Universitetet
i Stavanger

FACULTY OF SCIENCE AND TECHNOLOGY

MASTER'S THESIS

Study program/specialization: **Master of Science in Biological Chemistry**	Spring / Autumn semester, 2019 Open Access
Author: **Stine Grude Telstø**	Stine Grude Telstø (signature of author)

Faculty supervisor: **Hanne Røland Hagland, PhD**

External supervisor(s): **Dr. Aasmund Fostervold, Eva Bernhoff, PhD**

Title of master's thesis:

Investigating the molecular epidemiology of Norwegian blood culture isolates of *Klebsiella pneumoniae* using whole genome sequencing.

Credits: **60 sp**

Keywords: ***Klebsiella pneumoniae*, antimicrobial resistance, whole genome sequencing, phylogeny, pulse field gel electrophoresis.**	Number of pages: 103 + supplemental material/other: 5 Stavanger, 15.06.2019

Investigating the molecular epidemiology of Norwegian blood culture isolates of *Klebsiella pneumoniae* using whole genome sequencing.

By

Stine Grude Telstø

A thesis submitted in fulfilment with the
degree of Master of Science in Biological Chemistry

at the

Faculty of Science and Technology

Department of Mathematics and Natural Science

Acknowledgements

I would like to express my deepest appreciation for the major opportunity it has been to write my master's thesis at the Department of Medical Microbiology at Stavanger University Hospital. The experience has been rewarding, challenging and full of learning and new knowledge. I would especially like to acknowledge and thank my supervisors, Eva Bernhoff and Aasmund Fostervold for great advice, mentoring and patience along the way.

I would also like to thank Ragna-Johanne Bakksjø and Marit Andrea Klokkhammer Hetland; for all the support, advice and help in the lab and on the bioinformatics, you have truly been priceless. Many thanks goes to the entire department, especially Iren Høyland Lörh, for being including and welcoming during this year, as well as for help and assistance.

I am also very grateful to my family and friends, who have been great supporters and cheered me on through this process.

Stine Grude Telstø, Stavanger 2019

Abstract

Antimicrobial resistance (AMR) is a rapidly increasing threat to public health, and was in 2019 listed as one of top ten global health threats by the World Health Organization (WHO). The gram-negative *Enterobacteriaceae Klebsiella pneumoniae* (*K. pneumoniae*) has as a ranking of "Priority 1: Critical" on WHOs list of pathogens, due to its development of multi-drug resistance (MDR) towards last-line antibiotics.
The development of new effective antibiotics, as well as research on this bacteria is much needed to limit and ultimately reduce the spread and evolvement of resistant and highly pathogenic strains. Strengthened prevention and infection control and increased surveillance as well as utilization of the "One Health" approach are all actions to regard for an effective strategy against AMR.
Whole genome sequencing (WGS) can aid in characterization, detection, tracking and surveillance of the evolution and emergence of AMR related strains. WGS enables accurate characterization of transmission and outbreaks by allowing comparison of clinical isolates with an accuracy of a single nucleotide difference. WGS also provides knowledge of the presence resistance- and virulence genes, and the number of SNPs in the whole genome, as well as information on conserved and variable genes in various lineages. Phylogenetic analysis, such as core genome MLST, can uncover specific sequence types (STs) associated with resistance- or virulence genes, and genetic relations between clinical isolates.
PFGE was and still is considered the "gold standard" for bacterial strain typing in many laboratories. This method was therefore compared to a WGS based strain typing method to see if there is correlation or large differences between effectiveness, reliability and resolution in the two methods.

In this thesis, a collection of 722 *K. pneumoniae* strains provided by the NORKAB study was whole genome sequenced and analyzed to describe the genetic epidemiology between the isolates. Investigations were carried out with the phylogenetic analysis methods; core genome analysis and single nucleotide polymorphism (SNP) analysis. The phylogenetic analysis of the WGS data was compared with phylogenetic analysis of pulse field gel electrophoresis (PFGE) data.

In the investigated *K. pneumoniae* population, multilocus sequence typing (MLST) analysis revealed a diverse distribution, with a total of 378 different STs, with seven prevalent

variants, ST107 (n=67), ST20 (n=23), ST37 (n=20), ST45 (n=18) and ST307, ST25 and ST26 (n=12). Locus variants (LV) we detected in 178 isolates, distributed among 147 different LVs. In addition, three new STs were assigned to three of the isolates from the population.
No significant geographical differences were found in the distribution of STs, except for a few local build-ups of ST10 (8/8) and ST107 (49/67) in both the West and the South regions, and ST220 (5/6) and ST29 (5/8) in the East region, but the Middle region was under-represented with only one participation hospital.
The most prevalent ST, ST107, was investigated through PFGE and core genome SNP analysis. PFGE analysis indicated clonality in 33/37 isolates, based on a one band difference in the DNA fingerprints, and no positional differences. Core genome SNP analysis of the same isolates suggested close relations, based on an average SNP difference of ~15±6. The suggested SNP difference for indication of clonal isolates is ≤10. This showed that core genome SNP analysis has a stronger discriminative power and a higher resolution to differentiate isolates than that of PFGE. The core genome SNP analysis did not suggest clonality in the same number of isolates as the DNA fingerprint analysis did.

A large diversity of sequence types was identified among clinical isolates of *K. pneumoniae* from five geographical regions of Norway. The most prevalent, ST107, was found in all regions, with higher incidence in the South- and West regions. The majority of isolates were characterized as closely related, and some were characterized as clonal. Comparison of PFGE and core genome SNP analysis of the isolates suggested a higher discriminative power in the latter method.

Abbreviations

AMR – anti microbial resistance

ATM – amplicon tagment mix

BLAST – basic local alignment search tool

bp – base pair

CDC – Center for Disease Control

CE – capillary electrophoresis

clb - colibactin

CPS – capsular polysaccharide synthesis

DNA – deoxyribonucleic acid

ddNTP – dideoxynucleotide

ECDC – European Centre for Disease Prevention and Control

EtOH – ethanol

GB – gigabases

HGT – horizontal gene transfer

HS – high sensitivity

HT1 – hybridization buffer

ICU – intensive care unit

Indels – insertions and deletions

iro – salmochelin

iuc – aerobactin

Kbp – kilo base pair

K. africanesis – Klebsiella africanesis

K. pneumoniae – Klebsiella pneumoniae

K. quasipneumoniae – *Klebsiella quasipneumoniae subsp. smilipneumoniae/quasipneumoniae*

K. variicola – *Klebsiella variicola subsp. variicola/tropicalensis*

LNA1 – normalization buffer

LNB1 – bead mixture

LNS1 – normalization storage buffer

LNW1 – normalization wash buffer

Mb – mega base pair

MCS – MiSeq control software

MDR – multi-drug resistance

MGPs – magnetic glass particles

ML – maximum likelihood

ML STAR – (Hamilton) Microlab STAR

MLST – multilocus sequence type

NGS – next generation sequencing

NORKAB – the Norwegian *Klebsiella pneumoniae* Bacteremia Study

NPM – nextera PCR master mix

NT – neutralize tagment buffer

Oligos – oligonucleotides

OTU – operational taxonomic unit

PE – pair-end(ed)

PCR – polymerase chain reaction

PF – passing filter

RTA – real time analysis

RFID – radio frequency identification

SAV – sequence analysis viewer

SBS – sequencing by synthesis

SNP – single nucleotide polymorphism

ST – sequence type

TD – tagment DNA buffer

WGS – whole-genome sequencing

ybt - yersiniabactin

Table of contents

1. Introduction

1.1 Background

In the pre-antibiotic era, bacterial infections were a great threat and the cause of many deaths. Something as insignificant as a cut could potentially be fatal.
However, even though the development of antibiotics was a great step, bacterial infections remain a major problem, especially in already weakened people, such as elderly and neonates, even with functioning antibiotics.

The development of antimicrobial drugs has been called one of the greatest success stories in modern medicine. But, in recent years, AMR has become an unwelcome reality on a global scale, and has made it on the WHO top ten list of global health threats in 2019 [1]. *K. pneumoniae* is listed by WHO as a bacteria for which new antibiotics is urgently needed, with the ranking of "Priority 1: Critical" [2].
A continually increasing number of microorganisms are developing resistance towards antimicrobial agents in a frightening pace, seemingly too fast for society to keep up. AMR occurs naturally, but one significant reason for the alarming pace of its development is overuse and incorrect use of antimicrobial drugs, not just in humans, but also in animals, especially those used for food production, and the environment i.e. through pollution of rivers from antibiotic production waste. Other contributing factors can be a growing hesitancy towards vaccines and an increase in travelling, hence spreading resistant microbes [3-5].

A consequence of AMR is that the antimicrobial agents available become less effective, and infection prevention becomes more challenging. This causes infections that we today view as treatable, such as pneumonia and wound infections, to again pose a mortal threat. Moreover, life-saving and quality-of-life enhancing procedures, such as chemotherapy and immunosuppressive treatments, organ transplantations, caesarean sections and other surgeries will become high-risk procedures due to the danger of contracting a non-treatable infection.
K. pneumoniae is a bacteria naturally present in the intestinal system of most humans. However, it is considered an opportunistic bacteria, and is a major cause of hospital-acquired infections, such as pneumonia, sepsis, and infections in newborns and intensive-care units [6].

In addition to generating an increased mortality burden on society, AMR will cause a great economical loss, due to prolonged hospital stays, loss of workforce and excess healthcare system costs [1, 3, 4].

As a response to this ever growing threat, WHO initiated the "Global action plan on antimicrobial resistance" on May 15th 2015 [7]. Some of the goals of this plan is to improve the awareness and understanding of AMR, and to strengthen the knowledge of AMR through surveillance and research.

The "One Health Initiative" movement was started to create an all inclusive collaboration between professionals from different field, recognizing that human-, animal- and the ecosystem health of our planet is linked. The goal of One Health is to improve and defend the health and well-being of all species by promoting cooperation and collaboration between physicians, veterinarians, other scientific health professionals and environmental professionals [8]. An example of such a joint effort in Norway is the annual NORM/NORM-VET report for the Usage of Antimicrobial Agents and Occurrence of Antimicrobial resistance in Norway, encompassing both humans and animals. NORM is coordinated by the Norwegian Institute of Public Health, and NORM-VET by the Norwegian Veterinary Institute [9].

The Norwegian *Klebsiella pneumoniae* bacteremia study (NORKAB) is a prospective multicenter cohort study, which will work towards determining if *K. pneumoniae* blood culture isolates, from patients all over Norway, has STs, clonal groups or phylo-groups associated with a 30-days case fatality rate for *K. pneumoniae* bacteremia. The study is approved by an ethics committee [10].

An important part of winning in the race against AMR is the surveillance of the spread of pathogenic strains, and the detection of novel pathogenic lineages. NORKAB contributes to this in Norway by mapping national epidemiology on strain and genomic levels. A part of this puzzle can be found through WGS by detecting and mapping STs. Phylogenetic analysis can be performed to possibly detect crucial phylogenetic relations between dominating STs, and to determine if there are any geographical differences of significance in certain STs on a national level.

1.2 *K. pneumoniae* – general characteristics

K. pneumoniae is a bacteria of the *Enterobacteriaceae* family, first described by Carl Friedländer in 1882, after being isolated from the lungs of patients deceased from pneumonia [11]. *K. pneumoniae* is a gram-negative, non-motile, usually encapsulated rod-shaped bacteria. The bacteria is ubiquitous in the environment, such as surface water, soil and plants. It also colonizes mucosal surfaces of several mammals, such as humans, horses and swine, thereby being a natural part of the flora. In humans, the bacteria can be present in nasopharynx and the intestinal tract [6].

Although *K. pneumoniae* often is a commensal, it is also an opportunistic pathogen. It is a common cause of nosocomial infections, and most often affects the respiratory and urinary tract, but is also known to cause soft tissue infections and septicemia. The source of hospital-acquired infections are often contaminated hospital equipment and blood products, as well as the gastrointestinal tract of patients and the hands of hospital personnel [12].

1.2.1 Virulence factors and resistance genes of K. pneumoniae

K. pneumoniae has several virulence factors for evading the host's immune system. An example of such a factor is the capsular polysaccharide synthesis (CPS) locus, which synthesizes capsules, such as rmpA, which will help the bacteria avoid phagocytosis by macrophages. RmpA, as well as rmpA2 are associated with hypermucoidity [13]. Another factor is secretion of several siderophores, such as yersiniabactin (ybt), aerobactin (iuc) and salmochelin (iro). Nutritional immunity is a host defense mechanism based on decreased concentration of minerals, such as iron and zinc. CPS and the mentioned siderophores are the best described virulence factors of *K. pneumoniae* [14, 15].

These virulence factors are associated with both community-acquired and nosocomial infections, such as pneumonia, pyogenic liver abscess, meningitis and other invasive infections, such as bacteremia [13].

Certain sequence types has been identified as high-risk associated to resistance and different virulence factors. Examples of STs associated with resistance are ST11, ST15, ST37 and ST147, and examples of STs associated with virulence factors are ST23, ST65, ST86, ST163 and ST375 [16].

The spread of virulent and multi-drug resistant *K. pneumoniae* strains in hospitals and throughout the environment are of great concern. All *K. pneumoniae* has the chromosomal *AmpH* gene, encoding narrow spectrum beta-lactamase [17], but some *K. pneumoniae* strains can i.e. through horizontal gene transfer (HGT) acquire extended-spectrum-beta-lactamase (ESBL) genes, conferring resistance to extended-spectrum cephalosporins, as well as penicillin and monobactams. Common ESBL genes in *K. pneumoniae* are bla_{SHV}, bla_{TEM}, and $bla_{CTX\text{-}M}$ [18]. HGT, or lateral gene transfer, refers to nonsexual transmission of genetic material across species boundaries, and usually involves mobile genetic elements such as phages, plasmids or transposons. HGT has a great impact on the evolution of the bacteria, and is heavily involved in the dissemination of virulence factors and resistance genes [19, 20]. Acquisition of ESBL-genes are facilitated by mobile genetic elements, usually plasmids or transposons. Plasmids in particular often carry resistance genes affecting other drug classes as well, resulting in bacterial strains resistant to several classes of antimicrobials. This limits options for treatment. The last resort treatment in such cases has usually been carbapenems, which are therefore called last-line antibiotics. But in recent years some strains has also developed resistance to certain carbapenems. A continued spread and development of these strains may result in few or no available therapeutic options [6, 21].

1.2.2 *K. pneumoniae* species

Through studies conducted by Bialek-Davenet *et. al* (2014) [22], which was later confirmed by Holt *et. al* [13], *K. pneumoniae* sensu lato was distinguished into specific, but phylogenetically related species: KpI – *K. pneumoniae* sensu stricto, KpII – *K. quasipneumoniae* (encompassing KpII-A subspecies *quasipneumoniae* and KpII-B *similipneumoniae*) and KpIII – *K. variicola*, shown in Figure 1. The term *K. pneumoniae* hence can be used to describe all four phylogroups sensu lato, or specifically to the KpI phylogroup *K. pneumoniae* sensu stricto. In this thesis, the term *K. pneumoniae* will be used in general to describe all phylogroups, sensu lato, unless otherwise stated.

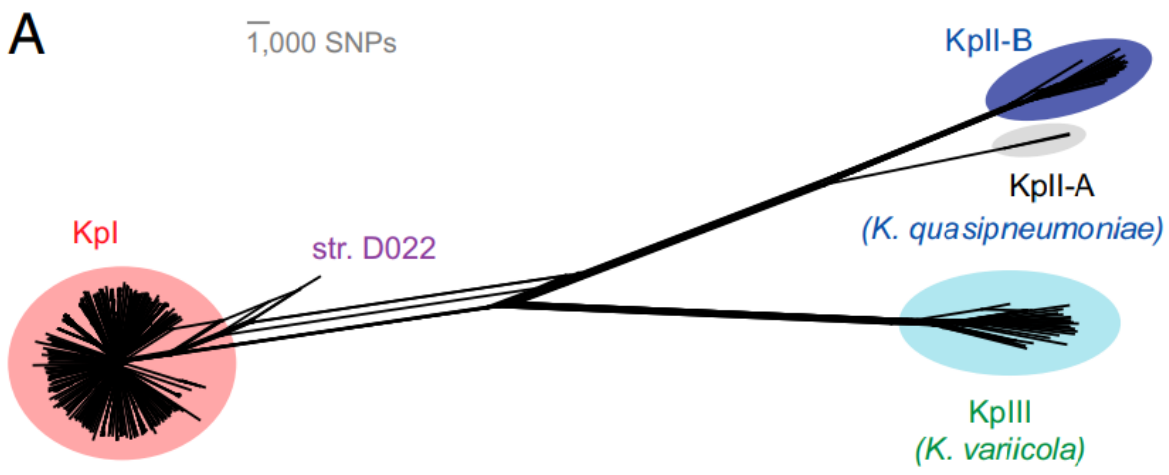


Figure 1: The distinction of K. pneumoniae into phylogroups KpI, KpII-A, KpII-B, and KpIII.
Adapted from: [13]

In a recent study, two other phylogenetic groups of *K. pneumoniae*, KpV – *K. variicola subsp. tropicalensis* and KpVII – *K. africanesis*, has been described. KpV is closely related to KpIII, and KpVII is closely related to KpI [23].

1.3 *K. pneumoniae* epidemiology

Epidemiology is the study how often diseases and other health issues occur in different groups of people and why, in other words "who, when and where" [24].
K. pneumoniae is naturally present in the environment, as well as being a part of the commensal microbial flora of the nasopharynx, skin and intestinal tract in some humans. The bacteria is most prevalent in the intestinal tract, and the detection rate of *K. pneumoniae* in feces from adults ranges from 5-38%. The carrier frequency depends on many factors, such as the duration of a previous of on-going hospital stay or international travel [6, 25, 26]. *K. pneumoniae* is considered a nosocomial opportunistic pathogen, where patient groups, such as elderly, neonatal, immunosuppressed and intensive care patients are especially at risk [6, 25]. In Norway *K. pneumoniae* is responsible for ~8-9% of all detected bacteremia's [9].

An outbreak of a disease occurs when individuals are infected from a common source, such as contaminated medical equipment or water supply, or when the infection can be transmitted between individuals directly, or by vector borne transmission [27]. There has been a few *K. pneumoniae* outbreaks reported in Norway, one in an intensive care unit (ICU) where the outbreak source was a fiber optic intubation endoscope [28], one in a neonatal ICU associated with contaminated breastmilk [29], and one in another ICU, where the outbreak was maintained and prolonged due to contaminated sinks [30]. In all of these outbreaks, the strains were MDR, in two of the cases they were ESBL producing, and in the third case they were carbapenemase producing. In many countries, multi-resistance to antibiotics is rapidly increasing and has become an everyday occurrence. The European Antimicrobial Resistance Surveillance System Network (EARS-Net) reported that 22.3% of all *K. pneumoniae* invasive isolates were resistant to at least three antimicrobial classes, and that resistance towards carbapenems increased from 8-15% over a period of 5 years [12, 31]. The occurrence of MDR strains in Norway has been low, but increasing [9].
While traditional views has been that *K. pneumoniae* is a nosocomial infection, recent studies suggests that as many as ~40-50% of all invasive *K. pneumoniae* infections are community acquired [32, 33].

1.4 Methods for determining epidemiology in *K. pneumoniae*

Bacterial genomes of the same species shares a set of common genes that are present in all isolates, but differences among bacterial genomes of the same species may occur. Examples of differences are single nucleotide polymorphisms (SNPs), deletions and insertions (indels), which are all point mutations. Transfer of genetic material between different species, HGT, which causes recombination or insertions, may also be a cause of variability in the common genes in a species. Different methods of bacterial strain typing differ greatly concerning cost, effort, and reliability and the capacity to discriminate between strains of bacterial pathogens, as well as reproducibility and repeatability. Many methods are also very organism specific, so no technique is optimal for all types of investigations [34, 35].

1.4.1 Multilocus sequence typing

Multilocus sequence typing is a technique based on identifying alleles from the nucleotide sequence of normally seven so-called housekeeping genes (~450-500 bp), that are assumed present in all strains of a species. The housekeeping genes has been chosen specifically for different pathogenic species. For each isolate of a species, the alleles of usually seven housekeeping genes define the allelic profile, or sequence type (ST). Hence, each isolate of a species can be unequivocally characterized by a series of seven numbers in the order of discovery, which corresponds to the alleles at the seven housekeeping loci. In MLST, the sequences are assigned as different alleles whether they differ at a single nucleotide site or many sites. This is done because it is not possible to determine whether differences in one or many nucleotide sites are caused by point mutations or recombinational replacement.
An illustration of MLST can be seen in Figure 2. A great benefit of MLST is that the sequence data for each individual isolate is unique, and the allelic profiles of isolates can be compared to those in a database, such as The Klebsiella PasteurMLST sequence definition database [35-39].

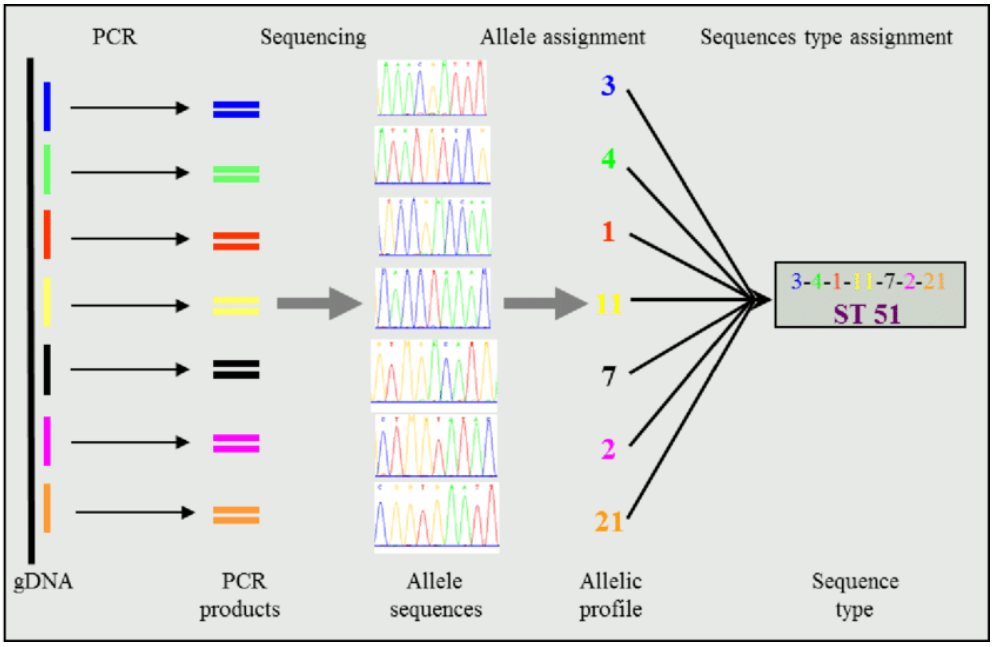


Figure 2: Scheme for multilocus sequence typing.
Adapted from: [40]

1.4.2 Pulse field gel electrophoresis

Pulse field gel electrophoresis (PFGE) was invented in 1984 by Schwartz and Cantor to resolve the issue of comigrating of larger fragments in conventional electrophoresis. As opposed to conventional electrophoresis, where there is only one electrical field, PFGE applies spatially distinct pairs of electrodes, with alternating electrical fields [41]. The technique is currently the "gold standard" method for creating DNA fingerprints, which is used to determine the genetic kinship between two or more bacterial isolates belonging to the same species. Nevertheless, the method also has disadvantages. Different laboratories may apply different protocols and use different parameters for conditions and electrophoresis settings, hence making the DNA fingerprints incomparable.
The method is time-consuming and usually takes 3-4 days, depending on the protocol being used.
A standardized bacterial cell suspension is casted into agarose plugs to improve DNA stability throughout the procedure. The bacterial cells are lysed and treated with a restriction enzyme, which cuts the bacterial DNA at specific recognition sites unique to the enzyme. This creates DNA fragments of various size, depending on the genome of the microbe being analyzed. The

fragments are then separated during the PFGE process, creating specific band patterns on the gel [42, 43]. Figure 3 illustrates the entire PFGE process.

The restriction enzyme commonly used for *K. pneumoniae* PFGE is *XBaI*, which has the following recognition site:

5' T↓C T A G A 3'
3' A G A T C↑T 5'

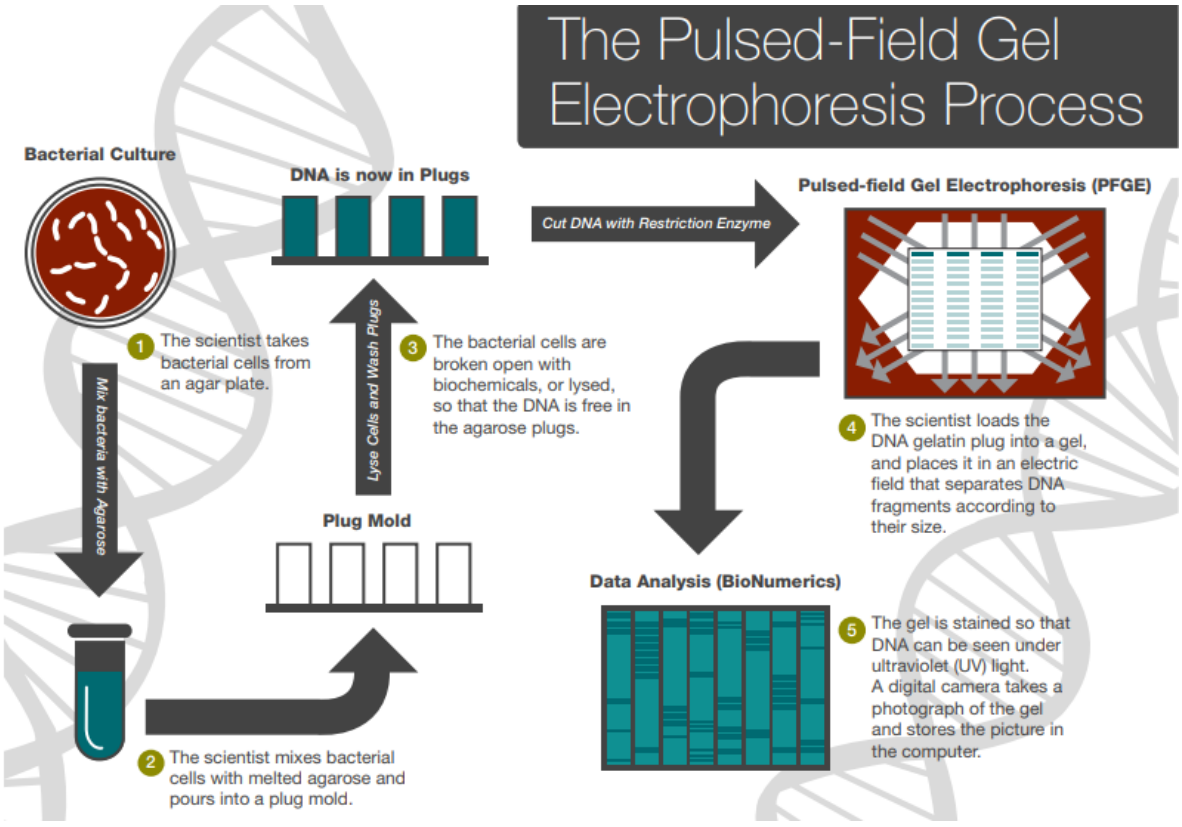


Figure 3: An illustration of all the steps of the PFGE process [44].

Throughout the PFGE process, the DNA molecules are elongated upon application of an electric field, and return to an unelongated state when the electric field is removed. The relaxation rate correlates with the size of the DNA fragment. When the orientation of the

electric field is changed, the DNA will return to the elongated state prior to reorientation, thereby influencing the migration rate of the fragment. This effect allows for separation of larger sized DNA fragments [41]. After the electrophoresis, the fingerprints generated are stained and can be imaged using ultraviolet light. By comparing the number of bands for each isolate, it can be determined whether the isolates are clonal; hence belonging to the same strain, or if they are genetically unrelated. To be considered clonal, isolates need to have a band difference of ≤1 [45].

1.4.3 Phylogenetic analysis

The understanding of the genetic relationship between bacteria is essential for researchers and clinicians to be able to give better treat and stay ahead of the constantly evolving bacteria. Genomes of most bacterial species are highly adaptable and undergoing constant change. Bacteria pass down their genome vertically, but also has the ability to gain genetic material from the environment and other species and organisms through HGT. In addition to this, genes of bacteria are often duplicated or lost. All mechanisms contributing to recombination can make phylogenetic analysis of bacterial genomes complicated and challenging to interpret [46].

A cornerstone in the phylogenetic analysis is the phylogenetic tree. It is a diagram that displays lines of evolutionary descent of different species, organisms or genes from a common ancestor. A phylogeny provides a helpful structure for organizing knowledge of biological diversity, either within an entire group of higher organisms, or just within a single species of i.e. bacteria. A phylogenetic tree is composed of nodes and branches. The nodes represent a common ancestor shared by two or more terminal taxa, and terminal taxa are connected by the branches. The branch corresponds to the common ancestor of all species included in the tree [46, 47]. An example of a phylogenetic tree is shown in Figure 4.

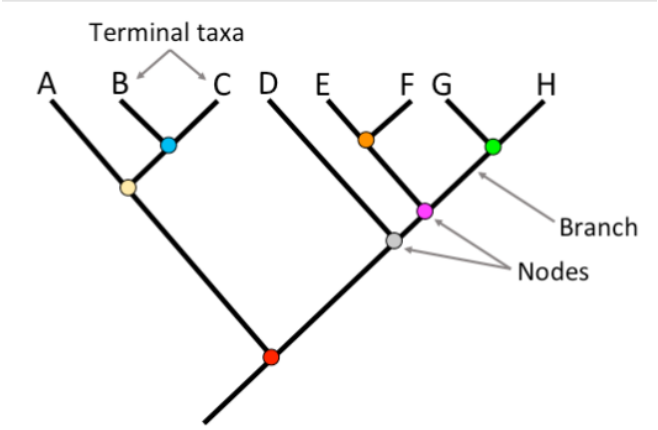


Figure 4: An example of a phylogenetic tree, illustrating the branch, nodes and the terminal taxa. Adapted from: [48]

A tree represents the genetic relationship between i.e. bacterial isolates, and can be made using a number of different mathematical and statistical techniques. In this study, an Unweighted Pair Group Method with Arithmetic Mean (UPGMA) tree, and a Maximum Likelihood (ML) tree was used to display the genetic relationship between the bacterial isolates of the population.

An UPGMA tree is constructed by the use of a simple clustering method that assume a constant rate of evolution. The method requires a distance matrix of the analyzed taxa, which can be calculated from i.e. the multiple alignment of the DNA fingerprints of the isolates. The phylogenetic tree is calculated by taking the two terminal taxa with the smallest genetic distance, A and B, and cluster them together to form a new operational taxonomic unit (OTU), which is referred to as the composite OTU (AB). A new smaller distance matrix is created, with OTU AB, instead of taxa A and B. Subsequently from the new group of OTUs, the terminal taxa with the highest similarity is identified, and this process is repeated until there is only two OTUs left [49, 50]

1.4.4 Whole genome phylogeny

Whole genome (WG) phylogenetic analysis is a direct measure for genetic sequences.
A method for WG phylogenetic analysis is to create a maximum likelihood tree. In the construction of a ML tree, a method of statistical interference is applied. Likelihood provides probabilities of the DNA sequence given a model of their evolution on a particular tree. The more probable the sequence given the tree, the more favorable the tree is [51].
When creating an ML tree, multiple bioinformatics tools will be needed to perform the different steps of the process. Individual tools can be used for each step, or a bioinformatics pipeline can be applied. An example of the process of creating an ML tree from NGS data is using the RedDog pipeline. The program implements a workflow pipeline for short read length sequencing analysis, including mapping of reads, variant detection and analysis of SNPs [52]. The raw reads from a number of sequencing runs are first aligned and mapped against a provided reference, such as a curated reference genome, by Bowtie [53]. An ML tree can be generated based on i.e. SNPs in the core genome, and in the RedDog pipeline, the SNPs are identified by another bioinformatics tool, SAMtools [54]. The tree can then be generated on the basis of the identified SNPs by FastTree [55] to generate the ML tree for the aligned samples.

Advantages of using SNP based approaches, such as ML, for phylogenetic analysis of bacterial isolates, is that it can give a very high resolution, provided that the reference genome used is closely related to the samples. The chances of mismapping is reduced, and the regions present in the reference genome to which the reads are mapped against will increase. When analyzing a set of samples that are closely related, i.e. in the case of an outbreak, it is favorable to use an assembled set of closely related sample as reference.
A disadvantage of SNP-based approaches is that there is low comparability between studies, especially when using different reference genomes. There may also be a difference in threshold settings, such as the parameters for SNP identification. This makes it difficult to reproduce the analyses, unless its reads, settings, used reference genome and pipeline is made publicly available [34].

1.5 Antimicrobial resistance

“Antimicrobial resistance (AMR) is the ability of a microorganism (like bacteria, viruses, and some parasites) to stop an antimicrobial (such as antibiotics, antivirals and antimalarials) from working against it.” This is WHOs definition of antimicrobial resistance [56].
Antimicrobial resistance is becoming a larger and larger threat to global health through contributing to making microorganisms more persistent and less susceptible to standard treatments [56].
The need for knowledge on how to defeat this threat and finding new effective treatments is urgent.

1.5.1 General features of AMR

The global antimicrobial crisis is determined by three factors:
1. The increased incidence of AMR phenotypes amongst microbes is an evolutionary response to the extensive use of antimicrobial agents.
2. The human population is large and globally connected, allowing pathogens access to all of humanity in all environments.
3. The widespread and often superfluous use of antimicrobials by humanity enables the strong selective pressure driving the evolutionary response in the microbial world.
Hence, AMR actually occurs naturally, but the dissemination and emergence of novel resistance mechanisms may have been sped up due to overuse and improper use of antimicrobials in both humans, animals and plants [57].
Two of these factors can be influenced, giving humanity a chance to manage the AMR crisis. Reducing the applied selective pressure, by extensive reduction of the use of antimicrobial agents, may slow down the rapid evolution of virulence factors and AMR [58].

1.5.2 AMR in *K. pneumoniae*

K. pneumoniae is considered a reservoir and source of AMR genes, and several major families of these genes, such as bla_{SHV-1} which is a known ESBL encoding gene [59] and *K. pneumoniae* carbapenemases (KPC) KPC1-KPC7, which confer resistance or decreased susceptibility to most beta-lactams, including carbapenems, which are last-line antibiotics [60].

ESBLs are defined as enzymes produced by specific bacteria, enabling them to hydrolyze extended-spectrum cephalosporins, making the bacteria less susceptible or resistant to beta-lactam antibiotics such as ceftazidime and cefotaxime [61].

ESBLs can be classified according to different classification systems. One divides ESBLs into three classes, $ESBL_A$ which are class A ESBL, $ESBL_M$, which are miscellaneous ESBLs and $ESBL_{CARBA}$, which hydrolytic activity against carbapenems [62]. Table 1 shows an overview of the classes with the most prevalent ESBLs.

Table 1: An overview of ESBL classes, A, M and CARBA.

	$ESBL_A$	**$ESBL_M$[1]**	**$ESBL_{CARBA}$**
Beta-lactamase classes	CTX-M TEM SHV VEB PER	CYM FOX MIR MOX DHA	KPC GES[2] NML SME IMI-1,2
Operational definition	Non-susceptibility to extended-spectrum cephalosporins	Non-susceptibility to extended-spectrum cephalosporins	Non-susceptibility to extended-spectrum cephalosporins and at least one carbapenem

1: $ESBL_{M-C}$ Plasmid mediated AmpC; class C

2: GES-2, -4, -5, -6, -8

Adapted from: [62]

ESBLs can also be classified into four classes , A-D, based on amino acid sequences [63].

1.5.3 Prevalence of AMR in *K. pneumoniae*

AMR and the emergence of MDR *K. pneumoniae* isolates has been recognized as a worldwide problem. AMR reports from the CDC, the UK Department of Health and WHO has identified MDR in *K. pneumoniae* as an urgent threat to public health, due to a high incidence of resistance toward carbapenems and broad-spectrum beta lactams [64-67]. According to a study, the prevalence of AMR in 43 countries in 2015 was 66.9% for third-generation cephalosporin resistant *K. pneumoniae* and 23.4% for carbapenem-resistant *K. pneumoniae.* If this trend in AMR follows the same pace, it is estimated that by 2030 the prevalence of third-generation cephalosporin resistant *K. pneumoniae* will decrease to 58.2%, while carbapenem-resistant *K. pneumoniae* will increase to 52.8% [68].
The European Center for Disease Prevention and Control (ECDC) issued a report in 2017, which revealed that more than one third of *K. pneumoniae* isolates reported to EARS-Net for 2017 was resistant to at least one of the antimicrobial groups under regular surveillance, and that 87.8% of third-generation cephalosporin resistant isolates were ESBL positive. Several countries also reported a carbapenem resistance percentage >10% [31]. Figure 5 reveals the proportion of *K. pneumoniae* isolates resistant to carbapenems in 2017, in 29 European countries.

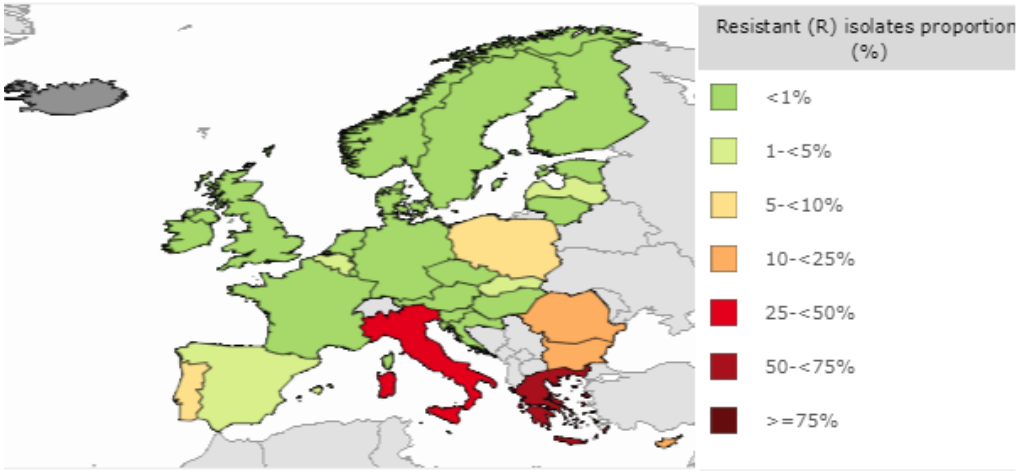


Figure 5: The proportion of *K. pneumoniae* isolates resistant to carbapenems in European countries 2017. Adapted from: [69]

1.5.3.1 In *K. pneumoniae* in Norway

The prevalence of AMR in *K. pneumoniae* in Norway relatively low, but increasing. Figure 6 shows the development of resistance in *K. pneumoniae* in Europe towards third-generation cephalosporins from 2015-2017.

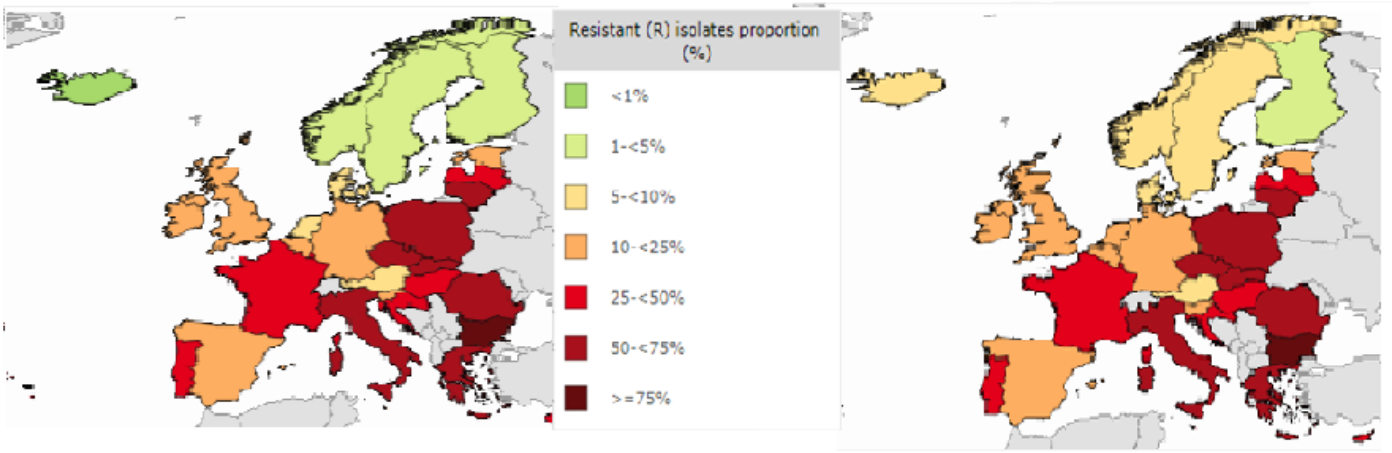


Figure 6: The development of resistance in *K. pneumoniae* towards third-generation cephalosporins 2015 (left) to 2017(right).
Adapted from: [69]

From the data in the ECDC surveillance atlas [69], it can be seen that the proportion of resistance in Norway has increased from 1-<5% to 5-<10%. According to the NORM/NORM-VET report from 2017, resistance towards both third-generations cephalosporins and the emergence of ESBLs has been steadily increasing, but no genetic determinants for carbapenemases was detected. From 2016 to 2017, the proportion of isolates resistant towards ceftazidime and ceftotaxime has increased from 4.8-5.1% and 4.7-5.4% respectively, while the proportion of isolates harboring ESBL-encoding genes has increased from 4.6-5.3%. The most prevalent ESBL group was $bla_{CTX\text{-}M}$ [9].

1.6 Background for DNA sequencing

The history of DNA sequencing is not a very long one, but the timid beginning of an adventure that is DNA sequencing, started when Watson and Crick discovered the three-dimensional structure of DNA in 1953, with the help of Rosalind Franklin and Maurice Wilkin's crystallography data. This field of science has come a very long way since, and is still in continuous development. The major breakthrough for DNA sequencing came in 1977, when Sanger developed the "chain termination" or dideoxy technique, which today is called Sanger sequencing. For several years this method was the most commonly used when sequencing DNA. [70].

Since then, DNA sequencing has come a long way. Next generation sequencing (NGS) provides the ability to sequence millions of small DNA fragments in parallel, as opposed to only a single gene of interest at a time with the Sanger method. In contrast to the Human Genome Project, which applied capillary electrophoresis-based (CE) Sanger technology, took nearly 10 years and cost almost 3 billion dollars, NGS makes whole-genome sequencing (WGS) easy and practical to use, as well as not being extremely costly. Using NGS, a whole human genome can be sequenced in a single day [71, 72].

The use of NGS in microbiology is useful because the genome of a pathogen can provide information that could not be obtained through only phenotypical and visual characterization techniques of the pathogen itself. The genome may hold information about sensitivity to drugs and virulence factors, and phylogenetic relations between bacterial isolates. Contrary to phenotypical methods, SNPs and new genes can be detected using NGS methods. This information can be helpful in tracing the source of an infection outbreak, by determining if the outbreak originates from one bacteria spreading, or if there are several different bacteria that has spread. NGS can be crucial for clinicians in order to provide rapid and efficient care, as well as limiting and stopping an ongoing outbreak based on information of resistance toward different antibiotics and virulence factors, as well as determining if the source of the outbreak is different strains or a clonal outbreak [72].

1.6.1 Illumina Sequencing

Illumina provides a platform for NGS with a high throughput, scalability and speed. Illumina sequencing applies a method called paired-end (PE) sequencing. In PE sequencing, both ends of the DNA fragment are sequenced, and the forward and reverse reads are aligned, creating read pairs. This provides twice the number of reads, as well as providing more accurate read alignment, and the ability to detect indels, which cannot be done with single-read data [73].

The Illumina workflow is divided into four major steps, which are shown in Figure 7. Step 1 is performed manually or by a pipetting robot and steps 2 and 3 are performed on the MiSeq instrument, which is the NGS platform from Illumina used in this study, while step 4 is performed partially on the MiSeq instrument, as well as on computational software.

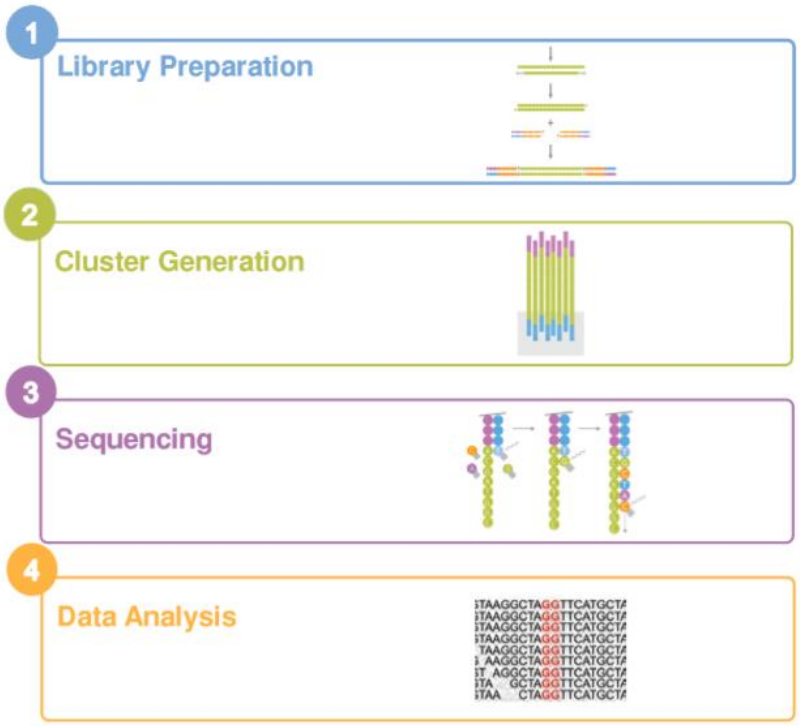


Figure 7: An illustration of the major steps in the Illumina workflow.
Adapted from: [74]

There are many different sequencing kits available for the MiSeq. The sequencing kits mainly differ in read length and amount of DNA input into the reagent cassette. For WGS of bacterial isolates, as performed in this study, the Nextera XT kit was the recommended procedure for the DNA library preparation.

1. DNA library preparation

DNA library preparation is divided into several sub-processes; tagmentation, amplification, clean-up and normalization.

During tagmentation, transposones cut the genomic DNA into fragments, sizing the fragments to a desired length. The size of the inserts normally range between 200-500 bp, but can also be as large as 1000 bp. Simultaneously with the cutting of the DNA, oligonucleotide adapters that contain binding regions for sequencing primers are ligated to the both ends of the DNA fragment. Figure 8 shows an illustration of what happens during tagmentation.

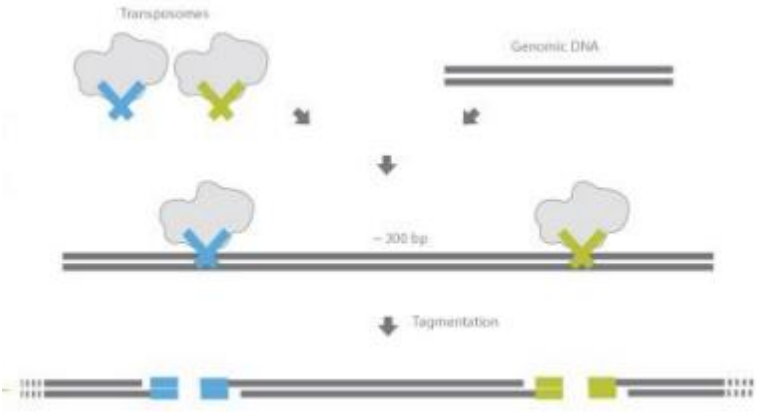


Figure 8: Tagmentation – The transposones fragment the genomic DNA and attach adapters
Adapted from: [75]

After tagmentation, the adapter-ligated DNA fragments are amplified by PCR, creating a DNA library. The adapters being attached are P5 to the 5' end of the fragment, and P7 to the 3' end of the fragment. These adapters are complementary to the oligonucleotides (oligos) on the surface of the flow cell, enabling the fragments to attach to the flow cell during the sequence run, hence playing an essential role for the clustering process of the sequencing run. In the same process, a pair of unique index sequences are added to each DNA fragment. This is called multiplexing. The index sequences makes identification of a single bacterial isolate among a pool of isolates possible. This feature enables for large number of libraries to be pooled and sequences simultaneously, thereby reducing both the cost and time of sequencing multiple samples. The final size of the entire adapter is ~126 bp, ~63 bp on each end of the insert.

Figure 9 illustrates a) the addition of indexes to a fragment and b) a sequencing-ready fragment.

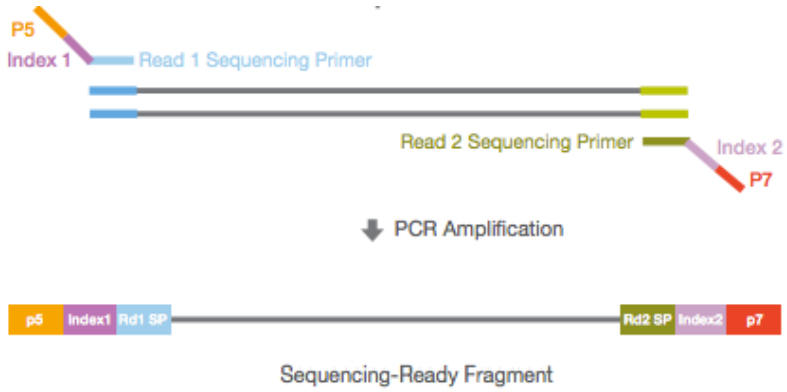
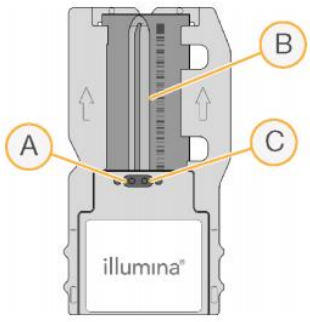


Figure 9: Library amplification – the unique index sequences and P5 and P7 adapters are attached to the partial adapters on each side of the DNA fragment during PCR amplification, creating a sequencing-ready, easily identifiable library.
Adapted from: [75]

When library amplification is completed, the DNA library is cleaned by the help of magnetic beads. During library clean up, fragments of a size larger than 1000-1500 bp and smaller than 200 bp will be removed from the library. Adapters and primer dimers are usually < 200 bp, and will therefore be removed as well. Removing the smaller fragments, such as adapter dimers is especially important because the oligos covering the surface of the flow cell has an affinity to smaller fragments, and adapter dimers form clusters very efficiently. Hence, if such artefacts are present during the sequencing run they will occupy valuable space on the flow cell without creating any useful data. [73-76].

2. Cluster generation by bridge amplification

The process of Illumina sequencing happens on a flow cell. A flow cell is a thin glass slide that is applied to the MiSeq instrument. (The MiSeq flow cell is a random flow cell, which means the lawn of surface bound oligos (P5 and P7) are randomly placed, enabling the clusters to form randomly on the flow cell. The flow cell allows for variable insert sizes and ensures less duplicates, preferably below 10%, and ~ 2% for a very successful DNA library preparation. This is due to the chemistry and imaging of the random flow cell, which gives every strand the same possibility to form a cluster). An illustration of a MiSeq flow cell is shown in Figure 10.



A Outlet Port
B Imaging Area
C Inlet Port

Figure 10: An Illumina MiSeq flow cell
Adapted from: [77]

The entire process of bridge amplification is illustrated in Figure 11. The first step of bridge amplification is the loading of a single-stranded sequencing library, with complementary adapters, into the flow cell. As the libraries "flow" across the lawn of oligos, individual molecules will hybridize to complementary oligos. The bound libraries are then extended at the 3' by polymerases, creating a double-stranded DNA. The double stranded DNA will then be denatured and the original template will be discarded, while the newly synthesized strand will be anchored to the flow cell surface through a covalent attachment (1, Figure 11). A process called priming will then take place as the opposite end of a ligated fragment bends over and "bridges" to another complementary oligo on the surface of the flow cell. The hybridized primer will then be extended by a polymerase, giving rise to a double stranded bridge (2, Figure 11). This cycle will be repeated; thereby creating what is called clusters (3, Figure 11). A cluster is a clonal grouping of template DNA bound to the surface of the flow cell, and the process of bridge amplification will continue until the cluster has ~ 1000 copies. The double-stranded DNA bridges will subsequently be denatured resulting in two copies of covalently bound single-stranded DNA templates (4, Figure 11). The revers strands will be discarded, leaving only forward DNA strands (5, Figure 11). The free 3' of the DNA strand and the oligo primers are blocked to avert unwanted primer annealing, and the forward strand is now ready for sequencing.

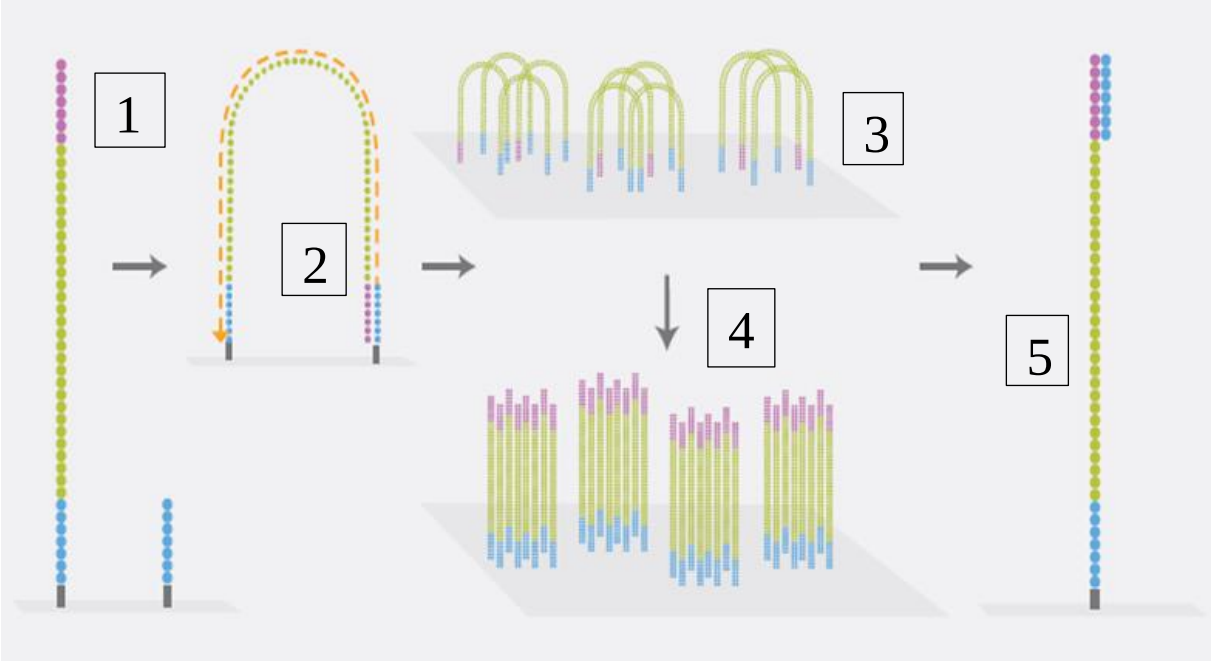


Figure 11: Cluster generation by bridge amplification
Adapted from: [78]

3. Sequencing by synthesis

The MiSeq platform applies a four-channel sequencing by synthesis (SBS) chemistry for the sequencing process. Each of the four bases of the DNA are assigned a fluorescent-labeled terminator-bound dideoxynucleotide (ddNTP); G blue, C is green, A is yellow and T is red. These, along with primers and polymerase are introduced into to the flow cell. The primer will bind to the adapter sequence-annealing site on the forward single-stranded DNA, and by this, the SBS can start. A polymerase will start incorporating a fluorescent nucleotide, the ddNTP, to the template strand. The terminator on the nucleotide will act as a "reversible terminator", preventing further polymerization of the template strand. After ddNTP incorporation, the fluorescent dye will be identified through laser excitation. A camera images the emitted light, and the base will later be called on the unique wavelength of the light that was emitted. After base calling, the fluorophore and terminator will be enzymatically cleaved off allowing the incorporation of a new ddNTP and a new base calling. This process continues in a cycle until all bases are called. Figure 12 illustrates the wavelengths of the emitted light of each fluorescent label, and an image of what the final base-calling image could look like for each base. The images are actually only shown in black and white, but is here illustrated with the respective colors of the fluorescent labels.

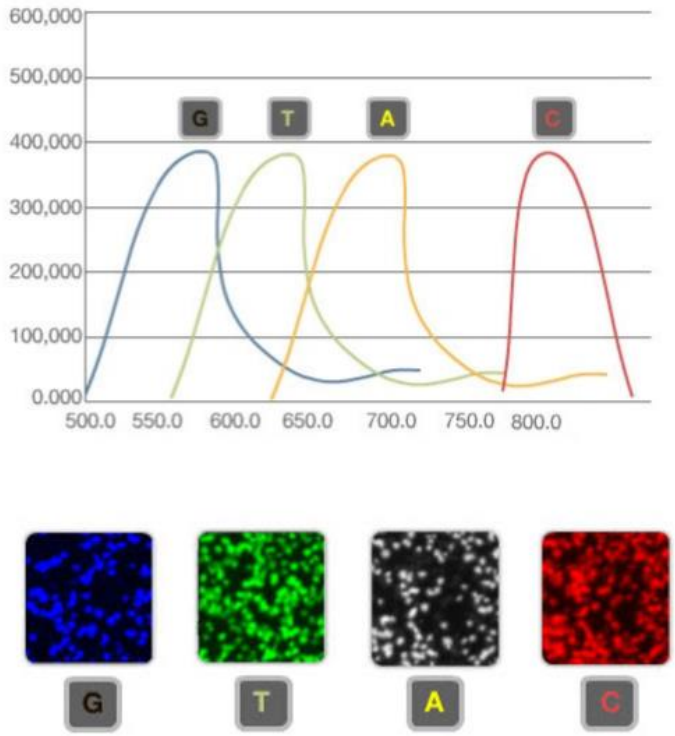


Figure 12: Each DNA base emits a unique wavelength of light during base calling. During each cycle each clusters appears in one of four images.
Adapted from: [74]

As mentioned, the Illumina MiSeq system uses PE sequencing. Both ends of the DNA library are sequenced and aligned in read pairs (R1 and R2). After a strand has been sequenced, the sequence product (R1) is removed, leaving a forward strand on the flow cell surface, ready for a new cycle of bridge amplification. However, after denaturing of the bridge, the original forward template is cleaved off and discarded, leaving a reverse strand to be sequenced in a new round of SBS, generating read 2 (R2). During the sequencing run, the indexes are read after R1 and before R2 [73, 74].

4. Data analysis by Illumina

Illumina provides different types of software for handling and analyzing the data generated throughout a sequencing run. In this study, the Control Software on the MiSeq instrument was used to view flow cell images, intensities and base calls. The Sequence Analysis Viewer Software was also used, and provides metrics of extraction, quality, error, the tiles and control, as well as files on run info, run parameters and thumbnails [79].

2. Aims of the study

The main aim of this study was to describe genetic epidemiology of human blood culture isolates of *K. pneumoniae* from a selected number of Norwegian hospitals between March 2017 and August 2018.

Research questions:

- What is the multilocus sequence type (ST) distribution in the study population of ~ 1000 blood culture isolates?
- Are there significant geographical differences in ST-distribution with a particular view on Helse-Vest region comprising Sogn og Fjordane, Hordaland and Rogaland County?
- What are phylogenetic relationships within the dominating sequence type?
- How does multilocus sequence typing compare with PGFE and whole genome based phylogeny methods such as core genome analysis or SNP analysis?

3. Materials and methods

3.1 Materials

3.1.1 Collection of bacterial isolates

In this study 722 *K. pneumoniae* blood culture isolates collected through the

NORKAB study from March 2017-August 2018 were included.

NORKAB criteria:

Inclusion:

- Patient to be ≥ 18 years with *Klebsiella* non-oxytoca-bacteremia, identified with local laboratory routine methods, like MALDI-TOF MS or VITEK.

Exclusion:

- Under age (<18 years).

- *Klebsiella* non-oxytoca-bacteremia within the past 8 weeks; where the strain had the same phenotypical characteristics.

- Request to be reserved from participation in the study [10].

Participating laboratories are shown in Table 2. Some laboratories serve more than one hospital.

Table 2: A list of participating institutions and # of samples analyzed throughout this study.

Laboratories	Region	# of samples analyzed
The University Hospital of Northern Norway	North	23
Nordland Hospital	North	4[1,4]
St. Olav's Hospital	Middle	39
Helse Førde	West	15
Haukeland University Hospital	West	97
Stavanger University Hospital	West	82
Sørlandet Hospital	South	12
Vestfold Hospital	South	115
Akershus University Hospital	East	97
Oslo University Hospital	East	116[2]
Innlandet Hospital	East	65
Vestre Viken HF	South	46[3]
Østfold Hospital	South	11[4]

1: Lofoten Hospital was not included

2: Lovisenberg Diaconal Hospital was not included

3: Vestre Viken – Bærum Hospital was not included

4: Isolate collection is not complete

Adapted from [10].

The data and strain collection for the NORKAB study was not complete when this thesis was finished, so only 722 isolates were included in this thesis.

The laboratories at Levanger Hospital and Møre og Romsdal Hospital, both region Middle, and Fonna Hospital from region West does not participate in the NORKAB study.

3.1.2 Commercial kits

Table 3: A list of commercial kits used, and their purposes

Commercial kit	Function	Supplier	City, country
MagNA Pure 96 DNA and Viral NA Small Volume Kit [80]	Purify bacterial DNA	F. Hoffmann-La Roche AG	Basel, Switzerland
Quant-iT™ 1X dsDNA HS Assay Kit[81]	DNA quantification	Thermo Fischer Scientific	Waltham, MA, USA
Qubit™ 1X dsDNA HS Assay Kit[82]	DNA quantification, using Qubit Fluorometer	Thermo Fischer Scientific	Waltham, MA, USA
Nextera® XT Library Preparation Kit[83]	Prepare sequencing libraries for small genomes	Illumina	San Diego, CA, USA
Nextera® XT Index Kit v2 Set A/Set B [83]	Provides unique identification to each sample	Illumina	San, Diego, CA, USA

AMPure XP Beads[84]	Library cleanup, removing contaminants	Beckman Coulter	Brea, CA, USA
PhiX Control Kit v3[85]	Control library for sequencing runs	Illumina	San Diego, CA, USA
MiSeq Reagent Kit v2/v3[86]	Sequencing reagents in pre-filled, ready-to-use cartridges	Illumina	San Diego, CA, USA
Agilent High Sensitivity DNA Kit for 2100 Bioanalyzer System[87]	Sizing and quantification of DNA	Agilent Technology	Santa Clara, CA, USA

3.1.3 Solutions for PFGE

Table 4: Reagents used for *XBaI*-PFGE

Solution	Contents	Origin
TE-buffer	10 mM Tris, 1 mM EDTA, pH 8.0	In-house
Cell suspension buffer	100 mM Tris, 100 mM EDTA, pH 8.0	In-house
Cell lysis buffer	50 mM Tris, 50 mM EDTA, pH 8.0, 1% Sarcosyl	In-house
10xTBE-buffer	89 M Tris, 89 M Borate, 32 mM, EDTA, pH 8.0	In-house

Table 5: *XBaI* restriction enzyme master mix used in PFGE of *K. pneumoniae*, in accordance with the PulseNet central Standard Operating Procedure for PulseNet PFGE of various Enterobacteriaceae.

Reagent	Volume (per sample)	Supplier
dH_2O	173.0 µL	In-house
CutSmart Buffer, (restriction enzyme buffer) (10x) [88]	20.0 µL	New England Biolabs, Ipswich, UK
BSA[1] (10 mg/mL) [89]	2.0 µL	Promega, Madison, WI, USA
***XBaI* (10 U/µL) [90]**	5.0 µL	New England Biolabs, Ipswich, UK
Total	200 µL	

1: Bovine serum albumin – BSA

3.2 Methods

3.2.1 Cultivation/over-night inoculation of bacterial isolates

Microbank freezing vials containing *K. pneumoniae* were collected from a -80ºC freezer and thawed in room temperature. Blood agar plates were barcoded. A glass bead coated with bacteria was collected from the Microbank vial using a 10-µL sterile inoculation loop, and was dispersed onto the blood agar. A new 10-µL sterile inoculation loop was then used to streak the bead on the agar in a four-quadrant dilution pattern. An illustration of the method is shown in Figure 13 [91].

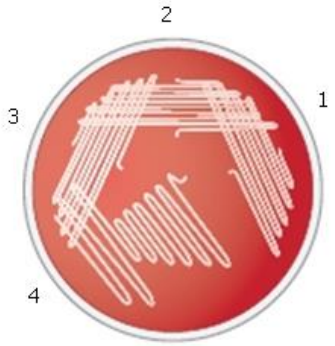


Figure 13: An illustration of the four-quadrant streak method.
Adapted from: [91]

The seeded plates were incubated overnight at 35ºC.

3.2.2 Extraction of bacterial DNA

For sequencing of a bacterial genome, *K. pneumoniae*, the DNA had to be extracted. This was accomplished by the use of the MagNA Pure 96 system (Roche, Basel, Switzerland). The instrument is fully automated and high-throughput, providing DNA purification by the use of magnetic glass particles (MGPs). It can purify a broad range of sample materials, and up to 96 samples in one run [92]. The kit used for this study was the MagNA Pure 96 DNA and Viral NA Small Volume Kit (Table 3).

Bacterial colonies from the over-night blood agar plates were collected and dispersed in ID-labeled Eppendorf tube containing 500 µL sterile saltwater, using a sterile loop. The tubes were vortexed thoroughly to obtain a homogenous bacterial suspension, and 200 µL of each sample was pipetted into a MagNA Pure Processing Cartridge. The Pathogen Universal 200 3.1 protocol was used. The complete protocol is included in Appendix A.

3.2.3 Measurement of DNA concentration

To determine the concentration of purified DNA from the *K. pneumoniae* isolates the fluorescence-based Quant-iT^{TM} 1X dsDNA HS Assay Kit (Table 3) was used with the Spark® Multimode Microplate Reader (Tecan, Männedorf, Switzerland), and the SparkControl Magellan data analysis tool (Tecan, Männedorf, Switzerland) to analyze the results [93]. For smaller numbers of samples the Qubit^{TM} 1X dsDNA HS Assay Kit (Table 3) was used with the Qubit® 4 Fluorometer (Thermo Fischer Scientific, Waltham, MA, USA).

Before measuring the DNA concentration, 50 µL of 10 mM Tris-HCl, pH 7.5, was added to all samples for dilution, and to obtain a sample volume and DNA concentration that was in a range optimal for the Hamilton Microlab STAR pipetting robot to accept.

The Quant-iT^{TM} 1X dsDNA HS:

The assay was equilibrated to room temperature, before 200 µL of the Quant-iT 1X dsDNA working solution was pipetted into all wells to be used in a black microplate (Tecan, Männedorf, Switzerland) using a multichannel pipette. After being vortexed and spun down, 10 µL of each of the eight Quant-iT 1X dsDNA HS Standards were pipetted into 16 out of 96 wells to create a duplicate of the standards. Lastly, the plate containing the DNA sample was placed in a magnetic stand to avoid pipetting out magnetic glass particles sustained from the DNA extraction, and 10 µL of each DNA sample was pipetted into separate wells of the microplate. The microplate was sealed and shaken for ~ 1 minute at 800 rpm using a plate shaker (Eppendorf, Hamburg, Germany). After shaking, the microplate was incubated for ~2 minutes at room temperature, and placed into the Spark Multimode Microplate Reader for measurement.

The following steps were executed:

1. A sample sheet was created by pressing **"Create/Edit a sample ID list",** and scanning the barcodes for each sample, as well as selecting a blank and 16 standards. The sample sheet was save as with a unique name (i.e. date_name_cons).

2. From the start page of the software: Press **"Start Measurement".**

3. The method **"Quant-iT ds DNA High-Sensitivity single"** was chosen.

4. Press **"Insert Sample List"**, and choose the sample list created.

5. Press **"Start".**

The results were transferred to an Excel sheet. An overview of the workflow from the protocol is shown in Figure 14. The complete protocol can be found in Appendix A.

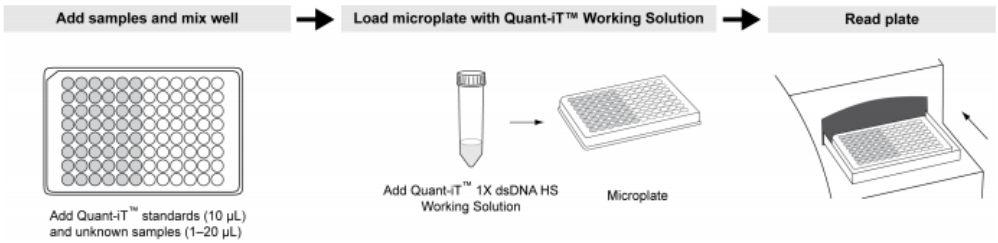


Figure 14: The workflow of the Quant-iT^{TM} 1X dsDNA HS Assay Kit
Adapted from: [81]

Qubit^{TM} 1X dsDNA HS Assay Kit:

The assay was equilibrated to room temperature, before 10 µL of Qubit standard was added to labeled, thin wall, clear, 0.5 mL PCR tubes (Thermo Fischer Scientific, Waltham, MA, USA) equal to the number of DNA samples to be measured + two standards. From the plate on the magnetic stand, 10 µL of DNA samples were then transferred to each tube, before adding the Qubit 1X dsDNA 1X buffer to all tubes, obtaining a final volume of 200 µL. All tubes were vortexed and incubated at room temperature, covered from daylight, for ~ 2 minutes, before being measured on the Qubit® 4 Fluorometer. The results were plotted into an Excel sheet. An overview of the workflow from the protocol is shown in Figure 15. The complete protocol can be found in Appendix A.

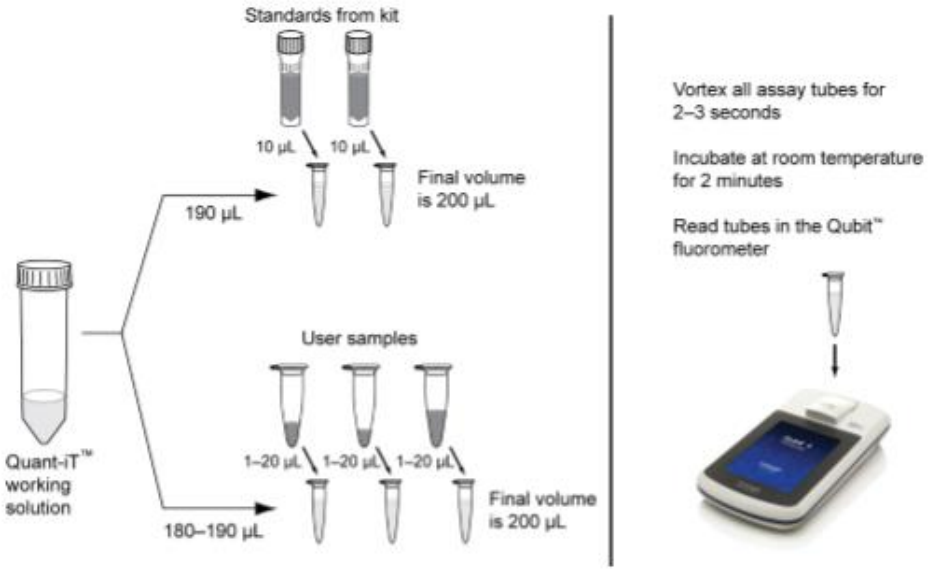


Figure 15: The workflow of the Qubit™ 1X dsDNA HS Assay Kit
Adapted from:[82]

Normalization of DNA concentration:

To perform a library preparation of the bacterial DNA using the Nextera XT protocol from Illumina, all samples needed to have a concentration of 0.2 ng/µL. The normalization was performed on the Hamilton Microlab STAR pipetting robot using 10 mM Tris-HCl, pH 7.5 as a dilution buffer. The Excel file containing the measured concentrations from the Spark Multimode Microplate Reader was formatted to a .csv file and put into the software. ~ 100 µL of each DNA samples was pipetted into a midi-plate (Thermo Fischer Scientific, Waltham, MA, USA) and placed into the pipetting robot.

3.2.4 Nextera XT library preparation using Hamilton Microlab STAR

Library preparation with the Nextera XT protocol on the Hamilton Microlab STAR pipetting robot:

The entire process of the library preparation of the DNA was performed by the Hamilton Microlab STAR (ML STAR) pipetting robot, except for the PCR, which was completed on a thermal cycler (Eppendorf, Hamburg, Germany). In each library preparation, 32 samples were prepped. For the library preparation, the Nextera® XT Library Preparation Kit (Table 3) was

used. Library preparation of the DNA sample occurs according to the following steps shown in the flowchart in Figure 16. The entire process can be completed in one sitting, or a partially prepared library can be stored at -25°C to -15°C for up to seven days, after the different safe stopping points, also shown in Figure 16.

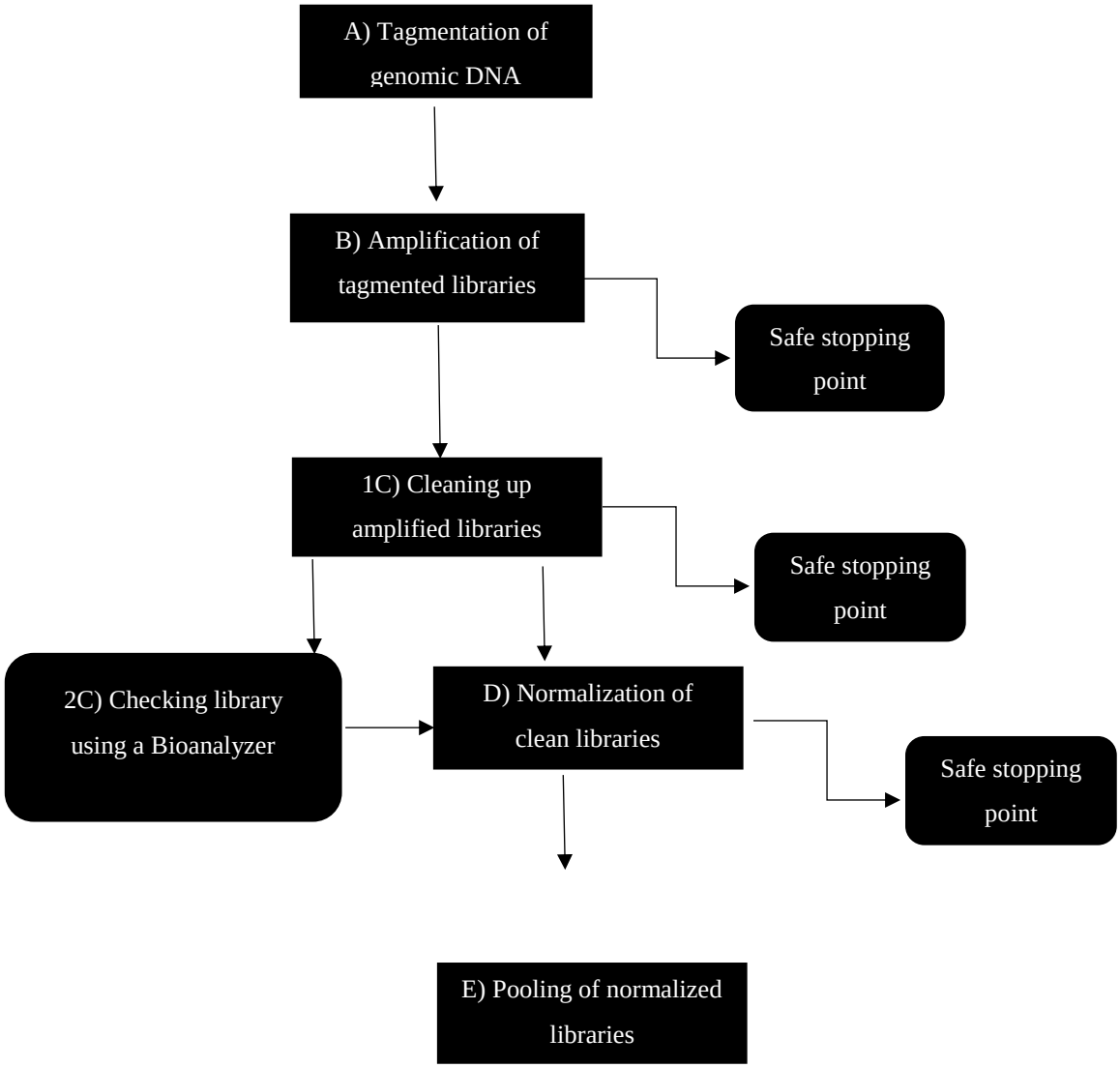


Figure 16: A flowchart of the different steps of the Nextera XT library preparation of bacterial DNA on the Hamilton Microlab STAR.

For this study the library preparation method was done in two parts, first tagmentation of the genomic DNA and amplification of the tagmented DNA, then cleanup of the amplified library, normalization and pooling of the library consecutively.

A detailed description of the various steps performed in the Nextera XT library preparation is included in Appendix B.

3.2.5 Sequencing using the Illumina MiSeq System

Before the pooled libraries could be added to the reagent cartridge for sequencing, they had to be diluted and denatured, and 1% PhiX was added.

Diluting and denaturing libraries:

The hybridization buffer (HT1) was thawed in a water bath, or overnight in at 4^{o}C. The total volume of a Nextera XT library pool was 600 μL, so depending on the DNA input used, X μL were removed from the 600 μL of HT1, and X μL of vortexed, pooled libraries was added in a new tube. The mixture was gently vortexed and centrifuged at 280 x g for 1 minute before being put in a 98^{o}C preheated incubator for exactly 2 minutes. The tube was directly placed on ice for a minimum of 5 minutes. 6 μL of 1% PhiX was then added to the DNA library pool as a control. The complete protocol is included in Appendix A.

Prior to sequencing of the pooled DNA libraries, the MiSeq instrument had to be prepared. This includes creating a sample plate and a sample sheet and setup of the different components necessary for a successful run. The complete protocol is included in Appendix A.

Creating a sample plate:

A sample plate is a map of each well of a plate, containing sample information, as well as the library preparation type applied and the positions of the unique dual index pair for each sample in the plate. A unique experiment name was given to each sample plate.

To create a sample plate the Illumina Experiment Manager Software (IEMS), version 1.15.1, was applied. A layout of the software is shown in Figure 17.

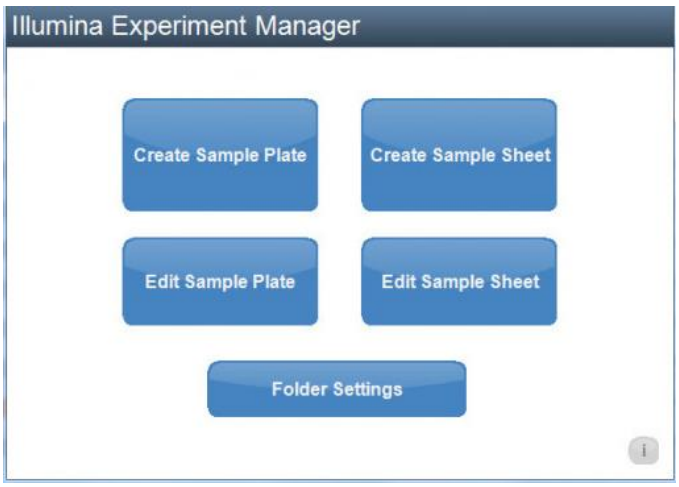


Figure 17: A layout of the IEMS [94].

According to the Illumina Experiment Manager Software guide [95], the following steps were performed:

1. Select “Create Sample Plate” (Figure 17).

2. Select index kit – Nextera XT Index Kit; Kit A and B were used in this study.

3. Name sample plate: A unique name, (i.e. date_initials_run#)

4. Select index scheme-2-libraries are dually indexed.

5. Select plate tab: A display of the layout of a 96-well plate with columns A-H and rows 1-8.

6. Insert sample IDs, select the correct i5 and i7 indexes.

7. Select Finish and save.

Creating a sample sheet:

A sample sheet contains all the information needed to correctly analyze the data from the sequencing run.

According to the Illumina Experiment Manager Software guide [95] (Appendix A), the following steps were performed:

1. Select “Create Sample Sheet” (Figure 18).

2. Select Instrument: “MiSeq”.

3. Select appropriate application; in this study, we applied “FastQ Only”.

The steps are shown in Figure 18.

Figure 18: The layout of the steps to create a sample sheet in Illumina sequencing systems [96].

4. Insert the kit ID into the Reagent Cartridge Barcode field by scanning the reagent cartridge or the kit box.

5. Select applied library preparation method: Nextera XT

6. Select index adapters – Nextera XT Index Kit (Kit A or Kit B). The software has a default setting of dual index.

7. Enter the Experiment name, Investigator name, Description and Date.

8. Select read type: Paired end.

9. Choose the number of Cycles read: Enter wanted number of reads.

- For the MiSeq Reagent Kit v2, 24 samples: 151 reads.

- For the MiSeq Reagent Kit v3, 32 samples: 351 reads.

10. To select samples to include in the sample sheet: Press “Select Plate”, and upload the previously made sample plate for the specific sequencing run.

11. Press “Add Selected Samples” to transfer the samples form the sample plate to the sample sheet. Double check that all samples are present and given the correct indexes.

12. Complete and save the sample sheet by pressing “Finish”.

Run set-up using MiSeq Control Software:

The MiSeq Control Software (MCS) is pre-installed on the MiSeq sequencing system, and guides the user through pre-sequencing set-up. In addition, MCS quality statistics throughout the entire sequencing run, allowing for constant quality monitoring. The preparation of the instrument consists of three main parts:

1. Cleaning and loading the flow cell:

The flow cell was taken out of the container by the base of the plastic cartridge, and gently rinsed with filtered 18 Ω water to remove excess salts. The flow cell was carefully dried using lens paper until completely dry. The previously used flow cell was removed from the MiSeq, and the new was carefully loaded into place. When the flow cell was approved, the MCS would move on to the next step.

2. Loading reagents:

The PR2 bottle was taken out of the refrigerator and inverted to mix. The hatchet to the reagent compartment of the instrument was opened and the zipper-handle was gently raised. The wash bottle was removed and the PR2 bottle was loaded into the MiSeq, before the zipper-handle was lowered again. The instrument would now read the PR2 radio frequency identification (RFID) to ensure the reagent is compatible with the selected kit, before moving on to the next step.

3. Preparation and loading of the reagent cartridge.

The reagent cartridge was thawed either in a water bath for ~ 1 hour, or at 4^{o}C overnight. It was inverted 10 times to ensure that the reagents were thawed sufficiently and was free from precipitates. The foil covering the reservoir labeled “Load Samples” was carefully pierced using a sterile 1 ml pipette tip, before the 600 µL of diluted and denatured DNA libraries containing 1% PhiX was loaded into the reagent cartridge. The MCS would now control the barcode on the reagent cartridge and match it to the sample sheet previously created, as well as the flow. When all the parameters were approved, the sequencing process was started by pressing “Start Run”.

3.2.6 Quality control post sequencing

In Illumina sequencing, the MCS provides real time analysis (RTA) throughout the larger parts of the sequencing run. The 25 first cycles of the sequencing run align to the PhiX reference genome, and the MCS can provide quality statistics for important parameters for the run. These parameters include Cluster density, % Passing filter (% PF), % Q30 score, estimated yield and error rate. The v2 and v3 MiSeq reagent kits have slightly different specifications of acceptable values for the various parameters [97], which are given in Table 6.

Table 6: Specifications of acceptable values for the various quality parameters in MiSeq v2 and v3 reagent kits, respectively. The v2 kit has a read length of 2x150 bp, and the v3 kit has a read length of 2x300 base pairs (bp).

Quality statistics	Optimal value	Description
Cluster density	865-965 k/mm^2 / 1200-1400 k/mm^2	The number of clusters per square millimeter on the flow cell (k/mm^2).
% PF, pair-ended reads	As high as possible, ~ 24-30 million / 44-50 million	Percentage of clusters passing the Illumina chastity filter.
% Q30	>80% bases higher than Q30 / >70% bases higher than Q30	The average percentage of bases greater than Q30[1].
Estimated yield	4.5-5.1 gigabases (Gb) / 13.5-15 Gb	Projected number of bases called for the sequencing run.
Error rate	No specifications	The rate of mis-matches between sequencing data and PhiX, the reference genome [2]

1: Q30 = prediction probability of an error in base calling 1:1000 = the chance that a base calling is correct is 99.9%.

2: PhiX contains three SNP's, which will always give an error-rate; the error rate can never be 0.

Quality assessment using Sequence Analysis Viewer (SAV) post sequencing:

When the sequencing run is completed an output directory can be opened in the SAV [79, 98]. The SAV uses several files as input, but the ones saved and used in this study was the FastQ files, RunInfo.xml, RunParameters.xml, Sample Sheets and InterOpt. InterOpt contains the following files: Extraction metrics, Quality metrics, Error metrics, Tile metrics, Extended tile metrics, Corrected intensity metrics, Image metrics, Index metrics and Empirical Phasing metrics. The SAV has an Analysis Tab, which includes six panels, displaying different plots. Four of the panels was used in this study: the Flow Cell Chart, the Data by Cycle Plot, the

Data by Lane plot and the Q-Score Distribution Plot.

An overview of the charts is shown in Figure 19.

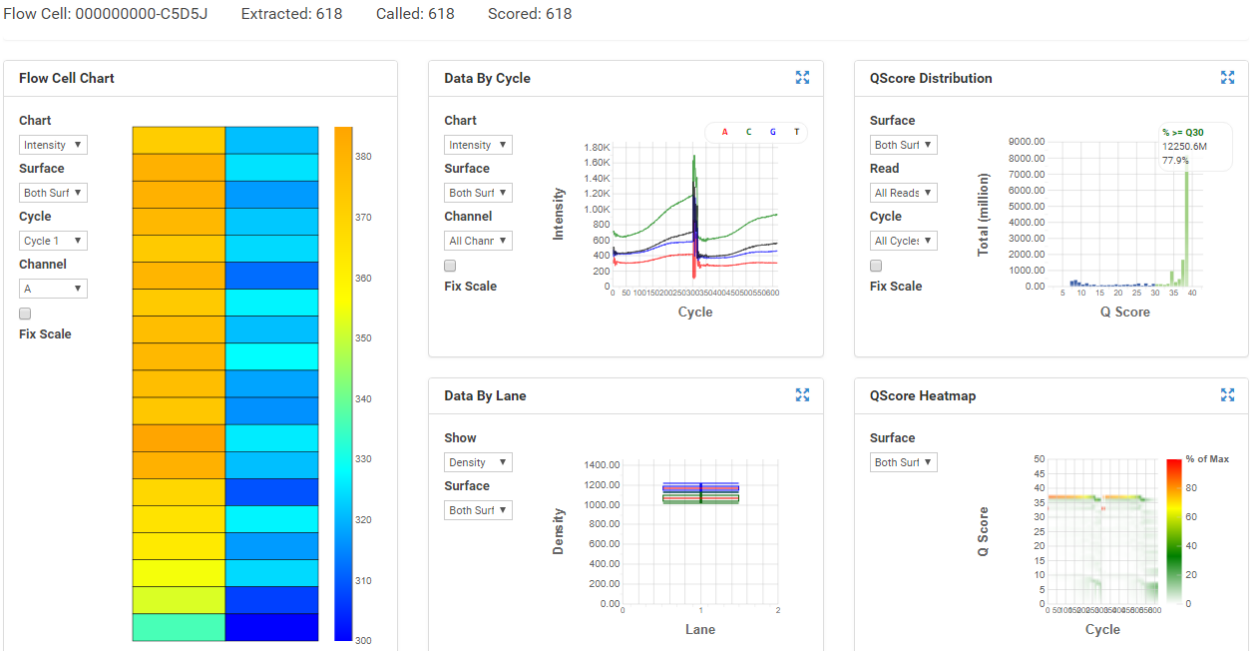


Figure 19: A screenshot of the overview of the analysis charts from SAV.

The Flow Cell Chart: The flow cell chart displays color-coded quality metrics per tile for the entire flow cell, and can oversee different color metrics. An example of a flow cell chart can be seen to the left in Figure 19.

The Data by Cycle Plot: The data by cycle plot depicts the progression of quality metrics throughout a run. The most important metrics for this study was % base and %Q>30.

- The %Base shows the percentage of clusters for which the selected base has been called.
- The %Q>30 shows the percentage of bases with a quality score of 30 and higher. This is one of the charts created after the 25th cycle.

An example of a %Q>30 can be seen in Figure 20.

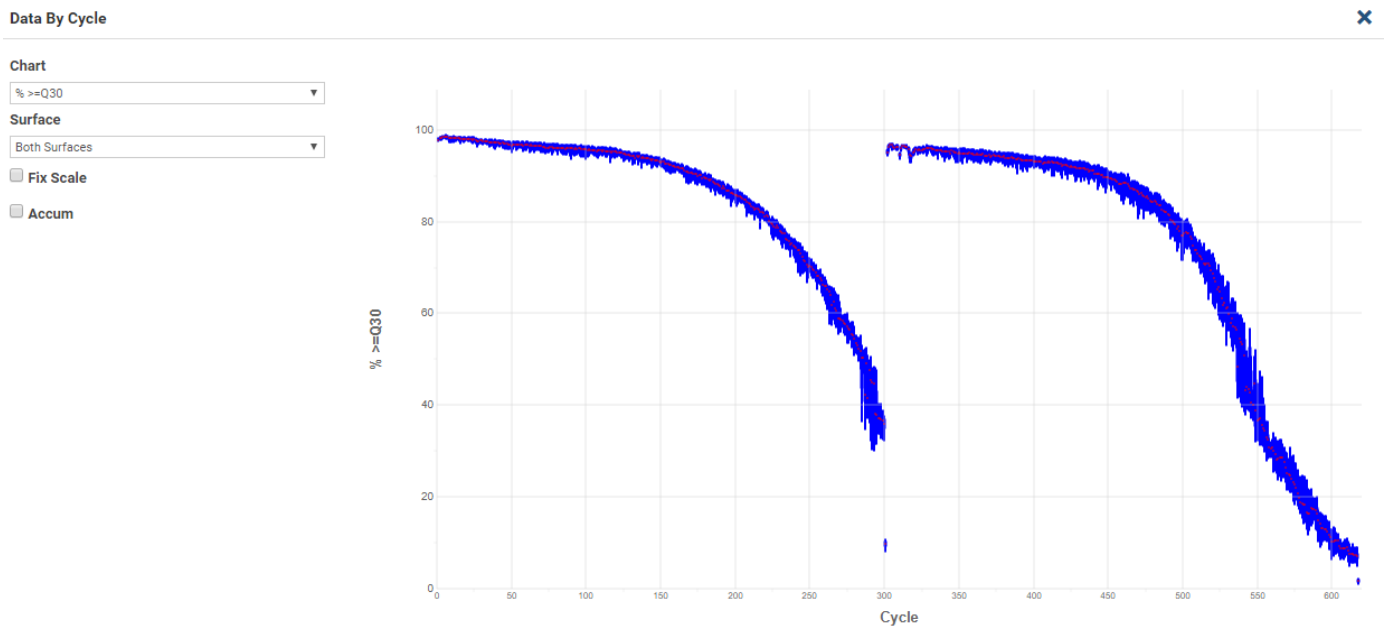


Figure 20: A screenshot form the SAV showing the %Q>30 plot. The curve is satisfactory.

The Data by Lane Plot: The data by lane pane shows plots that allow viewing of quality metrics per lane. In the data by lane pane, the important plot for this study was the density plot. The density plot shows the density of clusters for each tile in thousands per mm^2. The cluster density was monitored by comparing the raw cluster density and the clusters passing filter. An example of a density plot can be seen in Figure 21.

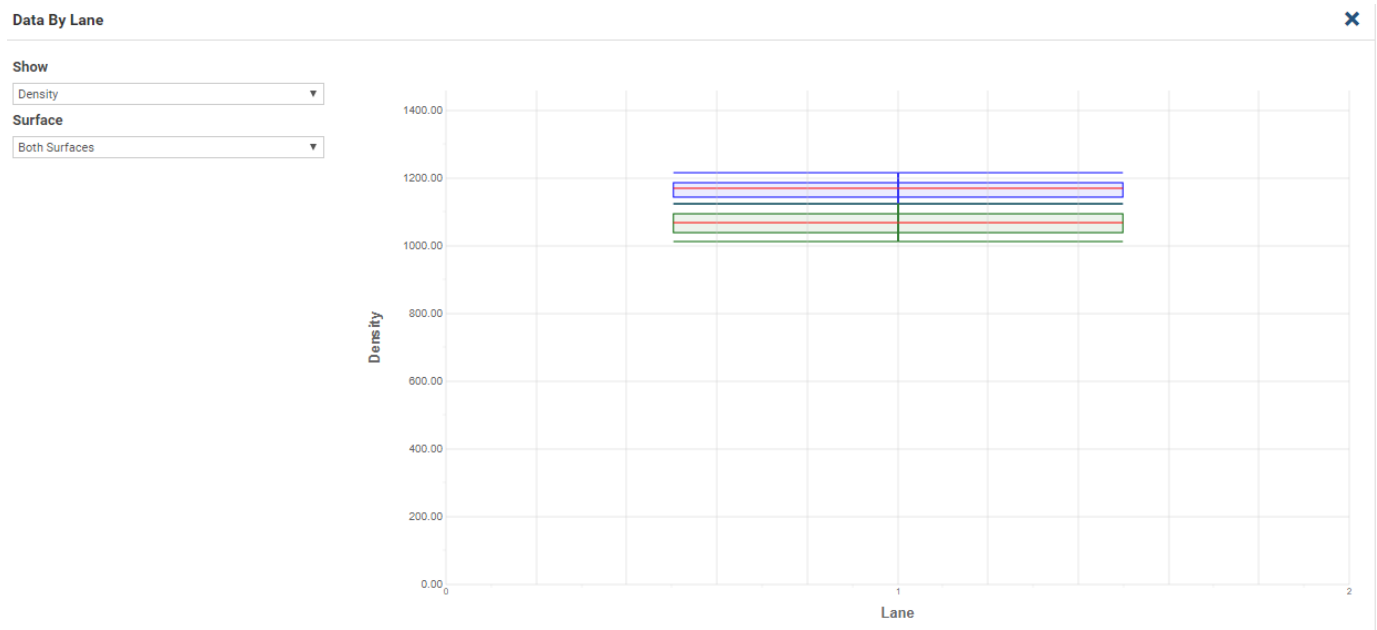


Figure 21: A screenshot from the SAV showing the density plot. The plot is satisfactory, and the cluster density is here 1167, which is satisfactory for a v3 kit.

The Q-Score Distribution Plot: The Q-score distribution plot allows the viewing of the number of reads by quality score. Only reads that pass the quality filter are included.
An example of a Q-Score distribution plot can be seen in Figure 22.

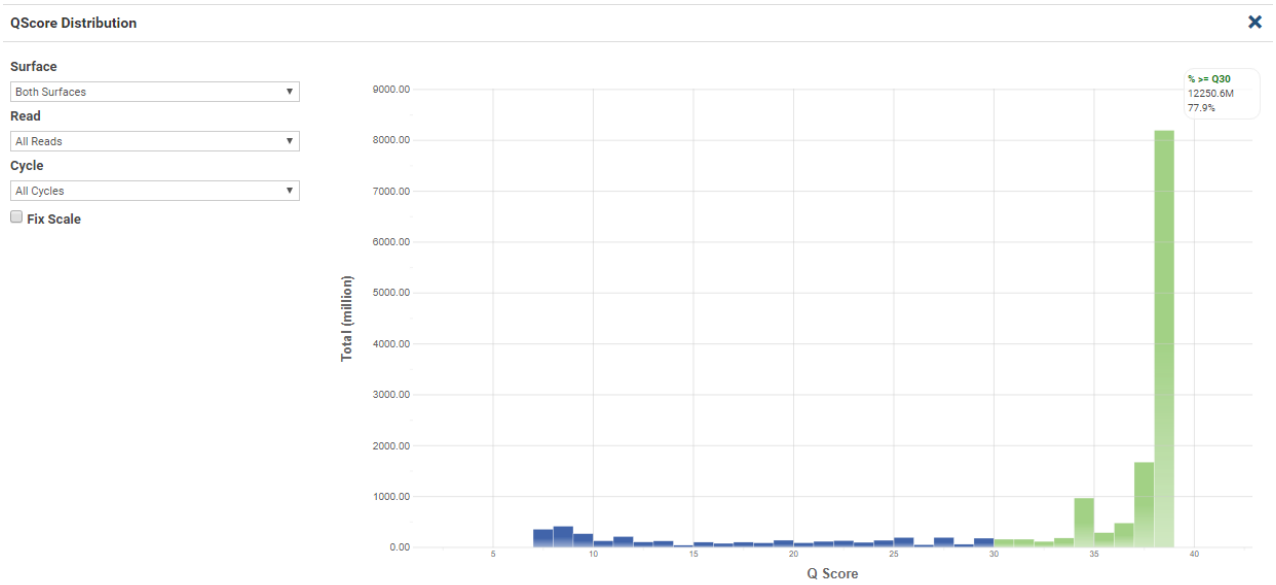


Figure 22: A screenshot form the SAV showing a Q-Score Distribution plot. The Q30 score is here over 70%, which is satisfactory for a v3 kit.

Table 7 shows the different metrics and analysis results needed for a run to be acceptable in this study.

Table 7: An overview of satisfactory results and metrics used for quality assessment in this study, with the expected outcome and observations.

Analysis plot	**Metrics**	**Expected result/observation**
Flow cell chart	Intensity	A range within 200
Data by Cycle plot	%Q>30	The Q30 per cycle will decrease; significantly after cycle 250
	%Base	Initially the four bases should be stable throughout the cycles, not crossing each other
Data by Lane plot	Density	The raw cluster box (blue) and the cluster passing filter box (green) (Figure 21) should be as close as possible. Each box should also be as narrow as possible
The Q-Score Distribution Plot	%Q30	An overall percentage of Q30> 70% for a v3 kit and Q30> 80% for a v2 kit

3.2.7 Computational analyses

1: MiSeq output:

The raw data generated on a MiSeq during a sequencing run are stored in FASTQ format, which consist of short-read sequences, sequence identifiers and quality scores. These files are ultimately stored in a result output folder on the MiSeq instrument.

2: Quality assessment of raw data

The quality of the short-read sequences was evaluated with FastQC v0.11.7 [99]. A quality check was run on the raw sequence data, generating an overall quality assessment of the sequence run, highlighting any issues that may have occurred during library preparation, the sequencing process or on the samples themselves. Sample problems may be duplicates, adapters, PCR primers, low-quality reads or additional contaminants. The software gave an output report on each individual sequenced isolate, indicating whether the results were within or outside the set parameters, such as base statistics, per base GC content, per sequence GC content, per base N content, sequence length distribution, sequence duplication level, overrepresented sequences and Kmer content.

In addition to FastQC, MultiQC version 1.4 [100]was used. The MultiQC software gives an overview necessary to detect failing samples, and identify groups of samples behaving in an irregular manner. There are certain specifications in FastQC for *K. pneumoniae*, which are shown in Table 8.

Table 8: Specification for different features in FastQC.

Feature	Specification
%GC content	56-57, can accept 58 if the rest of the parameters are overall optimal
N content	0
Adapter content	0

Trim Galore v0.6.1 [101] was used to trim the raw reads. Trim Galore uses FastQC and CutAdapt to apply quality metrics and adapter trimming to FASTQ files. Low-quality bases (<Q score 20), adapter sequences and 1 bp are all trimmed off the 3’ end. Lastly, read-pairs that are less than 20 bp long are removed.

3: De novo assembly

To reconstruct the genome from the raw data output from the Illumina sequencing, *de novo* assembly was performed with Unicycler v0.4.4 [102], which uses SPAdes v3.13.0 [103], an assembly tool for prokaryotic genome assembly, for assembly, and Pilon polishing v1.22 [104] for assembly optimization. The assembly output contains FASTA-files with multiple contigs of varying length.

During genome assembly, the sequence reads are assembled into contigs; a stretch of continuous sequence created by aligning overlapping sequencing reads [73]. When there are no overlapping reads, "gaps" of unknown length will occur. This is why there are several contigs instead of only one per replicon. An example of *de novo* assembly is shown in Figure 23.

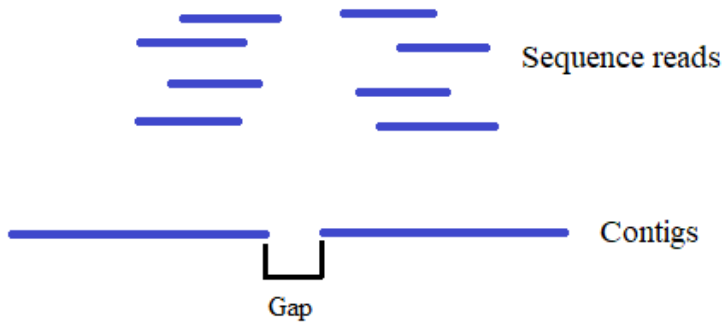


Figure 23: An illustration of *de novo* assembly; short-reads are aligned creating contigs.

4: Quality assessment of the genome assembly

To assess the quality of the genome assembly, Quast v4.6.3 [105] was applied, providing an output statistic of several metrics.

SeqDepth v1 [106] was used to align the trimmed FASTQ reads to their assembled contigs. SeqDepth uses Burrow-Wheler Alignment v0.7.17-r1188 (BWA-MEM) [107] and Picard Tools v2.17.8 [108] to create a Binary Alignment Map (BAM), SAMtools v1.7 [54]is subsequently used to calculate the general sequence depth. The depth of each position, as well as mean sequence depth is determined by SAMtools by dividing the mean of all positions by the total genome size.

The most important metrics for this study is shown with a description and specification in Table 9 [109].

Table 9: Quast metrics and sequencing depth with a description and specification for this study.

Metric:	Description:	Specification
# of contigs (>= 0bp)	The total # contigs in the assembly	In this study <700 was used. Few but long contigs are favorable. <1000 contigs indicate good quality, but <100 contigs are generally realistic for organisms with a genome of 5-6 Mb [110]
Total length of sequence (>=0 bp)	The total number of bases in the assembly	~ 5-6 Mb
Largest contig	The length of the largest contig in the assembly	Keep 135 000 and larger, with the proviso that the rest of the parameters are overall optimal
N50	The length of which the collection of all contigs of that length or longer cover at least half the assembly	As high as possible, >15 000 bp normally indicates good quality, but a minimum size of 30 000 bp is generally preferred [110]
L50	The total number of contigs equal to or longer than N50	As low as possible, 40 or higher was excluded
%GC	Percentage of nitrogenous bases on a DNA that is guanine or cytosine	For *K. pneumoniae* the %GC is 56-58 [111]
Sequencing depth	Sequencing depth is calculated as the average read depth of each position across the total genome	A minimum depth of 30x is usually preferred [110], but depends on usage[112]. In this experiment 16 was the lowest accepted value.

5: Phylogenetic analyses for the sequenced *K. pneumoniae* population

To investigate the phylogenetic relationships between the 722 sequences, and to visualize the collection of sequenced strains, the RedDog v1beta.10.3 [52] pipeline was applied with the raw reads as input, and the well documented and manually curated *K. pneumoniae* (HS11286) strain (GenBank accession: NC_016845.1 [113]), as the reference genome. A subpopulation of ST107 isolates (n=36) was selected for further investigation with RedDog, and due to a lack of public ST107 genomes – a local isolate (NK-07) was used as the reference genome.

RedDog applies the read alignment tool Bowtie2 v2.2.5 [53] (with the setting: sensitive-local mapping) to map the isolates' reads against the provided reference. It then uses SAMtools

v1.7 [54] to identify the SNPs. RedDog then runs FastTree v2.1.7 [55] to generate an ML tree for the aligned isolates. To visualize the core chromosomal SNP tree, mcs18_main.Rmd (https://github.com/marithetland/msctrees/blob/master/mscs18_main.Rmd) was used for the 722 isolates. To visualize the core genome SNP trees, msc18_st107.Rmd (https://github.com/marithetland/msctrees/blob/master/mscs18_st107.Rmd) was used for the 66 ST107 isolates, msc18_st107_pfge.Rmd, (https://github.com/marithetland/msctrees/blob/master/mscs18_st107_pfge.Rmd) was used for the 35 ST107 isolates subjected to PFGE, and msc18_speciesmatch.Rmd, (https://github.com/marithetland/msctrees/blob/master/mscs18_speciesmatch.Rmd) was used for the weak species match tree.

6: Species identification and multi locus sequence typing.

The software Kleborate v 0.3.0 [114] was in this study used to screen *K. pneumoniae* genome assemblies for species and STs, as well as AMR genes.
Kleborate reports the species with the closest match by using Mash [115] to compare the assembly to a set of curates *K. pneumoniae* assemblies from NCBI [116]. A Mash distance of ≤0.1 gives a strong species match, and a Mash distance between > 0.1 – ≤ 0.03 gives a weak species match.

Multilocus sequence typing (MLST) was used to determine the allelic profiles of the population of *K. pneumoniae* isolates. The technique is a tool to characterize isolates of microbial species with the help of internal sequences of seven housekeeping genes. The various alleles of each individual housekeeping gene present in a bacterial isolate determines the distinct allelic ST for that exact isolate. Imprecise ST calls are indicated with –nLV in Kleborate, where n is the number of loci that does not match with the reported ST. Kleborate uses the STs as defined by Institut Pasteur BIGSbd scheme [36], with the assembled contigs as input.

Isolates not assigned to an ST were submitted to the Institut Pasteur to be checked against The Klebsiella Pasteur MLST sequence definition database [36] to assign new STs.

7: Analysis of resistance genes using Basic Local Alignment Search Tool (BLAST).

Kleborate uses BLAST to query its provided database (https://github.com/katholt/Kleborate/blob/master/kleborate/data/ARGannot_r2.fasta). If resistance genes were indicated with "*" (imprecise nucleotide match) or "?" (incomplete coverage) in Kleborate, they were further analyzed using BLAST [117], to see if the sequences could be found with complete matches in other databases.

3.2.8 Pulse field gel electrophoresis

Thirty-six isolates of ST107 were selected from the population of *K. pneumoniae* isolates, and subjected to PFGE for genotyping. The 36 isolates were selected from large hospitals, which were preferably complete with regard to inclusions in the study.

The isolates were analyzed using the operating procedure for PulseNet PFGE of various Enterobacteriaceae [118]. All solutions applied are listed in Table 4 and Table 5.

1. *XBaI* - PFGE

1. Inoculate 12 isolates on blood agar and incubate for 20-24 hours T 35ºC.

2. Suspend bacterial isolates in 2.0 mL cell suspension buffer at a concentration of 3-4 McFarland.

3. Mix 200 µL of cell suspension with 10 µL of 20 mg/mL Proteinase K (Roche, Mannheim, Germany), then add 200 µL of 1% SeaKem Gold agarose (Lonza, Basel, Switzerland), mix briefly and transfer to PFGE plug molds.

4. After solidification, lyse the plugs at 55ºC for 1.5-2.0 hours with agitation in 5.0 mL cell lysis buffer, with 25 µL of Proteinase K added.

5. Wash the plugs twice with 10-15 mL pre-heated dH_2O (55ºC) for 10-15 minutes at 55ºC, then four times with 10-15 mL pre-heated TE-buffer (55ºC) for 10-15 minutes at 55ºC.

6. Store washed plugs at 4ºC in TE-buffer.

7. After storage, wash plugs twice with pre-heated TE-buffer before use.

8. Cut 1.5 mm pieces of the plugs and incubate in a 200 µL of a 1:10 diluted Cutsmart™ restriction buffer at room temperature for 10-15 minutes.

9. Incubate the plugs at 37ºC for 1.5-2.0 hours in XBaI restriction enzyme master mix (see Table 5).

10. Prepare 1.0% SeaKem Gold agarose gel and equilibrate at 55ºC.

11. Pour the equilibrated agarose into a gel frame and place the gel comb. After solidification, remove the gel comb.

12. Fill the electrophoresis chamber with 2.0-2.2 L of 0.5 x TBE-buffer. Calibrate the pump to circulate one liter per minute, and set temperature to 14^{o}C.

13. Incubate the plugs in 200 µL 0.5 x TBE-buffer for five minutes at room temperature.

14. Place the plugs into the well of the gel and add lambda ladder (New England Biolabs, Ipswich, UK). Seal wells with agarose.

15. Run the gel with the settings presented in Table 10.

16. Dye the gel with GelRed solution (Biotium, Fremont, CA, USA) for 30-40 minutes, and then de-stain with dH_2O for 1.0-1.5 hours with water change every 20 minutes.

17. Depict the gel.

Table 10: PFGE program parameters for *K. pneumoniae* used at Stavanger University Hospital, provided by PulseNet Central for various Enterobacteriaceae.

Parameters	Value
Pulse time	1-20 s
Total run time	21 H
Voltage	6.0 v/cm (200 V)
Angle	120^{o}
Buffer temperature	14^{o}C
Buffer	0.5 x TBE

The gel was imaged in ChemiDoc XRS+ (BioRad, Hercules, CA, USA).

2. Analyzing the PFGE gels using the BioNumerics Software

The gel images was analyzed using the BioNumerics software (Applied Maths, St. Martens-Latem, Belgium) [119]. The software allows for linking fingerprint data to isolates, as well as comparing and calculating dendrograms for several different experiments at the same time.

1. Strips: The lanes on the gel are defined and adjusted to each fingerprint. Background subtraction and spot removal is performed to improve the quality of the image.
2. Curves: Width is set to extract densitometric curves used to filter the data to exclude background and noise (smoothing).
3. Normalization: The fingerprints are aligned by pattern recognition using internal reference bands.

4. Bands: First assigns bands by an automatic band search, then manually remove or assign new band where that is needed.
5. Comparison: The three normalized gels are compared in the Comparison Window. A dendrogram was created from the band relations by using the Dice coefficient with 4.0% position tolerance. Clustering analysis was performed by the unweighted pair group method with arithmetic mean (UPGMA).

2. Criteria for evaluating PFGE restriction patterns

The analysis of the PFGE fingerprint patterns in this study was based on the guidelines given in "Interpreting Chromosomal DNA Restriction Patterns Produced by Pulse-Field Gel Electrophoresis: Criteria for Bacterial Strain Typing" [45].

To determine if the isolates subjected to PFGE were genetically related, the generated DNA fingerprint patterns were interpreted after the criteria in Table 11.

Table 11: Criteria for interpreting PFGE patterns

Category of genetic relatedness	Typical # of band differences compared with other isolates
Indistinguishable	0-1
Closely related	2-3
Possibly related	4-6
Unrelated	≥7

Adapted from: [45]

The criteria are reliable if there are at least 10 bands detected. Isolates with indistinguishable patterns are considered clonal.

4. Results

4.1 Quality assessment of sequencing raw data output and assembly

The raw data output generated from the sequencing of the 722 isolates was assessed with FastQC. After *de novo* assembly, the contigs were consecutively quality checked using Quast. Isolates with raw data displaying a low quality in both FastQC and Quast were re-sequenced. See Appendix C for further details. The results from the final quality assessment is shown in Table 12.

Table 12: The results from the post-assembly quality assessment of the 722 sequenced bacterial isolates.

Parameter	Range	Mean value	Standard deviation	Median value
# of contigs (>=0bp)	36-331	83	51.5	90
Largest contig (Mb)	0.13-1.9	0.58	0.24	0.54
N50 (Mb)	0.03-0.98	0.23	0.10	0.23
L50	2-45	9.9	6	8
Sequence depth	16.4-115.6	58.5	18.4	61
GC content (%)	56-58	57.4	0.24	57.4
Total length (Mb)	4.9-6.0	5.4	0.17	5.4

As seen in Table 9 in Methods, the preferred minimum size of N50 is 30 000 bp. < 1000 contigs indicates good quality, and with a genome of 5-6 Mb size, < 100 contigs is generally realistic.

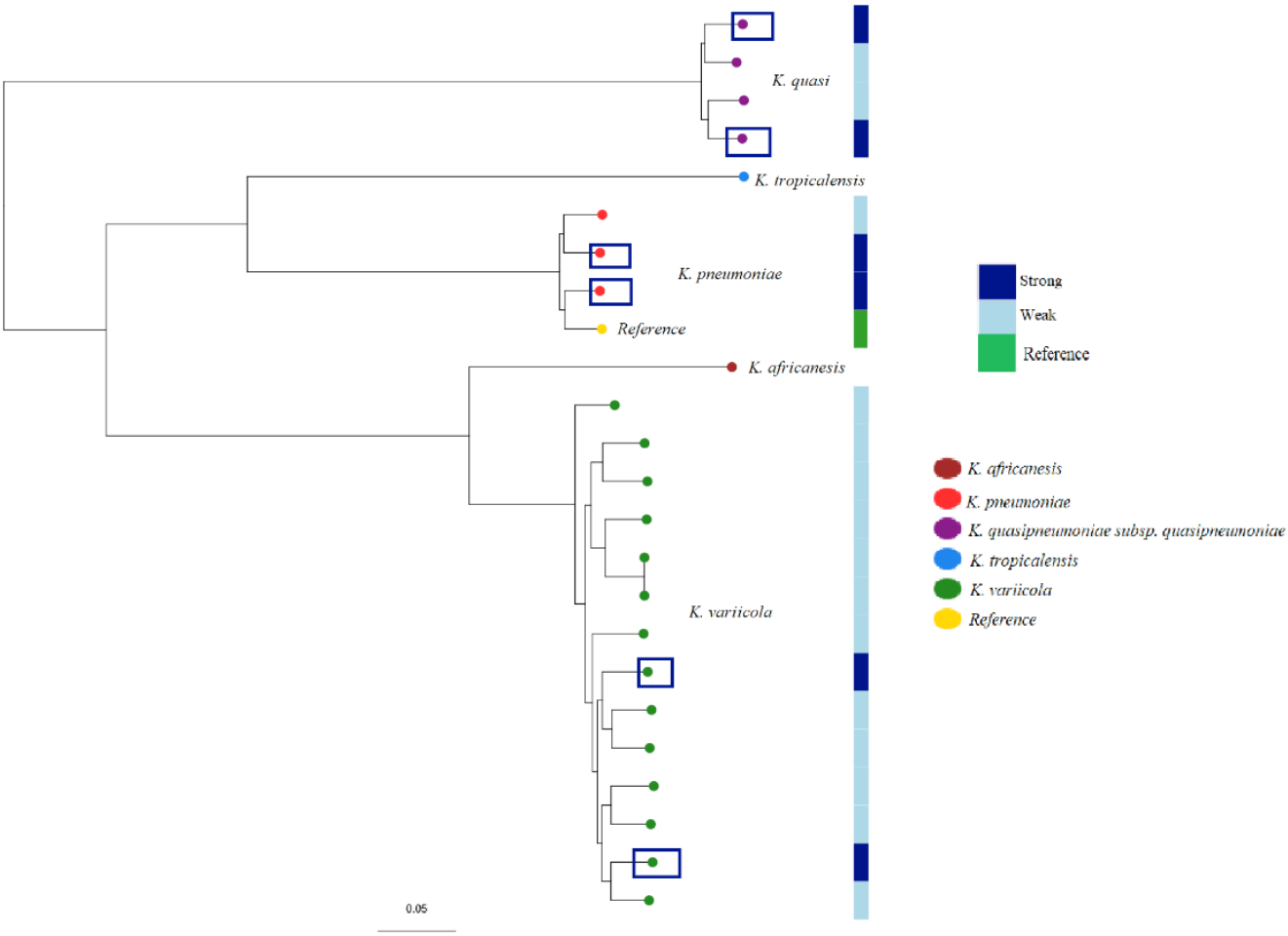
4.2 *K. pneumoniae* population description

Species identification of the *K. pneumoniae* sensu lato isolates (n=722) was determined using Kleborate. Table 13 displays the species distribution and strong/weak species identification matches in the population. The most prevalent species in the population was *K. pneumoniae* sensu stricto (n=566), and *K. variicola* (n=120) was the second most prevalent species.

Table 13: Species distribution in the *K. pneumoniae* sensu lato population.

Species	Strong match	Weak match	Total #
***Klebsiella pneumoniae* sensu stricto**	565	1	566
Klebsiella variicola	107	13	120
Klebsiella quasipneumoniae subsp. similipneumoniae	24	0	24
Klebsiella quasipneumoniae subsp. quasipneumoniae	10	2	12
Total number	**706**	**16**	**722**

For 16 strains Kleborate made a weak species identification match, which may indicate a novel lineage or a hybrid between multiple *Klebsiella* species. Figure 24 shows a core genome tree of the weak matches, with a strain reference for each species for comparison, including the recently described phylogroups *K. variicola subsp. tropicalensis* and *K. africanesis* [23].



nucleotide substitutions per site.

The reference strains for each species are illustrated in dark blue tiles and dark blue outlining; the weak matches are illustrated by light blue tiles and the reference by a yellow tile. All species with a weak match are cluster with the corresponding species references.

4.3 Detected antimicrobial virulence genes and resistance determinants

All virulence genes and genetic resistance determinants reported were identified using Kleborate.

The siderophore combinations and capsular types of the 722 isolates is shown in Table 14. RmpA was present in 17 strains, and rmpA2 in 9 strains in total. Eight isolates had both rmpA and rmpA2, and 4/8 was the high-risk associated ST23 or ST86.

Table 14: An overview of isolates with siderophore combinations and strains with rmpa1 or rmpa2.

Siderophore combinations	Number of isolates	Strains with rmpA	Strains with rmpA2
No virulence loci	570	0	0
Only yersiniabactin	124	2	0
Yersiniabactin or colibactin, or colibactin only	2	0	0
Aerobactin and/or salmochelin only (without yersiniabactin or colibactin)	11	3	4[1]
Aerobactin and/or salmochelin with yersiniabactin (without colibactin)	4	4	2[2]
Yersiniabactin, colibactin and aerobactin and/or salmochelin	11	8	3[1]
SUM	722	17	9

1: Three isolates had both rmpA and rmpA2

2: Two isolate had both rmpA and rmpA2

ESBL-encoding genes were found in 50 isolates. All belong to either the bla_{SHV} or the $bla_{CTX\text{-}M}$ families. Two isolates harbored both a bla_{SHV} and a $bla_{CTX\text{-}M}$ allele, with the combinations of $bla_{SHV\text{-}1}$ and $bla_{CTX\text{-}M\text{-}15}$, and $bla_{SHV\text{-}11}$ and $bla_{CTX\text{-}M\text{-}15}$. An overview of the prevalence of the ESBL-encoding genes can be seen in Table 15.

Table 15: Overview of the prevalence of the ESBL-encoding genes in the population

bla_{SHV}	**$bla_{CTX\text{-}M}$**
$bla_{SHV\text{-}1}$ (n=1)	$bla_{CTX\text{-}M\text{-}1}$ (n=1)
$bla_{SHV\text{-}2}$ (n=1)	$bla_{CTX\text{-}M\text{-}3}$ (n=1)
$bla_{SHV\text{-}2a}$ (n=2)	$bla_{CTX\text{-}M\text{-}14}$ (n=2)
$bla_{SHV\text{-}40}$ (n=2)	$bla_{CTX\text{-}M\text{-}15}$ (n=34)
$bla_{SHV\text{-}11}$ (n=8)	**SUM: 38**
SUM: 14	

Among the fifty-two detected ESBL-encoding genes, three genes were listed with a "?", which indicates a partial match. Twenty-four genes were listed with a "*", which indicates an imprecise allele match. These particular genes were further analyzed with BLAST. The results from the BLAST analysis is shown in Table 16.

Table 16: Results from BLAST analysis of the genes indicated with incomplete coverage, imprecise matches, or both in Kleborate.

Gene	Explanation	Conclusion
$bla_{CTX\text{-}M\text{-}15?}$ (n=1)	Allele divided over two contigs	$bla_{CTX\text{-}M\text{-}15}$
$bla_{SHV\text{-}2?}$ (n=1)	100% nucleotide identity and gene coverage to accession #[1]: MF402903.1	$bla_{SHV\text{-}2}$
$bla_{SHV\text{-}12*?}$ (n=2)	100% nucleotide identity and gene coverage to accession # MF402908.1	$bla_{SHV\text{-}2a}$
$bla_{SHV\text{-}101*}$ (n=8)	1: 1/8 100% nucleotide identity and gene coverage to accession # CP014123.1.	1: $bla_{SHV\text{-}1}$
	2: 7/8 Are identical. No 100% match to any given reference in BLAST. 1 SNP from both $bla_{SHV\text{-}101}$ and $bla_{SHV\text{-}1}$ in various reference-genomes.	2: A bla_{SHV}-variant, but could not be specified.
$bla_{SHV\text{-}13*}$ (n=15)	1: 8/15 100% nucleotide identity and gene coverage to accession # CP032175.1	1: $bla_{SHV\text{-}11}$

	2: 3/15 No 100% match to any given reference in BLAST or Kleborate. 2 SNP from bla_{SHV-13} in Kleborate, 1 SNP from bla_{SHV-11} in BLAST, accession # CP032175.1	2: A bla_{SHV}-variant, but could not be specified
	3: 1/15 No 100% match to any given reference in BLAST or Kleborate. 2 SNP from bla_{SHV-13} in Kleborate, 1 SNP from bla_{SHV-11} in BLAST, accession # CP032175.1, but another SNP than the three above.	3: A bla_{SHV}-variant, but could not be specified
	4: 1/15 No 100% match to any given reference in BLAST or Kleborate. 2 SNP from bla_{SHV-13} in Kleborate, 1 SNP from bla_{SHV-11} in BLAST, accession # CP032175.1, but with another SNP than the four above	4: A bla_{SHV}-variant, but could not be specified
	5: 2/15 No 100% match to any given reference in BLAST or Kleborate. 4 SNP from bla_{SHV-13} in Kleborate, 3 SNP from $bla_{SHV-225}$ in BLAST, accession # NG062297.1	5: A bla_{SHV}-variant, but could not be specified

1: Accession numbers are given as an identifier to any public sequence, here obtained from GenBank.

4.4 *K. pneumoniae* sensu lato phylogeny

4.4.1 Multilocus sequence typing

Sequence types of the *K. pneumoniae* isolates in the population were determined using Kleborate. The diversity of STs found in this population is high, with 378 different STs among the 722 isolates. Table 17 shows the ST distribution and the most prevalent STs in the population. Three hundred and thirty STs occur only one or 2 times. The seven most dominating STs (n≥12) are listed individually in Table 17. The remaining ST types are given in ranges from 1-4 isolates per ST and 5-9 isolates per ST in Table 17.

Table 17: An overview of the STs found in the population and their prevalence.

	Number of isolates	**% in the population**
ST107	67	9%
ST20	23	3%
ST37	20	3%
ST45	18	2%
ST307	12	2%
ST25	12	2%
ST26	12	2%
STs with 5-9 isolates (13 different STs)	81	11%
STs with < 5 isolates (207 different STs)	296	41%

Novel STs and ST locus variants

Kleborate detected 178 isolates with LVs, with 147 different LVs, constituting 25% of the population. LV are imprecise ST calls, and a given number indicates the number of loci that does not match with the reported ST. The range of LVs in this population was 1LV-4LVs. Three isolates were not assigned to an ST by Kleborate, indicating novel loci combinations. These were submitted to the Insitut Pasteur to be assigned new STs. The new STs were ST4009, ST4010 and ST4011.

4.4.2 Core chromosomal SNP analysis

The core chromosomal SNP tree, shown in Figure 25, was made to illustrate the distribution of species, the most prevalent STs and the ESBL-encoding gene within these, that were identified in the *K. pneumoniae* population investigated in this study. The most prevalent STs with >5 isolates are colored in grey the tree.

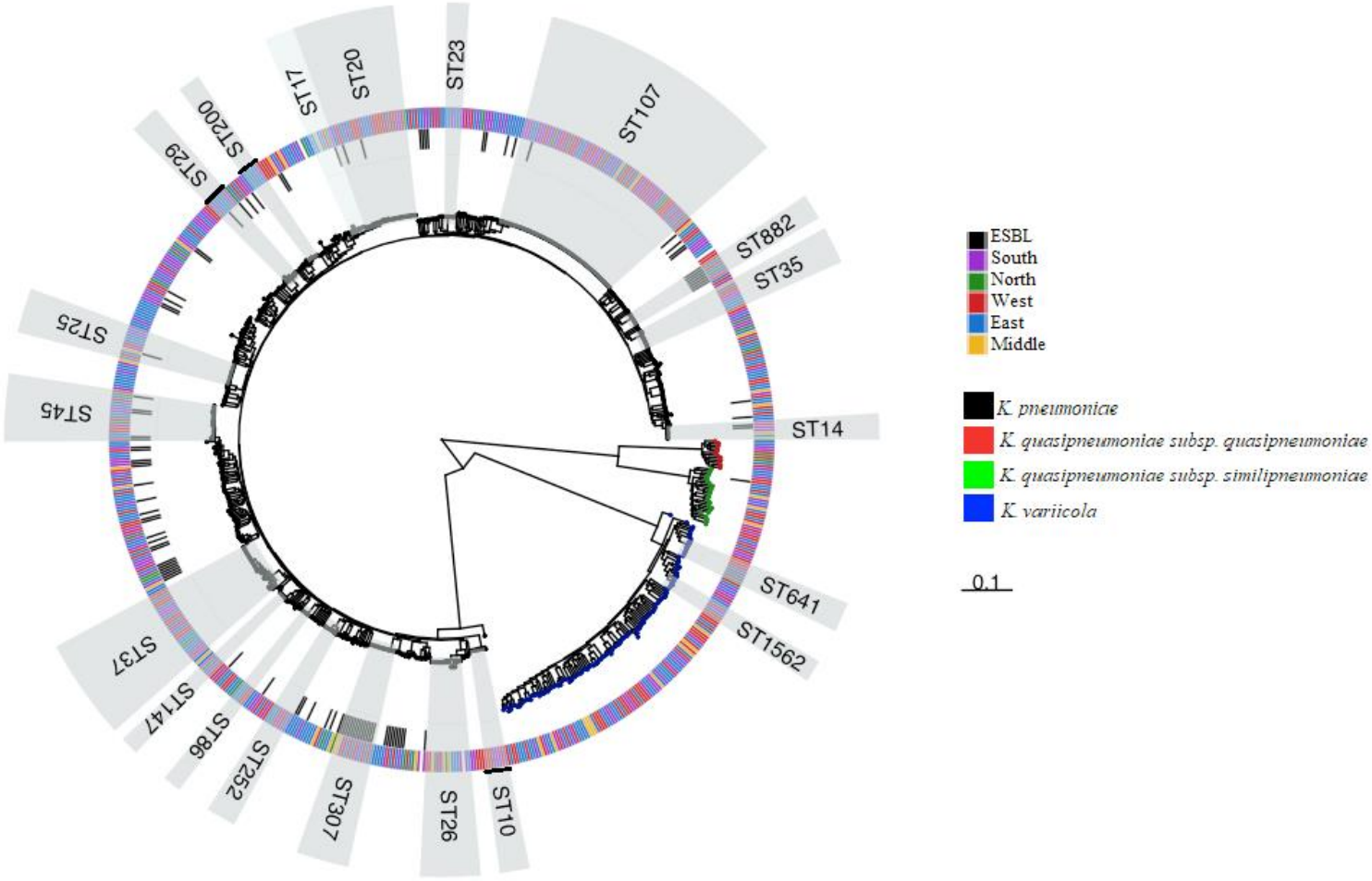


Figure 25: A core chromosomal SNP tree of the 722 *K. pneumoniae* isolates investigated in this project. 0.1 nucleotide substitutions per site.

The core chromosomal SNP tree illustrates the four *K. pneumoniae* species that are present in this population. As seen in Figure 25, there is no buildup of isolates with ESBL-encoding genes in a specific geographical region. Sequence type ST307 (n=12) containing ESBL-encoding genes in 11/12 isolates is found in all regions, except for the North. All ESBL-encoding genes are found in *K. pneumoniae* sensu stricto apart from a single one, which is found in an isolate with ST1525 of the species *K. quasipneumoniae* subsp. *similipneumoniae*.

There are a few local buildups of ST200 (5/6) and ST29 (5/8) in the East region. All ST10 isolates are from the two neighboring regions, West and South. The majority of ST107 (49/67) are also located in the West or the South region. Apart from these local buildups, the STs are distributed in all regions, but it should be noted that region Middle is under-represented, as only one hospital is participating in NORKAB. ST641 and ST1562 are STs on the *K. variicola* branch. ST17 and ST20 are closely related, which creates some overlap in the tree, since the STs are highlighted after the most recent common ancestor. Five isolates were assigned to ST458, but these are not highlighted in the tree due to a high diversity besides the seven housekeeping genes, resulting in the strains being scattered throughout the tree.

4.4.3 Investigation of ST107 by PFGE and core genome SNP analysis

ST107 was the most prevalent ST in the population (n=67/722). Virulence- and resistance genes are only detected in two isolates. One isolate harbored an ESBL-encoding gene, $bla_{CTX\text{-}M\text{-}1}$, and one isolate harbored an ybt variant. None of the isolates harbored siderophores rmpA or rmpA2.

1. ST107 PFGE analysis

36 ST107 isolates were selected for PFGE from large hospitals, which were preferably complete with regard to inclusions in the study. The PFGE fingerprint results with an UPGMA dendrogram are displayed in Figure 26.

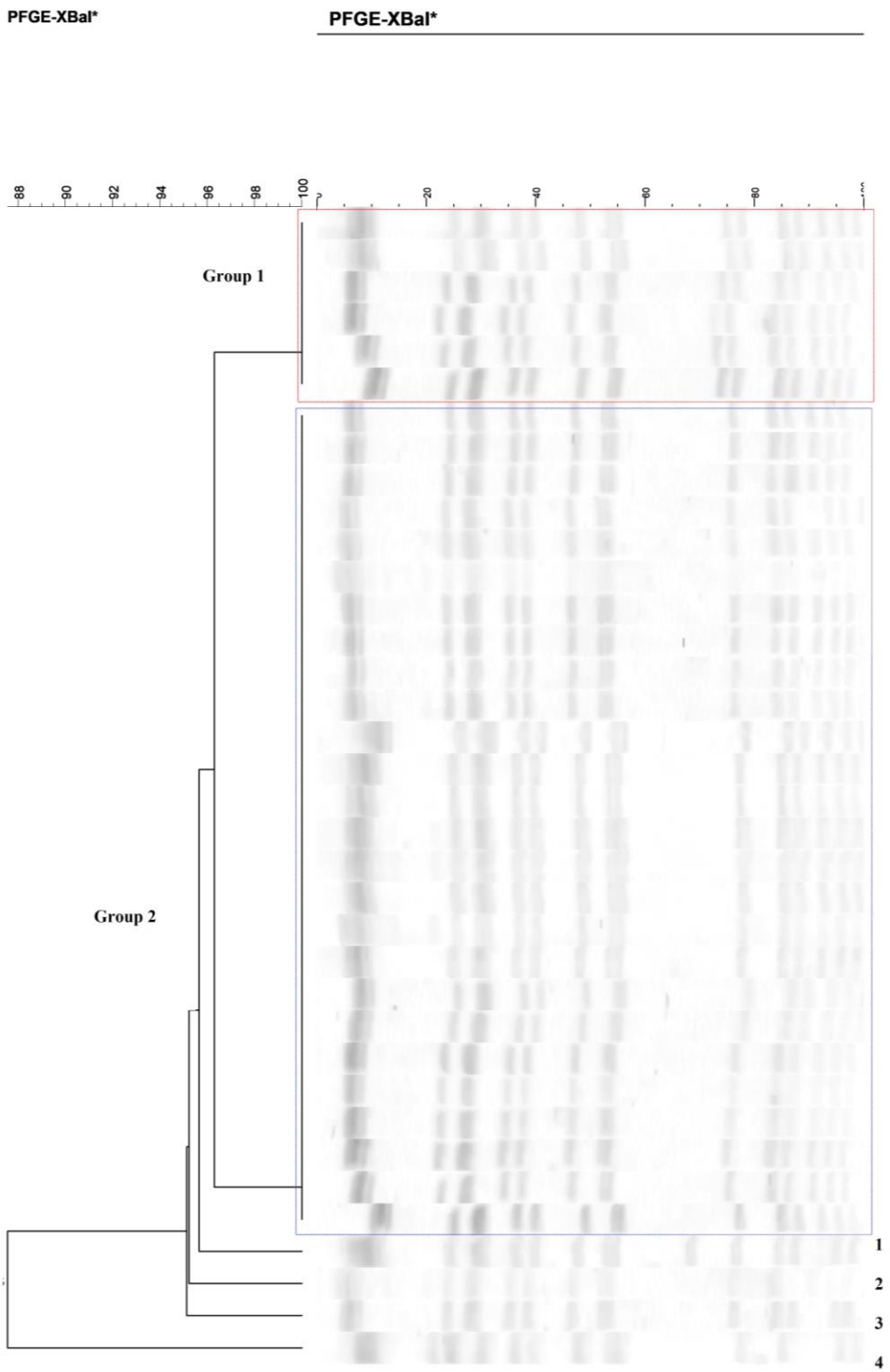


Figure 26: The comparison of fingerprint data from three rounds of *XBaI* PFGE, with a UPGMA dendrogram.

Figure 26 shows that the samples subjected to PFGE have clustered into two major groups and four distinct isolates. Group 1 (n=6) and Group 2 (n=26) are clonal according to the PFGE criteria set by Tenover *et al.* (1995) [45]. Table 18 shows an overview of the number of bands, band positions and genetic relations between all 36 isolates. Table 19 shows the number of bands in difference between the two groups, and the three distinct isolates.

Table 18: An overview of the number of bands, band positions and genetic relations between the 36 isolate subjected to PFGE.

	# of isolates	**# of bands**	**# of different band positions**	**Genetic relations[1]**
Group 1	6	14	Same band positions as group 2	Indistinguishable to group 2
Group 2	26	13	Same band positions as group 1	Indistinguishable to group 1
Isolate 1	1	13	Different band position in 1/13 bands	Closely related to all isolate
Isolate 2	1	12	Different band position in 1/12 bands	Closely related to all isolates
Isolate 3	1	12	Different band positions in 1/12	Closely related to all isolate
Isolate 4	1	11	Different band position in 1/11 bands	Closely related to all isolates

1: Genetic relations between isolates are according to Tenover *et al*. [45], divided into four categories; indistinguishable (clonal), closely related, possibly related and unrelated.

Table 19: The # of bands in difference between Group 1 and 2, and the four distinct isolates.

	Group 1	**Group 2**	**Isolate 1**	**Isolate 2**	**Isolate 3**
Group 1					
Group 2	1				
Isolate 1	1	0			
Isolate 2	2	1	1		
Isolate 3	2	1	1	0	
Isolate 4	3	2	2	1	1

2. ST107 Core genome SNP analysis

A core genome SNP tree, shown in Figure 27, was made for 66 ST107 isolates, with an in-group reference. One isolate, with the ybt variant, was excluded due to a large difference in SNPs, ~ 1400 SNP, compared to the other isolates, which had a range of 0-61 SNPs.

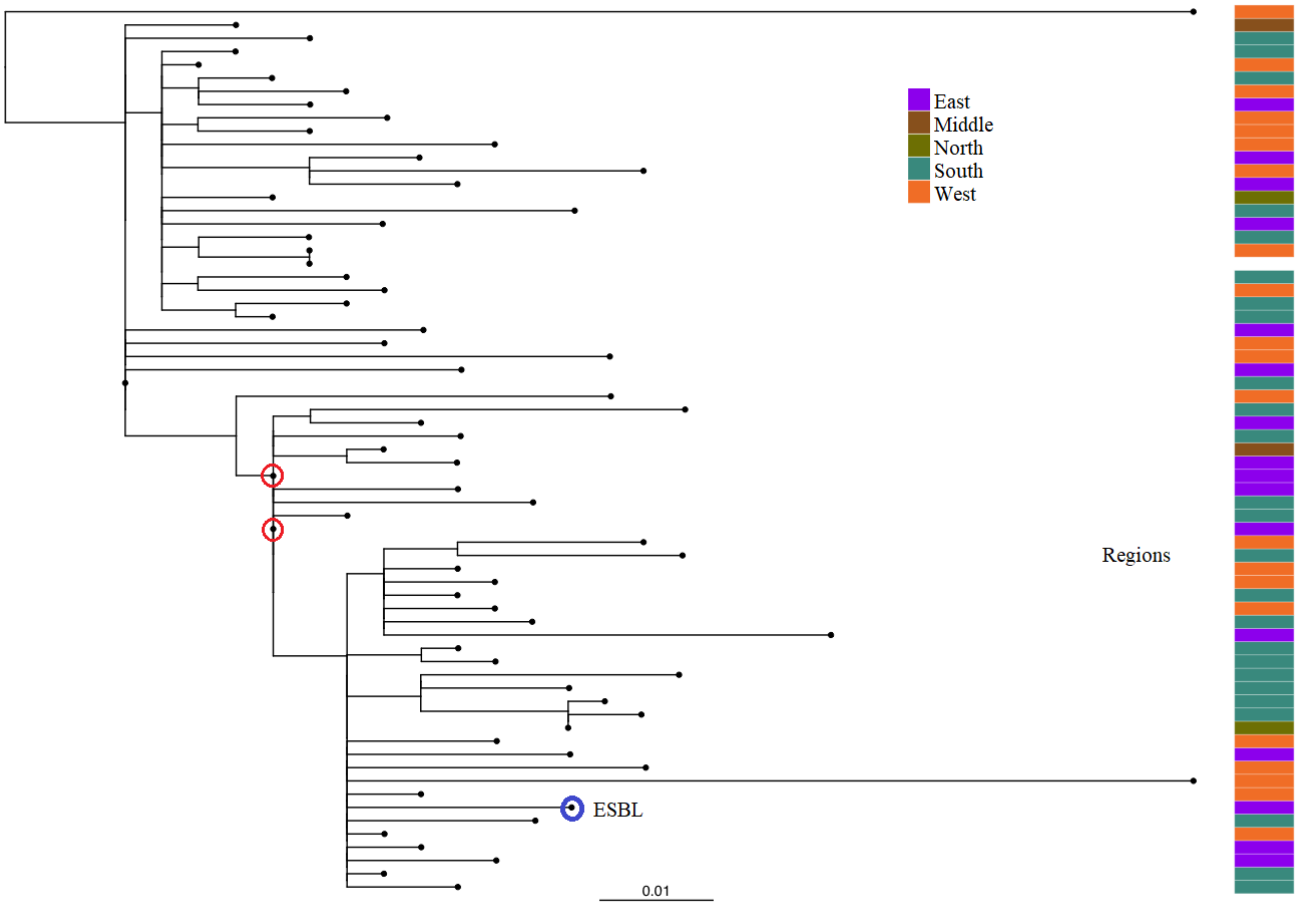


Figure 27: A core genome SNP tree of 66 ST107 isolates in the population. 0.01 nucleotide substitutions per site.

As seen from the color-panel in Figure 27, ST107 is more prevalent in some regions, the South region and the West region, with 49/66 isolates. The average SNP difference for the 66 isolates is 15.1±7.7 SNPs. Two isolates stands out from the others, these were checked in Kleborate, but no significant differences in AMR or virulence genes were found between these and the other isolates. Two isolates from the East region have 0 SNP difference, marked in Figure 27 with red circles.

3. Comparison of PFGE and core genome SNP results

One isolate harboring a ybt variant, was excluded from the core genome tree due to a SNP difference of ~1400 SNPs, whereas the rest of the isolates that were submitted to PFGE had a SNP difference of 2-41. This large difference created a misleading and out of scale tree (Appendix D). Figure 28 shows the core genome tree of the 35 ST107 isolates which were submitted to PFGE, with an in-group reference (NK-07).

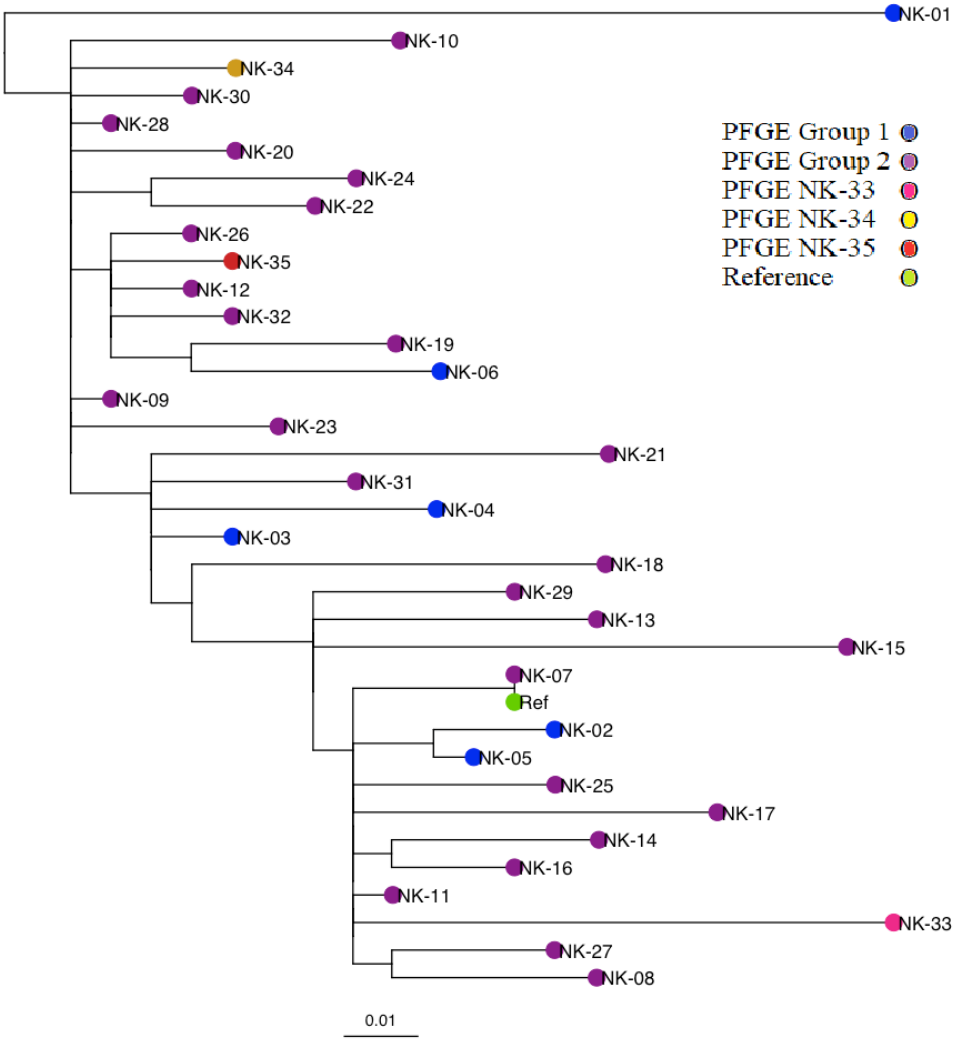


Figure 28: A core genome tree of the 35 ST107 isolates subjected to PFGE. NK-33-NK-35 corresponds Isolate 1-3, respectively, in the PFGE analysis. 0.01 nucleotide substitutions per site.

The core genome tree in Figure 28 shows an indication of a larger diversity in the relationship between isolates than what the PFGE results do. The isolates of group 1, group 2 and the three

distinct isolates do not cluster together in the core genome tree as they did in the PFGE dendrogram. Table 20 shows the range of SNP differences within the different groups, and the average SNP differences. The SNP distance matrix for the 36 isolates can be seen in Appendix E.

Table 20: An overview of the ranges of SNP differences within the different groups and average SNP differences.

	Range of SNPs	Average SNP difference
Group 1	4-33	15.5±7.5
Group 2	2-28	15.3±5.9
Group 1+2	2-33	15.3±6.2
Isolates 1-3[1]	8-24	16.3±7.8
All isolates	2-41	17.1±6.4

1: Isolates 1-3 corresponds to NK-33-NK-35 respectively.

Isolate 4 was excluded from the core genome tree was also excluded the average SNP differences were calculated.

5. Discussion

5.1 Discussion methods

5.1.1 Wet lab Challenges

1. Wet lab Challenges for NGS

One challenge encountered during the sequencing was an inconsistency in the cluster density. Both underclustering and overclustering occurred. Underclustering will maintain a high quality, but a lower data output, while overclustering may cause the run to fail, poor run performance, lower Q30 scores, and a total lower data output due to fewer clusters passing the chastity filter. The optimal cluster density for Illumina MiSeq V3 reagents are 1200-1400 k/mm^2, values below this was considered underclustering and values above was considered overclustering[120].

For Illumina Nextera XT library preparation, the initial DNA concentration was very important because transposones used for tagmentation of genomic DNA are end-point, which means they only cut the fragment once. The Nextera XT kit is optimized for 1 ng double-stranded genomic DNA. Using > 1 ng can lead to undertagmentation, resulting in large fragments, which result in low cluster density and low sequence output. Using < 1 ng can lead to overtagmentation, overclustering and low sequence quality [121].

To achieve concentrations of ~1 ng the DNA concentration of each sample was quantified after purification, using the Quant-iT^{TM} 1X dsDNA HS kit, and was then normalized to 0.2 ng/μL by Hamilton pipetting robot.

Since large parts of the library preparation was performed on the Hamilton Microlab STAR pipetting robot, there is also a possibility of pipetting error by the robot. To counteract this, a daily and weekly maintenance was performed on the pipetting robot after use. A fluctuation in room temperature and humidity could also influence the library preparation.

Due to the fact that the initial DNA concentration was thoroughly measured and the samples diluted by a pipetting robot, the amount of cleaned DNA library was suspected to be the reason for the inconsistency of the cluster density. Several of the cleaned DNA libraries were

quantified and quality controlled using the Agilent High Sensitivity DNA Kit. The High Sensitivity DNA kit detects possible PCR artifacts, impurities and excess adapters. The results of these analysis gave satisfactory results, and did not reveal any errors that would explain the varying cluster densities [122].
The concentration of the pooled DNA libraries loaded to the flow cell was also adjusted to try to obtain the optimal cluster density, and this was partially successful in some of the runs. In the first run, 15 μL of pooled DNA libraries were added, giving a low cluster density. The amount was then gradually increased, and an input of ~22 μL of pooled DNA libraries gave satisfactory cluster density in most runs.

Another possible cause of the difference in cluster density was the denaturation of the DNA libraries. For the most optimal conditions, the diluted NaOH was always freshly made, and 5 mL was prepared each time to prevent small-volume pipetting errors.

2. Wet lab Challenges for PFGE

One of the challenges faced during the PFGE procedure was to obtain evenly sized plug slices, especially the ladder. This caused the bands of some of the ladders to be larger than wanted.

There were cases of “wavy” bands. This might be to the use of an old scalped which might have been uneven or jagged, which would cause the plug slices to be damaged. It could also be due to debris in the gel or the plugs. Some dust from the surroundings was observed when casting the agarose gel, and on top of the gel after it was set. Possible solutions to this problem is to use a new scalped, and not wipe the scalpel with alcohol between plug slices. The gel should also be casted in more appropriate surroundings, protecting the gel from dust and particles [123].

There was also occurrence of “shadow” bands. The cause of this is often incomplete digestion, which could be caused by a number of reasons. The plugs could be of poor quality due to remaining proteinase K after washes, that the enzyme inhibitor was not properly washed out or that the cell concentration was too high. To avoid too high cell concentration, each sample was measured on a densitometer, and five samples from each round of PFGE

was also measured on a spectrophotometer. To counteract residual proteinase K or enzyme inhibitor, the plugs were washed 4x with TE-buffer, instead of 2x [123].

"Shadow" bands could also be caused by poor enzyme or buffer quality. The buffer was freshly made, but the enzyme was not new, and it is not known how many times the vial has been thawed and opened before this experiment was conducted [123].

The enzyme digestion conditions are also important for proper digestion. Old BSA, too few or too many units of enzyme, too long or too short incubation time and incorrect incubation time will influence the digestion. The digestion conditions followed in this experiment was according to the operating procedure for PulseNet PFGE of various Enterobacteriaceae [118]. A possible solution may be to purchase new enzyme, and to use concentrated enzyme (40 U/μL) as opposed to diluted enzyme (10 U/ μL), which was done for this experiment. Foreign particles or bubbles embedded in the plugs may also cause uneven bands [123].

There was also the issue of one gel being slanted. Reasons for this may be the problems with the power supply, pump flow rate, humidity, air temperature, ventilation and neighboring equipment. Possible reasons for a change in the flow rate may be kinks or air bubbles in the tubing, or leftover agarose pieces. Between each round of PFGE, the pump flow was reversed to flush out any agarose pieces.

Lane curvature could also be a result of the gel not being level when it was poured, the electrophoresis chamber not being level, the gel was not completely casted against the black platform, the buffer was not flowing evenly, fluctuation in temperature during the run ($14^{o}C \pm 2^{o}C$), broken electrodes or the electrophoresis chamber needs to be serviced [123].

To ensure that the gel and the electrophoresis chamber was level, a leveler was used when assembling the casting frame, and when adjusting the feet of the electrophoresis chamber. The equipment used is old, and had not been in use for some time.

5.1.2 Dry lab Challenges

1. Dry lab Challenges for NGS

Due to the inconsistent cluster density, some isolates were given poor quality scores during sequencing. This resulted in equally poor generated sequence data, which was visualized using FastQC. Twenty isolates in total was re-sequenced.

For the analysis of the sequencing data, Kleborate was applied. Kleborate was chosen to do this because it is a tool created exclusively for screening of *K. pneumoniae* genome assemblies. Kleborate identifies species, virulence factors and resistance genes, as well as perform MLST analysis and assigning STs.

2. Dry lab Challenges for PFGE

The DNA fingerprints generated from PFGE was analyzed using BioNumerics by Applied Maths. Since some DNA fingerprints had lane curvature, the normalization of the gel image by an internal reference was difficult. Some fragments also had different sizes and/or waves due to possible damages of the plug slices.

The best way to resolve this issue would most likely be to redo the PFGE of the isolates affected by these problems.

Another issue was the presence of “shadow” bands. This made the assignment of bands challenging.

This problem would also most likely be resolved with repetition of the PFGE with new buffers and enzymes, as well as completing a service of the electrophoresis chamber, or possibly replacing it with a new one.

5.2 Discussion results

1. Quality assessment of sequence data

K. pneumoniae has a median GC content of 57.2 and a median total length of 5.58 Mb [111]. In this population, the GC content ranged from 56.6-58.4% with an average of 57.4±0.23, and the total length ranged from 4.9-6.0 Mb, with an average of 5.4±0.17 Mb. The GC content of this population is within the expected range, but with slight variations. The isolates with a GC content ≥58% were only accepted with the proviso that the other parameters were satisfactory. There are various possible causes for variation in GC content within a species, such as genome size, the length coding sequences and environmental pressure [124], as well as accessory genes acquired from various bacterial taxa [13].

There is a slight intra-species variation in genome size, although the mean total length of the genome is within the expected range.
It has not yet been established whether variation in genome size is coincidental, coevolutionary or causative, but there are theories that suggest this happens due to mutational pressure, or that non-coding DNA enable the ability to expand for the organisms own benefit. Examples of this is i.e. bacterial plasmids and transposable elements, such as transposons [125].

Currently, there is no defined minimum performance standards for WGS data [110], only estimated recommendations. The overall quality parameters were within the optimal ranges. The total number of contigs should be low, since it is optimal with large but few contigs. In this study all contigs >700 bp were excluded, and an organism with the genome size of 5-6 Mb, <100 contigs is realistic. The total number of contigs ranged from 36-331.

The optimal value for sequence depth depends on the application. Authors of SRST2 reports that a sequence depth of only 10x is needed to obtain a >90% call rate [112], but a sequencing depth of 30x is usually recommended for bacterial genomes [110]. In this study, the sequence depth ranged from 16-115x, with 66 isolates having a sequencing depth between 16-29x. For the 35 ST107 isolates subjected to PFGE and core genome SNP analysis the average sequencing depth was 61±15.7, with a range of 29-97x, where only one isolate had a sequencing depth of <30x. There may be several reasons for a low sequencing depth.

Short sequencing reads gives rise to several possible problems in assembly. Short reads enable several valid alignments, but only one possibility corresponds to the target genome. In

addition, short reads do not necessarily have enough sequence context to determine the relative position of the read in the genome, hence distinguishing between near-exact copies of the same repeat in different parts of the genome may be impossible. In *de novo* assembly there is also created "gaps" of unknown sequence in the contigs. Errors also occur during sequencing, and the error rate especially increase towards the end of the sequencing [126, 127].

A way to improve the sequence depth would initially be to re-sequence the isolates with the lowest values, to try to obtain a more satisfactory result, or do a hybrid assembly with long-read inserts. In addition to this, the sequencing software, as well as assembly tools is continually being improved to decrease error rates and generate more uniform sequence read length, and to create assemblies with a higher level of confidence [126, 128].

2. Bacterial population description

The bacterial population identified in this study consisted of four distinct *Klebsiella* species, with *K. pneumoniae* sensu stricto as the most prevalent species (n=566/722). This was expected, since *K. pneumoniae* sensu stricto represents ~80% of invasive isolates, being responsible for most cases of hospital-acquired disease by *Klebsiella* [129]. *K. variicola* is the second most prevalent species (120/722) [130].

N=16/722 isolates were given a weak species match by Kleborate, *K. pneumoniae* sensu stricto (n=1), *K. quasipneumoniae subsp. quasipneumoniae* (n=2) and *K. variicola* (n=13). A core genome analysis of the isolates with weak species matches, a pair of isolates with strong species matches for each species, and *K. variicola subsp. tropicalensis* and *K. africanesis* as references was performed. The core genome tree revealed that the isolate with weak matches clearly clustered together with the isolates with strong species matches of the same species.

Within a bacterial species, a set of genes is present in all members; this is considered the core genome. The genome of *Klebsiella* has ~5000-6000 genes, and the core genome of *Klebsiella* consist of ~ 2000 genes, and is present in >95% of all isolates [13]. This means that a major part of the genome is comprised of what is called accessory genes. Different species can be identified by variations in their core genome, but also by the content of their accessory genes [129].

A weak species match may indicate a novel lineage, or a hybrid between multiple *Klebsiella*

species [131]. The sequence depth for the isolates in question was checked to determine if the reason for the weak species matches was low sequencing depth. The majority of the isolate with weak species matches (n=11/16) had a sequence depth ≥ 30x, which is the recommended value [110], the remaining ranged from 17-29x.

A possibility of why the majority of isolates given a weak species match was *K. variicola,* is that the diversity within the isolates is larger than the available reference material for the species used by the Kleborate software.

3. Detected antimicrobial resistance genes and virulence genes

Kleborate determines resistance genes against the ARG-ANNOT database of acquired resistance genes (SRST2 version [112]) .N=63/722 isolates had ESBL-encoding genes, and the most prevalent was $bla_{CTX\text{-}M}$, ($bla_{CTX\text{-}M\text{-}15}$ n=31/63) which was expected since this family is considered the most prevalent beta-lactamase enzyme [132, 133].

Bla_{SHV} alleles were detected in 28 isolates, but only 50% (n=14/28) could be identified as a specific variant. Fourteen alleles were given an imprecise allele match to $bla_{SHV\text{-}13*}$ by Kleborate, and was subjected to a BLAST analysis, but with no conclusive results. There was a SNP difference of 1-4 between the isolates and different annotated public references for various bla_{SHV}-variants.

All isolates (n=14) had a sequencing depth of ≥30x, except one which had a value of 21x.

Using only WGS may not be the best option for detection of new resistance genes, because it may overlook these. Genotyping prediction of genes rely on highly curated databases for known resistance determinants, hence, it cannot predict or identify mechanisms that are not yet defined. A combination of WGS and phenotypic methods such as broth micro dilution or minimum inhibitory concentration could aid in detection and characterization of new genes encoding resistance. The “gaps” between contigs also opens up for the possibility that resistance genes may go undetected. [134, 135].

4. Multilocus sequence typing

MLST is a nucleotide sequence-based technique for unambiguous characterization of i.e. bacterial strain types. It is based on the comparison of the sequence of seven housekeeping genes, with previously identifies alleles at that locus for each strain of a particular species.

The alleles of each of the seven loci are assigned to a number, which then constitutes the allelic profile of that strain. Different allelic profiles are assigned as a particular ST [136].

Three hundred and seventy-eight different STs were found in the population, and 330 STs only occurred one or two times. One of the most dominant STs in the population was ST307 (n=12). ST307 is associated with the ESBL encoding gene $bla_{CTX\text{-}M\text{-}15}$ [137], and 11/12 ST307 isolates in this population had the gene. Some STs are high-risk associated with resistance or virulence. Some high-risk resistance associated STs present in this population is ST11 (n=1) which was an ESBL, ST15 (n=2) where one was an ESBL, ST37 (n=20) where 16 isolates had resistance genes ranging from 1-17, and ST147 (n=5) where all had resistance genes ranging from 1-7.

High-risk virulence STs, associated with pyogenic liver abscess were also found in the population. For ST23 (n=5), all isolates had the siderophores, aerobactin (iuc), salmochelin (iro) and colibactin (clb). 4/5 also had ybt, and 3/5 had rmpA and rmpA2. In ST86 (n=5), all had iuc, iro and clb, one had ybt and 2/5 had rmpA and/or rmpA2 [16].

MLST is a good typing method for *K. pneumoniae* because there is a large MLST database available online, The Klebsiella Pasteur MLST sequence definition database [36]. The database is continually updated with new STs, and three new STs were assigned from isolates in this study. In addition, 147 different LVs were detected in 178 isolates.

The method is also highly reproducible and portable, as well as giving consistent results [138]. It also has a higher discriminatory power for determining relatedness between strains compared to i.e. PFGE [139].

Disadvantages of using MLST is that is only uses ~0.1% of the genomic sequence to assign STs, and therefore lacks the discriminatory power to differentiate between bacterial strains [140], and the method does not provide information of absence or presence of genes relevant for i.e. virulence or antimicrobial resistance [141].

5. Core chromosomal SNP analysis

There are several obstacles with using phylogenetic trees in the analysis of bacterial species. HGT is observable in most complete genomes, which makes it a challenge, if not impossible to try to define a single phylogenetic tree that represents all the evolutionary history of the bacteria. HGT will confound relationships between bacteria by indicating different relationships within a set of taxa. Relationships derived from a single gene will be a

combination of vertical and horizontal evolution of that gene [142]. However, complete genomes can be separated into what is chromosomal and what is plasmids. The chromosomal part of the genome is more stable, with the exception of insertions from mobile genetic elements, such as phages [143].

Therefore, when creating the phylogenetic tree for all isolates in the population (n=722), we used the chromosome of strain HS11286 (GenBank accession: NC_016845.1) was used as a reference. The plasmid-content will most likely be very varied in such a diverse population of 722 isolates distributed among 377 different STs.

6. Investigation of ST107 by PFGE and core genome SNP analysis

ST107 was the most prevalent ST in this study, and was investigated by both PFGE and core genome SNP analysis. The PFGE analysis indicated that there were two clusters of clonal isolates, with only a one-band difference on the DNA fingerprints. The core genome SNP analysis indicated more diversity and that the isolates were more distantly related than what the PFGE indicated.

PFGE used to be called the "gold standard" for bacterial strain typing. The method is said to have a high concordance with epidemiological relatedness, as well as creating stable and reproducible DNA restriction patterns. It can also be applied to many different types of bacteria by choosing restriction enzyme and electrophoresis conditions optimized for the specific species in question. However, the method also has many disadvantages. It is time consuming, the DNA restriction patterns may vary between technicians, the separation of fragments cannot be optimized in every part of the gel, bands of the same size may not originate from the same part of the chromosome and changes in one restriction site may result in more than one band change. Hence, the degree of relatedness determined from PFGE DNA fingerprints can only be used as a "guideline", and not a true phylogenetic measure [144].

In contrast, WGS differentiates isolates to a much higher degree. SNP-based analysis, gives more epidemiologically correct results at a higher resolution, even at low coverage [144, 145]. Core genome SNP analysis is also easily reproducible, and the impact of differences in analysis steps and parameter settings is not as considerable as it is with PFGE. But compared to PFGE, which has standardized guidelines for assessing isolate relatedness, there are no standardized measure for relatedness in the context of SNP differences. The degree of relatedness must be established based on the type of bacteria and epidemiological context, as well as each individual case [34, 145].

A study has suggested SNP relatedness criteria for *K. pneumoniae*, with emphasis on them only being guidelines. The suggested relatedness threshold for isolates to be clonal is an SNP difference of ≤10 [34].

The SNP analysis of the isolates form PFGE groups 1 and 2 gave an average SNP difference of ~15±6, both in within both groups and when the two groups were compared. According to the suggested relatedness threshold, the number of isolates that are indicated to be clonal is lower than what the PFGE results indicated.

5.2.1 Conclusion

The MLST distribution within the population (n=722) was diverse, with 378 STs, with seven dominating STs (n≥10). One hundred and seventy-eight isolates had LV detection, distributed among 147 different LVs. In addition, three new STs were assigned to three isolates from the population.

There were no significant geographical differences in the STs, except for a few local build-ups of ST10 (8/8) and ST107 (49/67) in the West and the South regions, and ST220 (5/6) and ST29 (5/8) in the East region, but the Middle region was poorly represented.

The PFGE analysis of ST107 indicated that 33/36 isolates were clonal. Core genome SNP analysis suggested that a lower number of isolates were clonal. Hence, the two different methods of bacterial strain typing gave slightly different results giving an indication that core genome SNP analysis has a stronger discriminative power and a higher resolution to differentiate isolates than what PFGE has.

5.3 Future perspectives

The findings from this project has aided in the survey of the genetic diversity of clinical *K. pneumoniae* isolates in Norway, therefore being of relevance for infection control and the surveillance of dissemination, and possibly having a positive impact on public health and patient treatment.

Future research should include further investigation of the species *K. variicola* and *K. quasipneumoniae*. *K. pneumoniae* sensu stricto is by far the most prevalent pathogen, but since *K. variicola* and *K. quasipneumoniae* was recognized as distinct species it has been discovered that these species has chromosomal and mobile genes encoding resistance mechanisms and virulence factors found *K. pneumoniae* sensu stricto. Carbapenemase carrying strain has also been identified in both *K. variicola* and *K. quasipneumoniae*. It is estimated that ~2% of human infection, previously attributed to *K. pneumoniae* sensu stricto is actually caused by *K. variicola* or *K. quasipneumoniae* [146]. In this population *K. variicola* constituted 17% of the population, and *K. quasipneumoniae* 5% of the population.

Re-sequencing of isolates from this population having sequencing depth below the recommended optimal value and undetermined bla_{SHV}-variants should also be prioritized, to possibly gain a greater insight into the genetic diversity of this population. The isolates with LVs detected should also be submitted to The Institut Pasteur for further investigation.

It would also be interesting to continue this research with a more complete population seen from a geographical point of view. The Middle region is heavily underrepresented in this study, due to a lack of participating laboratories. For further investigations, it would be interesting to try to include a larger number of isolate also from this region.

The most dominant ST in this study was ST107 (n=67/722). All isolates except two are very similar. Except for these two isolates, where one has an ybt variant and one has an ESBL-encoding gene, $bla_{CTX\text{-}M\text{-}1}$, there are no known virulence factors or resistance genes in any other isolates. Despite this, ST107 is present in isolates spread across four regions, West, South, East and Middle, to a greater or lesser degree. It would be interesting to further investigate this ST to try to determine why it is so prevalent, and disseminated in Norway, and to find out if this is the case internationally as well.

Another dominant sequence type in the population was ST307 (n=12), where all isolates but one contained the ESBL-encoding $bla_{CTX\text{-}M\text{-}15}$ gene. The sequence type is also disseminated in all regions of Norway apart from the North region in this study. A recent study suggests that the clone is emergent and spreading rapidly not just in Norway, but on a global scale, and bla_{KPC} genes has been found in ST307 isolates [137]. It would therefore be reasonable to further investigate and monitor ST307.

It was assumed that the plasmid-content in the population was diverse, therefore it would be interesting to investigate the plasmid-population. Since short-read approaches such as the one applied in this thesis generates incomplete and fragmented plasmid assemblies, it would be necessary to apply another method to obtain more accurate plasmid assemblies.
A newcomer to the wide sequencing technology is the Oxford Nanopore MinION sequencing device. The device is pocket-sized and allows for rapid, real-time, long read sequencing of nucleic acids, based on nanopore sequencing technology. An ionic current is passed through the nanopores, and the MinION measures the alterations in the current as biological molecules pass through the nanopore or near it. The changes in the current enables identification of that specific molecule [147]. Assemblies of long nanopore reads generated by MinION has obtained coverage of 98% of all plasmids and accurately identified AMR genes in a known *K. pneumoniae* isolate, demonstrating that nanopore sequencing would deliver a more accurate plasmid assembly and a more accurate detection of AMR genes [148].

List of references

1. Organization, W.H. *Ten threats to global health in 2019* 2019 19.02.2019]; Available from: https://www.who.int/emergencies/ten-threats-to-global-health-in-2019.
2. Organization, W.H. *WHO publishes list of bacteria for which new antibiotics are urgently needed* 2017 19.02.2019]; Available from: https://www.who.int/news-room/detail/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed.
3. Jansen, K.U. and A.S. Anderson, *The role of vaccines in fighting antimicrobial resistance (AMR).* Human vaccines & immunotherapeutics, 2018. **14**(9): p. 2142-2149.
4. Organization, W.H. *Antimicrobial Resistance* 2018 19.02.2019]; Available from: https://www.who.int/en/news-room/fact-sheets/detail/antimicrobial-resistance.
5. Marathe, N.P., et al., *Functional metagenomics reveals a novel carbapenem-hydrolyzing mobile beta-lactamase from Indian river sediments contaminated with antibiotic production waste.* Environment International, 2018. **112**: p. 279-286.
6. Podschun, R. and U. Ullmann, *Klebsiella spp. as nosocomial pathogens: epidemiology, taxonomy, typing methods, and pathogenicity factors.* Clinical microbiology reviews, 1998. **11**(4): p. 589-603.
7. Organization, W.H. *Global action plan on antimicrobial resistance*. 2015 20.02.2019]; Available from: https://www.who.int/antimicrobial-resistance/global-action-plan/en/.
8. Initiative, O.H. *One Health Initiative Mission Statement*. 04.06.2019]; Available from: http://www.onehealthinitiative.com/mission.php.
9. 2017, N.N.-V., *Usage of Antimicrobial Agents and Occurrence of Antimicrobial Resistance in Norway*. 2018.
10. Fostervold, A., *NORKAB - The Norwegian Klebsiella pneumoniae bacteremia study.* 2017: p. 13.
11. Friedländer, C., *Ueber die Schizomyceten bei der acuten firbösen Pneumonie.* Arch. für Pathol. Anat, und Physiol. und für Klin. Med. , 1882. **87**(2): p. 319-324.
12. Tacconelli, E., et al., *ESCMID guidelines for the management of the infection control measures to reduce transmission of multidrug-resistant*

Gram-negative bacteria in hospitalized patients. Clinical Microbiology and Infection, 2014. 20: p. 1-55.
13. Holt, K.E., et al., *Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health.* Proceedings of the National Academy of Sciences, 2015. 112(27): p. E3574-E3581.
14. Hennigar, S.R. and J.P. McClung, *Nutritional Immunity: Starving Pathogens of Trace Minerals.* American journal of lifestyle medicine, 2016. 10(3): p. 170-173.
15. Bengoechea, J.A. and J. Sa Pessoa, *Klebsiella pneumoniae infection biology: living to counteract host defences.* FEMS Microbiology Reviews, 2018. 43(2): p. 123-144.
16. Yan, J.J., et al., *Associations of the major international high-risk resistant clones and virulent clones with specific ompK36 allele groups in Klebsiella pneumoniae in Taiwan.* New microbes and new infections, 2015. 5: p. 1-4.
17. Heinz, E., et al., *Resistance mechanisms and population structure of highly drug resistant Klebsiella in Pakistan during the introduction of the carbapenemase NDM-1.* Scientific reports, 2019. 9(1): p. 2392-2392.
18. Bajpai, T., et al., *Prevalence of TEM, SHV, and CTX-M Beta-Lactamase genes in the urinary isolates of a tertiary care hospital.* Avicenna journal of medicine, 2017. 7(1): p. 12-16.
19. Raghavendra, P. and T. Pullaiah, *Chapter 8 - Future of Cellular and Molecular Diagnostics: Bench to Bedside*, in *Advances in Cell and Molecular Diagnostics*, P. Raghavendra and T. Pullaiah, Editors. 2018, Academic Press. p. 203-270.
20. Choudhuri, S., *Chapter 2 - Fundamentals of Molecular Evolution**The opinions expressed in this chapter are the author's own and they do not necessarily reflect the opinions of the FDA, the DHHS, or the Federal Government*, in *Bioinformatics for Beginners*, S. Choudhuri, Editor. 2014, Academic Press: Oxford. p. 27-53.
21. Neuner, E.A., et al., *Treatment and outcomes in carbapenem-resistant Klebsiella pneumoniae bloodstream infections.* Diagnostic microbiology and infectious disease, 2011. 69(4): p. 357-362.
22. Suzanne, B.-D., et al., *Genomic Definition of Hypervirulent and Multidrug-Resistant Klebsiella pneumoniae Clonal Groups.* Emerging Infectious Disease journal, 2014. 20(11): p. 1812.
23. Rodrigues, C., et al., *Description of Klebsiella africanensis sp. nov., Klebsiella variicola subsp. tropicalensis subsp. nov. and Klebsiella variicola subsp. variicola subsp. nov.* Res Microbiol, 2019.

24. Coggon, D., G. Rose, and D. Barker, *Epidemiology for the uninitiated*, ed. F. edition. 1979.
25. Calfee, D.P., *Recent advances in the understanding and management of Klebsiella pneumoniae.* F1000Research, 2017. **6**: p. 1760-1760.
26. Östholm-Balkhed, Å., et al., *Travel-associated faecal colonization with ESBL-producing Enterobacteriaceae: incidence and risk factors.* Journal of Antimicrobial Chemotherapy, 2013. **68**(9): p. 2144-2153.
27. Korsman, S.N.J., et al., *Epidemiology*, in *Virology*, S.N.J. Korsman, et al., Editors. 2012, Churchill Livingstone: Edinburgh. p. 22-23.
28. Jørgensen, S.B., et al., *Heat-resistant, extended-spectrum β-lactamase-producing Klebsiella pneumoniae in endoscope-mediated outbreak.* Journal of Hospital Infection, 2016. **93**(1): p. 57-62.
29. Rettedal, S., et al., *First outbreak of extended-spectrum ß-lactamase-producing Klebsiella pneumoniae in a Norwegian neonatal intensive care unit; associated with contaminated breast milk and resolved by strict cohorting.* APMIS, 2012. **120**(8): p. 612-621.
30. Tofteland, S., et al., *A Long-Term Low-Frequency Hospital Outbreak of KPC-Producing Klebsiella pneumoniae Involving Intergenus Plasmid Diffusion and a Persisting Environmental Reservoir.* PLOS ONE, 2013. **8**(3): p. e59015.
31. Control, E.C.f.D.P.a. *Surveillance of Antimicrobial Resistance in Europe 2017*. 2017 22.05.2019]; Available from: https://ecdc.europa.eu/en/publications-data/surveillance-antimicrobial-resistance-europe-2017.
32. Laupland, K.B. and D.L. Church, *Population-based epidemiology and microbiology of community-onset bloodstream infections.* Clinical microbiology reviews, 2014. **27**(4): p. 647-664.
33. *Changing trends in clinical characteristics and antibiotic susceptibility of Klebsiella pneumoniae bacteremia FAU - Hyun, Miri FAU - Noh, Chang In FAU - Ryu, Seong Yeol FAU - Kim, Hyun Ah.* Korean J Intern Med, 2018. **33**(3): p. 595-603.
34. Schürch, A.C., et al., *Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches.* Clinical Microbiology and Infection, 2018. **24**(4): p. 350-354.
35. Foxman, B., et al., *Choosing an appropriate bacterial typing technique for epidemiologic studies.* Epidemiologic perspectives & innovations : EP+I, 2005. **2**: p. 10-10.
36. Pasteur, I. *Klebsiella Sequence Typing Home Page*. 2019 15.01.2019]; Available from: https://bigsdb.pasteur.fr/klebsiella/klebsiella.html.

37. Chan, M.-S., M.C.J. Maiden, and B.G. Spratt, *Database-driven Multi Locus Sequence Typing (MLST) of bacterial pathogens.* Bioinformatics, 2001. **17**(11): p. 1077-1083.
38. Maiden, M.C., et al., *Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(6): p. 3140-3145.
39. PubMLST. *Multilocus Sequence Typing (MLST)*. 2019 23.05.2019]; Available from: https://pubmlst.org/general.shtml.
40. Ruppitsch, W., *Molecular typing of bacteria for epidemiological surveillance and outbreak investigation / Molekulare Typisierung von Bakterien für die epidemiologische Überwachung und Ausbruchsabklärung*. Vol. 67. 2016.
41. Bio-Rad. *Pulse Field Gel Electrophoresis*. 11.02.2019]; Available from: http://www.bio-rad.com/en-no/applications-technologies/pulsed-field-gel-electrophoresis?ID=LUSORPDFX.
42. PulseNet. *Pulse-Field Gel Electrophoresis*. 2018 11.02.2019]; Available from: https://www.cdc.gov/pulsenet/pathogens/pfge.html.
43. Maths, A. *Pulse-field gel electrophoresis (PFGE) typing*. 11.02.2019]; Available from: https://www.cdc.gov/pulsenet/pathogens/pfge.html.
44. PulseNet, *Pulse-Field Gel Electrophoresis Infographic*. 2018.
45. Tenover, F.C., et al., *Interpreting Chromosomal DNA Restriction Patterns Produced by Pulse-Field Gel Electrophoresis: Criteria for Bacterial Strain Typing.* Journal of Clinical Microbiology, 1995. **33**(9): p. 2233-2239.
46. Ding, W., F. Baumdicker, and R.A. Neher, *panX: pan-genome analysis and exploration.* Nucleic acids research, 2018. **46**(1): p. e5-e5.
47. Baum, D., *Reading a Phylogenetic Tree: The Meaning of Monophyletic Groups.* Nature Education, 2008.
48. Life, D.A.o.A.
49. Weiß, M. and M. Göker, *Chapter 12 - Molecular Phylogenetic Reconstruction*, in *The Yeasts (Fifth Edition)*, C.P. Kurtzman, J.W. Fell, and T. Boekhout, Editors. 2011, Elsevier: London. p. 159-174.
50. Borris, R., et al. *Taxonomy of Prokaryotes: Phylogenetic trees*. 2011 18.02.2019]; Available from: https://www.sciencedirect.com/topics/medicine-and-dentistry/phylogenetic-tree.
51. NCBI. *Maximum Likelihood* 18.02.2019]; Available from: https://www.ncbi.nlm.nih.gov/Class/NAWBIS/Modules/Phylogenetics/phylo15.html.

52. Edwards, D.J., B.J. Pope, and K.E. Holt. *RedDog: comparative analysis pipeline for large numbers of bacterial isolates using high-throughput sequences*. 2015 15.01.2015]; Available from: https://github.com/katholt/RedDog/wiki/1.-Instruction-Manual.
53. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature Methods, 2012. **9**: p. 357.
54. Genome Project Data Processing, S., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.
55. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree 2--approximately maximum-likelihood trees for large alignments.* PloS one, 2010. **5**(3): p. e9490-e9490.
56. Organization, W.H. *Antimicrobial resistance* 2019 01.06.2019]; Available from: https://www.who.int/antimicrobial-resistance/en/.
57. Hay, S.I., et al., *Measuring and mapping the global burden of antimicrobial resistance.* BMC medicine, 2018. **16**(1): p. 78-78.
58. Michael, C.A., D. Dominey-Howes, and M. Labbate, *The antimicrobial resistance crisis: causes, consequences, and management.* Frontiers in public health, 2014. **2**: p. 145-145.
59. Chaves, J., et al., *SHV-1 beta-lactamase is mainly a chromosomally encoded species-specific enzyme in Klebsiella pneumoniae.* Antimicrobial agents and chemotherapy, 2001. **45**(10): p. 2856-2861.
60. Nordmann, P., G. Cuzon, and T. Naas, *The real threat of Klebsiella pneumoniae carbapenemase-producing bacteria.* The Lancet, 2009. **9**(4): p. 228-236.
61. Ghafourian, S., et al., *Extended Spectrum Beta-lactamases: Definition, Classification and Epidemiology.* Current Issues in Molecular Biology 2015. **17**: p. 11-22.
62. Giske, C.G., et al., *Redefining extended-spectrum beta-lactamases: balancing science and clinical need.* The Journal of antimicrobial chemotherapy, 2009. **63**(1): p. 1-4.
63. Bush, K. and G.A. Jacoby, *Updated functional classification of beta-lactamases.* Antimicrobial agents and chemotherapy, 2010. **54**(3): p. 969-976.
64. Care, U.D.o.H.a.S., *UK Five Year Antimicrobial Resistance Strategy 2013 to 2018*. 2013.
65. Prevention, C.f.D.C.a., *Antibiotic Resistance Threats in the United States*. 2013.
66. Organization, W.H., *Central Asian and Eastern European Surveillance of Antimicrobial Resistance* 2014.
67. WHO, *Antimicrobial resistance: Global report on surveillance 2014*. 2014.

68. Alvarez-Uria, G., et al., *Global forecast of antimicrobial resistance in invasive isolates of Escherichia coli and Klebsiella pneumoniae.* International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases, 2018. **68**: p. 50-53.
69. ECDC, *ECDC Surveillance Atlas*. 2019.
70. Heather, J.M. and B. Chain, *The sequence of sequencers: The history of sequencing DNA.* Genomics, 2016. **107**(1): p. 1-8.
71. Institute, N.H.G.R. *The Cost of Sequencing a Human Genome* 2016 02.04.2019]; Available from: https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/.
72. Behjati, S. and P.S. Tarpey, *What is next generation sequencing?* Archives of disease in childhood. Education and practice edition, 2013. **98**(6): p. 236-238.
73. Illumina. *An Introduction to Next-Generation Sequencing Technology*. 2017 30.01.2019]; Available from: https://www.illumina.com/documents/products/illumina_sequencing_introduction.pdf.
74. Illumina, *Illumina Sequencing Overview*. 2018.
75. Illumina. *Nextera XT Library Prep: Tips and Troubleshooting* 2015 31.01.2019]; Available from: https://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/nextera-xt-troubleshooting-technical-note.pdf.
76. Head, S.R., et al., *Library construction for next-generation sequencing: overviews and challenges.* BioTechniques, 2014. **56**(2): p. 61-passim.
77. Illumina. *MiSeq System Guide*. 2018 31.01.2019]; Available from: https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/miseq-system-guide-for-miseq-reporter-1000000061014-00.pdf.
78. cegat.de, *Clonal amplification*.
79. Illumina. *Sequence Analysis Viewer User Guide* 2014 11.11.2018]; Available from: http://emea.support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/sav/sequencing-analysis-viewer-user-guide-15020619-f.pdf.
80. Lifescience, R. *MagNA Pure 96 DNA and Viral NA Small Volume Kit* 2018 08.11.2018]; Available from: https://lifescience.roche.com/en_no/products/magna-pure-96-dna-and-viral-na-small-volume-kit.html#overview.

81. Scientific, T.F. *Quant-ITTM 1x dsDNA HS Assay* 2018 08.11.2018]; Available from: https://www.thermofisher.com/order/catalog/product/Q33232.
82. Scientific, T.F. *QubitTM 1x dsDNA HS Assay Kit*. 2018 08.11.2018]; Available from: https://www.thermofisher.com/order/catalog/product/Q33230.
83. Illumina. *Nextera XT Library Preparation Kit*. 2018 08.11.2018]; Available from: https://emea.illumina.com/products/by-type/sequencing-kits/library-prep-kits/nextera-xt-dna.html?langsel=/no/.
84. Coulter, B. *AMPure XP for PCR Purification*. 2018 28.11.2018]; Available from: https://www.beckman.com/reagents/genomic/cleanup-and-size-selection/pcr.
85. Illumina. *PhiX Control Kit v3*. 2018 08.11.2018]; Available from: https://emea.illumina.com/products/by-type/sequencing-kits/cluster-gen-sequencing-reagents/phix-control-v3.html?langsel=/no/.
86. Illumina. *MiSeq Reagent Kit v3*. 2018 08.11.2018]; Available from: https://emea.illumina.com/products/by-type/sequencing-kits/cluster-gen-sequencing-reagents/miseq-reagent-kit-v3.html?langsel=/no/.
87. Technologies, A. *Agilent High Sensitivity DNA Kit for 2100 Bioanalyzer System*. 2017 08.11.2018]; Available from: https://www.agilent.com/cs/library/datasheets/public/5991-7889EN.pdf.
88. Biolabs, N.E. *Cutsmart Buffer*. 2019 17.02.2019]; Available from: https://international.neb.com/products/b7204-cutsmart-buffer#Product%20Information.
89. Promega. *Bovine Serum Albumin, Actylated*. 2019 17.02.2019]; Available from: https://no.promega.com/products/biochemicals-and-labware/biochemical-buffers-and-reagents/bovine-serum-albumin_-acetylated/?catNum=R3961.
90. Biolabs, N.E. *XBaI*. 2019 17.02.2019]; Available from: https://international.neb.com/products/r0145-xbai#Product%20Information.
91. Pharmaceuticalmicrobiology, *Four quadrant streak*. 2016.
92. Lifescience, R. *MagNA Pure 96 Instrument*. 2018 08.11.2018]; Available from: https://www.lifescience.roche.com/en_no/products/magna-pure-96-instrument-382411-1.html.
93. Tecan. *Spark Multimode Microplate Reader* 2017 08.11.2018]; Available from: https://lifesciences.tecan.com/multimode-plate-reader?p=tab--5.
94. Bioscientific, *Illumina Experiment Manager*. 2108.
95. Illumina, *Illumina Experiment Manager Software Guide.* 2018
96. Illumina, *Local run manager overview*. 2018.

97. Illumina. *MiSeq Specifications*. 2018 11.11.2018]; Available from: https://emea.illumina.com/systems/sequencing-platforms/miseq/specifications.html?langsel=/no/.
98. Illumina. *Sequence Analysis Viewer Software Guide*. 2017 11.11.2018]; Available from: http://emea.support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/sav/sequencing-analysis-viewer-v-2-4-software-guide-15066069-03.pdf.
99. Andrews, S. *Babraham Bioninformatics - FastQC: A Quality Control Tool* 2018 15.01.2019]; Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
100. Ewels, P., et al., *MultiQC: summarize analysis results for multiple tools and samples in a single report.* Bioinformatics, 2016. **32**(19): p. 3047-3048.
101. Bioinformatics, B. *Taking approperiate QC measures for RRBS-type or other - Seq applications with Trim Galore*. 2019 26.05.2019]; Available from: https://github.com/FelixKrueger/TrimGalore/blob/master/Docs/Trim_Galore_User_Guide.md.
102. Wick, R.R., et al., *Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads.* PLoS computational biology, 2017. **13**(6): p. e1005595-e1005595.
103. Bankevich, A., et al., *SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.* Journal of computational biology : a journal of computational molecular cell biology, 2012. **19**(5): p. 455-477.
104. Walker, B.J., et al., *Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement.* Plos One, 2014. **9**(11): p. e112963.
105. Gurevich, A., et al., *QUAST: quality assessment tool for genome assemblies.* Bioinformatics, 2013. **29**(8): p. 1072-1075.
106. Hetland, M. *SeqDepth*. 2019 27.05.2019]; Available from: https://github.com/marithetland/SeqDepth.
107. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows–Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-1760.
108. Institute, B. *Picard* 2019 27.05.2019]; Available from: http://broadinstitute.github.io/picard/.
109. Quast. *QUAST 5.0.2 manual*. 2018 13.02.2019]; Available from: http://quast.bioinf.spbau.ru/manual.html.
110. Ellington, M.J., et al., *The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee.* Clinical Microbiology and Infection, 2017. **23**(1): p. 2-22.

111. NCBI. *Klebsiella pneumoniae*. 19.02.2019]; Available from: https://www.ncbi.nlm.nih.gov/genome/?term=Klebsiella%20pneumoniae[Organism]&cmd=DetailsSearch.
112. Inouye, M., et al., *SRST2: Rapid genomic surveillance for public health and hospital microbiology labs.* Genome Medicine, 2014. **6**(11): p. 90.
113. Liu, P., et al., *Complete genome sequence of Klebsiella pneumoniae subsp. pneumoniae HS11286, a multidrug-resistant strain isolated from human sputum.* Journal of bacteriology, 2012. **194**(7): p. 1841-1842.
114. Lam, M.M.C., et al., *Genetic diversity, mobilisation and spread of the yersiniabactin-encoding mobile element ICEKp in Klebsiella pneumoniae populations.* bioRxiv, 2018: p. 098178.
115. Ondov, B.D., et al., *Mash Screen: High-throughput sequence containment estimation for genome discovery.* bioRxiv, 2019: p. 557314.
116. NCBI. *Assembly*. 2019 26.05.2019]; Available from: https://www.ncbi.nlm.nih.gov/assembly.
117. Information, N.C.f.B. *BLAST: Basic Local Alignment Search Tool*. 2019 12.05.2019]; Available from: https://blast.ncbi.nlm.nih.gov/Blast.cgi.
118. PulseNet. *Standard Operating Procedure for PFGE of various Enterobacteriaceae*. 2017 16.02.2019]; Available from: https://www.cdc.gov/pulsenet/pdf/ecoli-shigella-salmonella-pfge-protocol-508c.pdf.
119. Maths, A. *Bionumerics Quick Guide*. 2019 12.05.2019]; Available from: https://download.applied-maths.com/sites/default/files/download/bn_quickguide_0.pdf.
120. Illumina. *Cluster Optimization*. 2019 24.05.2019]; Available from: https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/cluster-optimization-overview-guide-1000000071511-00.pdf.
121. Illumina. *Nextera XT Library Prep: Tips and Troubleshooting*. 2015 24.05.2019]; Available from: https://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/nextera-xt-troubleshooting-technical-note.pdf.
122. Technologies, A. *Performance characteristics of the High Sensitivity DNA kit for the Agilent 2100 Bioanalyzer*. 2009 24.05.2019]; Available from: http://hpst.cz/sites/default/files/attachments/performance-characteristics-high-sensitivity-dna-assay-agilent-2100-bioanalyzer.pdf.
123. PulseNetInternational. *PFGE Troubleshooting* 24.05.2019]; Available from: http://www.pulsenetinternational.org/assets/PulseNet/uploads/pfge/PFGEPresentation6-PFGETroubleshooting.pdf.

124. Wu, H., et al., *On the molecular mechanism of GC content variation among eubacterial genomes.* Biology direct, 2012. **7**: p. 2-2.
125. Gregory, T.R., *Coincidence, coevolution, or causation? DNA content, cellsize, and the C-value enigma.* Biological Reviews, 2001. **76**(1): p. 65-101.
126. Haiminen, N., et al., *Evaluation of Methods for De Novo Genome Assembly from High-Throughput Sequencing Reads Reveals Dependencies That Affect the Quality of the Results.* PLOS ONE, 2011. **6**(9): p. e24182.
127. Desai, A., et al., *Identification of optimum sequencing depth especially for de novo genome assembly of small genomes using next generation sequencing data.* PloS one, 2013. **8**(4): p. e60204-e60204.
128. Paszkiewicz, K. and D.J. Studholme, *De novo assembly of short sequence reads.* Briefings in Bioinformatics, 2010. **11**(5): p. 457-472.
129. Martin, R.M. and M.A. Bachman, *Colonization, Infection, and the Accessory Genome of Klebsiella pneumoniae.* Frontiers in cellular and infection microbiology, 2018. **8**: p. 4-4.
130. Guo, Y., et al., *Complete Genomic Analysis of a Kingdom-Crossing Klebsiella variicola Isolate.* Frontiers in microbiology, 2018. **9**: p. 2428-2428.
131. Kleborate. *Kleborate*. 2019 15.01.2019]; Available from: https://github.com/katholt/Kleborate.
132. Castanheira, M., et al., *Contemporary diversity of β-lactamases among Enterobacteriaceae in the nine U.S. census regions and ceftazidime-avibactam activity tested against isolates producing the most prevalent β-lactamase groups.* Antimicrobial agents and chemotherapy, 2014. **58**(2): p. 833-838.
133. Cantón, R., et al., *Prevalence and spread of extended-spectrum β-lactamase-producing Enterobacteriaceae in Europe.* Clinical Microbiology and Infection, 2008. **14**: p. 144-153.
134. Zhao, S., et al., *Whole-Genome Sequencing Analysis Accurately Predicts Antimicrobial Resistance Phenotypes in Campylobacter spp.* Applied and environmental microbiology, 2016. **82**(2): p. 459-466.
135. Köser, C.U., M.J. Ellington, and S.J. Peacock, *Whole-genome sequencing to control antimicrobial resistance.* Trends in genetics : TIG, 2014. **30**(9): p. 401-407.
136. Aanensen, D.M. and B.G. Spratt, *The multilocus sequence typing network: mlst.net.* Nucleic acids research, 2005. **33**(Web Server issue): p. W728-W733.

137. Wyres, K.L., et al., *Emergence and rapid global dissemination of CTX-M-15-associated Klebsiella pneumoniae strain ST307.* The Journal of antimicrobial chemotherapy, 2019. **74**(3): p. 577-581.
138. Ahmed, N., et al., *Multilocus sequence typing method for identification and genotypic classification of pathogenic Leptospira species.* Annals of clinical microbiology and antimicrobials, 2006. **5**: p. 28-28.
139. Kotetishvili, M., et al., *Multilocus sequence typing has better discriminatory ability for typing Vibrio cholerae than does pulsed-field gel electrophoresis and provides a measure of phylogenetic relatedness.* Journal of clinical microbiology, 2003. **41**(5): p. 2191-2196.
140. Jordan, K. and O. McAuliffe, *Chapter Seven - Listeria monocytogenes in Foods*, in *Advances in Food and Nutrition Research*, D. Rodríguez-Lázaro, Editor. 2018, Academic Press. p. 181-213.
141. van der Vossen, J., et al., *12 - A comparison of molecular technologies and genomotyping for tracing and strain characterization of Campylobacter isolates*, in *Tracing Pathogens in the Food Chain*, S. Brul, P.M. Fratamico, and T.A. McMeekin, Editors. 2011, Woodhead Publishing. p. 263-274.
142. Haggerty, L.S., et al., *Gene and genome trees conflict at many levels.* Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 2009. **364**(1527): p. 2209-2219.
143. Oliveira, P.H., et al., *The chromosomal organization of horizontal gene transfer in bacteria.* Nature communications, 2017. **8**(1): p. 841-841.
144. PulseNet. *Pulse-field Gel Electrophoresis (PFGE)*. 2019 31.05.2019]; Available from: https://www.cdc.gov/pulsenet/pathogens/pfge.html.
145. Saltykova, A., et al., *Comparison of SNP-based subtyping workflows for bacterial isolates using WGS data, applied to Salmonella enterica serotype Typhimurium and serotype 1,4,[5],12:i.* PloS one, 2018. **13**(2): p. e0192504-e0192504.
146. Long, S.W., et al., *Whole-Genome Sequencing of Human Clinical Klebsiella pneumoniae Isolates Reveals Misidentification and Misunderstandings of Klebsiella pneumoniae, Klebsiella variicola, and Klebsiella quasipneumoniae.* mSphere, 2017. **2**(4): p. e00290-17.
147. Tyler, A.D., et al., *Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications.* Scientific Reports, 2018. **8**(1): p. 10931.
148. Lemon, J.K., et al., *Rapid Nanopore Sequencing of Plasmids and Resistance Gene Detection in Clinical Isolates.* Journal of clinical microbiology, 2017. **55**(12): p. 3530-3543.

149. Lifescience, R. *MagNA Pure 96 DNA and Viral NA Small Volume Kit*. 2018 28.11.2018]; Available from: https://www.n-genetics.com/products/1295/1023/16530.pdf.
150. Scientific, T.F. *Quant-iT 1X dsDNA HS Assay Kit* 2017 28.11.2018]; Available from: https://www.thermofisher.com/document-connect/document-connect.html?url=https://assets.thermofisher.com/TFS-Assets/BID/manuals/MAN0017526_Quant_iT_1X_dsDNA_HS_Assay_Kit_UG.pdf&title=User%20Guide:%20Quant-iT%201X%20dsDNA%20HS%20Assay%20Kit.
151. Scientific, T.F. *QubitTM 1X dsDNA HS Assay Kit*. 2017 28.11.2018]; Available from: https://www.thermofisher.com/document-connect/document-connect.html?url=https://assets.thermofisher.com/TFS-Assets/BID/manuals/MAN0017455_Qubit_1X_dsDNA_HS_Assay_Kit_UG.pdf&title=User%20Guide:%20Qubit%201X%20dsDNA%20HS%20Assay%20Kit.
152. Illumina. *Hamilton Microlab STAR Reference Guide*. 2015 28.11.2018]; Available from: http://emea.support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/mlstar/hamilton-ml-star-reference-guide-15070074-a.pdf.
153. illumina. *Nextera XT Library Prep Kit Reference Guide* 2018 28.11.2018]; Available from: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_nextera/nextera-xt/nextera-xt-library-prep-reference-guide-15031942-03.pdf.
154. Technology, A. *Agilent High Sensitivity DNA Kit Guide*. 2013 28.11.2018]; Available from: https://www.agilent.com/cs/library/usermanuals/Public/G2938-90321_SensitivityDNA_KG_EN.pdf.
155. Illumina. *MiSeq Systems: Denature and Dilute Libraries Guide*. 2018 28.11.2018]; Available from: http://emea.support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/miseq-denature-dilute-libraries-guide-15039740-06.pdf.
156. Illumina, *Nextera XT Library Prep Reference Guide*. 2018.

APPENDIX A

List of protocols for all kits and instruments used:

- **MagNA Pure 96 DNA and Viral NA Small Volume Kit: [149]**
- **Quant-iT™ 1X dsDNA HS Assay Kit: [150]**
- **Qubit™ 1X dsDNA HS Assay Kit: [151]**
- **Hamilton Microlab Star Reference Guide for Illumina: [152]**
- **Nextera XT Library Prep Kit Reference Guide: [153]**
- **Agilent High Sensitivity DNA Kit Guide: [154]**
- **MiSeq System: Denature and Dilute Libraries Guide: [155]**
- **Illumina Experiment Manager: Software Guide: [95]**
- **Sequencing Analysis Viewer Software Guide and User Guide: [79, 98]**
- **Standard Operating Procedure for PulseNet PFGE of various Enterobacteriaceae [118]**
- **BioNumerics Quick Guide [119]**

APPENDIX B

Performing the Nextera XT protocol on the Hamilton Microlab STAR pipetting robot:

1. Turn on the ML STAR
2. Turn on the instrument control computer and enter your user name and password
3. From the ML STAR desktop, open Hamilton App Launcher
4. Select **Maintenance and Verification**
5. Select appropriate program
6. Select appropriate method
7. Enter your user name and password, and then click **OK**
8. Follow the on-screen instructions to load the ML STAR carriers, and select the sample number. Click **OK** after loading each carrier
9. Click **OK** to verify all lab-ware positions and begin the run

Different reagents and procedures are needed to ensure a successful library prep; these are described in the following paragraphs.

A) Tagmentation of genomic DNA:

During tagmentation, the DNA is simultaneously fragmented and tagged with specialized adapter sequences by the Nextera transposome.

1. 5 µL of normalized gDNA with a concentration of 0.2 ng/µL was transported from the midi plate to a hard-shell PCR plate (Bio-Rad Laboratories, Hercules, CA, USA).
2. 10 µL of Tagment DNA Buffer (TD) was added and mixed.
3. 5 µL of Amplicon Tagment Mix (ATM) was added to each well and mixed.
4. The plate was shaken on the plate shaker at 280 x g at 20^{o}C for 1 minute.

5. The plate was held at 55ºC for 5 minutes before being held at 10ºC, which is the tagmentation program.

6. 5 µL of Neutralize Tagment Buffer (NT) was added to each well and mixed.

7. The plate was subsequently shaken at 280 x g for 1 minute prior to incubation at room temperature for 5 minutes.

B) Library amplification:

The DNA libraries are amplified by adding indexes and full adapter sequences to each tagmented DNA fragment, which are essential for cluster generation.

The Nextera® XT Library Preparation Kit applied dual indexing and an illustration of the positions of index 1 and index 2 is shown in Figure A1.

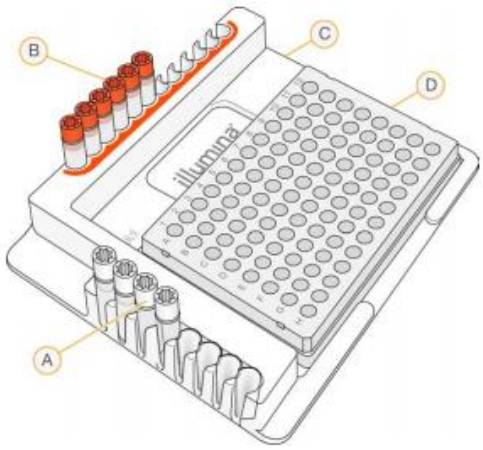


Figure A1: An alternative set-up of indexes 1 and 2. Rows A-D illustrate index 2 (i5) adapters, columns 1-6 illustrate index 1 (i7) adapters.
Adapted from: [156]

In this study, 32 samples were prepped in each library preparation, resulting in a setup of four i5 indexes and four i7 indexes.

1. 5 µL of each index was added to their respective rows and columns. All caps were replaced after use to avoid cross contamination of the indexes.

2. 15 µL of Nextera PCR Master Mix (NPM) was added to all wells before the plate was sealed and removed from the pipetting robot and placed in a thermal cycler, and run on the program "Library Amplification" with the following settings:

- 72°C for 3 minutes
- 12 cycles of:
  - 95°C for 10 seconds
  - 55°C for 30 seconds
  - 72°C for 30 seconds
- 72°C for 5 minutes
- Hold at 10°C.

1C) Library clean up:

To remove short library fragments, non-attached indexes and adapters, AMPure XP Beads (Table 3) are used to purify the DNA libraries.

1. The plate with the amplified libraries was shaken at 280 x g at 20°C for 1 minute.
50 µL of PCR product was transferred from each well of the PCR plate to a new midi plate.

For the concentration of AMPure XP Beads to be correct, the beads were thoroughly vortexed before ~ 36 µL were pipetted by hand, and placed into the pipetting robot. A low concentration of beads would give larger fragments, and the smaller fragments would disappear.

2. The beads were pipetted into each well of the midi plate.

3. The midi plate was shaken at 1800 rpm for 2 minutes and incubated at room temperature for 5 minutes

4. The plate was placed on a magnetic stand for ~ 2 minutes, until the liquid was clear.

5. The supernatant was removed from each well.

6. 200 µL of freshly made 80% ethanol (EtOH) was added to each well.

7. The plate was incubated on a magnetic stand for 30 seconds, before the supernatant was removed and discarded. This wash was performed two times.

8. The residual 80% EtOH was removed from each well of the plate, which was then left to air-dry on a magnetic stand for 15 minutes.

9. After being removed from the magnetic stand, ~ 52.2 μL of RSB was added to each well, and the plate was shaken at 1800 rpm for 2 minutes.

10. The plate was again placed on a magnetic stand for ~ 2 minutes for the liquid to clear.

11. Lastly, 50 μL of supernatant was transferred from the midi plate to a new hard-shell PCR plate.

1 μL of undiluted, clean library could now be quantified as a quality check using the Agilent High Sensitivity DNA Kit for 2100 Bioanalyzer System.

2C) Agilent High Sensitivity DNA Kit for 2100 Bioanalyzer Systems:

To quantify and quality check the cleaned library, an Agilent High Sensitivity DNA Chip (Table 3) was used together with the 2100 Bioanalyzer System. For this procedure the Agilent High Sensitivity DNA Kit Guide [154] (Appendix A) was followed.

D) Library normalization:

To ensure an equal library representation in the pooled library, the quantity of each DNA library is normalized; a process called library normalization. The method of normalization applied is called bead normalization, which ensures that an equal amount of DNA library binds to DNA binding normalization beads and elutes at approximately equal concentration for each DNA sample.

1. 20 μL of supernatant from the hard-shell PCR plate containing the cleaned library was transferred to a new midi plate.

2. A mixture of 44 μL normalization buffer (LNA1) and 8 μL of bead mixture (LNB1) was mixed by pipetting by hand in a conical tube, before being placed in the pipetting robot. LNB1 was vortexed before pipetting.

3. 45 μL of the mixture was added to each well, and the plate shaken at 1800 rpm for 30 minutes.

4. The plate was then moved to a magnetic stand for the liquid to clear for ~ 2 minutes before the supernatant was removed and discarded.

5. 45 µL of normalization wash buffer (LNW1) was added to each well, the plate was shaken at 1800 rpm for 5 minutes.

6. The plate was moved to a magnetic stand for the liquid to clear for ~2 minutes.

7. The supernatant was then removed and discarded, before the wash was performed one more time.

8. 30 µL of freshly prepared 0.1 M NaOH and the plate was shaken again at 1800 rpm for 5 minutes.

9. 30 µL of normalization storage buffer (LNS1) was added to each well of a new hard-shell PCR plate, to neutralize the NaOH.

10. After the 5 minute elution, the supernatant was transferred to the new hard-shell PCR plate, which was then shaken at 1000 x g for 1 minute.

The plate containing the clean library and the plate containing the normalized library could now be stored at -25^{o}C to -15^{o}C for up to seven days.

E) Pooling of the libraries:

During the pooling process, equal amounts of the normalized libraries were combined in one tube. 5 µL from each well of the hard-shell PCR plate containing the normalized libraries was transferred to a new tube. The tube with the unused pooled libraries could be stored at -25^{o}C to -15^{o}C for up to seven days. The complete protocol for library preparation can be found in Appendix A.

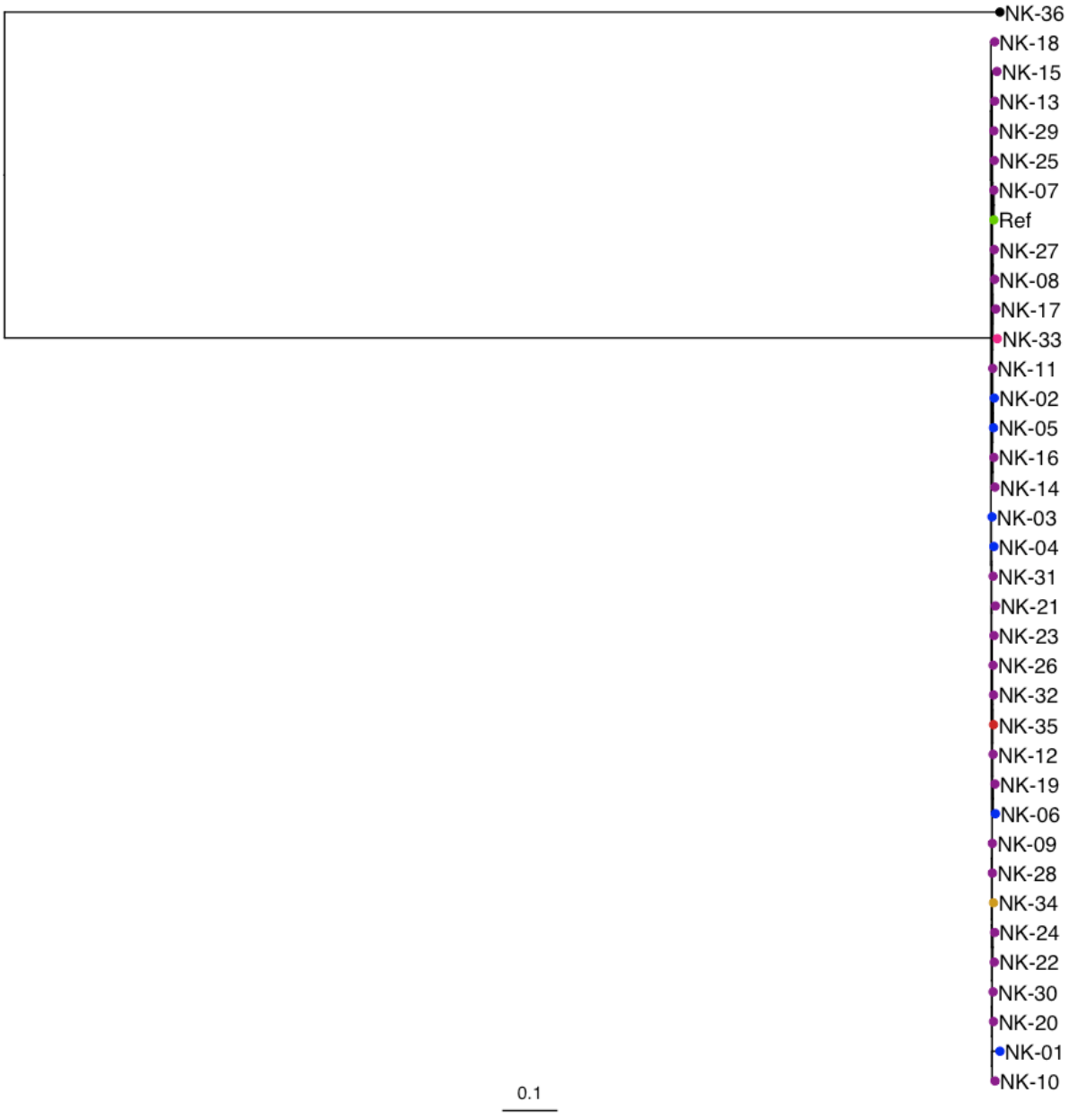
APPENDIX C

Number of samples re-sequenced, with the reason for re-sequencing and the outcome of the re-sequencing

Sample number	Reason for re-sequencing	After re-sequencing
NK-37	No contigs over 10 000 bp, total length = 7279, little Kleborate output	Largest contig = 653126 bp, total length = 5259112 bp
NK-38	# of contigs = 1001, L50 = 53	# of contigs = 73, L50 = 10
NK-39	# of contigs = 1102	# of contigs = 83
NK-40	ST107 missing wzi gene	wiz74
NK-41	ST107 missing wzi gene	wiz74
NK-42	Incomplete ESBL-gene: SHV-2?	Same result, determined SHV-2 by BLAST analysis
NK-43	L50=51	L50 = 7
NK-44	Incomplete ESBL-gene: SHV-12*?	Same result, determined SHV-2a by BLAST analysis
NK-45	L50 = 60	L50 = 8
NK-46	# of contigs = 404 (a little high) L50 = 51	# of contigs = 120, L50 = 7
NK-47	L50 = 48	L50 = 11
NK-48	L50 = 51, %GC = 58.03	L50 = 11, %GC = 57.57
NK-59	No ST given (alleles are given, so maybe novel ST?) and capsule wzi missing, but QC looks good.	Same result, later assigned to a new ST, ST4009
NK-50	No ST given (alleles are given, so maybe novel ST?) and capsule wzi missing, but QC looks good.	Same result, later assigned to new ST, ST4010
NK-51	No ST given (alleles are given, so maybe novel ST?) and capsule wzi missing, but QC looks good.	Same result, later assigned to new ST, ST4011
NK-52	Overall good, but N50 below 100k	N50 = 79700, still below 100k
NK-53	Low N50 (ST107)	N50 = 269061
NK-54	L50 = 41	L50 = 16
NK-55	L50 = 41, largest contig 100574	L50 = 9, largest contig = 796026 bp
NK-56	Overall good, but N50 below 100k	N50 = 94764, still below 100k

APPENDIX D

ST107 core genome tree including Isolate 4/NK-36, 0.1 nucleotide substitutions per site.



APPENDIX E

SNP distance matrix for 36 ST107 isolates subjected to PFGE.