgMLST Seqsphere	3	8	24	2,5	[2, 6]	0,987	0,961
chewBBACA	1	2	70	35	[22, 48]	0,437	0,727
chewBBACA	2	8	22	2	[2, 5]	0,990	0,875
chewBBACA	3	7	27	3	[2, 9]	0,977	0,859
Stable cgMLST Seqsphere	4	8	38	2.5	[2, 18]	0,927	0,948

1 The cutoff at the lowest threshold where all epidemiologically related cases are genetically related
2 The cutoff at the highest threshold where all epidemiologically unrelated cases are genetically unrelated
3 The cutoff at the maximum accuracy
4 The current cutoff (29 alleles) used for the stable cgMLST scheme by Kluytmans et al, 2016

Table 3 shows that clustering based on cutoff 1 forms large clusters, and has a high number of isolates within a cluster and a low Simpson's diversity index. The other two cutoffs form smaller clusters, a lower number of isolates within a cluster, a high Simpson's diversity index and a high Silhouette score. At least 84.3% of the samples at cutoff 1 are considered part of a cluster at cutoff 1, while it is 22.9% - 40.0% for cutoffs 2 and 3. In other words, samples within clusters based on cutoffs 2 and 3 generally have a high similarity between samples in their own cluster and a low similarity between samples outside their cluster. Samples within clusters based on cutoff 1 are similar to samples within their own cluster, but also somewhat similar to samples outside their cluster.

The stable *E. coli* cgMLST seems to form well-defined clusters as implied by a relatively high Simpson's diversity index and Silhouette score, but approximately half the number (54.3%) of samples are considered part of a cluster. This indicates that this typing method is less discriminating compared to the other typing tools at cutoffs 2 & 3.

For a more visual representation, an example of clustering based on cutoffs 1 and 2 using PopPUNK is shown in the figure below (figure 12).

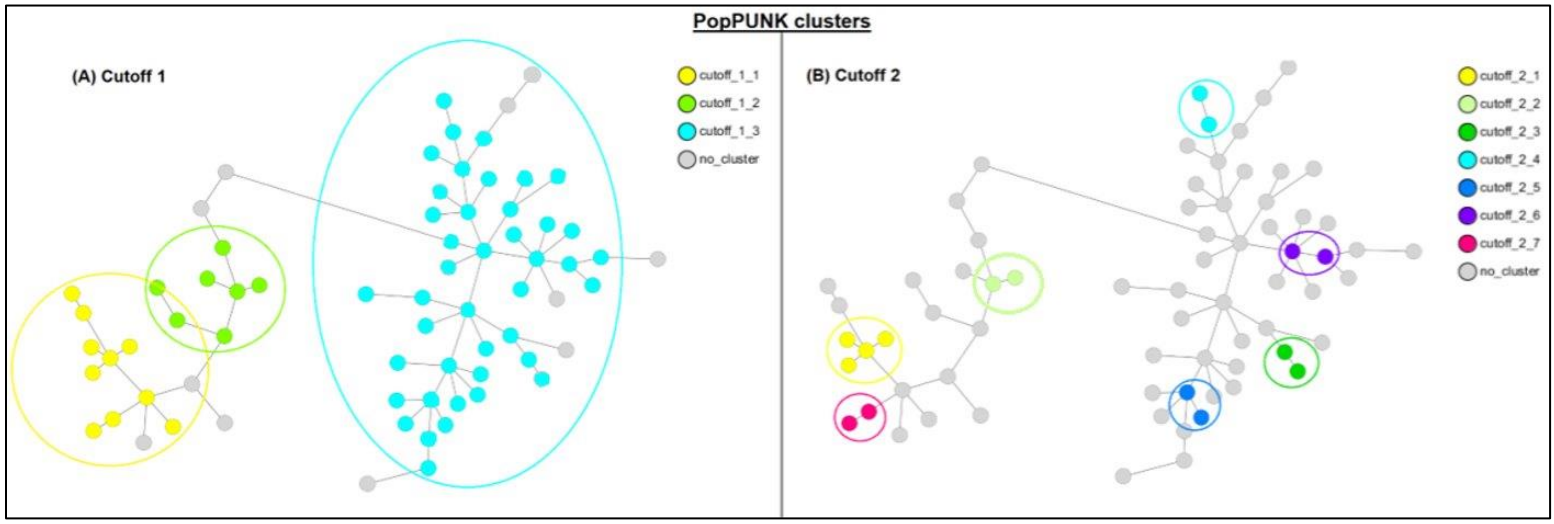


Figure 12, illustration of clusters defined by PopPUNK at cutoff 1: The cutoff at the lowest threshold where all epidemiologically related cases are genetically related and cutoff 2: The cutoff at the highest threshold where all epidemiologically unrelated cases are genetically unrelated. The figure is based on the distances of the specific E. coli ST131 wgMLST scheme and the MST is generated by Ridom SeqSphere.

Three clusters were defined using PopPUNK based on cutoff 1 (see figure 12A) and 7 clusters based on cutoff 2 (see figure 12B). Cutoff 1 has less well-defined clusters compared to the clusters based on cutoff 2. Clusters based on cutoff 1 are fewer in number and contain more samples within a cluster for cutoff 1 compared to cutoff 2. This results in a lower Simpson's diversity index for the clusters based on cutoff 1, because a random sample is more likely to be part of a cluster when using cutoff 1 compared to cutoff 2. Furthermore, there can be more genetic distance between the samples in larger clusters, so samples might be genetically closer related to samples outside the cluster compared to the average distances of samples within its own cluster. Thus, the Silhouette score is also lower for clusters in cutoff 1.

Discussion & conclusion

The current typing method based on a stable cgMLST *E. coli* scheme lacks the discriminatory power to distinguish epidemiologically related from unrelated *E. coli* ST131. In this study, to differentiate between related and unrelated *E. coli* ST131, we compared the discriminatory power of 5 different whole genome typing tools: a stable *E. coli* and specific *E. coli* ST131 wgMLST scheme in Ridom SeqSphere, a specific *E. coli* ST131 wgMLST scheme in chewBBACA, Snippy and PopPUNK at three cutoffs: (1) a cutoff at the lowest threshold where all epidemiologically related cases are genetically related, (2) a cutoff at the highest threshold where all epidemiologically unrelated cases are genetically unrelated and (3) a cutoff at the maximum accuracy. The whole genome typing tools Snippy and PopPUNK were expected to be best capable of disseminating epidemiological-related and unrelated *E. coli* ST131.

Contrary to the expectation, the whole genome typing methods compared in this study have enough discriminatory power to distinguish strictly related from strictly unrelated *E. coli* ST131 isolates. The Mann-Whitney U test shows that there is a significant difference between the genetic distances between strictly epidemiologically related and unrelated *E. coli* ST131. Spearman's rank correlation coefficient is larger than 0.86, indicating a high correlation between the whole genome typing methods. A correlation between the typing methods cg/wgMLST and Snippy has been described in earlier studies of wgMLST and core SNP alignment (Blanc et al, 2020) and for core SNP alignment and PopPUNK (Kalizang'oma et al, 2022).

The discriminatory power of the typing methods was determined at the three cutoffs. When cutoff 1 was used, the misclassifications were between 35.9% - 47.1% for the wgMLST schemes in SeqSphere and chewBBACA, 18.3% for PopPUNK and 3.3% for Snippy. This indicates that the typing methods except for Snippy will contain a high percentage of false positive related cases at cutoff 1. The misclassifications at cutoffs 2 and 3 were 8.0% for the typing methods and for chewBBACA, it was 20.0% at cutoff 2, 16.0% misclassified as related and 0.7% misclassified as unrelated at cutoff 3. Thus, typing methods have a relatively lower percentage of false negative-related cases compared to the false positives at cutoff 1. However, cutoff 1 was heavily influenced by two outlier-related cases, the large genetic distance between related cases might be attributed to multiple carriages of *E. coli* ST131 in a patient (Christansson et al, 2011).

Samples seem to be well clustered at the cutoff at which epidemiologically unrelated are genetically unrelated and the cutoff at max accuracy for all typing methods, as implied by a high Simpson's diversity index (> 0.9) and Silhouette coefficient (> 0.85). The cluster at which the cutoff includes all epidemiologically related as genetically related has a poorer clustering performance, as indicated by the corresponding Simpson's diversity index (0.43 - 0.75) and Silhouette coefficient (0.66 - 0.93). This is also reflected by fewer, but much larger clusters and a total number of samples within a cluster at cutoff 1 compared to the small clusters formed at cutoffs 2 and 3. The current standard typing practice, allele typing using a stable cgMLST *E. coli* scheme in Ridom SeqSphere, has a relatively poor performance compared to the typing methods at cutoffs 2 and 3. The number of samples within a cluster is higher, 54.3% compared to 22.9% - 40.0%.

Clusters formed at cutoffs 2 and 3 are cohesive and well separated from other clusters, but due to the lack of detailed epidemiological data, the result of the clusters cannot be compared to the true clusters with a measure such as the adjusted Rand index (Rand, 1971). The relatively high number of samples within a cluster can also be attributed to the genetic closeness of *E. coli* ST131, a previous study suggests it is genetically more conserved (Clark et al, 2012). This is unexpected because *E. coli*

ST131 is a pathogenic strain and pathogenic strains usually have an accelerated rate of mutation and recombination (Wirth et al, 2006).

This study has several limitations. Samples part of the *BRMO surveillance dataset* have limited epidemiological information since only the data of the clinical setting is known but epidemiological links outside the clinical setting are unknown. This is important to take into consideration because previous studies suggest a high prevalence of *E. coli* ST131 in community settings as well (Torres et al, 2018) Additionally, the study is based on *E. coli* ST131 collected within a time frame of one year and collected from hospitals and nursery homes in the South-Western region of the Netherlands and thus limited to one geographic setting, a study that encompasses a larger area or other sites may find a different epidemiology of *E. coli* ST131. Although this factor probably has limited influence due to the rapid spread of *E. coli* ST131 worldwide. Lastly, Snippy used a representative *E. coli* ST131 genome as a reference genome for finding SNPs, but this can be optimized by using a pangenome of *E. coli* ST131 instead of a single *E. coli* ST131 genome to include more genomic positions.

For alternate cutoff threshold determination, the mutation rate should be taken into consideration so that the time frame of transmission can be estimated. A future study can take a similar approach to the study by *Coll et al, 2020* and apply it to *E. coli* ST131. By following a cohort of patients during a set time period, samples can be taken at regular intervals to determine the mutation rate of *E. coli* ST131 using a linear mixed model. A Python script has been written to analyze the mutation rate (see Data availability). Other limitations can be adjusted for in a future study by collecting more epidemiological data to evaluate clusters with the adjusted Rand index, including a wider variety of samples from different locations and optimizing Snippy by using a pangenome as a reference genome for identifying SNPs. A pangenome could be constructed using representative *E. coli* ST131 genomes from different institutions/locations using Roary (Page et al, 2015) or Panaroo (Tonkin-Hill et al, 2020).

To conclude, the whole genome typing methods compared in this study can differentiate between strictly epidemiologically related and unrelated *E. coli* ST131 based on genetic differences, of which Snippy showed the highest degree of discriminatory power and chewBBACA the lowest degree of discriminatory power. Clustering with a single-linkage algorithm can be best performed using the cutoff at the maximum accuracy which closely resembles the cutoff at the highest threshold where epidemiological unrelated are genetically unrelated. Any typing method compared in this study has a better performance at this cutoff compared to the current typing method, a stable *E. coli* cgMLST scheme in Ridom SeqSphere at a cutoff at 29 alleles, which has a higher total number of samples within a cluster. Based on the results of this study, it is recommended to use Snippy with cutoff 3 based on performance. Although for implementation in the routine analysis, using a stable *E. coli* wgMLST scheme in Ridom SeqSphere with cutoff 3 is recommended due to its usability for non-bioinformaticians.

Data availability

The scripts used for the analysis during this study are available online on GitHub: https://github.com/Freekdek/E_coli_ST131_Bsc_thesis

The data that support the findings of this study are available from Microvida but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data is however available from the authors upon reasonable request and with permission from Microvida.

Acknowledgements

I am grateful to the microbiology technicians at Microvida for their contributions to the collection of microbiological data.

Bibliography

1. Kaper, J., Nataro, J. & Mobley, H. (2004) Pathogenic Escherichia coli. Nat Rev Microbiol 2, 123–140. https://doi.org/10.1038/nrmicro818
2. Poolman, J. T., & Wacker, M. (2016). Extraintestinal Pathogenic Escherichia coli, a Common Human Pathogen: Challenges for Vaccine Development and Progress in the Field. The Journal of infectious diseases, 213(1), 6–13. https://doi.org/10.1093/infdis/jiv429
3. Nicolas-Chanoine, M. H., Bertrand, X., & Madec, J. Y. (2014). Escherichia coli ST131, an intriguing clonal group. Clinical microbiology reviews, 27(3), 543–574. https://doi.org/10.1128/CMR.00125-13
4. Sabat, A. J., Budimir, A., Nashev, D., Sá-Leão, R., van Dijl, J.m, Laurent, F., Grundmann, H., Friedrich, A. W., & ESCMID Study Group of Epidemiological Markers (ESGEM) (2013). Overview of molecular typing methods for outbreak detection and epidemiological surveillance. Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin, 18(4), 20380. https://doi.org/10.2807/ese.18.04.20380-en
5. Gemma Clark, Konrad Paszkiewicz, James Hale, Vivienne Weston, Chrystala Constantinidou, Charles Penn, Mark Achtman, Alan McNally, Genomic analysis uncovers a phenotypically diverse but genetically homogeneous Escherichia coli ST131 clone circulating in unrelated urinary tract infections, Journal of Antimicrobial Chemotherapy, Volume 67, Issue 4, April 2012, Pages 868–877, https://doi.org/10.1093/jac/dkr585
6. Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics, 107(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003
7. Clark, D. P., Pazdernik, N., & McGehee, M. (2019). Molecular Biology Third edition. Chapter 8 – DNA Sequencing, pages 240 - 269.
8. Khan, A. R., Pervez, M. T., Babar, M. E., Naveed, N., & Shoaib, M. (2018). A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective. Evolutionary bioinformatics online, 14, 1176934318758650. https://doi.org/10.1177/1176934318758650
9. Baker, M. De novo genome assembly: what every biologist should know. Nat Methods 9, 333–337 (2012). https://doi.org/10.1038/nmeth.1935
10. van Belkum, A., Tassios, P. T., Dijkshoorn, L., Haeggman, S., Cookson, B., Fry, N. K., Fussing, V., Green, J., Feil, E., Gerner-Smidt, P., Brisse, S., Struelens, M., & European Society of Clinical Microbiology and Infectious Diseases (ESCMID) Study Group on Epidemiological Markers (ESGEM) (2007). Guidelines for the validation and application of typing methods for use in bacterial epidemiology. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases, 13 Suppl 3, 1–46. https://doi.org/10.1111/j.1469-0691.2007.01786.x
11. Tankeshwar, A., & Tankeshwar, A. (2021b, May 30). Bacterial Typing Methods: Aim, Attributes, Types. Microbe Online. https://microbeonline.com/bacterial-typing-methods-aim-attributes-and-types/
12. Kluytmans-van den Bergh, M. F., Rossen, J. W., Bruijning-Verhagen, P. C., Bonten, M. J., Friedrich, A. W., Vandenbroucke-Grauls, C. M., Willems, R. J., & Kluytmans, J. A. (2016). Whole-Genome Multilocus Sequence Typing of Extended-Spectrum-Beta-Lactamase-Producing Enterobacteriaceae. Journal of clinical microbiology, 54(12), 2919–2927. https://doi.org/10.1128/JCM.01648-16
13. Katrien De Bruyne, Bruno Pot & Hannes Pouseele. Whole genome MLST analysis | Scientist Live. (n.d.). https://www.scientistlive.com/content/whole-genome-mlst-analysis

14. Jünemann, S., Sedlazeck, F. J., Prior, K., Albersmeier, A., John, U., Kalinowski, J., Mellmann, A., Goesmann, A., von Haeseler, A., Stoye, J., & Harmsen, D. (2013). Updating benchtop sequencing performance comparison. Nature biotechnology, 31(4), 294–296. https://doi.org/10.1038/nbt.2522
15. Silva, M., Machado, M. P., Silva, D. N., Rossi, M., Moran-Gilad, J., Santos, S., Ramirez, M., & Carriço, J. A. (2018). chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. Microbial genomics, 4(3), e000166. https://doi.org/10.1099/mgen.0.000166
16. Rasko, D.A., Myers, G.S. & Ravel, J. Visualization of comparative genomic analyses by BLAST score ratio. BMC Bioinformatics 6, 2 (2005). https://doi.org/10.1186/1471-2105-6-2
17. Abril, J. F., & Castellano, S. (2019). Genome Annotation. Elsevier EBooks, 195–209. https://doi.org/10.1016/b978-0-12-809633-8.20226-4
18. Lees, J. A., Harris, S. R., Tonkin-Hill, G., Gladstone, R. A., Lo, S. W., Weiser, J. N., Corander, J., Bentley, S. D., & Croucher, N. J. (2019). Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome research, 29(2), 304–316. https://doi.org/10.1101/gr.241455.118
19. Seemann T (2015), snippy: fast bacterial variant calling from NGS reads, https://github.com/tseemann/snippy
20. Kaushik, Manju & Mathur, Bhawana. (2014). Comparative Study of K-Means and Hierarchical Clustering Techniques. Internatioonal journal of Software and Hardware Research in Engineering. 2. 93-98.
21. Simpsons Diversity Index. (n.d.). http://www.countrysideinfo.co.uk/simpsons.htm
22. Schwengers, O., Jelonek, L., Dieckmann, M. A., Beyvers, S., Blom, J., & Goesmann, A. (2021). Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microbial genomics*, *7*(11), 000685. https://doi.org/10.1099/mgen.0.000685
23. Wick RR, Judd LM, Gorrie CL, Holt KE (2017) Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13(6): e1005595. https://doi.org/10.1371/journal.pcbi.1005595
24. Blanc, D. S., Magalhães, B., Koenig, I., Senn, L., & Grandbastien, B. (2020). Comparison of Whole Genome (wg-) and Core Genome (cg-) MLST (BioNumericsTM) Versus SNP Variant Calling for Epidemiological Investigation of Pseudomonas aeruginosa. Frontiers in microbiology, 11, 1729. https://doi.org/10.3389/fmicb.2020.01729
25. Kalizang'oma, A., Kwambana-Adams, B., Chan, J. M., Viswanath, A., Gori, A., Richard, D., Jolley, K. A., Lees, J., Goldblatt, D., Beleza, S., Bentley, S. D., Heyderman, R. S., & Chaguza, C. (2023). Novel Multilocus Sequence Typing and Global Sequence Clustering Schemes for Characterizing the Population Diversity of Streptococcus mitis. Journal of clinical microbiology, 61(1), e0080222. https://doi.org/10.1128/jcm.00802-22
26. Christiansson, M., Melin, S., Matussek, A., Löfgren, S., & Söderman, J. (2011). MLVA is a valuable tool in epidemiological investigations of Escherichia coli and for disclosing multiple carriage. Scandinavian journal of infectious diseases, 43(8), 579–586. https://doi.org/10.3109/00365548.2011.568953
27. William M. Rand (1971) Objective Criteria for the Evaluation of Clustering Methods, Journal of the American Statistical Association, 66:336, 846-850, DOI: 10.1080/01621459.1971.10482356
28. Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L. H., Karch, H., Reeves, P. R., Maiden, M. C., Ochman, H., & Achtman, M. (2006). Sex and virulence in Escherichia coli: an evolutionary perspective. Molecular microbiology, 60(5), 1136–1151. https://doi.org/10.1111/j.1365-2958.2006.05172.x

29. Torres, E., López-Cerero, L., Morales, I., Navarro, M. D., Rodríguez-Baño, J., & Pascual, A. (2018). Prevalence and transmission dynamics of Escherichia coli ST131 among contacts of infected community and hospitalized patients. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases, 24(6), 618–623. https://doi.org/10.1016/j.cmi.2017.09.007
30. Coll, F., Raven, K. E., Knight, G. M., Blane, B., Harrison, E. M., Leek, D., Enoch, D. A., Brown, N. M., Parkhill, J., & Peacock, S. J. (2020). Definition of a genetic relatedness cutoff to exclude recent transmission of meticillin-resistant Staphylococcus aureus: a genomic epidemiology analysis. The Lancet Microbe, 1(8), e328-e335. https://doi.org/10.1016/S2666-5247(20)30149-X
31. Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics (Oxford, England), 31(22), 3691–3693. https://doi.org/10.1093/bioinformatics/btv421
32. Tonkin-Hill, G., MacAlasdair, N., Ruis, C. et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol 21, 180 (2020). https://doi.org/10.1186/s13059-020-02090-4

Figures

1. Whole Genome Sequencing | Bio Basic Asia Pacific Pte Ltd | Home. (n.d.). https://biobasic-asia.com/services/next-generation-sequencing/whole-genome-sequencing/
2. Lu, Y., Shen, Y., Warren, W., & Walter, R. (2016). Next Generation Sequencing in Aquatic Models. InTech. doi: 10.5772/61657
3. The Sequencing Center. (2022, September 26). What is de novo assembly? - The Sequencing Center. https://thesequencingcenter.com/knowledge-base/de-novo-assembly/
4. Society, M. (n.d.). Uncovering the fungal pangenome. Microbiology Society. https://microbiologysociety.org/blog/uncovering-the-fungal-pangenome.html
5. Whole genome MLST analysis | Scientist Live. (n.d.-b). https://www.scientistlive.com/content/whole-genome-mlst-analysis
6. Lees, J. A., Harris, S. R., Tonkin-Hill, G., Gladstone, R. A., Lo, S. W., Weiser, J. N., Corander, J., Bentley, S. D., & Croucher, N. J. (2019). Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome research, 29(2), 304–316. https://doi.org/10.1101/gr.241455.118
7. Ali, A. (2021, December 7). Clustering(K-Mean and Hierarchical) with Practical Implementation. Medium. https://medium.com/machine-learning-researcher/clustering-k-mean-and-hierarchical-cluster-fa2de08b4a4b
8. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Attachments

Duurzaamheidsanalyse van methoden voor het typeren van het hele genoom om onderscheid te maken tussen epidemiologisch gerelateerde en niet-gerelateerde E. coli ST131-isolaten

De toenemende verspreiding van multiresistente bacteriën is wereldwijd een groot probleem voor de volksgezondheid. Het vermogen om snel epidemiologisch verwante en niet-verwante bacteriestammen te differentiëren is essentieel om de verspreiding van deze infecties te voorkomen. Dit kan efficiënt gedaan worden met gebruik van Next Generation Sequencing (NGS) en vervolgens de genomen te analyseren met Whole Genome Typing methodes (WGT). Het is echter belangrijk om te onderzoeken of het gebruik van WGT-methoden voor het identificeren van verspreiding van multiresistente bacteriën bijdraagt aan een duurzame samenleving.

Het gebruik van WGT-methoden voor het identificeren van bacteriële overdracht draagt bij aan een duurzame samenleving door drie duurzame ontwikkelingsdoelen (SDG's) van de Verenigde Naties aan te pakken: goede gezondheid en welzijn (SDG 3), verantwoorde consumptie en productie (SDG 12) en partnerschap om doelstellingen te bereiken (SDG 17).

Het gebruik van WGT-methoden voor het identificeren van transmissie van infectieziekten kan bijdragen aan het verbeteren van de kwaliteit van leven en het bevorderen van een goede gezondheid en welzijn. Door patiënten met multiresistente bacteriële infecties te identificeren en te isoleren op een snelle en effectieve manier, kan de verspreiding van deze infecties worden verminderd. Hierdoor wordt verdere transmissie voorkomen en andere patiënten beschermd van een infectie van een multiresistente bacterie, wat behandeling lastiger maakt en de behandeling verlengt.

Ook kan het gebruik van WGT-methoden voor het identificeren van transmissie bijdragen aan verantwoorde consumptie en productie. Door op tijd patiënten met multiresistente bacteriële infecties te identificeren en te isoleren, wordt voorkomen dat meer patiënten besmet raken. Hierdoor kan onnodig gebruik van antibiotica worden verminderd, waardoor het risico op antibioticaresistentie wordt geminimaliseerd. Dit kan ook leiden tot een duurzamer gebruik van antibiotica, waardoor de algehele consumptie van antibiotica afneemt. Een nadeel is dat het sequensen van bacteriën kostelijk is en de materialen die hiervoor gebruikt worden zijn niet duurzaam. Echter, de kosten van isolatie zijn erg hoog en veel afvalproducten in de vorm van persoonlijke beschermingsmiddelen worden geproduceerd, hier wordt dus ook op bespaart. Het is daarom belangrijk om selectief DNA te sequensen en zo min mogelijk onnodig NGS te gebruiken. Door bijvoorbeeld alleen isolaten van opgenomen patiënten in ziekenhuizen of andere zorginstellingen eenmaal in het jaar te analyseren.

Ten slotte, het gebruik van WGT-methoden voor het identificeren van bacteriële overdracht kan bijdragen aan de partnerschap om doelstellingen te bereiken. Multidisciplinaire samenwerking tussen microbiologen, deskundige infectiepreventie en gezondheidswerkers is essentieel voor een effectieve implementatie van WGT-methoden voor het identificeren van transmissie en vervolgens actie te ondernemen om verspreiding in te perken. Het delen van kennis en middelen kan de ontwikkeling van duurzame oplossingen vergemakkelijken om het probleem van multiresistente

bacteriële infecties aan te pakken. Een nadeel van een brede samenwerking is dat gevoelige patiëntgegevens beschermd moeten blijven, om de patiënten te beschermen tegen bevooroordeling van het dragen van een multiresistente bacterie. Dit kan worden voorkomen door zorgvuldig delen van gegevens en deze te anonimiseren waar nodig.

Het gebruik van WGT-methoden voor het identificeren van transmissie van multiresistente bacteriën is essentieel om de verspreiding van multiresistente bacteriële infecties te voorkomen. Deze technologie draagt bij aan een duurzame samenleving door de SDG's van goede gezondheid en welzijn, verantwoorde consumptie en productie en partnerschap om doelstellingen te bereiken aan te pakken. Het is echter belangrijk om rekening te houden met de potentiële kwetsbaarheden die aan deze technologie zijn verbonden, zoals gegevensbeveiliging en kosten en vervuiling van het sequensen van DNA, om ervoor te zorgen dat het gebruik van WGT-methoden voor het identificeren van bacteriële overdracht duurzaam blijft.