UiT The Arctic University of Norway

Faculty of Health Sciences

Investigation into the presence of CRISPR-Cas- and R-M systems in Klebsiella pneumoniae and correlation with antibiotic resistance, virulence and plasmids

Miriam Kristine Nilsen

Master of Science in Biomedicine MBI-3911 June 2020

Acknowledgements

This master`s thesis is the result of 2 years of intense study and good support of family, friends, co-workers, connections in other projects and of course my supervisors.

First and foremost, I would like to thank my supervisors Arnfinn Sundsfjord, Erik Hjerde, Ørjan Samuelsen and Niclas Peter Raffelsberger for endless patience, availability, guidance and the possibility to be a part of the KLEB-GAP project. I would also like to thank everyone involved in the KLEB-GAP network for their willingness to share experiences, data and to share contributions like protein profiles, scripts and more. It is fair to say that they have contributed tremendously through the project, so thank you!

Secondly, I would like to give a special thanks to Marit Hetland, Espen Åberg and Espen Mikal Robertsen for all the help in bioinformatics! The results would not have been possible without the guidance, help and additional scripts. Thank you so much for teaching me commands, coding, scripting and debugging, so I could work independent with bioinformatics!

I have also received good help and advices from Lotte Leonore Eivindsdatter Andreassen that did a great job in preparing, screening and documenting the information about the carrier strains. I would also like to thank Aasmund Fostervold for the national ST307 strains from both NORM and NORKAB collection. I also give my thanks to João Pedro Alves Gama for sharing documentation, giving advices and being of great help in my project.

My gratitude also goes to my fellow students that encouraged me through the 2 years of study. I especially have to thank Marianne Kjersem, Arja Arnesen Løchen and Lennart Van Ligtenberg, thank you for the support and encouragement thought the project!

Needless to say, I am so grateful for the support from my family through the years! The moral support, understanding and the encouragements have been endless. And a special thanks goes to Arne-Endrè Karlsen for the constant support, understanding and encouragements through the years. And last, but not least my little rabbits providing snuggles and support in my home office.

Abstract

The aims were to perform a bioinformatic-statistical analysis of CRISPR-Cas- and R-M systems in a diverse whole genome sequenced *K. pneumoniae* population, and their correlations to virulence score and AMR -/ plasmid content. The strains (n=999) consisted of Norwegian fecal carrier (n=484) and clinical (NORM; ESBL and non-ESBL producing) (n=414), and national-international clinical ST307 strains (n=101).

Structural complete CRISPR-Cas systems were found in 26% of the strains; carrier (30%), NORM non-ESBL (26%), NORM-ESBL (29%), and ST307 (0%). R-M systems were found in 48% of the strains; carrier (43%), NORM (44%) and ST307 (90%). The presence of CRISPR-Cas- and R-M systems seems to be equally distributed between carrier and clinical strains. The systems distributions had ST-specific profiles as illustrated with the ST307 strains.

Some significant cross-population correlations were observed between the presence/absence of CRISPR-Cas-/R-M systems in terms of MGE acquisition, represented by virulence score, AMR - and plasmid content. CRISPR-Cas systems strains were associated with a higher virulence score and a lower AMR-/plasmid load. The R-M systems strains were associated with lower virulence score and a higher AMR-/plasmid load.

Future studies should include analysis of CRISPR spacer content and specificity, overall phage content and utilize more advanced comparisons and statistical analyses.

Table of Contents

Keywords: *Klebsiella pneumoniae*, Horizontal gene transfer, CRISPR Cas systems, Restriction- Modification systems, Antimicrobial resistance, Tromsø7, NORM, NORKAB, ST307

Abbreviations

AMR = antimicrobial resistance

AST = antimicrobial susceptibility testing

BSI = bloodstream infections

CAI = community acquired infections

CRISPR- Cas = clustered regularly interspaced short palindromic repeats -CRISPR associated proteins

ESBL = extended spectrum β-lactamases

EUCAST = European Committee on Antimicrobial Susceptibility Testing

HAI = hospital acquired infections

HGT = horizontal gene transfer

KPC = *Klebsiella pneumoniae* carbapenamase

KpSC = *Klebsiella pneumoniae* species complex

MDR = multidrug resistance

MGE = mobile genetic elements

NCBI = National Centre for Biotechnology Information

RS = restriction sites

R-M = restriction- modification

WHO = World Health Organization

UTI = urinary tract infections

VGT = vertical gene transfer

1 *Klebsiella pneumoniae* - a global opportunistic multidrug resistant pathogen

The first known description of *K. pneumoniae* was done by Carl Friendländer documenting the bacteria as a cause of pneumonia in 1882 (1)(2). Since then *K. pneumoniae* has been established as a major opportunistic pathogen causing infections primarily in hospitalised patients (1). The most common infections caused by *K. pneumoniae* are urinary tract infections (UTIs), bloodstream infections (BSIs) and pneumonia (3)(4)(5)(6).

Due to the ability of *K. pneumoniae* to acquire antimicrobial resistance (AMR) genes and develop multidrug resistance (MDR), the World Health Organization (WHO) has now acknowledged *K. pneumoniae* on their critical pathogen priority list "*Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics*" (7). Moreover, *K. pneumoniae* is also included in the ESKAPE pathogens (*Enterococcus faecium*, *Staphylococcus aureus*, *K. pneumoniae*, *Acinetobacter baumannii, Pseudomonas aeruginosa and Enterobacter*) that is known to cause difficult-to-treat MDR and/or hypervirulent nosocomial infections (1)(3).

In addition to its ability to acquire MDR, *K. pneumoniae* has also the capability to disseminate AMR genes within and across species through horizontal gene transfer (HGT) (8). However, there are mechanisms limiting the acquisition and adaptation of foreign DNA protecting the host from unwanted DNA. These mechanisms include Restriction- Modification (R-M)- and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)- CRISPR associated proteins (Cas) systems (9)(10). Previous publications indicate that the CRISPR-Cas systems are not equally distributed in *K. pneumoniae* populations, but highly associated to certain ST types (11). A recent review on the population genomics of *K. pneumoniae* underscores knowledge gaps in understanding the potential relatedness of CRISPR-Cas and R-M-systems on plasmid and phage diversity (1). The association between the presence/absence of R-M- and CRISPR-Cas systems and the presence/absence of defined accessory genome elements (plasmids, AMR- and virulence genes) in different *K. pneumoniae* strain collections is the topic of this thesis.

1.1 Relevant characteristics of *Klebsiella pneumoniae*

K. pneumoniae is a rod-shaped, gram-negative, facultative anaerobic, non-spore forming opportunistic bacteria (3). In humans, *K. pneumoniae* is mostly found as part of the normal mucosal flora, particularly in the lower intestines (3).

1.1.1 Taxonomy

Klebsiella pneumoniae forms a group of bacteria within the order of '*Enterobacteriales*' (synonym: *Enterobacterales* ord. nov) consisting of seven families; *Enterobacteriaceae*, *Erwiniaceae* fam. nov., *Pectobacteriaceae* fam. nov., *Yersiniaceae* fam. nov., *Hafniaceae* fam. nov., *Morganellaceae* fam. nov., and *Budviciaceae* fam. nov. (12). The *Klebsiella* genus is located within the family of *Enterobacteriaceae* (12). The order includes 60 different genera (by 2016) and over 250 species (12). However, most of the species is within the family *Enterobacteriaceae,* resulting in a highly taxonomically diverse family (12).

The *K. pneumoniae* species complex (KpSC) includes seven phylogroups (Kp1-Kp7); *K. pneumoniae* sensu stricto (Kp1), *K. quasipneumoniae* subsp. *quasipneumoniae* (Kp2) and subsp. *similipneumoniae* (Kp4), *K. variicola* subsp. *variicola* (Kp3) and subsp. *tropica* (Kp5), "*K. quasivariicola*" (Kp6), and *K. africana* (Kp7) (13). Kp5 and Kp7 do not have a formal taxonomic status yet (1). The seven phylogroups (species) are closely related and share 95-96% average nucleotide identity (1). There are significant gaps in knowledge concerning the ecology of *K. penumoniae* phylogroups, and their main reservoirs and distributions may be distinct (14). In this study, *K. pneumoniae* sensu stricto, referred as *K. pneumoniae* will be the main focus as the dominant species associated with infections in humans (13).

It is a challenge in the clinical microbiology laboratory to phenotypically distinguish between the seven phylogroups (1)(13). Thus, most strains will technically be identified as KpSC (1)(13). In the clinical setting, *K. pneumoniae sensu stricto* is the most frequent identified species by ~85% (1).

1.1.2 KpSC ecology and distribution

In brief, KpSC is not only found in humans, but also broadly in nature. The bacteria has been associated with plants, water, soil and a variety of animals (1)(3)(15). Spread of the bacteria itself, can therefore happen across many niches (3). However, the extent of transmission and the lines of dissemination between the different reservoirs need to be further examined (15). We also lack knowledge on the occurrence, relative abundance and characteristics of KpSC in different environments due to the absence of systematic studies (1).

The faecal carrier rate in humans varies between geographical locations and differences in study populations (8). Human faecal colonisation rates for *K. pneumoniae* have been reported to range from 6% in one Australian study, up to 62% in healthy Chinese adults and up to 88% in the Chinese population in Malaysia (16)(17). The culture-based KpSC detection in the seventh Tromsø population study (T7) in 2015-2016 performed in community based adults (≥ 40 years n=3000), revealed an overall faecal carrier rate of 16,5% with a relative abundance of *K. pneumoniae* sensu stricto (61%), *K. variicola* (28%) *K. quasipnemoniae subsp. quasipneumoniae* (7%) and *K. quasipneumoniae subsp. similipneumoniae* (4%)(18).

Pangenome studies have revealed a large genetic diversity within KpSC and raises the question of what gene repertoire are associated with the dissimilar locations (1)(19). A metabolic profiling study of various clonal lineages within KpSC has revealed a highly diverse set of biochemical properties for carbon and nitrogen metabolic capacities that could contribute to broad ecologic distribution of KpSC (19). In the same study, core metabolism features suggested an adaption to plant associated environments (19).

1.1.3 *K. pneumoniae* genomics and population structure

K. pneumoniae possesses a large and diverse genome (3). The genome has in average ~5-6 Mbp and ~5000-6000 protein coding genes (1). In comparison the *Escherichia coli* encompass ~5,1 Mbp and ~4915 protein coding genes (3)(8). The core genome consists of ~1700 genes that are present in all strains regulating basic functions for survival in different environments (1)(3)(8). In addition, *K. pneumoniae* host ~3300-4300 accessory genes that varies between strains illustrating a large adaptive capacity (Figure 1) (1)(3)(8). This results in a diverse pangenome (core genome + all accessory genes detected in various *K. pneumoniae* strains) emphasizing a high capacity for horizontal gene transfer (HGT), including chromosomal recombination, plasmids and bacteriophages (3)(8). In the latest genomic review of *K. pneumoniae* it was estimated that the increasing pangenome now surpass 100.000 open reading frames (1).

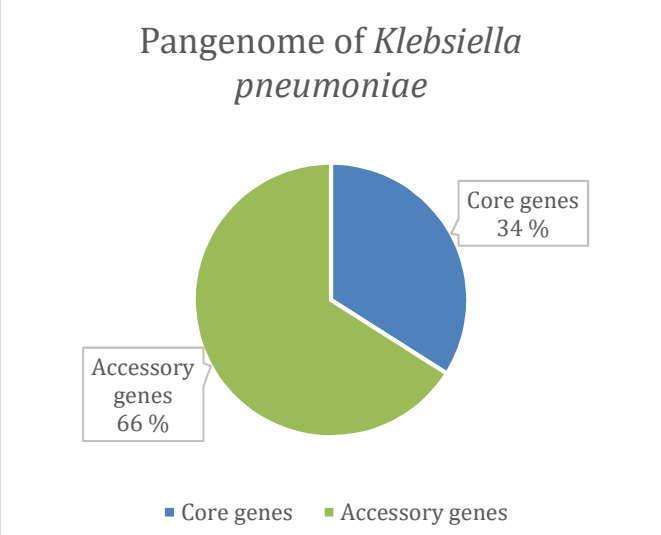


Figure 1: Pangenome of Klebsiella pneumoniae and its average distribution between core genes and accessory genes (1)(3)(8). In the example, the pangenome was set at a total 5Mbp

There are 100s of independent phylogenetic lineages or "clones" in *K. pneumoniae* based on core genome analyses (8). They only differ by ~0,5% nucleotide divergence (1)(8). A clone is genetically defined as a group of genetically closely related strains sharing a recent common ancestry of the genome (1)(20). Clonal groups (CG) or clones of *K. pneumoniae* are defined by a core genome multi locus sequence typing (cgMLST) scheme that consists of 694 different alleles and/or by the traditional seven core-genes MLST system (21)(22). *K. pneumoniae* cgMLST CGs are a groups of cgMLST profiles having <100/694 allelic mismatches (i.e., 14,4% of the 694 alleles) with at least 1 other member of the group (22). The low level of genomic diversity among *K. pneumoniae* strains as defined by traditional MLST classification has made it difficult to line up boundaries between clones (21). However, the nearly 700 allelic cgMLST system provide a 100x more sequence information than the seven-gene MLST (21)(22). Clones can also be separated by looking into the accessory genome (8).

The diverse ST types may possess differences in pathogenicity and some are spreading globally (2). Some clones have been more associated with MDR and others with hypervirulence causing serious community acquired infections (CAI) (3)(8). It is today still not fully understood why this distinction occurs, but it might be connected to the physical availability of either MDR or hypervirulence in the specific niche, lower fitness costs for some ST carrying plasmids, antibiotic induced pressure or possibly the absence/presence of systems restricting HGT (3)(8)(15)(23).

MDR clones can be defined as resistant to ≥3 antimicrobial classes in addition to their intrinsic resistance to ampicillin (1)(24). The presence of MDR is favourable in hospital environments where the bacteria may need an antibiotic resistant phenotype for its survival (1)(3)(15). Some MDR-phenotypes like those expressing extended-spectrum β-lactamases (ESBLs) or carbapenamases have been shown to be spreading globally causing hospital outbreaks and are of particular concern (1)(3). Local (25) and global comparative genomic studies (3) of *K. pneumoniae* have revealed a diversity of MDR-clones and some distinct successful epidemic clones that have become defined as "global MDR clones", including CGs 258, 15, 20, 29, 37, 101, 147, and 307 (1)(5)(26).

Hypervirulent clones have mostly been reported spreading from Taiwan and Southeast Asia since the mid-1980s and pose a concern because of their ability to cause serious CAIs (2)(27). As opposed to MDR-clones, hypervirulent clones are dominated by a limited number of CGs (1). It is of great clinical concern that recent research have documented convergence of MDR

and virulence at the strain level (1). The mechanisms involved in convergence have recently been reviewed (1) and seem to include the acquisition of plasmids encoding virulence factors (virulence plasmids) or hybrid AMR-virulence plasmids by MDR clones (1)(8). Global MDR- and hypervirulent clones are named high-risk clones.

Nosocomial infections in humans have often been associated with high-risk clonal types like CG23, CG25, CG65 (including ST65 and ST375), CG66, CG86, ST258 and the more recent ST307 (1). The situation is dynamic, and in the last 5 years the high-risk ST307 clone has displayed a significant increased prevalence (26). ST307 is also represented by *Klebsiella pneumoniae c*arbapenamase (KPC) producing strains emerging from the globally disseminated ESBL CTX-M -producing parental strain (26). This CG is associated with frequent nosocomial outbreaks caused by MDR phenotypes (1)(8). Clinically, patients infected with KPC producing ST307 displayed over 50% mortality and longer hospital stay, compared to patients infected with other strains (26).

1.1.4 Antimicrobial resistance and development of MDR

K. pneumoniae is well known for the ability to develop and spread AMR genes through HGT within and across species (8). This is an important feature in the development of MDR lineages such as ST258 and ST307 (8). In Europe hospital acquired infections (HAIs) like the high-risk MDR-associated STs, including ST11, ST15 and ST258, are seen frequently (1). In addition to these ST types, ST70 and ST323 alongside CG20 (CG17), CG29, CG37, CG147, CG101 (CG43) and MDR- CG307 are most often observed causing nosocomial outbreaks (1).

By 2018 over 400 AMR genes had been identified in available genomes of *K. pneumoniae*, the majority being plasmid borne (3). In the latest population genomics review of *K. pneumoniae*, several hundreds of distinct acquired AMR alleles are referred (1).

AMR-genes in *K. pneumoniae* can be divided into intrinsic and acquired determinants (1). *K. pneumoniae* species complex (KpSC) carry intrinsic class A β-lactamase genes (bla_{SHV} in *K. pneumoniae* sensu stricto, bla_{LEN} in *K. variicola*, bla_{OKP} in *K.quasipneumoniae*) that confer clinical resistance to penicillins including ampicillin and piperacillin (1)**.** Moreover, *oqxAB* and *fosA* are core genes in *K. pneumoniae* that mediate reduced susceptibility, but not clinical resistance, to fluoroquinolones or fosfomycin, respectively (1)(3)(28)(29). There is also numerous core genes participating in AMR development by point mutations increasing resistance, particularly genes associated with lipopolysaccharide (LPS) production and efflux or membrane permeability (1)**.**

The acquired AMR alleles can be grouped by which antibiotic drug class they might affect using different bioinformatic tools. The most clinically relevant classes of antibiotics in the treatment of *K. pneumoniae* infections include aminoglycosides, β -lactams (e.g. cephalosporins and carbapenems), β-lactam β-lactamase inhibitor combinations (e. g piperacillin-tazobactam and ceftazidime-avibactam), fosfomycin, fluoroquinolones, polymyxins (colisitin), sulphonamides, trimethoprim and tigecycline (1)(3)(28)(29). A recent review have summarized studies that have explored the distribution of acquired AMR genes in different *K. pneumoniae* populations (1)**.** The review revealed a bimodal distribution represented by global MDR clones with a high genetic AMR load affecting a number of drug classes (≥6), in contrast to hypervirulent clones hardly carrying any acquired AMR-determinants at all. Several factors that could contribute to the irregular distribution of AMR between different genetic lineages, including differences in host plasmid maintenance mechanisms, have been suggested (1). The spread and distribution of β -lactamase-genes encoding ESBL are of particular concern and will be discussed further.

ESBL producing *K. pneumoniae*. ESBL mediate resistance to the oxyimino-cephalosporins (cefotaxime, ceftazidime, ceftriaxone and cefepime) and monobactams (aztreonam) (30).

ESBLs are classified in different ways, but for this study the Giske et al. definition will be utilized (24). This classification divides ESBLs into three main groups; $ESBL_A$, $ESBL_M$ and $ESBL_{CARBA}$-, all acquired by HGT (6)(31). The $ESBL_A$ encoding genes includes $bla_{CTX\text{-}M}$, and some allelic variants of bla_{SHV} and bla_{TEM} (6)(31). $ESBL_A$ type β -lactamases hydrolyses all penicillin, monobactams and cephalosporins, leaving only cephamycin, carbapenems and penicillin + β -lactam inhibitors as potential useful antimicrobial agents (6)(31). The $ESBL_M$ group is a diverse group of β -lactamases where the plasmid-mediated AmpCs are most prevalent, including CMY and DHA (6)(31). $ESBL_M$ hydrolyse most cephalosporins, all monobactams and penicillins, leaving only carbapenems and 4th generation cephalosporins to work. Most $ESBL_{CARBA}$ enzymes (except members of the OXA-family) hydrolyse all β -lactams, including the carbapenems and include Ambler class B metallo β-lactamases (ex. VIM and NDM), class A serine (ex. KPC) and class D (ex. OXA-48 family) carbapenamases (6)(31).

The ESBL-genes are most often located on plasmids, often carrying additional AMR-genes towards important commonly used antibiotics such as aminoglycosides, trimethoprim-sulfamethoxazole and fluoroquinolones (6)(32). This contributes to development of MDR, defined as acquired resistance to three or more different antibiotic classes (33). Transmission

of ESBL -producing bacteria (*Enterobacteriaceae)* most often occur through faecal-oral contamination in the community, but also involves various transmission lines in hospital-environments (32).

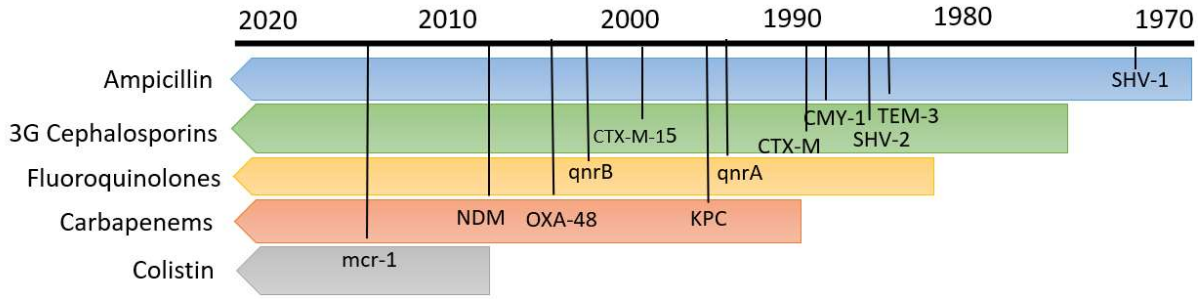


Figure 2: The timeline of clinically important classes of AMR-genes detected in Klebsiella pneumoniae from 1970 to 2020 (3)(27). The picture is modified from reference (3).

The historical acquisition of different classes of ESBL-genes in *K. pneumoniae* is outlined in Figure 2 (3). Throughout 1980-1990s ESBL-variants of TEM and SHV emerged at least partly as a response to the use of third generation cephalosporins (3)(30). In the 1990s the CTX-M-type ESBLs started to dominate and also causing CAIs with CTX-M producing *E. coli* (28)(30). The CTX-M are divided into five groups; 1, 2, 8, 9 and 25 based on amino acid homology (34). CTX-M-14 and CTX-M-15, have been the most frequently appearing ESBLs in worldwide surveillance studies (6)(28)(35)(36). Some CTX-Ms are mostly bond to geographical locations (6)(28)(35)(36).

The plasmid mediated genes *qnrA* and *qnrB* mediating resistance to quinolone were later detected and has disseminated worldwide in the high-risk MDR-clones including ST258 (3)(37). The OXA, CTX-M-14/15, CMY-1, NDM-1 and KPC β -lactamases have increased significantly globally since 2005 (3). The *mcr-1* gene is often plasmid borne and mediates colistin resistance by modifying LPS as a response to colistin exposure (2)**.**

1.1.5 Pathogenicity

Pathogenicity is defined as the ability to cause disease (38). A virulence factor by definition enhances pathogenicity (38). Virulence factors can be either intrinsic or acquired (38). Intrinsic virulence factors are encoded by loci present in all *K. pneumoniae* making it generally capable of causing infections (1). *K. pneumoniae* is defined as an opportunistic pathogen in general, but some strains have acquired more virulence factors making them a true pathogen also causing CAIs (1). See Table 1 for an overview of the most common and validated virulence factors.

Table 1: Important intrinsic and acquired virulence factors in *K. penumoniae*			
Virulence factor	Variants associated with increased pathogenicity (encoding genes)	Associated feature	Biological function
K-antigens (Capsule polysaccharide; CPS)	K1, K2 & K5 especially	Increased survival in serum and tissue	Antiphagocytic effect preventing phagocytosis and intracellular killing by macrophages and neutrophils (2)(39). The lack of specific mannose residue repeats recognised by the immune system of the host (2)(39). Slightly increased release of reactive oxygen species by neutrophils (2)(39). Host-specific monosaccharide sialic acid on the surface mimicking host cells (2)(39)
O-antigen (lipopolysaccharide, LPS)	O1, O2 & O3 especially	Vascular survival by altering outer membrane properties?	Little to no documentation, just assumptions that the biosynthesis associated O3 might be associated with survival in blood (39). Decreased local proliferation and dissemination suppressing the inflammatory response (40)
Regulator of the mucoid phenotype A gene (rmpA)	*RmpA* and *rmpA2*	Hypermucoid phenotype	Increasing capsule production creating a hypermucoid "sticky" bacteria (8)(39)
Siderophores	Aerobactin (*icu*), Salmochelin (*iro*), Enterobactin (*ent*) and Yersiniabactin (*ybt*)	Increased survival in the vascular system & ascites fluid. Necessary for clinical infection (1)	Enhances iron sequestration and thus promotes survival in the vascular system (*icu, iro* and *ent*)(8). *Ybt* provides an iron scavenging system and has the ability to escape Lcn-2 binding and avoids the inflammatory response enhancing survival in the spleen (41). It can also bind other heavy metals avoiding toxicity (41). *Iro* modifies enterobactin to escape Lcn2 binding (41). *Icu* scavenge iron from transferrin (41)
Colibactin genotoxin	Located in the KPHP1208 pathogenicity island (39)	Increases the invasive potential	Eukaryotic cell death by DNA damage and promotes invasion from the intestine to the vascular system (1)(8)(39).
Fimbriae	Type 1 & 3	Increase colonization	Increases biofilm production and thus contribute to colonization (1). Adherence to environmental, biological- and non-biological surfaces (40)
Glycogen production	*gly*	Increased survival in the urinary system	Higher survival in nutrient deficient environments (26)

Intrinsic virulence factors found in all *K. pneumoniae* include enterobactin (*ent*), fimbriae (*fim* and *mrk*), lipopolysaccharide LPS (O-antigen) and capsular polysaccharide (K-antigen) (1)(39). These intrinsic factors comprise ~10% of the genome coding capasity and can be found

in many variations (1). Combinations of virulence factors may result in increased pathogenicity (8)(39). These combinations are often found in certain CG types, hypervirulent clones (e.g. CG23, CG65 an CG86 (1)(8)(39).

Previously *K. pneumoniae* has been serotyped using antisera distinguishing between different surface polysaccharides, K- and O- antigens (39). In total, there are at least 78 capsular serotypes for *K. pneumoniae*, but only K1 and K2 are strongly associated with hypervirulence and a hypermucoid phenotype (2)(8)(39). However, the K5 variant has also been associated with liver abscesses, but is not commonly seen (1). O serotype diversity is also connected to K1/K2 providing further mechanisms to invade the host (2)**.** However, the O-antigen is less understood and today there are mostly assumptions that it contributes to increased survival in the vascular system (39).

Acquired virulence factors enhance pathogenicity (1). Colibactin endotoxin was formerly only seen in *E. coli*, but has emerged in ~10% of *K. pneumoniae* and is associated with liver abscess clones (1). Regulatory genes like r*mpA* and *rmpA2* are associated with a hypermucoid phenotype as well as other siderophore gene clusters playing important roles (5)(37). The combination of K1- or K2 capsule types and the r*mpA* genes has shown a proven ability to enhance the pathogenicity (1)(2)(39).

Acquired siderophores have some similarities, but also mechanistic differences, and may therefore have an additive effect on pathogenicity in cooperation with the intrinsic Enterobactin (1). They are associated with hypervirulent strains causing invasive CAIs (1). Despite the many siderophores and their variants, aerobactin has been proven to play a greater role for hypervirulence and is the most prevalent acquired siderophore (2)**.** Aerobactin and salmochelin is often found co-located with the *rmpA/rmpA2* genes on virulence plasmids (1)(2)(8). However, the prevalence of these loci is still low in the *K. pneumoniae* population by only <10% (8).

1.1.6 Clinical perspectives

K. pneumoniae is an opportunistic pathogen most often associated with HAIs in immunosuppressed patients (1). UTIs, BSIs and pneumoniae are the classical infections (3)(4)(5)(6). *K. pneumoniae* thrive in the hospital environments, supported by the large genomic space available for adaptation, including AMR-genes and virulence determinants favouring survival within the antibiotic-exposed hospital environment (8)(26). Spread of *K. pneumoniae* can happen "silently" between patients or through health care workers and medical

equipment (3)(26)(15)(42)(43). Several studies have shown that the origin of the human *K. pneumoniae* infections is the patient's own gut microbiome (43). Studies have shown an overall prevalence of HAIs in Europe around 7 per 100 patients, increasing by the duration of patient-days in hospitals (15). Factors contributing to the prevalence of HAIs are access to resources, knowledge based hospital services, antibiotic restrictions and good basic hygienic practices (15). The situation in Scandinavia, compared to other European countries, displays a lower prevalence of nosocomial infections caused by MDR *K. pneumonia* (15). Previous HAIs was often linked to certain CGs with a MDR-phenotype, but recently CGs with a combined MDR- and hypervirulent phenotype have emerged, posing a clinical threat (1)(8).

2 Evolution of the bacterial genome- drivers and restrictions

Evolution of the bacterial genome is essential for survival and adaptation to new niches and selective pressures (44). There are many ways to achieve evolutionary adaptation including mutations and acquisition of new genetic material (44). However, unlimited access to new DNA is absolutely not in favour of the bacteria because it increases the fitness costs and may even be detrimental when it comes to bacteriophage infections (45). Therefore, mechanisms restricting, protecting and reviewing both new and already acquired genetic components are needed (44). The mechanisms are supposedly many, but the most studied and probably important ones are the CRISPR-Cas- and the R-M systems (9)(10). Although these system has been examined for decades, they are yet to be fully understood (46).

2.1 Horizontal gene transfer (HGT) contributing to evolution

HGT is the major driver for natural genetic diversity, evolution and bacterial survival including the dissemination of AMR genes, evolution of gene clusters encoding biochemical pathways and exchange of pathogenicity factors (46). Bacteria can share genomic elements through both HGT and vertical gene transfer (VGT), but for this study we will focus on HGT (20). HGT can be carried out in three principal mechanisms; transformation (uptake of free DNA), conjugation (cell-contact-mediated-transfer) and transduction (phage-mediated transfer) (20).

Transformation is a mechanism for bacterial uptake of free naked DNA from the environment shared by some bacterial species (44). Starvation, difficult growth conditions, low nutrient access or cell density induces a state of competence by expression of ~20-50 proteins making the bacteria accessible for naked DNA uptake (44). This process is not fully understood, but

the main mechanisms are recognized (44). The extracellular DNA often originates from decomposing cells, disrupted cells or viral particles (44). Translocation across the inner membrane is sometimes a more regulated process only allowing a certain length (44). If the sequence is highly similar to the receiving host genome, then the regions can initiate DNA pairing and exchange strands (44). The success rate of this process is highly varying between bacterial species and in general lower than other HGT mechanisms (44).

Plasmid transfer between bacteria occurs through bacterial conjugation (20). *K. pneumoniae* often harbours plasmids ranging in size containing a variety of accessory functions (47). Plasmids are extrachromosomal genetic elements, most often double-stranded (ds)DNA packages that can contain several core (encoding replication, mobility and potential transfer) and accessory genetic components varying in size (48). Usually, the size varies from a few kb and up to hundreds of kb (48). Current classification schemes use the backbone (core) loci for replication (replicon typing) or plasmid mobility (MOB typing) (49). Plasmids are easily shared through HGT, where the recipient can get access to essential mechanisms for survival in the environment (20)(47). However, fitness costs often increase by acquiring plasmids (45).

The conjugation process can be described by plasmid transfer, although only some plasmids are conjugative (i.e. encode transfer functions) (20). Others requires the help of another plasmid (helper-plasmids encoding conjugative functions) or other conjugative elements (i.e. conjugative transposons or integrative conjugative elements -ICE)(20). Conjugation is carried out by direct cell-to-cell contact (donor and recipient),by forming a relaxosome bridge, typically initiated by a pilus from the donor (20). This initiates synthesis of helicase/endonuclease nicking the donor DNA (20). Then the relaxosome forms a complex with other fertility factor proteins and unwinds the donors dsDNA before the transfer process can initiate (20). The single stranded (ss)DNA is transferred as a part of DNA-protein complex and DNA polymerase III is recruited for replication in the donor cell, pushing the rest of the strand through to the recipient (20). In the recipient cell, the strand quickly circulates and replicates (20). The conjugation complex falls apart and the membranes seal, leaving both cells as competent donors (20).

Bacteriophages possess the ability to move chromosomal DNA from a donor to another bacteria through transduction (20). Transduction has been divided into generalized- (the transfer of any gene) and specialized (transfer of only a few closely linked genes) transduction (20). Generalized transduction happens when the bacteriophage (phage) infects a cell, injects

viral DNA and makes subunit components for phage construction (20). Packing of phage- and/or host DNA into capsids and attachment to tails completes the formation of new phages (20). They are released through cell lysis and the new phages can inject a new host with DNA where it might initiative a new round of replication or recombine into the chromosome and create a lysogenic stage (20). Specialized transduction happens when improper excision of integrated phage DNA from the bacterial host chromosome takes place (20). This event is rare and results in the lack of a few viral genes where the host genes is inserted (20). Some of the phage can replicate themselves, but others heavily rely on a helper-phage harbouring the missing gene products (20). The product of this specialized transduction is a hybrid DNA also providing the receiving host with diploid host genes that may be a substrate for new recombination events (20).

2.2 CRISPR-Cas systems restricting evolution?

Clustered Regulatory Interspaced Short Palindromic Repeats (CRISPR) coupled with the CRISPR associated (Cas) proteins (CRISPR-Cas) are an adaptive immune system of prokaryotes protecting against foreign DNA (50). The repeated sequences of CRISPR-loci were first described in 1987, but their functional role as part of DNA-memory storage and a specific immune system was experimentally not proven until 2007 (50). Today this system provides a target for functional research for genetic engineering revealing many opportunities (27)(50). However, many CRISPR-Cas functions in the evolution of prokaryotes and their molecular mechanisms still remain unknown (50). Several studies have suggested that CRISPR-Cas systems also could be involved in biofilm formation, colonization and virulence regulation (11)(51).

A complex functional pathway has allowed the system to generate recognition of invading genetic elements through an evolving library of spacer sequences, generating memory. The system regulates exchange of genetic elements through HGT, induce removal of unnecessary genetic elements, regulates plasmid incorporation and virulence uptake in the bacteria harbouring these systems (11).

2.2.1 Structure, classification and basic characteristics

The CRISPR-loci, Cas proteins and CRISPR RNAs (crRNAs) are the basic functional parts of the system (51). A classic CRISPR-Cas locus consists of Cas genes, a leader sequence and a CRISPR array (Figure 3)(9)(10)(51). The CRISPR array consists of almost identical short direct repeats of ~21-47 nucleotides separating the spacers containing 20-60 nucleotides of

hypervariable sequences acquired from MGEs the system has been exposed to (45)(52). This sequence specific "memory" of the spacers, provides recognition of the same invading elements leading to destruction and thus preventing infection through the pathway (27)(45). The leader sequence is normally an AT-rich region usually ~100-500bp that possesses the ability to polarize potential new spacers using the proto-spacer adjacent motif (PAM) promoting transcription (27)(45). The Cas proteins vary in numbers and functions between CRISPR-Cas systems (27)(53)(54). The Cas proteins performs different enzymatic functions including nuclease, helicase or polymerase activity (11). Usually, the Cas genes are located close to the CRISPR arrays (10).

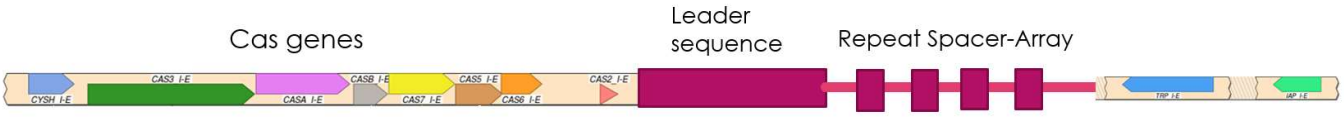


Figure 3: Structure of the CRISPR- Cas system displaying the Cas genes, the leader sequence, repeat spacer-array and the typical genes found upstream and downstream to the system. This figure was made using the SimpleSynteny output displaying a standard Class 1 type I-E system.

The Cas proteins and their functional pathways are most often divided into three functional steps; the adaptation (integration of spacers), expression (crRNA processing, maturation and target binding) and interference stage (target recognition and cleavage)(54)(55). Figure 4 displays the functional pathway for CRISPR-Cas type I systems. CRISPR- Cas systems in *Klebsiella* needs to be further studied, but so far *KpSC* is only found harbouring type I-E, I-V, I-F and the newly documented I-E* system (51).

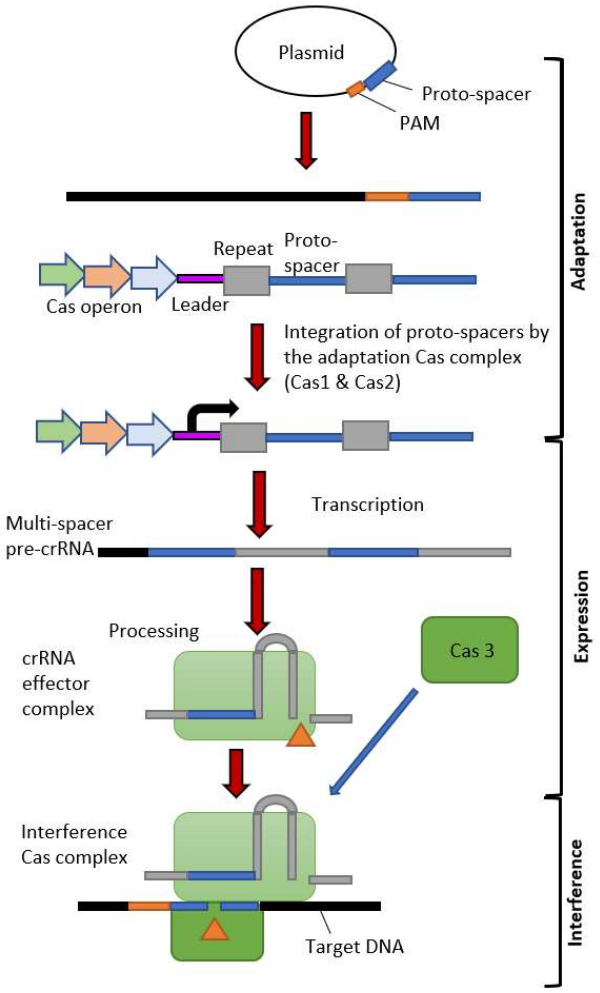


Figure 4: The functional pathway for CRISPR-Cas type I systems through adaptation (spacer integration), expression (crRNA processing and-maturation and the interference stage (target recognition, binding and cleavage). The figure is modified from (55).

The adaptation stage starts by recognition of the protospacer region by PAMs (27)(53). The protospacers are foreign DNA incorporated into the CRISPR arrays as small memory cassettes for future recognition of the same invading elements (54). The genetic fragments can be from DNA donors such as bacteriophages and plasmids (27). Some systems also have the ability to include RNA precursors after reverse transcription (27). The PAMs makes bonds with the adaptation Cas complex possible (27)(53). The adaptation Cas complex is formed by Cas1 and Cas2, for most known systems, which is necessary for spacer insertion (54). The specific mechanism and assembly of the adaptation Cas complex remains unknown (27)(53).

In the **expression stage** the temporarily spacers (protospacers) are transcribed from the leader, downstream to the CRISPR array and a multi-spacer pre-crRNA is formed (27)(54). Further the pre-crRNA is processed into a single spacer crRNA by either Cas9 (single multidomain protein) or a multi-subunit complex (for type I-E: Cas8 (CasA), Cas11 (CasB), Cas7, Cas5 and Cas6) forming the crRNA effector complex (type I systems), resulting in a more permanent mature spacer (27)(54)(56). Consequently, crRNA is processed into a shorter sequence of ~57 nucleotides by an endonuclease subunit of the multi-subunit effector complex (type I system)

or by an alternative mechanism performed by RNase III, an additional RNA species and tracrRNA (transactivating CRISPR RNA) (27)(53)(54).

The Cas proteins forms the **interference Cas complex** and binds to the mature crRNA and Cas9 or multi-subunit crRNA-effector complex (still bonded, but including Cas3) and this is where the interference stage starts (53). The Cas3 gene encodes a large protein with essential helicase and DNase activity (57). The interference complex works to prevent expression of foreign DNA, like phages, and leads to degradation by cleaving recognition sequences of DNA or RNA (53)(54). The spacer sequence can bind foreign DNA by perfect base pairing at the PAMs leading to recognition by the interference complex interfering with the foreign DNA causing degradation at the R-loop conformation through progressive hybridization (27)(53). Many functions and mechanisms are only slightly resolved (27)(53).

The constant need for evolution of this system has driven a modification of Cas genes resulting in multiple CRISPR-Cas systems with different mechanisms (Figure 5) (9). One driver that could partly explain the adaptive evolution of the Cas genes is the constant competition with virus-encoded dedicated anti-CRISPR proteins (9). Primarily, CRISPR-Cas systems are divided into two distinct classes based on the effector module (9)(58). Class 1 systems performs their functions through multi-subunit effector complexes (9). Whereas, Class 2 systems functions through single-protein effector modules, making the structure functions quite different (9)(58).

The two classes are also further divided into different subtypes based on phylogeny, function, gene arrangement, effector complexes and surrounding genes forming Type I, Type IV and Type III for the Class 1 systems and Type II, Type V and Type VI for the Class 2 systems (9)(51)(58)(59). See Figure 5.

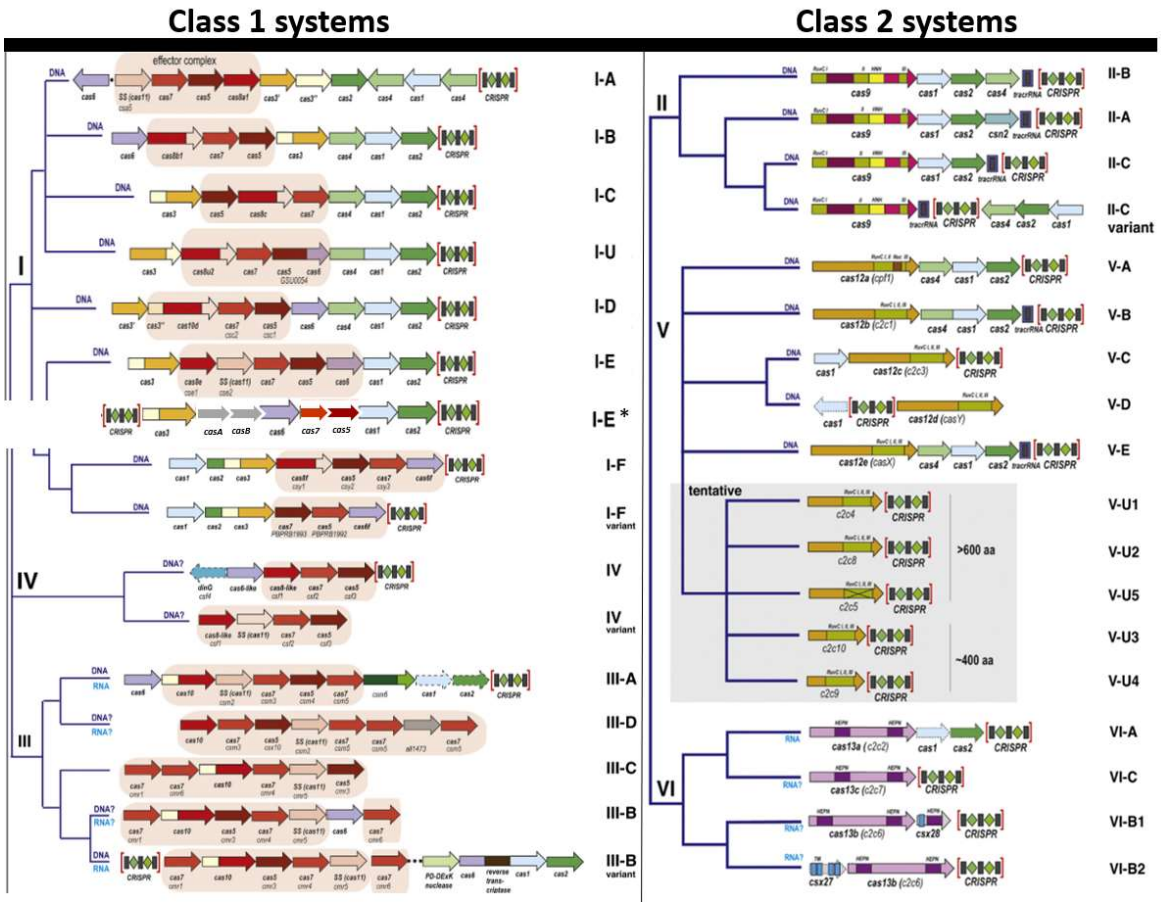


Figure 5: Class 1 and 2 CRISPR-Cas systems and their designated subtypes. Modified from(9)*. In addition, the new subtype I-E* was added displaying the gene arrangement, but not the correct phylogeny and marked effector complex because this still needs to be further examined* (51)*.*

Further subtyping (type + uppercase letters) is then done based on functional mechanisms and gene arrangements (9). The overall classifications nomenclature is best illustrated by an example from *K. pneumoniae* harbouring a Class 1 Type I system of subclass E, which is referred to a Class 1 Type I-E system (58).

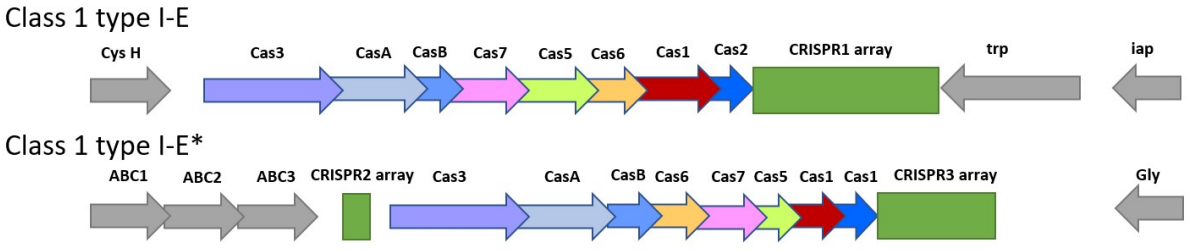


Figure 6: Genetic arrangements including adjacent genetic markers for the Class 1 type I-E and I-E subtypes found in K. pneumoniae* (10)(58)(59)*.*

The I-E and I-E* types are structurally different both in adjacent genetic markers located upstream or downstream, Cas gene arrangement and the number of CRISPR arrays, as seen in

Figure 6 (10)(58)(59). Signature genes, Cas 1 and Cas 3 both display homology within the subtype, but not between the subtypes (11).

Functional differences is seen as the subtype I-E* can integrate new spacers in both CRISPR arrays, however the system in *K. pneumoniae* seems to favour the second CRISPR array (labelled CRISPR3 array in the figure) adjacent to Cas1 (10). This could simply be a result of proximity or differences in AT content at the leader sequence for the arrays (10). The CRISPR2 marked for the I-E* system (Figure 6), has been found to have an AT content from ~46,8% in *K. pneumoniae* and the CRISPR3 an AT content of ~58-75% (10). A study of 40 CRISPR-Cas positive *K. pneumoniae* strains suggested that the I-E* subtype positive had less acquired spacers related to phages, plasmids and AMR-genes (including ESBLs), compared to the subtype I-E positive strains (11). The subtype I-E* has previously been associated with a lower number of plasmids and phage content and a higher susceptibility towards antimicrobial agents compared to the subtype I-E (11).

In addition to the chromosomally bound CRISPR-Cas subsystems I-E and I-E*, *K. pneumoniae* could also acquire ex. subtype IV only found in plasmids (60). The subtype IV has been detected exclusively on IncHI1b/IncFIB plasmids and is known to function in cooperation with other Cas-genes found in the chromosome (60). However, this system is not fully understood and the links to the KpSC must be further examined (60).

2.2.2 Distribution and genetic conservation

The CRISPR-Cas system evolved in bacteria as a protective response to invading genetic elements and is found with a prevalence of nearly ~50% in bacteria and ~80% archaea (9)(11)(20)(51). In total, the Class 1 systems are more prevalent then Class 2 systems and the former has been thought to be the ancestor (9). The Class 1 system has mostly been found in the chromosome for *K. pneumoniae*, but plasmid carriage is yet not sufficiently explored (45).

The Cas-genes display some resemblance, but low degrees of homology is observed between system types (58). The classes have already diverged in terms of structure and effector complex (58). The individual Cas-genes also display a varying degree of genetic conservation, making detection more difficult (58). For the Class 1 type I systems, the Cas1 is the most conserved gene with low resemblance to other genetic components (58). However the gene is prone to module shuffling and will thus cause difficulties in distinguishing some of the Class 1 systems (58)(57). Looking into the phylogeny between the Cas1 in Class 1 type I systems, it is clear that some of the subtypes is not easily distinguished (58). This co-evolution of the different types

of systems makes bioinformatical subtyping more difficult (58). Since the subtypes typically have well conserved PAMs and highly conserved repeats these becomes a good indicator alongside one signature Cas gene from the different types and subtypes (58). For the Class1 type I, Cas3 is the signature gene, despite being a multidomain protein and the close resemblance to other genetic components and low degree of conservation (58). However, some still argues that Cas1 is the best signature, for all Class 1 systems except type III (57). For the subtyping, other adjacent genetic components are required for a higher accuracy (10)(58). This is still under investigation.

2.2.3 Self-targeting spacers and riddance of the CRISPR-Cas system

A contradiction to the high value of harbouring a functional CRISPR-Cas system is that it often contains self-targeting sequences with potential to kill the host (51). The prevalence self-targeting spacers in CRISPR-Cas positive *K. pneumoniae* strains (n=18) was found to be 61% (51). In their next paper regarding *K. pneumoniae,* they concluded that out of all the spacers (31 of 550) only ~6% were self-targeting sequences towards the host-chromosome or plasmids (10). Self-target sequences towards the host prophages were also found (51). In conclusion, it could seem like self-targeting spacers in *K. pneumoniae* harbouring CRISPR-Cas systems might not be a rare phenomenon, but these components seems to be tightly regulated/tolerated and utilized in scrapping of the system (10). Despite having self-targeting sequences, there is no doubt that the main function of CRISPR-Cas system is focused on invading genetic elements and host protection (10).

2.3 Restriction- Modification systems in genetic abundance, but less restrictive?

Since the early 1950s Restriction- Modification (R-M) systems have been recognized as a simplistic enzymatic protection mechanism against invading foreign DNA (61)(62). R-M systems encode restriction endonucleases that can recognize invading DNA based on their restriction sites (RS) and cleaves the DNA in a nucleotide specific way (20)(63). Most of the R-M systems recognize dsDNA, while some can recognize ssDNA (63). However, the latter is uncommon (63).

2.3.1 Structure

The prototype R-M system encodes a restriction endonuclease (REase) and a corresponding DNA-methyl-transferase (MTase) (63). The REases recognise un-methylated specific short recognition sites (RS), binds to them and hydrolyse the DNA (61)(63). The RS are short and well defined palindromic sequenced at 4-8bp, often presented in multiple copies on the target

DNA (63)(64). The corresponding MTase is the opponent to the REase, and recognises the same sequences in the target DNA and methylate it to protect it from REase degradation (61)(63). Thus, the presence of a functional R-M system in a bacteria ensure the methylation and protection of the host DNA in contrast to the unprotected, unmethylated invasive DNA (63)(64).

2.3.2 Classification of R-M systems and basic characteristics

RM systems can be classified based on their origin, functions, subunit composition, biochemical properties, cofactor requirements and more (Table 2) (20)(61)(63). Generally the systems are divided into four types; type I, II, III and IV with a large variety of subtypes (61). The most predominantly seen R-M systems are type I, II and III (20). There are documented more than 600 subtypes of the systems, indicating an evolutionary arm race to keep up with MGEs (63).

Table 2: General characteristics of the Type I- IV R-M Systems

	Subunit recognition	**R-M activity**	**MTase + REase dependencies**	**Cleavage and modification sites**	**Co-factor dependencies**	**Possible to escape?**
Type I	*hsdM, hsdS* & *hsdR* (63)	Multifunctional protein (20)	Yes (61)	Translocate along the DNA strand RS and hydrolyse DNA by the meeting point of the analogous complex translocating in the opposite direction (61)	ATP & DNA-binding protein (63)	No (61)
Type II		Independent methylase and restriction endonuclease enzymes RS (20)	Yes (61)		C-protein or V-protein and transcription factors (63)	Yes (61)
Type III	*Mod* & *res* (63)	Multifunctional protein (20)	Yes (61)		ATP (63)	Rare (61)
Type IV	*McrA* & *McrB* or *Mrr* (63)	Modification dependent REase	No (61)	Only cleaves methylated DNA	?	Yes (61)
Type IIG (subtype of type II)	REase	REase uninterrupted protein chain function, could also include a tightly regulated MTase (65)	No (61)	Cleavage downstream at N10 and N14 (65)	Possibly S-adenosylmethionine (SAM) (65)	No (61)

The type I, II and III system shares many common factors and some differences (20)(61)(63). They are dependent on MTase and REase pairs with matching specificity (20)(61)(63). They share the same cleavage and modification sites, but they differ in subunit dependencies and the speed they can change specificity and recognise dsDNA (Figure 7) (20)(61)(63).

The complex formation that occur in example Type I and type III systems forms a multifunctional protein consisting of MTase (M), REase (R) and DNA-binding protein (S)(only type I) and thus can perform its functions (Figure 7) (63). The type II system is dependent on both MTase and REase recognizing the same RS, but they function separately and not in a complex (63).

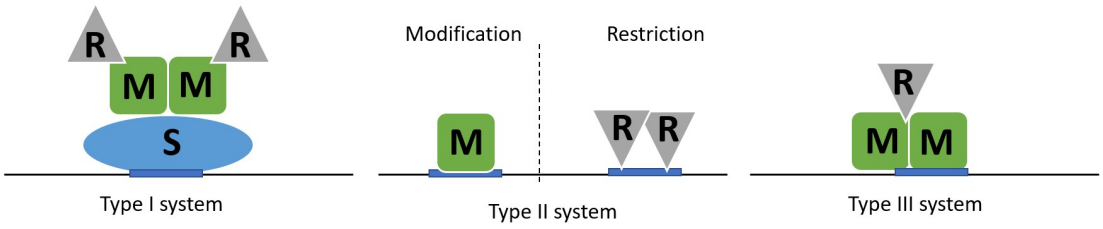


Figure 7: The multifunctional protein complex formation for MTase(M) and REase(R) independent type I, II and III RM systems. The S symbolises another DNA-binding protein. Modified from (63).

The type IV system contains only a REase, and seems to have evolved to enhance protection from phages with modified DNA trying to escape RM systems (63). In this system, the REase hydrolyse modified DNA with a low specificity and with different methylation patterns (63). Thus, the type IV system seems to be able to protect the host from a wide range of foreign DNA (61).

Type IIG is defined as a subtype, but it does also share biochemical properties with type III systems (61)(63). This type of system is more diverse and consists of a single gene encoding a protein with both REase- and MTase activities (63).

2.3.3 R-M systems- abundance and distribution

Fragments or complete R-M systems are found in most sequenced bacterial genomes in various numbers (65). The average number of R-M systems in prokaryotic genomes has been shown to be around 2.6, varying between 0-50 systems per genome in different bacterial phyla, but also within species (63). A study of 8500 prokaryotic genomes demonstrated only 385 genomes without R-M systems (63)(65).

The quantification and distribution of type I-IV R-M systems was also examined in 2261 fully sequenced prokaryotic (n=2261) and phage (n=831) genomes by Oliveira et al. (66). The authors identified a total of 4743 R-M systems in 2261 genomes and the relative abundance was Type II (42%), followed by type I (30%), Type IV (29%) and Type III (8%) (66). Moreover, the frequency of R-M systems was dependent upon genome size and the presence of MGEs, CRISPR-Cas systems, integrons and the ability to engage in natural transformation (66). The number of R-M systems rise with genome size up to 2 Mb, then saturated and declined in density (66). Concerning MGE, relative few R-M systems were observed in plasmids, some in prophages and very few in phages (66). However, all these MGEs contained incomplete R-M systems, dominated by solitary MTases (66). Solitary MTases may function as antidotes against host R-M systems supporting transfer and stabilization of MGEs in new hosts (66). With regard to R-M system types and MGE; the Type III systems were overrepresented in ICEs, Type IV underrepresented in all MGEs (consistent with their protective role against invading DNA methylation systems), and finally Type IIC were over-represented in all MGEs (66).

Oliveira et al. also observed significant co-occurrences of Type I- and IV-, Type I- and Type III-, and Type IIC- and Type IV R-M systems as well as CRISPR-Cas systems (66). The co-occurrence of Type I and IV may support the degradation of un-methylated DNA by Type I and of methylated DNA recognized by the Type IV system (66). Moreover, they reported a significant association between the co-presence of R-M and CRISPR-Cas systems also consistent with the very low frequency of spacers (0,01% out of 80685 CRISPR spacers) with sequence similarity to R-M systems (66).

The content of R-M systems varies between strains of the same species indicating rapid R-M gene loss and acquisition (63). HGT contributes to the spread and acquisition of R-M systems, but the relative contribution of the various transfer mechanisms are not clear (63). The study of Oliceira et al. revealed that R-M systems are over-represented in naturally competent procaryotes, indication that natural transformation is an important mechanism in HGT of intact R-M systems, rather solitary MTases, in plasmids (66). R-M systems can restrain HGT from donors, without the same R-M system, especially since longer unmethylated DNA fragments are more prone to REase cleavage (63).

Phage DNA evolves to avoid R-M systems. A natural response from phage DNA is to reduce the number of recognition sites to possibly escape the R-M systems (61). However, this response is not considered as a universal strategy (61). In addition to RS modification or

removal, there are also more specific, but rare, strategies for R-M system avoidance (61). Phages have displayed mechanisms like Orc protein (phage T7) inhibiting Type I systems by adapting the shape and charge of DNA and hydrolase of the essential ligand S-adenosylmethionine for Type I and III REases (phage T3) (61).

Host DNA regulation and degradation. Methylation of genes is an essential part of gene expression (63). Studies have shown that methylation contributes to the expression of virulence factors and thereby influence bacterial pathogenicity (63). These changes in gene expression could various effects, but has been associated with phenotypic switches, for example transition to a hypermucoid phenotype (63).

Genome content and rearrangements are also affected by the localisation of the R-M system (63). Because of the selfish nature of R-M systems, they often selectively keep MGEs like plasmids containing R-M systems, while cells missing the R-M system is more frequently eliminated by REase activity (63). R-M systems have also been shown to contribute differently to genome stability (63). Homologous recombination less frequently affects chromosomal fragment with R-M systems (63). However, R-M systems might induce recombination when making dsDNA breaks (63).

There are many upsides of harbouring functional R-M systems, but these systems are very primitive and could cause problems for the bacteria (61). The system is also mostly effective in initial environment adaptation and phage infection (63). Over time there are other factors, like surface receptor modification that will have a larger effect on bacterial ecology (63).

2.4 Whole genome sequencing in clinical microbiology

Computational analysis or *in silico* analysis has since 1990s been a useful way of working with DNA sequence data in a larger scale (67). Whole genome sequencing (WGS) is an important tool in uncovering the bacterial genome and allowing bioinformatical analyzation to take place (68). The method is based on assembling overlapping short reads into longer continuous reads (contigs) creating a good coverage of the genome (69). However these methods provides other challenges including competence, memory usage, speed, costs and working with sensitive data (69). In addition, software updates and critical software assessment is highly needed since new information is rapidly uncovered (70).

WGS and bioinformatic analyses can now be easily utilized in clinical microbiology laboratories for rapid detection and characterization (pathogenicity score, AMR and sequence

similarities) of clinical relevant bacteria (67)(69). This could have many purposes like choosing a beneficial treatment, identifying epidemic strains, outbreaks and tracking spread, all fitted into one software (67)(69).

3 The aim of the study

Molecular epidemiological studies of various *K. pneumoniae* strain collections have revealed a large genomic diversity associated with HGT (1)(3)(8). Population structure analyses of clinical strains have displayed that acquired AMR- and virulence genes are mostly separated in distinct subpopulations, MDR- and hypervirulent clones, respectively (1). While MDR clones show a large genetic diversity sharpened by HGT, the hypervirulent clones are dominated by a smaller subset of genetic lineages (1). To understand the dynamics in genome evolution and diversity, one must look at the systems and functions allowing and restricting HGT.

The CRISPR-Cas and the R-M systems are major participants in the evolution of and interactions between MGEs and bacterial hosts. We have a significant knowledge gap in the abundance, distribution and diversity of CRISPR-Cas and R-M systems in *K. pneumoniae* populations. Therefore, I have framed a primary aim and several secondary aims for this thesis.

Primary aim and the research question:

"Is the presence or absence of CRISPR-Cas- and R-M systems in different K. penumoniae populations associated with the presence and absence of MGE represented by acquired AMR- and virulence genes or plasmids?"

The secondary objectives:

1) Determine the prevalence and subtypes of structurally complete CRISPR-Cas- and R-M systems in different *K. pneumoniae* populations
2) Compare the content of MGEs represented by AMR-, virulence genes and plasmids, in *K. pneumoniae* with and without CRISPR-Cas systems
3) Compare the content of MGEs represented by AMR-, virulence genes and plasmids, in *K. pneumoniae* with and without R-M systems.

4 Materials & Methods

This study is based on bioinformatical analysis of 999 KpSC strains previously characterised by WGS and provided through The Norwegian *Klebsiella pneumoniae* network (NOR-KLEB-

NET) (http://www.nor-kleb.net/). Bioinformatic methods and programs were chosen based on literature and experience in collaboration with my supervisors and researchers within the network.

4.1 Bacterial strain collection

The strains were obtained from three different sources (Table 3). This research did not involve any patient sensitive data.

Table 3: Strain collection (n=999) and relevant characteristics

Categories	Origin and time of sampling	Number of strains	Clinical background	Collected material
Carrier strains (n=484)	Tromsø 7 study[1]	484	Carrier isolates	Faecal
NORM strains (n= 414)	NORM non-ESBL[2]	225	Hospital patients	Blood
	NORM -ESBL[3]	189	Hospital and community patients	Blood and urine
ST307 strains (n=101)	National ST307[4] (NORM (n=34) + NORKAB (n=19))	53	High risk clone Mostly HAI	Blood
	International ST307[5] (ENA)	48	High risk clone Mostly HAI	Blood, tissue, catheter, urine, respiratory, rectal ++

[1] Collected from 2015-2016 [2] Collected 2001-2015 [3] Collected between 2001-2015 [4] Collected from January 2017 to august 2019 [5] Collected 2010-2015. ENA-European Nucleotide Archive. NORM: Norwegian surveillance system for antibiotic resistance in microbes. NORKAB: Norwegian Klebsiella pneumoniae bacteraemia study.

4.1.1 Carrier strains from the Tromsø7 study

The faecal carrier strains (n=484) were collected in 2015-2016 from adults ≥40 years old participating in the seventh Tromsø (T7) population study (71).

Briefly, faecal samples were collected using a distributed kit by the participants, transported to and processed within the Norwegian National Advisory Unit on Detection of Antimicrobial

Resistance (K-res) University Hospital of North Norway (Rafaelsberger N. et al. unpublished (72)). The laboratory added 200µl 85% glycerol to the E-swab and stored the samples at -80°C. For identification and antimicrobial susceptibility testing (AST) a total of 100 µl was plated onto Simmons citrate with inositol agar plates and incubated for 48 hours at 37°C (18)(72). Colonies suspected to be *Klebsiella* were identified using MALDI-TOF MS (72). Colonies identified as *Klebsiella pneumoniae, Klebsiella quasipneumoniae* or *Klebsiella variicola* was selected for further characterization (72). The strains was tested according to EUCAST disc diffusion method and breakpoint table, for a total of 12 antimicrobial agents (amoxicillin-clavulanic acid, piperacillin-tazobactam, cefuroxime, cefoxitin, cefotaxime, ceftazidime, aztreonam, mecillinam, gentamicin, meropenem, ciprofloxacin, and trimethoprim-sufamethoxazole) (72)(73)(74). Whole genome sequencing (WGS) had been performed using the MagNA Pure 96 system for DNA extraction, Nextera Flex sample protocol and sequence library and sequencing was performed using Illumina MiSeq for 250 or 300bp paired end reads at Stavanger University Hospital (Hetland M. et al. unpublished (75)).

4.1.2 Clinical strains from the NORM study

The strains from the Norwegian surveillance system for antibiotic resistance in microbes (NORM) accounts for a total of 414 strains (76). NORM performs yearly surveillance studies by a common study protocol involving all Norwegian microbiology laboratories (76). These strains include both non-ESBLs and ESBLs collected mainly from infections in hospital patients. However, the collection of strains originating from urine samples (n=67) could be from non-hospitalised patients. The strains were selected from NORM studies performed during 2001-2015, creating a good diversity (Forstervold A. et al. unpublished (77)). The laboratories processed and determined the phenotype of the samples as a part of the normal routine by plating out the samples on non-selective agar media, identifying the strains using MALDI-TOF and then storing them at -80°C for further analyses. Susceptibility testing was performed at the designated laboratories following the EUCAST guidelines (73)(74). Antibiotic susceptibility testing was performed using gradient strip test up until 2007 when disk diffusion was introduced (77). Isolates with reduced susceptibility towards at least one broad-spectrum cephalosporin was systematically tested from 2002 (77). The strains from 2001 with reduced susceptibility to either cefpirome or ceftazidime were classified as ESBL producing (77). WGS had been performed by Stavanger University Hospital by the same procedure as the carrier strains (Hetland M. et al. unpublished (75)).

4.1.3 National and international high-risk clone ST307 strains

The ST307 (n=101) strains were retained from different studies. The Norwegian strains (n=53) originated from NORM (n=34) and the *Norwegian Klebsiella pneumoniae bacteraemia study* (NORKAB) (n=19) (76)(78)(79). The NORKAB study includes blood culture strains from nosocomial blood infections in patients over 18 years estimated to cover 90% of the Norwegian population (25)(78)(79). Strains were collected during 2017-2019 (25)(79).

The international ST307 (n=48) was collected from the public European Nucleotide Archive (ENA)(https://www.ebi.ac.uk/ena) (80). American strains without all datafiles available where excluded. This was because the absence of all file formats enables the original sequence data to be available without adjustments/improvements of the actual quality. The other international strains had all files, both assembled and the raw data, available providing a higher transparency. For these strains, the publicly available information is available in the supplementary table 1. The strain material is highly diverse for this collection.

4.2 Bioinformatic methods and tools

Multiple tools and programs were utilised in the analysis. Therefore, this section is divided into the different categories of interest. The commands used for running each tool, the tool database versions and additional scripts used can be found in the Appendix 1.

4.2.1 Sequencing, assembly and quality control

Prior to receiving the strains (all strains excluding the ones collected from ENA pre-assembled), they were all sequenced, assembled and controlled by the Department of clinical microbiology, Stavanger University Hospital.

Illumina MiSeq sequencing with Nextera Flex kit was performed on all Norwegian strains from the Tromsø 7 study, NORM and the NORKAB. The international ST307 was collected using enaget, a script for downloading ENA FASTQ files by their accession numbers https://github.com/stevenjdunn/enaget. All the strains, including the international ST307 strains where quality checked and assembled using the same pipeline; Asmbl available at https://github.com/marithetland/Asmbl. Asmbl is a pipeline using multiple programs for fast and easy quality control, trimming, assembly and another quality control for the final output. Quality parameters set additional to default parameters for Asmbl included GC% match for *K. pneumoniae*, total length match for *K. pneumoniae*, low number of contigs (ideally <700), coverage of 30X and a preference for fewer amounts of long contigs versus a high number of short contigs (Hetland M. et al. unpublished). In addition to Asmbl, Prokka version 1.12

(available at https://github.com/tseemann/prokka) was utilised for functional annotation of contigs >200nt.

4.2.2 Strain profile: Virulence score, AMR-profile, MLST and plasmid content

Virulence score, AMR-profile, MLST and plasmid content were detected using Kleborate version 0.4.0-beta (available at: https://github.com/katholt/Kleborate/blob/master/README.md). Kleborate is a complex pipeline utilising several databases and programs for the overall profile. Kleborate is specially developed for the *K. pneumoniae* species complex and was run using default parameter settings (81).

Virulence classification roughly categorised the siderophore profile for the strains. The virulence score ranged from 0-5 with the following profiles (81):

Virulence score 0 = none of the acquired virulence loci

Virulence score 1 = yersiniabactin only

Virulence score 2 = yersiniabactin and colibactin, or colibactin only

Virulence score 3 = aerobactin and/or salmochelin only (without yersiniabactin or colibactin)

Virulence score 4 = aerobactin and/or salmochelin with yersiniabactin (without colibactin)

Virulence score 5 = yersiniabactin, colibactin and aerobactin and/or salmochelin

AMR-gene prediction in Kleborate was performed through ARG-Annot database of acquired resistance genes with the software SRST2 (81)(69). Short Read Sequence Typing for Bacterial Pathogens (SRST2) is especially made for WGS Illumia sequencing data and performs searches for resistance genes with a match of 90% as default (69)(82).

In addition to annotating the resistance genes, according to the affected antibiotic drug class and the aggregating the total number of acquired AMR-genes, Kleborate will classify β -lactamase-encoding genes into six different categories; Bla (β -lactamases), Bla_broad (broad spectrum β -lactamases), Bla_broad_inhR (broad spectrum β -lactamases with resistance to β -lactamase inhibitors), Bla_Carb (carbapenemase), Bla_ESBL (extended spectrum β -lactamases) and Bla_ESBL_inhR (extended spectrum β -lactamases with resistance to β -lactamase inhibitors) (81). In order to be able to examine the association between the presence /absence of CRISPR-Cas systems and R-M systems and the number of AMR-genes we chose

a revised classification system. We used the Kleborate system giving us the total number of acquired AMR-genes, excluding mutational resistance mechanisms and intrinsic resistance genes such *bla*$_{SHV}$-, *bla*$_{OKP}$-, *bla*$_{LEN}$-, *fosA*- and *oqxAB*- alleles (1)(3)(28)(29)(48). Other gene variations were confirmed by a search at National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/). Moreover, we classified the strains AMR-gene profile in four categories, as shown in Table 4. The revised classification allowed a more straightforward analysis of associations between AMR-gene load and the presence or absence of CRISPR-Cas- /R-M- systems.

Table 4: Classification of AMR profiles based on the Giske et al. definition

Classification	Definition	Antimicrobial agent classes or genes included in the classification
AMR negative	No acquired AMR-genes	-
Non- MDR	Acquired AMR-genes to one or two classes of antibiotics	Aminoglycosides (AGly), β -lactams (third generation cephalosporins and/or carbapenems), Fluoroquinolones (Flq), Phenicols (Phe), Rifampin (Rif), Sulfonamides (Sul), Tetracyclines (Tet) and Trimethoprim (Tmt)
MDR	Acquired AMR-genes to three or more classes of antibiotics	
ESBLA, ESBLM and/or ESBLCARBA	Harbours one or more of the AMR-genes encoding ESBL as defined by Giske et al. (24)	*bla*$_{CTX-M}$, bla$_{OXA}$-, and *bla*$_{KPC}$- alleles

MLST examines seven different house-keeping genes and distinguish between different genetic strains within a species (83). By using MLST it is possible to identify seven known "MLST loci" (*rpoB, gapA, mdh, pgi, phoE, infB* and *tonB*) where each unique sequence is numbered to distinguish the unique allele numbers in the genome showing a sequence type (ST) for the *K. pneumoniae* species complex (81)(82)(83). The MLST typing is based on the SRST2 database (69)(82).

Plasmid detection was performed by Abricate version 0.8 with the PlasmidFinder function for plasmid detection using the large PlasmidFinder database to annotate the matching probes (https://bitbucket.org/genomicepidemiology/plasmidfinder_db/src/master/). PlasmidFinder

database is constructed from WGS collected from NCBI and whole-plasmid sequence data from *Enterobacteriaceae* and therefore well suited (84). The database identifies plasmids based on the probe similarity of the replicon sequence (49). In addition, the plasmids are annotated using the incompatibility (Inc) groups for *Enterobacteriaceae* with standardised nomenclature (8)(49)(84). Abricate will list the replicon types and summarize the number of sequence similarities. The program does not detect if the probes are located on different plasmids or the same, and thus the actual number of plasmids could be different than reported by other (49). All hits for plasmid replicons were included in the analysis.

In addition, the plasmid detection was performed without the minimum coverage and minimum id by mistake. This resulted in a total of 17% potential false positive nucleotide identity matches being accounted as present plasmids. The distribution of the potential false positive matches was 9.2% for the carrier-, 7.2% for the NORM- and 0.4% for the ST307 strain collection.

4.2.3 Phylogeny for ST307

Phylogenetic analysis was performed using Roary version 3.12.0 available at (https://github.com/sanger-pathogens/Roary)(85). Using the annotated GFF3-files from Prokka, Roary analysed the pangenome of the low diversity strains producing tree files. Unfortunately, five strains were not included in the phylogeny analysis. This was because of Prokka skipping annotation of these strains without notifying. This accounts for the strains named NORM_BLD_2014_101655, NORM_URN_2015_109059, NORM_BLD_2015_115990, NORM_BLD_2015_115991 and NORM_BLD_2015_116357, respectively. The accessory binary tree file was visualised using iTol available online at https://itol.embl.de/upload.cgi.

4.2.4 Detection of structural complete CRISPR-Cas system and control

CRISPR-Cas systems whereas primarily detected using the standalone version of CRISPRCasFinder software (available at: https://crisprcas.i2bc.paris-saclay.fr/). In addition, some strains were controlled using the online version because of unexpected low evidence level ranking from level 1-4, but mostly at level 1 (available at: -https://crisprcas.i2bc.paris-saclay.fr/CrisprCasFinder/Index). Level 1 being unlikely systems and level 4 highly promising systems based on the mathematical functions (86)(87). The CRISPR-Cas software was run on default with the following modifications; the minimum size of direct repeat was adjusted from 23 to 19 and the minimum spacer size was adjusted from 25 to 20 as advised by Roni Froumine and Kelly Wyres working in the KLEB-GAP project. Using the CRISPRCasFinder_Scripts

(https://github.com/rfroumine/CRISPRCasFinder_Scripts) all positive strains (evidence level 1-4) were summarised in one output file for further assessment. However, the script was adjusted to include all evidence levels, not only 3-4 (default) since *K. pneumoniae* could display some difficulties in detection of the systems if the software was to strict.

Because the software ranked all the strains at evidence level at 1 and the online version at level 4 for the controlled strains, all strains were assessed using additional software and manual assessment using following criteria (86):

- The *Cas*-gene array must be complete for the designated system
- There was one or more CRISPR array with over two spacers
- The system was detected on the same contig

The arguments for the given criteria were to assure that the CRISPR-Cas system was a structural complete system and thus potentially functional. Since the CRISPRCasFinder software does not separate the I-E and I-E* subtypes, manual assessment and additional programs were in order (88).

Since *Cas 1* has proven to distinguish between I-E and I-E* systems with conserved homology within the subtype, a quick BLAST nucleotide search (version 2.9.0+ available at https://anaconda.org/bioconda/blast) against several *Cas 1* genes from available *K. pneumoniae* strains from the NCBI database (https://www.ncbi.nlm.nih.gov/genome/) was performed on the positive strains from CRISPRCasFinder (Table 5). It could be debated if the *Cas 1* is the best signature-gene for the typing of subsystem, because *Cas 3* is often used as a signature (58)(57). However, displays homology to other helicase genes, and can often be wrongly annotated (seen in Artemis), *Cas 1* was selected for control selection (58)(57). The *Cas1* gene amino acid sequence from the control strains were manually extracted using Artemis version 18.1.0 available at https://www.sanger.ac.uk/science/tools/artemis. For *Cas 1* detection in Artemis, published primers were utilized (10). The extracted sequences were controlled using a nucleotide BLAST search at the NCBI database online version with a threshold of 80%, since the genes where poorly annotated in Artemis.

Quick subtyping of the extracted sequences using *Cas1*, revealed potential controls for further synteny examination and a more certain subtyping. The KPNIH27 and NTUH-K2044 (marked in bold in figure 5), displayed a good match also for the other primers of the CRISPR-Cas I-E

and I-E* system. Therefore, they were also used as a control in synteny analysis preformed in SimpleSynteny (10)(51).

Table 5: K. pneumoniae strains from NCBI database were subtyped using Cas1 as a test of PCR primers					
ID	Accession number	Species	System type	Cas1_Start	Cas1_End
KPNIH27	CP007731.1	*K. pneumoniae*	I-E	4180320	4181240
KPNIH31	CP009876.1	*K. pneumoniae*	I-E	4115411	4116331
CAV1344	CP011624.1	*K. pneumoniae*	I-E	3290664	3291584
CAV1193	CP013322.1	*K. pneumoniae*	I-E	3310810	3311730
Kp52.145	FO834906.1	*K. pneumoniae*	I-E	1042853	1043773
SB3432	FO203501.1	*K. pneumoniae (rhinoscleromatis)*	I-E	4237324	4238244
E718	CP003683.1	*K. michinganensis*	I-E	4905442	4906314
KCTC1686	CP003218.1	*K. michinganensis*	I-E	3414984	3415967
NTUH-K2044	AP006725.1	*K. pneumoniae*	I-E*	3002916	3003800
ATCC43816	CP009208.1	*K. pneumoniae*	I-E*	391954	392829
1084	CP003785.1	*K. pneumoniae*	I-E*	2327874	2328749
PMK1	CP008929.1	*K. pneumoniae*	I-E*	4767170	4768045
RYC492	APGM01000001.1	*K. pneumoniae*	I-E*	2106880	2107755
J1	CP013711.1	*K. pneumoniae*	I-E*	5073325	5074200
PittNDM01	CP006798.1	*K. pneumoniae*	I-E*	1171218	1172093
U25	CP012043.1	*K. pneumoniae*	I-E*	3266932	3267807
KP617	CP012753.1	*K. pneumoniae*	I-E*	564057	564932

Bold writing mark the selected references for SimpleSynteny

For the I-E system the NCBI strain KPNIH27 (accession number CP007731.1) full Gene Bank file was utilized in Artemis for collecting the nucleotide sequence as a fasta-file for all marker genes: *Cas1, Cas2, Cas3, Cas5, Cas6, Cas7, CasA, CasB, CysH, iap* and *Trp* (10)(51). The I-E* reference file was made in the same way using the NTUH_K2044 (accession number AP006725.1) isolate with the following reference genes: *ABC-Transporter 1, ABC-Transporter 2, ABC-Transporter 3, Cas1, Cas2, Cas3, Cas5, Cas6, Cas7, CasA, CasB* and *Glyoxalase* (10)(51).

The next criteria were that the systems must have one or more CRISPR array with multiple spacers. Because a functioning I-E and I-E* system needs a large array to insert spacers, and if its active it should have multiple spacers (9). CRISPRCasFinder did not include a summary the number of spacers or CRISPR arrays found and the script did not include it either. The spacers could only be found separate in one of the result files (json-file) for each individual strain. All the strains were manually controlled using SimpleSynteny, that produced an output including a physical space indicating the arrays positions and size. Additional testing and assessment

were done, controlling the strains to see how large the CRISPR arrays also had to be to display a distance in the SimpleSynteny analysis. On average, the CRSPR arrays was approximately the size of the *Cas* -array to be visual in SimpleSynteny. The spacers were also controlled for a random selection in the Jason-files for each strain, where they were listed.

SimpleSynteny was utilised for investigating the synteny, presence and placements of the system components. Also, checking the contigs to see if the systems had been distributed over one or several contigs. If they were fragmented, they could potentially be inactive. The software is available at: https://www.dveltri.com/simplesynteny/about.html. If the system was placed over different contigs, the png-file would display a clear breach and the system could possibly be dysfunctional because the location of the components could be anywhere. A functioning system relies on close contact of all the components (9). Also, to be able to predict the actual synteny of the genes and classify the systems, it is crucial for the systems are located on the same contig for higher bioinformatical accuracy.

Because of software limitations in SimpleSynteny and for the purpose of the png-file with the results, the samples were analysed in groups of 20 for the CRISPR-Cas positive strains and manually assessed (Appendix 2). Manual assessment ensured that all Cas-genes were present, the CRISPR-array was marked by a space, the signature genes upstream and downstream was located and that there were no breaches in the contigs. Cut-offs between the adjacent signature genes and an otherwise complete CRISPR-Cas system was allowed and counted as positive.

The program was performed on default settings with a chosen minimum coverage of 80% allowing small mutations, as would be expected for the individual systems (58). The minimum coverage was chosen based on experimenting with the parameters and seeing when positive controls of *Cas* genes no longer were present. Primarily the *Cas 3* revealed a tendency to fall out with a minimum coverage >80%. The other *Cas* genes displayed a much higher similarity and only fell out with a much higher minimum coverage.

4.2.5 Restriction-Modification system detection

Detection of R-M systems was performed using HMMer profiles available at https://github.com/EddyRivasLab/hmmer. The HMMer software mainly is made for mapping protein sequences for more advanced detection and production of databases. And therefore, the additional scripts provided for easy HMMer profiling were necessary for system detection, available at: https://github.com/pedrocas81/RMS?fbclid=IwAR2XonJfGM4UhBYZTQs-cTXTsA9UUVEVLYaRnJI-bb7aG6Sd8FAO2D9OMp4. The program was performed on

default settings. The script running the software did not include any form of instructions or examples, the hits-files where concatenated for each dedicated MTase or REase for the different types of systems and manually evaluated in Excel and counted. Criterions for systems was that they would need to have a matching subtype of MTase and REase for the co-dependent types (type I, II and III) and over one REase for the independent Type IV and IIG systems. In addition, only one pair of MTase and REase was set as one functional system, even though the system could have two MTases of the same subtype as the one REase. This manual assessment results in only possible functional systems being accounted as positive. The result files with the manual assessments can be found in Supplementary table 4,5 and 6.

4.3 PCA analysis

Visualisation of results, making heat plots and doing principal component analysis (PCA) analysis was done using R package FactorMineR. PCA plots are often used to explore relationships between datasets with multiple factors (89). This applies for both standardised data and the PCA plot allows them to be further standardised without losing information (89). The data analysed in this study were all categorised into smaller groups summarising the information and these data were the base of the PCA analysis (Supplementary table 2) (89).

The PCA plots are based on the largest difference observed in the Eigenvalues (linear representation of the largest to the smallest difference in the dataset) (89). The largest eigenvalues are represented on the principal component (PC) axes, making a 2D mapping of the results. By adding on concentration ellipses and the average point for the group, the definition of the selected groups becomes clearer (89).

4.4 Statistics

The Fishers Exact test online calculator was used to investigate if there was any statistical significance between the groups (https://www.socscistatistics.com/tests/fisher/default2.aspx). A p-value <0,05 was considered statistically significant.

5 Results

The result part is organized in six sections. Strain population structure (5.1), Main characteristics in the strain collections (5.2), Bioinformatic selection of structural complete CRISPR-Cas and R-M systems (5.3), PCA plot comparison of the strains collections (5.4), CRISPR-Cas and R-M systems: correlations with virulence profile, AMR classification and plasmid content (5.5), and Sub analysis of dominant STs and high-risk STs (5.6). The raw data

are available in Supplementary table 3. The overall results do not include fragmented and incomplete CRISPR-Cas- R-M systems, only structurally complete systems according to the criteria given in the Material and Methods section.

5.1 Strain population structure

The strain collections were grouped in three; carrier- (n=484), NORM- (n=414) and ST307 (n=101) strains. The NORM strain collection consisted of both ESBL (n=189) and non-ESBL (n=225) strains. However, the number of ESBL -producing strains in the NORM -collection were different than first assumed. Some strains, harbouring only variants of the intrinsic narrow spectrum SHV β -lactamase had simply been misidentified as ESBL-producing (n=45). Consequently, several strains were re-classified resulting in only 144 strains classified as ESBL- NORM strains. The rest of the strains were classified as the non-ESBL NORM (n=270).

The carrier strains included a total of 300 different ST types. Only 38 ST types had ≥2 strains (Figure 8). The most prevalent STs (n≥10) were ST20 (n=15), ST26 (n=13) and ST35 (n=10) (Figure 8).

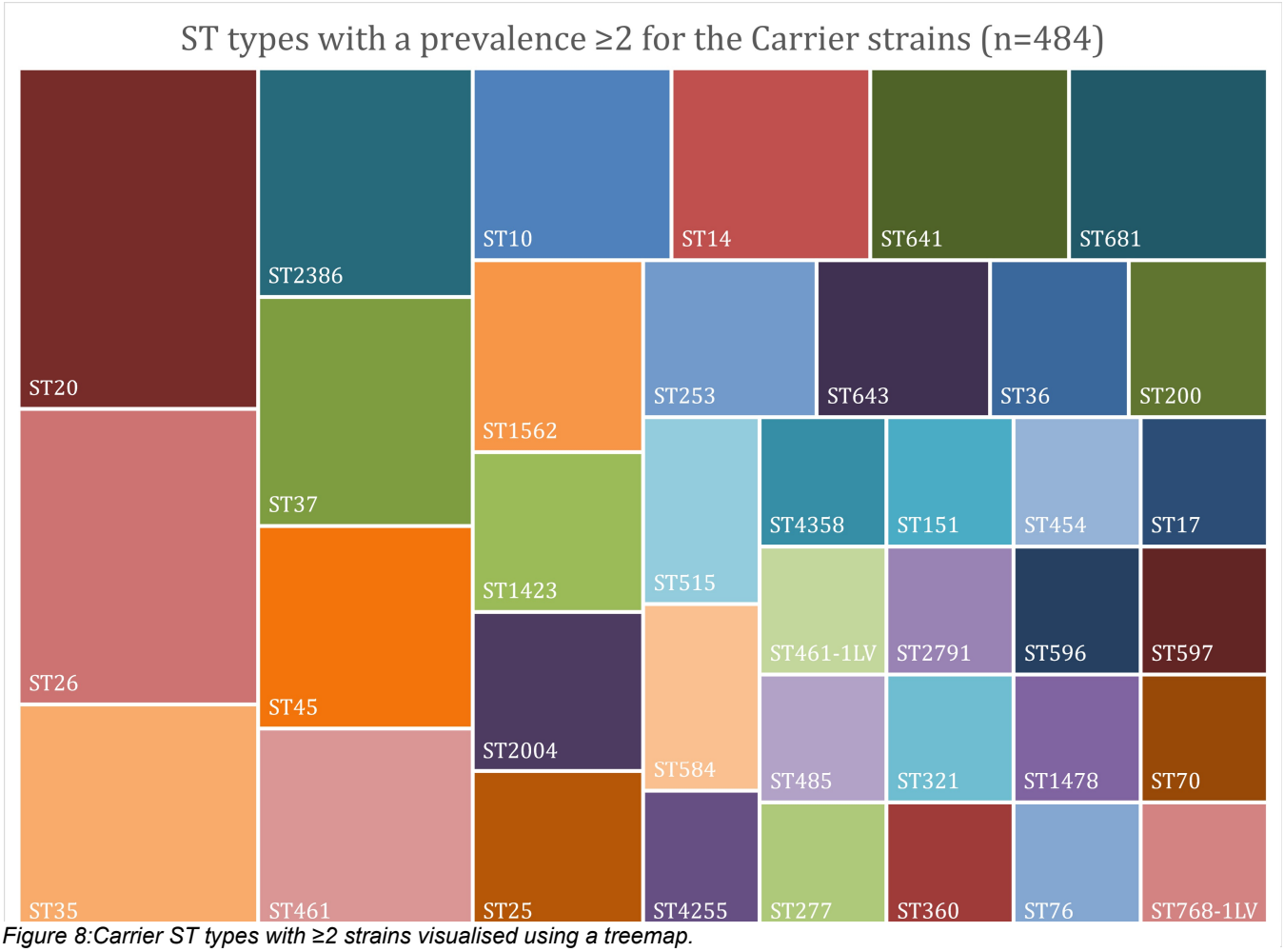


Figure 8:Carrier ST types with ≥2 strains visualised using a treemap.

In the NORM collection there was a total of 203 different ST types, where a total of 55 ST types had ≥2 strains including ST14 (n=18), ST20 (n=17), ST37 (n=15), ST70 (n=14), ST15 (n=13) and ST45 (n=13) (Figure 9).

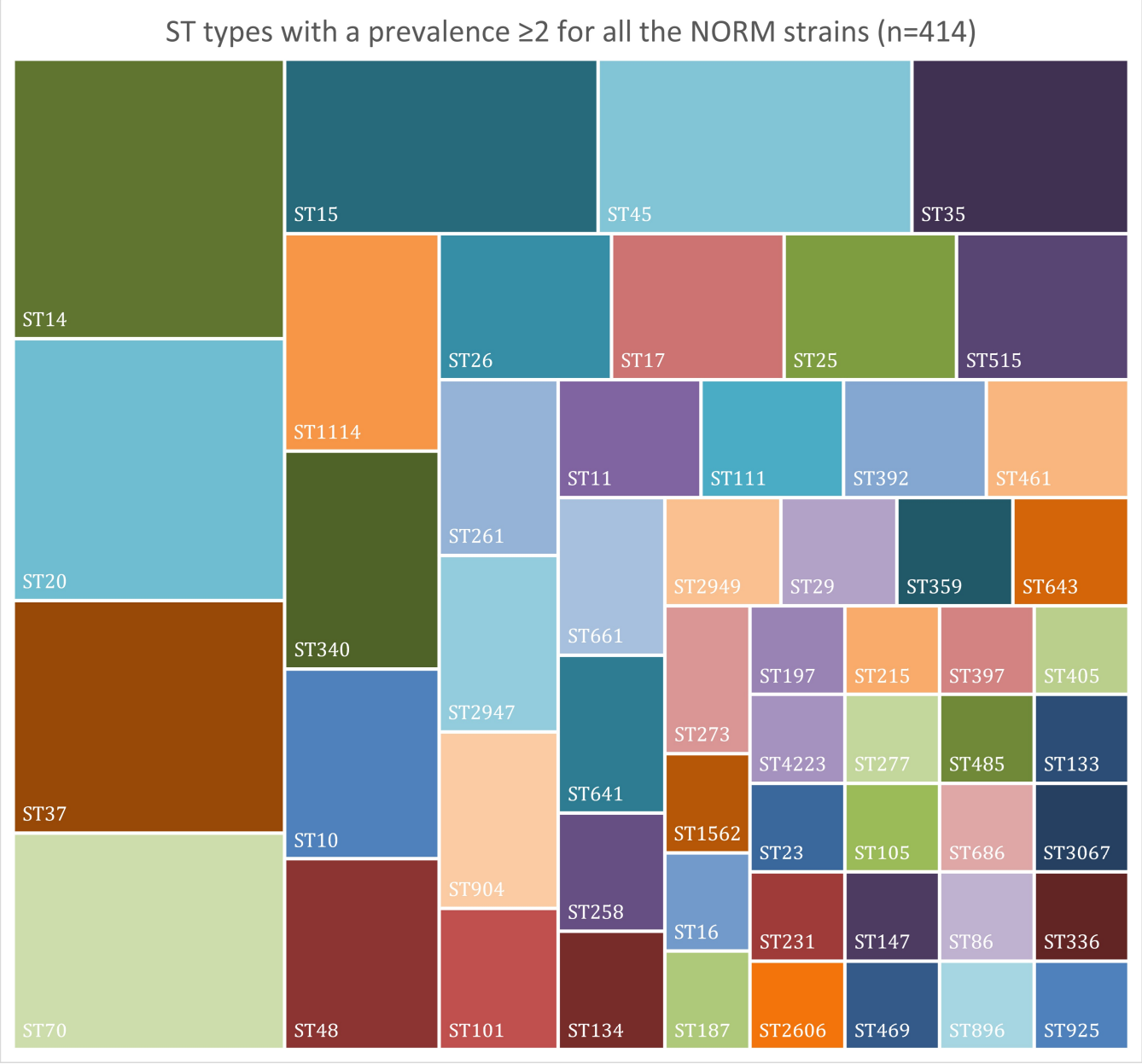


Figure 9: NORM STs with ≥2 strains visualised using a treemap

The ST307 strains are known for being a globally emerging MDR clone. Phylogeny of this strain population displayed a relatively similar core genome, but variances in terms of the accessory genome (Figure 10). Almost all strains were MDR, ESBL -producing, R-M Type I

positive, virulence score 0 and carrying 2-3 plasmids.

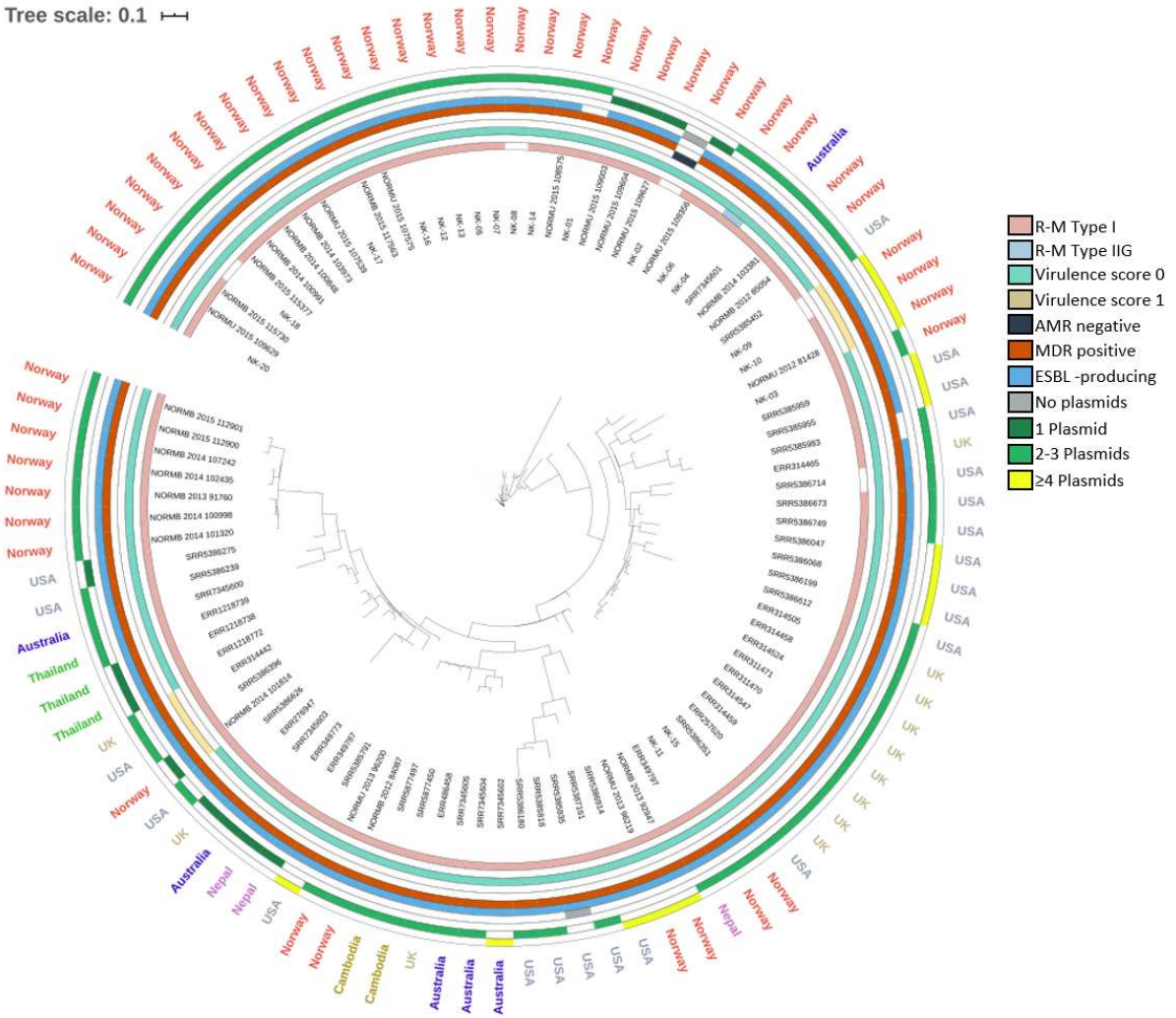


Figure 10: Phylogeny of ST307 based on the accessory genome. Made using Roary and iTol

The distances in the tree display diversity within the strain population. Moreover, some of the strains displayed similarities in the selected features marked in the rings which will be commented in result section 5.2. Most Norwegian clustered displaying similar features (figure 10).

5.2 Main characteristics in the strain populations

Table 6 summarises the main characteristics in carrier-, NORM (non-ESBL and ESBL) and ST307 strains including CRISPR-Cas- and/or R-M systems, virulence score, AMR classification and plasmid content. Statistical comparisons between groups, carrier- and NORM non-ESBL strains as well as NORM-ESBL- and ST307 strains, are also presented

Table 6: Result summary displaying the total number of positive strains and the percentage of the total							
Category	Element	Carrier strains n(%)	Statistics[1]	NORM strain collection		Statistics[2]	ST307 strains n(%)
				NORM non- ESBL n(%) 270	NORM ESBL n(%) 144		
CRISPR- Cas	CRISPR-Cas positive	144 (30)	NS	71 (26)	41(29)	0.00001	0 (0)
	CRISPR-Cas negative	**340 (70)**		**199 (74)**	**103 (72)**		**101 (100)**
	CRISPR-Cas Class 1 type I-E system*	**92 (19)**	NS	**39 (14)**	16 (11)	NS	0 (0)
	CRISPR-Cas Class 1 type I-E* system*	58 (12)	NS	33 (12)	**25 (17)**	NS	0 (0)
R-M systems	R-M system positive	213 (44)	NS	115 (43)	38 (26)	0.00001	**91 (90)**
	R-M system negative	**271 (56)**		**155 (57)**	**106 (74)**		10 (10)
	R-M type I system*	**116 (24)**	NS	**51 (19)**	**28 (20)**	NS	**91 (90)**
	R-M type II system*	0 (0)	NS	0 (0)	2 (1)	NS	0 (0)
	R-M type III system*	19 (4)	NS	14 (5)	1 (1)	NS	0 (0)
	R-M type IV system*	65 (13)	NS	22 (8)	8 (6)	0.00001	0 (0)
	R-M type IIG system*	51 (11)	NS	44 (16)	5 (4)	0.0029	1(0)
Virulence score	Virulence positive	55 (11)	0.0339	46 (17)	67 (47)	0.00001	6 (6)
	Virulence score 0	**429 (89)**		**224 (83)**	**77 (54)**		**95 (94)**
	Virulence score 1	48 (10)	NS	37 (14)	62 (43)	NS	6 (6)
	Virulence score 2	0 (0)	NS	2 (1)	0 (0)	NS	0 (0)
	Virulence score 3	3 (1)	NS	4 (2)	1 (1)	NS	0 (0)
	Virulence score 4	0 (0)	NS	1 (0)	4 (3)	NS	0 (0)
	Virulence score 5	4 (1)	NS	2 (1)	0 (0)	NS	0 (0)
AMR classification	AMR negative	**459 (95)**	NS	**205 (76)**	0 (0)	NS	1 (1)
	non-MDR	13 (3)	0.0032	21 (8)	11 (8)	0.031	1 (1)
	MDR	12 (2)	0.00001	44 (16)	**133 (92)**	NS	**99 (98)**
	ESBL-producing	0 (0)	NS	0 (0)	**144 (100)**	NS	**98 (97)**
Plasmid	none plasmids	75 (15)	NS	41 (15)	1 (1)	NS	2 (2)
	1 plasmid[3]	71 (15)	NS	39 (14)	11 (8)	NS	12 (12)
	2-3 plasmids[3]	**212 (44)**	NS	**121 45)**	**81 (56)**	0.0072	**74 (73)**
	over 4 plasmids[3]	127 (26)	NS	69 (26)	50 (35)	0.0001	13 (13)

[1]Fisher exact test displaying the difference between the carrier – and NORM non-ESBL collection. P<0,05 is considered statistically significant. [2]Statistically difference using Fisher exact test between the NORM ESBL- and ST307 strain collections. P<0,05 is considered statistical significance. NS= no statistical difference is (p >

0.05). Numbers marked in **bold** represents the highest percentage distribution(s) within the category. [3]Display ~16% false positives. * Co-occurrence of subsystems are not been taken in consideration.

In terms of **CRISPR-Cas systems**, absence of structural complete systems was observed in the majority of carrier- (70%), non-ESBL NORM (74%), and ESBL NORM (72%) strains whereas a complete absence was observed for the ST307 population. The difference in the prevalence of CRISPR-Cas systems between the NORM ESBLs and the ST307 was statistically significant. Both Class I-E and I-E* were detected and co-occurred in 7 strains. The occurrence of structural complete **R-M systems** varied between the groups; the carrier (44%), non-ESBL NORM (43%), ESBL-NORM (26%) and ST307 (90%). R-M Type I was the dominant system. Co-occurrent R-M systems were observed and will be commented below. Between the ST307 and the NORM ESBLs there was a statistically significant difference in the prevalence of R-M negative and positive strains. For all the strain populations, **virulence score 0** (no acquired virulence loci) were the most prevalent. A significant higher positive virulence score was observed in non-ESBL NORM strains compared to carrier strains. Likewise, a significant higher virulence score was observed in ESBL-NORM strains compared to ST307 strains. As expected, the carrier and NORM non- ESBL groups had the highest prevalence of **AMR** negative strains with no statistical significant difference between them ($p = 0.06$). NORM ESBL and ST307 had the highest prevalence of MDR and ESBL -producing strains, with no statistical significant difference between the groups. The **plasmid content** was similar between groups between carrier and non-ESBL NORM strains with a relative high proportion of strains (15%) with no plasmids compared to the ESBL-NORM and ST307 strains. There was a statistical significant difference between the ESBL-NORM and ST307 strains concerning the proportion of strains within the 2-3 plasmid and ≥ 4 plasmids categories.

Figure 11 present the main findings in a heatmap format. Overall, the strains displayed many features in common. However, the data also displayed small differences within distribution of e.g. the different R-M Types. ST307 strains displayed a more uniform overall profile with smaller differences within each category. The NORM ESBL strains displayed a different tendency by having the largest percentage of R-M system negative strains and also carrying a heavier plasmid load.

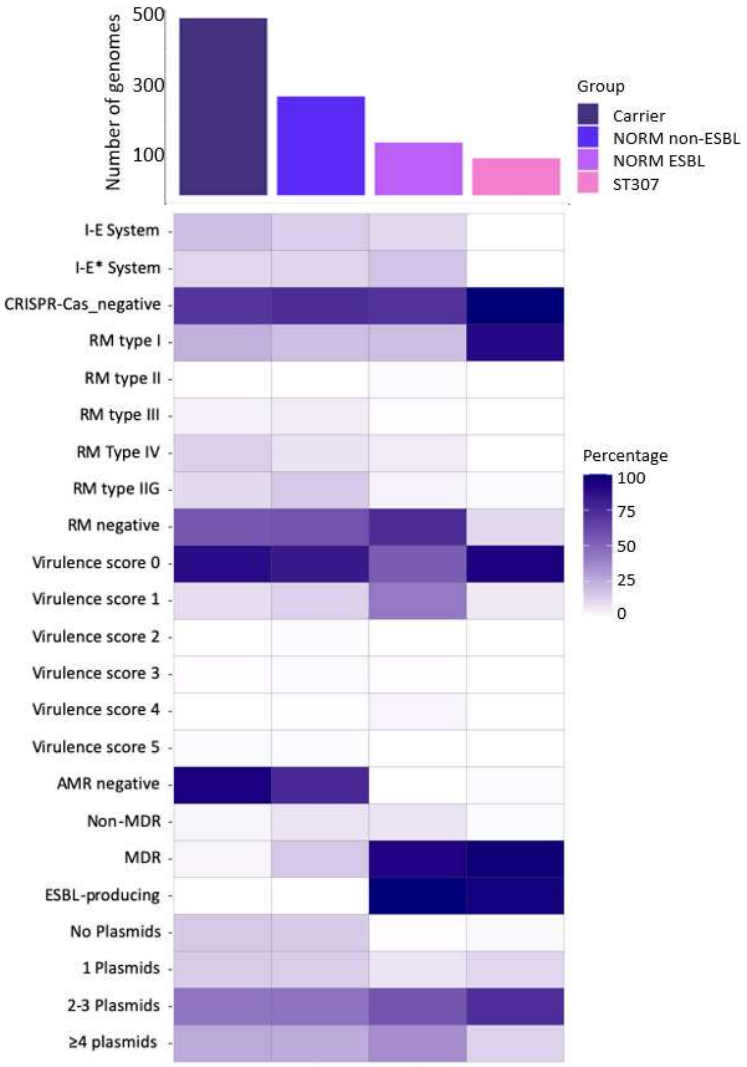


Figure 11: Percentage distribution of the selected features (CRISPR-Cas systems, R-M systems, virulence-, AMR profile and plasmid classification). This heatmap was made in R-studio using FactorMineR.

5.3 Bioinformatic selection of structural complete systems

Only structural complete and therefore potentially functional CRISPR-Cas- and R-M systems were included in the results. Methodical assessment of the systems using multiple programs were necessary to not only subtype the systems, but also to detect fragmented- and incomplete systems. Multiple deleted or lack of Cas-genes or non-matching MTase/REase were observed.

Using the strict criteria for CRISPR-Cas detection, a total of 24 CRISPR-Cas Class 1 Type I-E systems detected in CRIPSRCasFinder for the carrier strains was assessed as negative using BLASTn and SimpleSynteny (Figure 12). This was due to the lack of several Cas-genes, breaches in contigs or lack of Cas1. CRISPR-Cas Class 1 Type I-E* system in the carrier strains displayed either a complete presence or severe fragmentation. For the NORM strains, a total of

22 CRISPR-Cas Class 1 Type I-E and 2 I-E* systems were, for the same reasons, assessed negative. The ST307 did not harbour any complete or fragments of CRISPR-Cas systems.

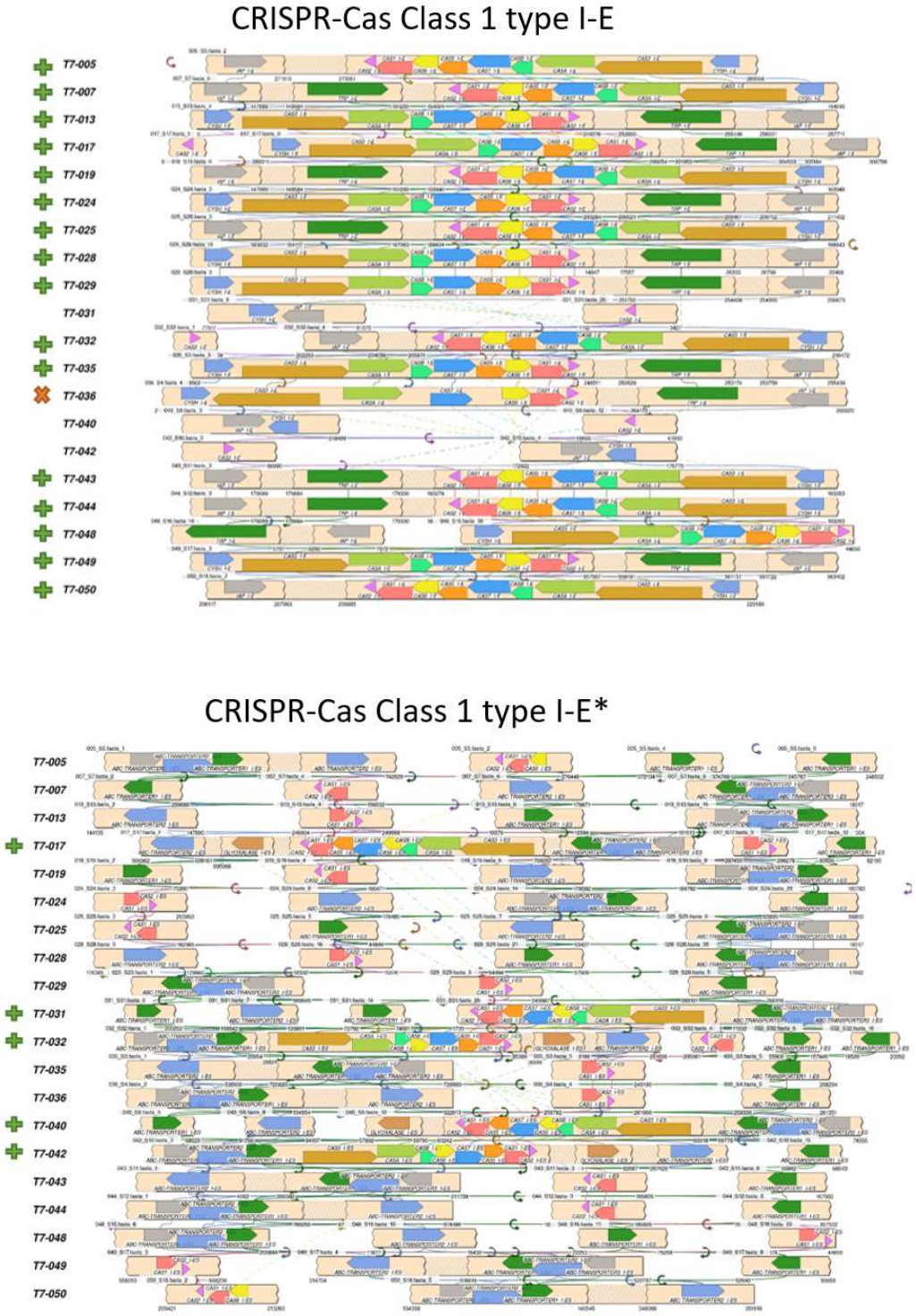


Figure 12: Results from SimpleSynteny analysis for both CRISPR-Cas Class 1 I-E and I-E systems for a selection of 20 strains. The green plus signs display structurally complete systems, the red x indicate one incomplete I-E system and the ones without any symbols are clearly incomplete I-E and I-E*I systems.*

The png-file from SimpleSynteny comparing the same strains for both CRISPR-Cas Class 1 Type I-E and I-E* displayed mostly clear boundaries between the presence or absence of

complete systems (Figure 12). The systems were classified and manual assessment ensured that all Cas-genes were present (labelled arrows), the CRISPR-array was marked by a significant space (beige with stripes), the signature genes upstream and downstream was located (labelled arrows) and that there were no breaches in the contigs (clear cut-offs in white). Cut-offs between the adjacent signature genes and otherwise complete CRISPR-Cas system was allowed (see T7-048 reference I-E in Figure 12).

To ensure that the R-M systems were structural complete according to the criteria given in the methods section, all systems were assessed to ensure that the strain had MTases and REases within the same type of R-M system and the same subfamily (valid for type I, II and III systems)(Figure 13). The other systems (IV and IIG) only needed a match for REase. The total prevalence of R-M systems was much lower than expected and potentially because of manual selection of only complete systems.

MT ases	i-Evalue	Best E-value	Subclass	MTase_count	RE ases	i-Evalue	Best E-va	Subclass	REase_count	Complete systems
T7-006_S6.fa Type_II_MTases.hmm	0.812968	3.8e-20&0.0(	FAM_22	5	T7-006_S6.fa Type_II_REases.hmm	0.863309	1.2e-86	FAM_6.einsi_trimmed	1	0
T7-006_S6.fa Type_II_MTases.hmm	0.85348	3.5e-08	FAM_24							
T7-006_S6.fa Type_II_MTases.hmm	0.98155	2.3e-86	FAM_4							
T7-006_S6.fa Type_II_MTases.hmm	0.977918	1.2e-117	FAM_2							
T7-006_S6.fa Type_II_MTases.hmm	0.838798	8.8e-21&8e-(	FAM_0							

Figure 13: Concatenation of one example strain from R-M Type II systems displaying a total of 5 MTases and one REase that does not match any of the MTases. This indicates fragments of R-M systems being present.

The prevalence and actual positive R-M systems (matches) found in the systems are displayed in Table 7. Overall there were more MTases, than REases in all the systems with co-dependence.

Table 7: Number of MTases and REases found in validation of complete R-M systems in all strain populations.

Strain population	R-M Type I			R-M Type II			R-M Type III			R-M Type IV	R-M Type IIG
	MTase	REase	Match	MTase	REase	Match	MTase	REase	Match	REase	REase
Carrier strains	373	321	**116**	2138	75	**0**	24	23	**19**	65	51
NORM strains	280	217	**51**	1442	69	**0**	22	20	**14**	49	30
ST307 strains	206	107	**91**	471	97	**0**	0	0	**0**	0	1

In total, if the R-M systems MTasse/REase subtypes had not been evaluated, the carrier strains would display a total of ~1.1 R-M systems per strain. With a proper manual evaluation, the average was found to be ~0.6 R-M systems per strain. Correspondingly the NORM strain collection would display a total of ~0.9 R-M systems per strain but were assessed to be ~0.4 R-M systems per strain. Likewise, ST307 displayed ~0.9 R-M systems per strain before evaluation and ~0-4 R-M systems per strain after evaluation.

5.4 PCA plot comparison of the strain collections

A PCA plot illustrates the grouping of the strain populations (carrier-, NORM non-ESBL-, NORM ESBL and ST307 strains), based on their selected features (Figure 14).

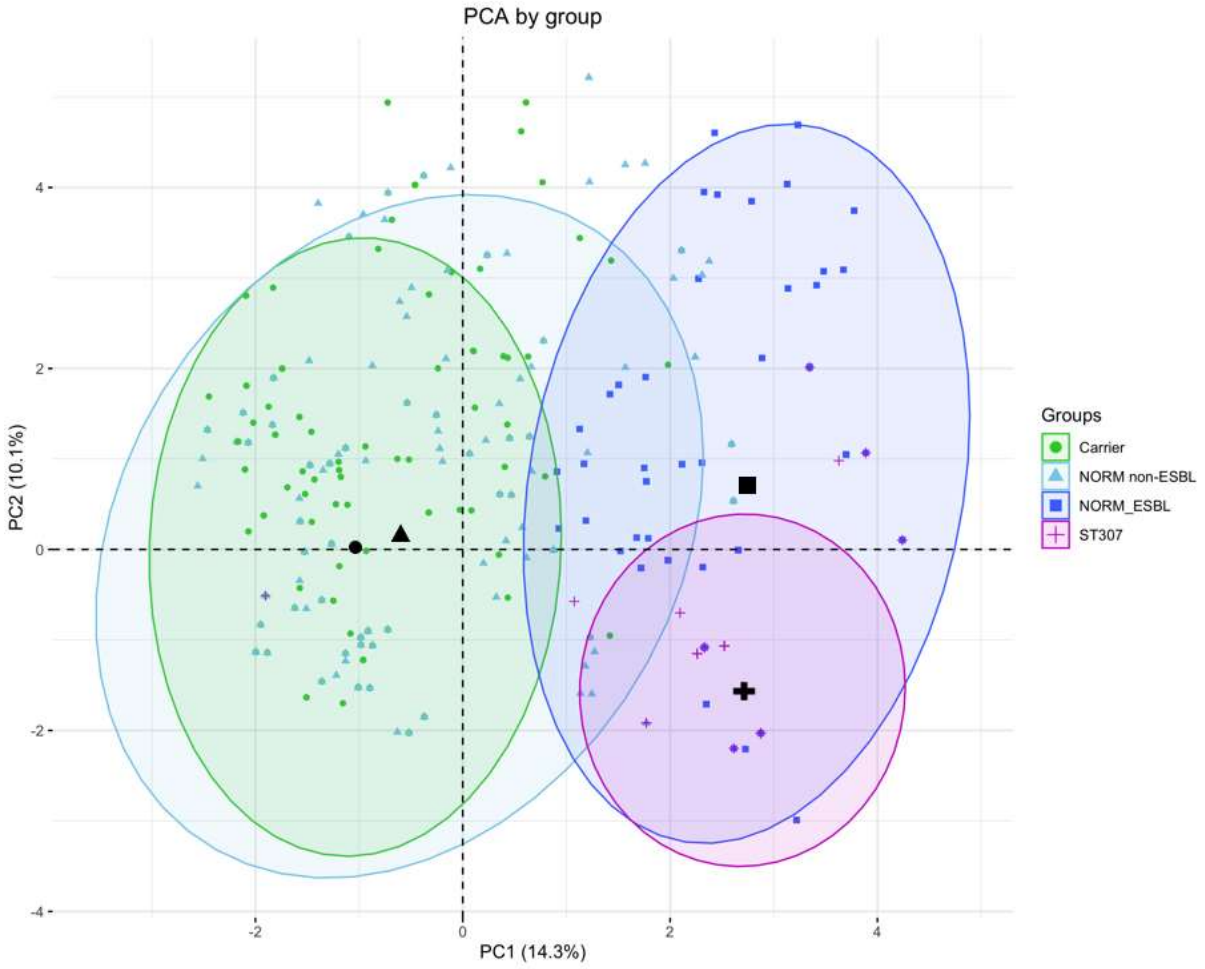


Figure 14: PCA plot displaying the distribution of the selected features (CRISPR-Cas, R-M systems, virulence score, AMR classification and plasmid content) grouped by the strain populations; Carrier- (n=484), NORM non-ESBL- (n=270), NORM ESBL- (n=144) an ST307 (n=101) strains. The ellipses display the concentration of the group and the average points (highlighted by added black symbols), display the average point for the group. This PCA scatter plot was generated using R studio and FactorMineR.

The largest difference seen by the separating concentration ellipses for the strain collections seemed to be affected by the presence or absence of ESBL alleles, in which they were classified by in the first place. However, the total eigenvalues in PC1 only displayed a total of 14,3% difference. The carrier strains and NORM non-ESBL strain population had the most in common with overlapping ellipses. Next, NORM ESBL and ST307 had many features separating them from the carrier and NORM non-ESBL population. The overlap between ST307 and NORM ESBL strains was almost complete.

The average point (the oversized black points) of the carrier- and NORM non-ESBL strains were placed a little apart both in the PC1 and PC2 axis, displaying differences in the eigenvalues. NORM ESBL- and the ST307 strains had a more defined separation between the average points in PC2, but a higher similarity in PC1. Interestingly, the distance between NORM ESBL and the ESBL negative strains (Carrier and NORM non-ESBL) in PC2 were quite small, compared to the ST307.

5.5 CRISPR-Cas and R-M systems and their correlation with virulence profile, AMR classification and plasmid content

The correlation between different combinations of presence-absence of CRISPR-Cas and/or R-M systems and their correlation to virulence profile, AMR -classification and plasmid content is given in this section.

5.5.1 CRISPR-Cas positive and negative strains: population comparisons, virulence score, AMR classification and plasmid content

Looking closer into the distribution of CRISPR-Cas systems in the strain populations revealed significant differences between ST307 and the other populations (Figure 15). ST307 displayed a complete CRISPR-Cas negative profile in contrast to the NORM ESBL population (Table 6). Moreover, no components or fragments of CRISPR-Cas components were detected in the ST307 strains. The other strain populations had a more similar CRISPR-Cas distribution and did not display any statistically significant difference between the presence/absence or system distribution between the carrier- and NORM non-ESBL strains (Table 6).

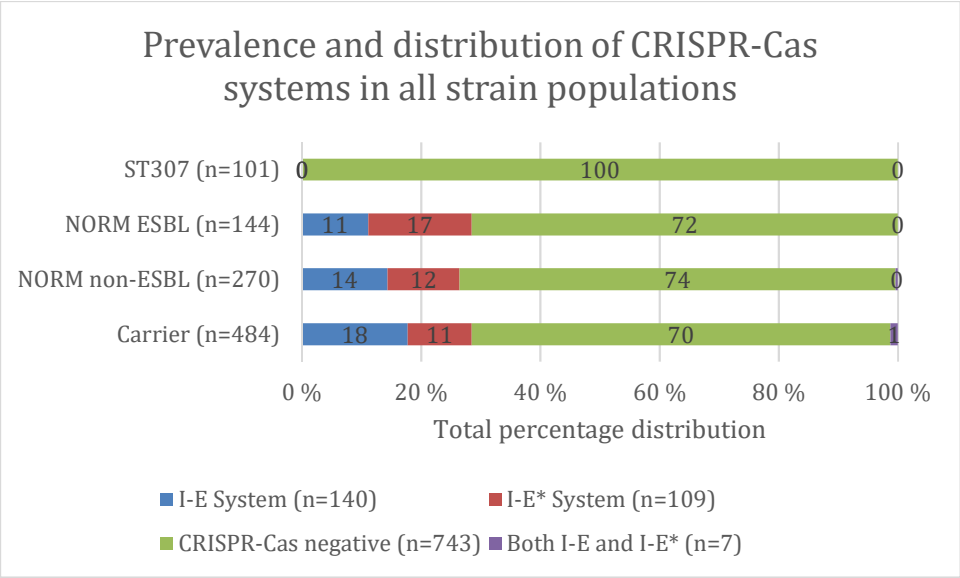


Figure 15: Prevalence and percentage distribution of CRISPR-Cas systems (n=256) in all strain populations.

A total of 26% (n=256) of the strains were defined as CRISPR-Cas positive (Figure 15). A total of 14% (n=140) were Class 1 Type I-E strains, 11% (n=109) Class 1 Type I-E* strains and 0,7% (n=7) harboured both subtypes. The MLST types associated with this co-occurrence was ST129 (n=2), ST790 (n=1), ST427 (n=1), ST1613 (n=1), ST1119 (n=1) and ST4290 (n=1). The distribution of subtypes was found to be a total of 59% for the I-E type and 41% I-E* amongst the CRISPR-Cas positive strains.

Further analysis of differences between the CRISPR-Cas presence/absence strains visualised in a PCA plot, revealed small differences between the groups (Figure 16). The PCA plot grouped all the strains based on their CRIPSR-Cas profile (absence, only subtype I-E, both subsystems and only subtype I-E*) separating them by their selected features. The presence/absence of R-M systems were not taken in consideration.

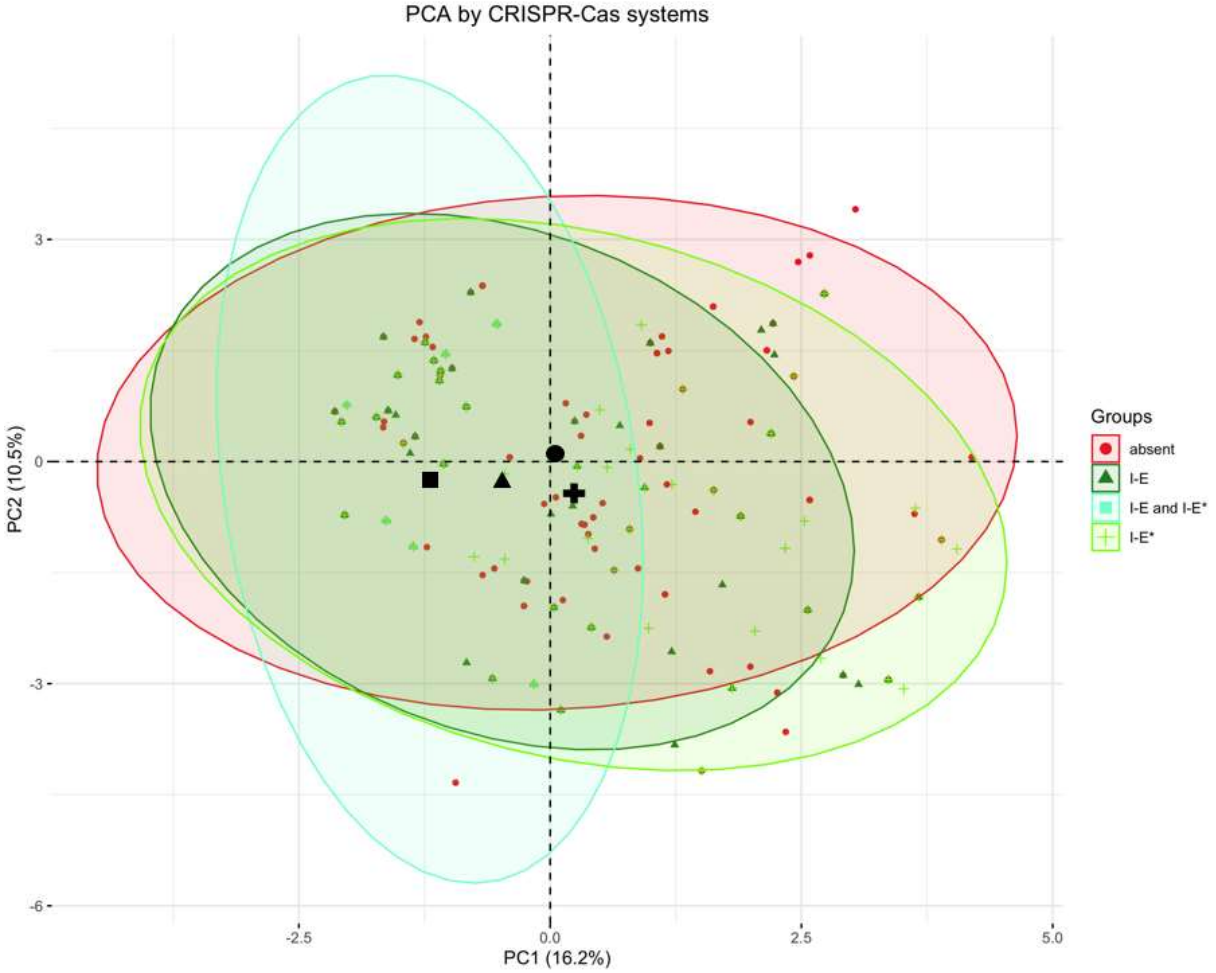


Figure 16: PCA plot displaying the distribution of the selected features (R-M systems, virulence score, AMR classification and plasmid content) grouped by the absence of CRISPR-Cas systems (n=743), CRISPR-Cas Class 1 Type I-E systems (n=140), CRISPR-Cas Class 1 Type I-E and I-E systems (n=7) or CRISPR-Cas Class 1 Type I-E* system (n=109). The ellipses display the concentration of the groups and the average points (highlighted by added black symbols), display the average point of the groups. This PCA scatter plot was generated using R studio and FactorMineR.*

The concentration ellipses in Figure 16 indicated a complex relationship between the absence and the presence of the different CRISPR-Cas subsystems in all strain populations. All the points for the specific strains were scattered across the plot with much overlap. Absence of CRIPSR-Cas in the strain populations had a large overlap with presence of CRISPR-Cas systems. Moreover, CRISPR-Cas Class 1 Type I-E* displayed only small differences compared to the strains with absence of the systems. CRISPR-Cas Class 1 Type I-E had an overall more concentrated ellipse with an almost complete overlap of the absent group. Strains harbouring both types of CRISPR-Cas subsystems (I-E and I-E*), seemed to be more defined in PC1 and had the largest difference compared to the strains with no CRISPR-Cas systems. Looking at the ellipse concentration points, the largest difference on PC1 was between strains harbouring both

Type I-E and I-E* versus only type I-E*. In the PC2 axis it can be debated witch of the absent and Type I-E* concentration points are located furthest away.

Further we investigated closer the actual features separating these groups. PCA analysis differentiating between CRISPR-Cas Class 1 System Type I-E or I-E*, compared to both systems, displayed differences between all populations across both axis (Figure 17).

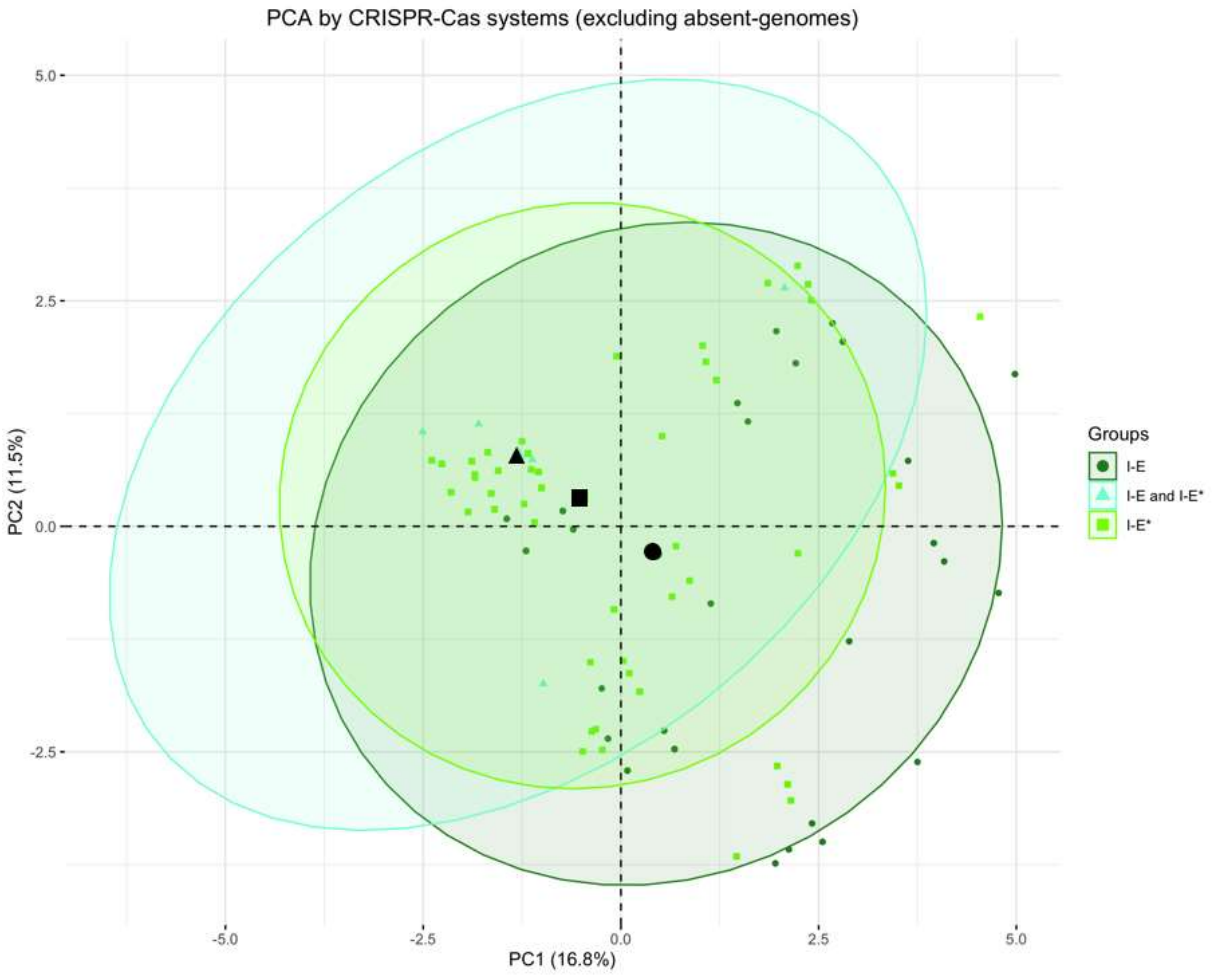


Figure 17: PCA plot displaying the distribution of the selected features (R-M systems, virulence score, AMR classification and plasmid content) sorted by their CRISPR-Cas Class 1 Type I-E (n=140), CRISPR-Cas Class 1 Type I-E and I-E (n=7) and CRISPR-Cas Class 1 Type I-E* systems (n=109). The ellipses display the concentration of the groups and the average points (highlighted by added black symbols), indicates the average point for the group. This PCA scatter plot was generated using R studio and FactorMineR.*

Compared to each other, the subtype I-E and I-E* displayed a different distribution compared to the previous PCA pot (Figure 17). In the total comparison, the Type I-E* were more defined and smaller than the distribution of Type I-E. This plot indicated differences between the CRISPR-Cas systems in both PC1 and PC2. Type I-E strains displayed the largest difference

to strains harbouring both systems and the Type I-E* strains were found in the middle of both groups. However, the biggest difference comprised a total of maximum 16,8% difference.

Analyses of the selected features (virulence profile, AMR classification and plasmid content) within the categories of strains used in the previous PCA plot, displayed small differences between the groups (Figure 18). The CRISPR-Cas I-E and I-E* category is included in the comparison although the number of strains (n=7) is very low.

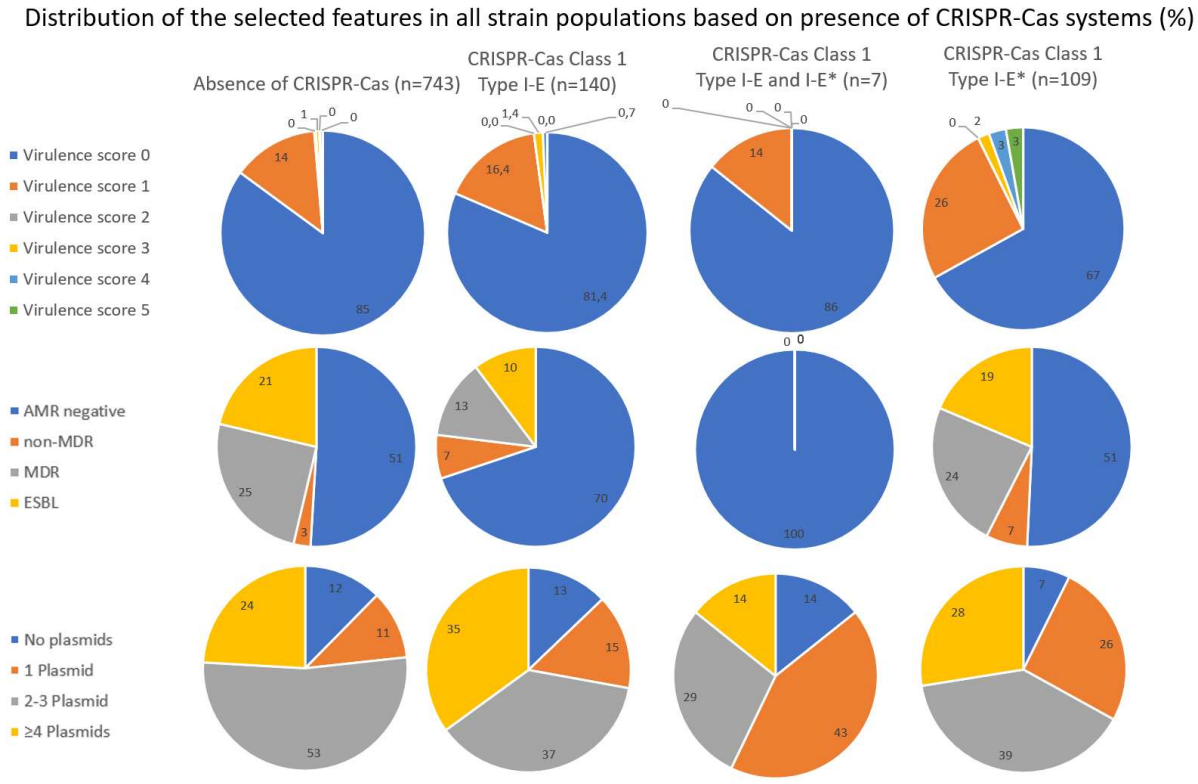


Figure 18:The percentage distribution of the selected features (virulence profile, AMR classification and plasmid content), grouped by the absence of CRISPR-Cas systems (n=743) or the presence of CRISPR-Cas Class 1 Type I-E (n=140), presence of both CRISPR-Cas Class 1 Type I-E and I-E systems (n=7) and presence of CRISPR-Cas Class 1 Type I-E* systems (n=109).*

The virulence profiles only differed by small percentages for all the categories, except CRISPR-Cas Class1 Type I-E* (n=109) displaying a total of 33% strains harbouring acquired virulence factors (virulence score 1-5) compared to 14-19% for the other groups. The presence of virulence factors was not significant between the Type I-E and I-E* systems (p=0.2128).

The AMR classification displayed the highest prevalence of AMR negative strains in CRISPR-Cas Class1 Type I-E and I-E* positive strains (n=7), by 100%. Next, were the CRISPR-Cas

Class 1 type I-E strains (n=140) with 70% AMR-negative and last, the strains with absence of CRISPR-Cas systems with 51% and CRISPR-Cas Class 1 Type I-E* strains with 51% AMR-negative strains. The number of AMR negative strains displayed a statistical significant difference between the Type I-E systems and strains harbouring both systems (p=0.00001). Non -MDR strains were found in a total of 3% for CRISPR-Cas negative strains, 7% for both the CRISPR-Cas Class 1 Type I-E strains and the CRISPR-Cas Class 1 Type I-E* systems. MDR strains were found in the highest percentage of 25% in the CRISPR-Cas negative strains, 13% in CRISPR-Cas Class 1 Type I-E strains and 24% in CRISPR-Cas Class 1 Type I-E * strains. The MDR content was significant different between the strains with absence of CRISPR-Cas systems and the strains with Type I-E* (p=0.00001). ESBL -production had the largest presence in the strain population without CRISPR-Cas (21%). This group included all ST307 strains (n=101). The second highest ESBL -production were found in 19% of the strains with CRISPR-Cas Class 1 Type I-E* systems and lastly, 10% amongst CRISPR-Cas Class 1 Type I-E strains. Overall, in decreasing order, the CRISPR-Cas positive strains harbouring both of the Class 1 Type I-E and I-E* subtype had the lowest AMR profile, CRISPR-Cas Class 1 Type I-E the second lowest, third lowest (based on prevalence of MDR and ESBL -producing strains) is the systems without CRISPR-Cas systems and last, the strains with CRISPR-Cas Class 1 Type I-E.

The plasmid profiles displayed similar proportions of absence of plasmids in the various strain categories. The categories of 2-3 plasmids and ≥4 plasmids were the highest in the CRISPR-Cas negative strains by 77%, Second highest in the CRIPSR-Cas Class 1Type I-E strains by 72%, thirdly a 67% carriage in the CRISPR-Cas Class 1 Type I-E* strains and lastly, 43% carriage in the CRISPR-Cas Class 1 Type I-E and I-E* strains.

5.5.2 R-M system positive and negative strains: population comparisons, virulence score, AMR classification and plasmid content

The distribution of R-M systems in the strain collections is presented in Figure 19. Briefly, contrary to the complete absence of CRISPR-Cas systems for the ST307 strain collection, a total of 90% were R-M Type I positive. The NORM ESBL population displayed a total of 26% R-M system positive strains which is lower compared to the NORM non-ESBL (57%) and carrier (50%) strains populations. A total of 19% R-M systems was classified Type I, 1% Type II, another 1% Type III, 6% Type IV and lastly 3% Type IIG. Statistically significant

differences between the R-M systems presence/absence and distribution of R-M Type IV and IIG systems between ST307 and NORM ESBL were observed (Table 6).

There was no statistical difference between the presence/absence of R-M systems between the NORM non-ESBL and the carrier strain collection (Table 6).

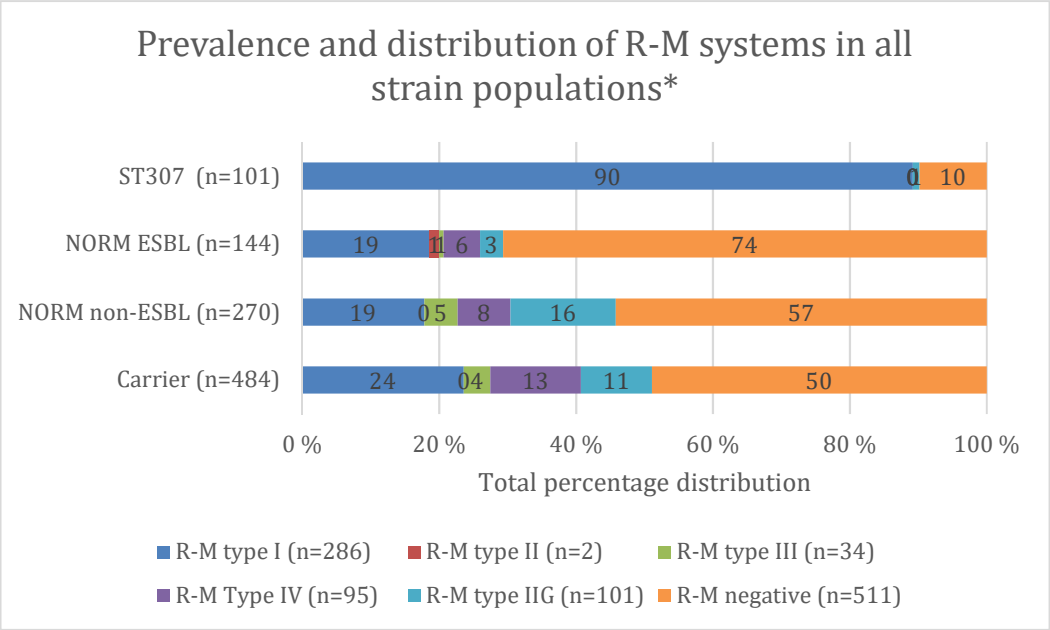


*Figure 19: Prevalence and percentage distribution of R-M systems (n=488) in all strain populations. *Including co-occurrence of R-M systems (n=60).*

Looking at the overall strain collections, a total of 51% (n=511) of the strains were R-M system negative and 49% (n=488) were R-M system positive. The most prevalent R-M Type I occurred in a total of 29% (n=286) of the total strain populations. The Type II was present in two strains, Type III 3% (n=34), Type IV 10% (n=95) and Type IIG 10% (101).

In addition to simply looking at the presence or absence of the R-M systems, a total of 35 strains had two occurrences of R-M Type I, two of them in co-occurrence with R-M Type IV and one in co-occurrence with R-M Type IIG. Additionally, one of the ST307 strains had three R-M Type I systems.

Co-occurrences of different R-M system types were found in a total of 60 strains. Their origins were from all the strain populations, carrier (n=38), NORM non-ESBL (n=15), NORM ESBL (n=6) and ST307 (n=1) (Figure 20).

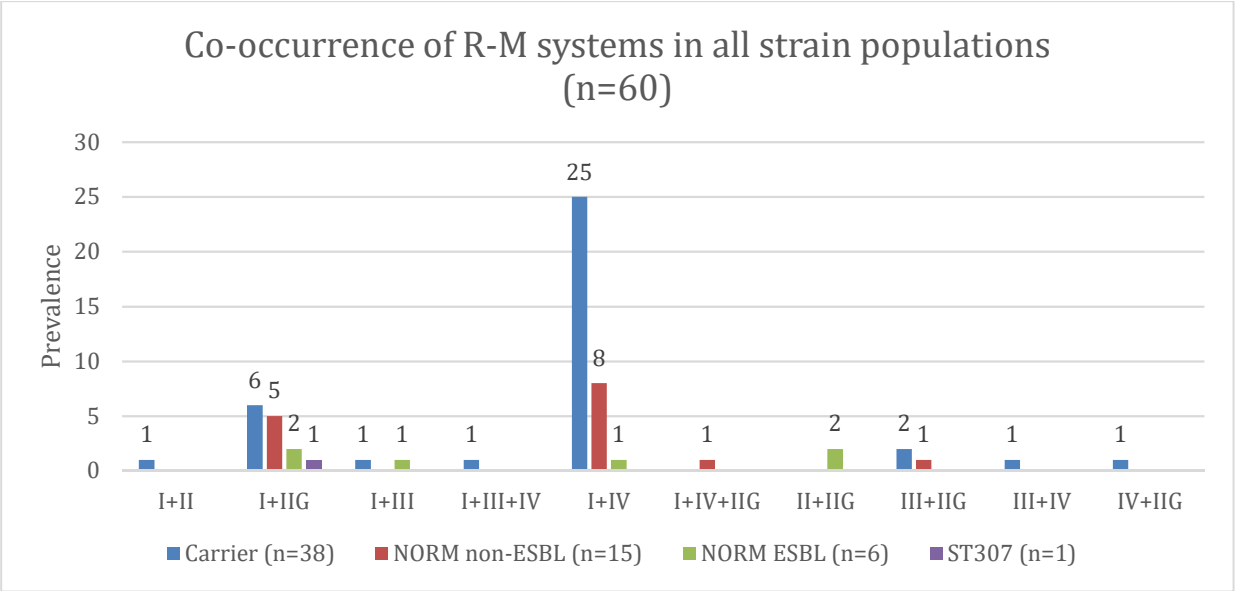


Figure 20: Number of strains found in ten different co-occurrences of R-M systems in a total of sixty strains.

The co-occurrence was seen in a total of 10 different combinations of both double and triple combinations (Figure 20). The most prevalent co-occurrence was the Type I and IV.

For the R-M systems, a PCA plots grouped the strains by the presence versus absence of R-M systems, based on their selected features (Figure 21).

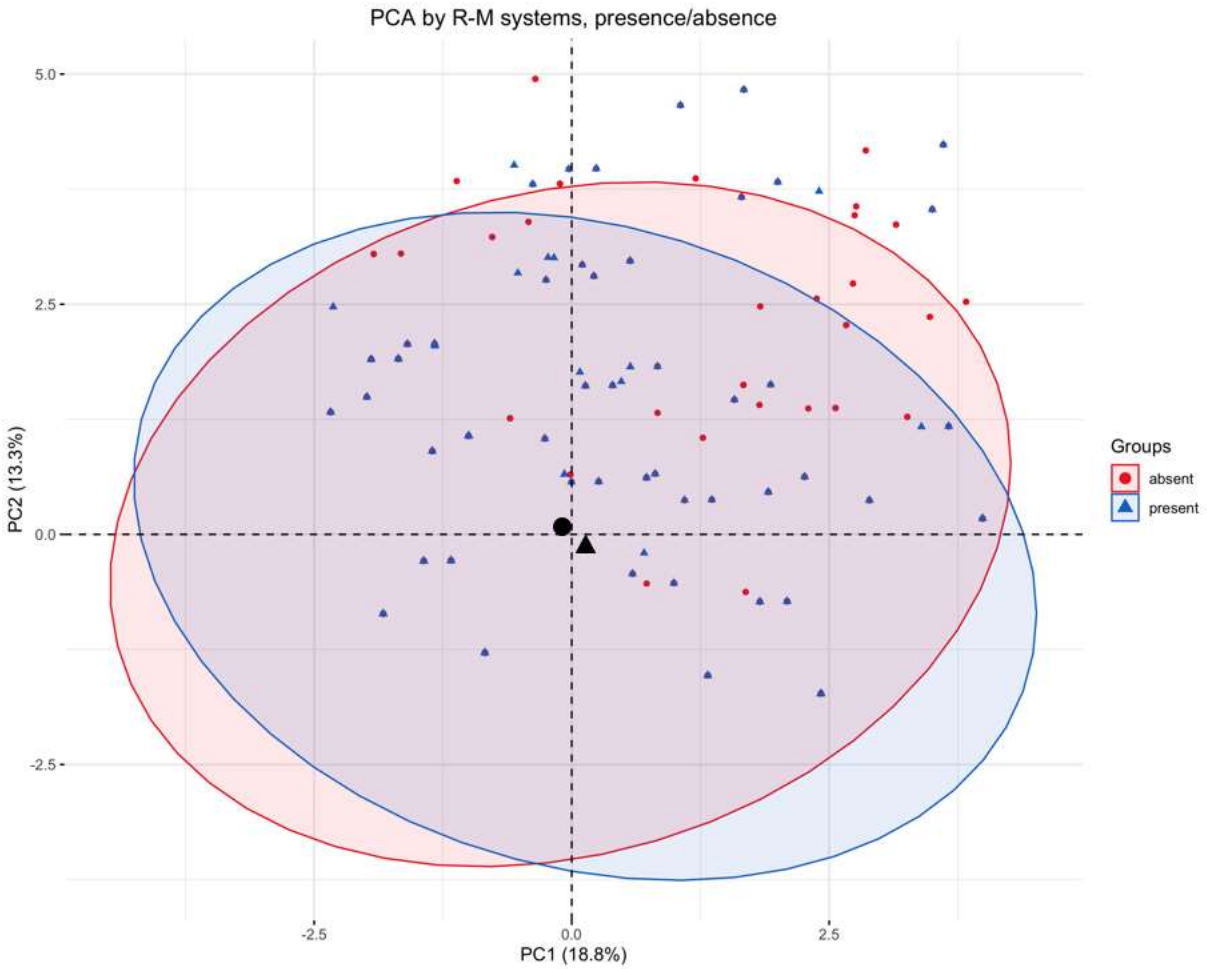


Figure 21: PCA plot displaying the distribution of the selected features (CRISPR-Cas, virulence profile, AMR-classification and plasmid content) sorted by the presence (n=457) or absence of R-M systems (n=542). The concentration ellipses display the average distribution of the groups and the average points (highlighted by added black symbols) display the group average point. This PCA scatter plot was generated using R studio and FactorMineR.

The overall spread PCA plot indicated only small differences between the presence/absence of R-M systems. The ellipses were mostly overlapping with only small differences separating the populations across the PCs. The individual points were mostly located in the same areas. The average points for the populations were close together and only revealing small differences between the groups.

PCA analysis of the total distribution displaying all single and multiple co-occurring R-M systems, were too complex and did not show a good distribution. Therefore, they were all categorised as single and multiple systems (Figure 22).

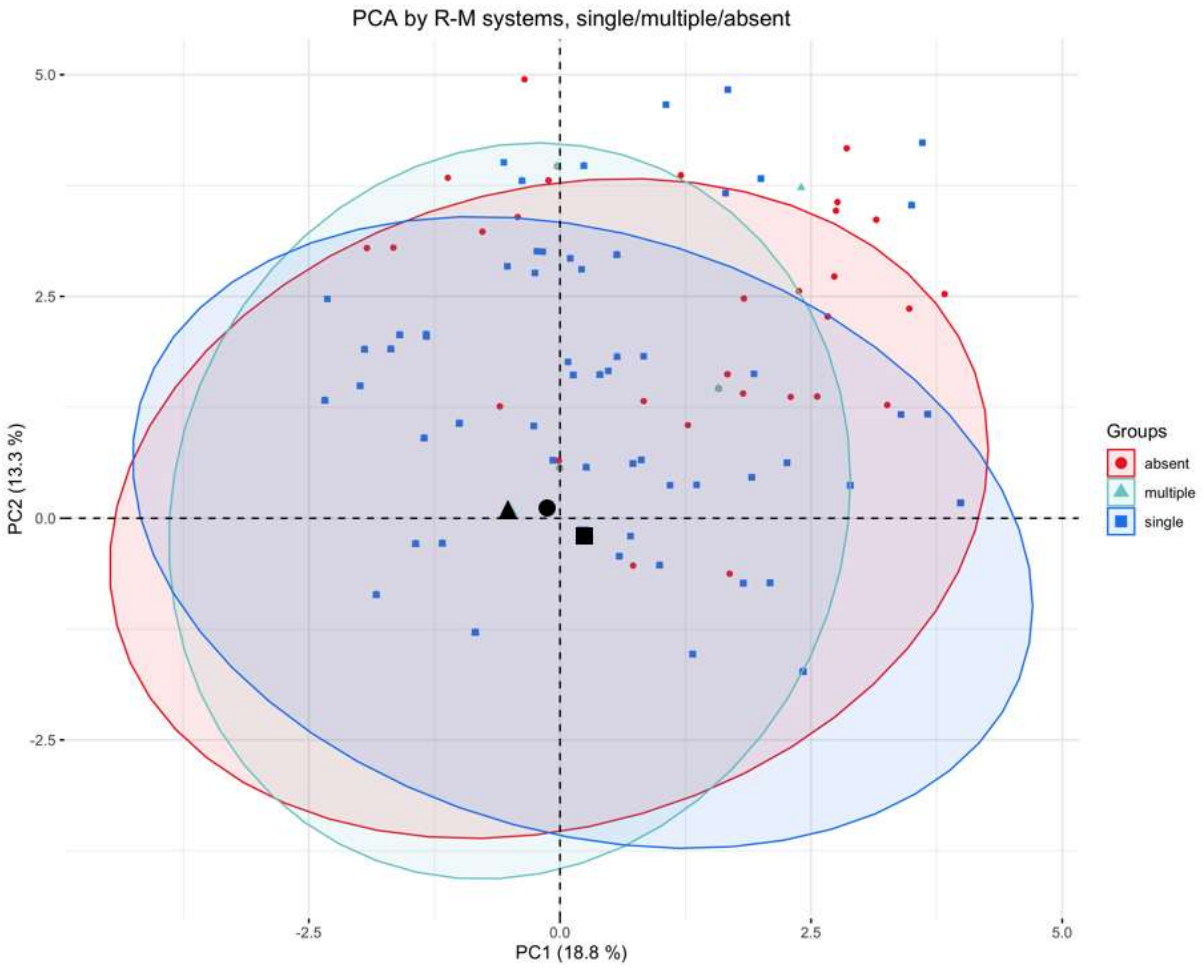


Figure 22: PCA plot displaying the distribution of the selected features (CRISPR-Cas, virulence score, AMR classification and plasmid content) sorted by the absence (n=541) or presence of single R-M systems (n=398) or multiple R-M systems co-occurring in the same strain (n=60). The ellipses display the concentration average of the groups and the average points (highlighted by added black symbols) display the group average point. This PCA scatter plot was generated using R studio and FactorMineR.

The difference between the previous presence/absence plot (Figure 21) and Figure 19, is mainly the location of the concentration ellipse of the strains with co-occurrence of multiple systems. The group of multiple systems displayed similarities with both groups, however it shared the most features with the group not harbouring any R-M systems. The concentration ellipses mainly overlapped, and the individual points were scattered in the same areas. In total, the largest difference indicated only ~19% difference in the features (PC1). However, the average points were slightly separated by the absence in the middle. Moreover, the total concentration ellipses displayed small variances between the populations.

Analysing their profiles (virulence score, AMR -classification and plasmid content), based on the occurrence of single systems, multiple systems or the absence of R-M systems, indicated small variances (Figure 23).

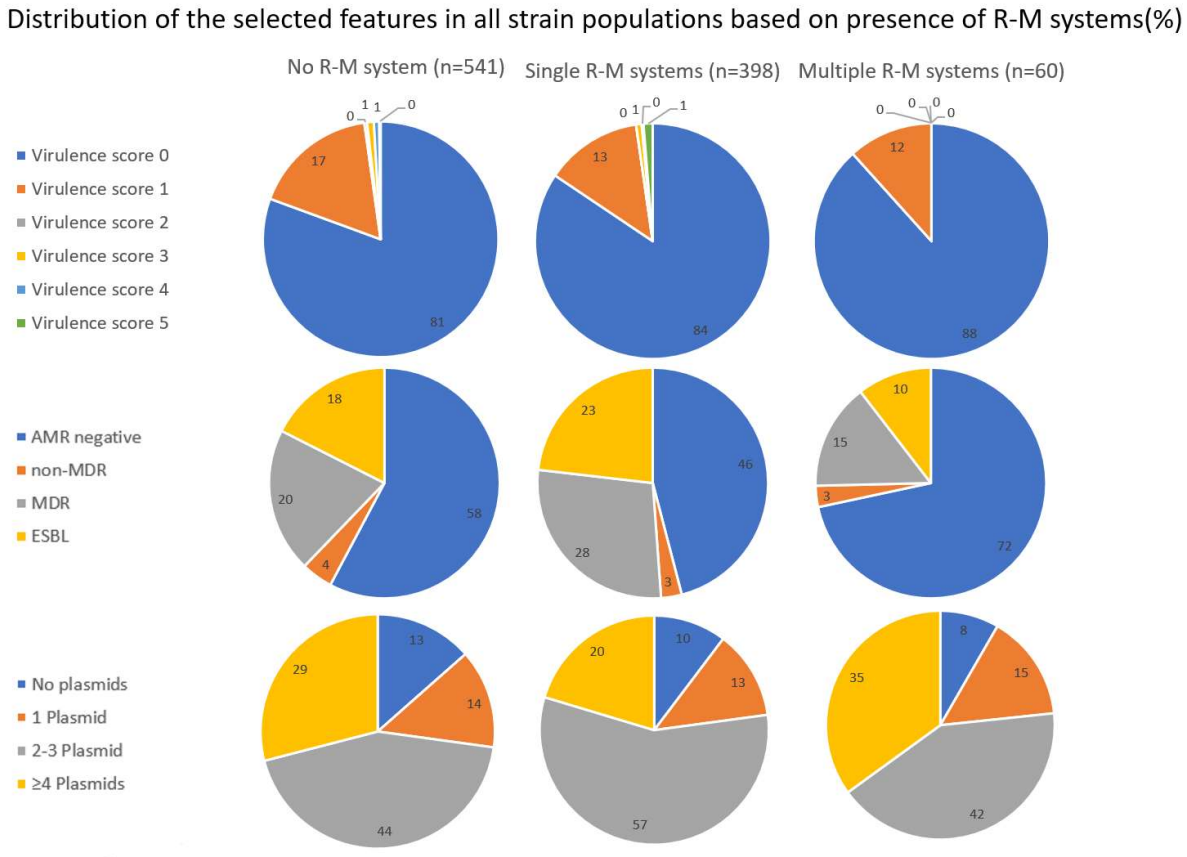


Figure 23: Percentage distribution of the selected features (virulence score, AMR classification and plasmid content) based on the absence of R-M systems (n=541), presence of single R-M systems (n=398) or presence of multiple R-M systems co-occurring in the same strains (n=60).

The virulence profiles for the different R-M categories was quite similar (Figure 23). In total, the strains with multiple R-M systems in co-occurrence displayed 88% of the strains without any acquired virulence factors (virulence score 0), 84% of the single R-M systems were virulence negative and last, the 81% of strains without any R-M systems without acquired virulence factors. There was a significant difference in virulence negative strains between the group without R-M systems and the group with single R-M systems (p=0.00001).

AMR profiles for the strains displayed a more complex relationship between the groups (Figure 23). The proportion of AMR negative strains were highest in the multiple R-M systems (72%), followed by strains without R-M systems (58%) and single R-M system strains (46%).

The prevalence of AMR negative strains was significant different for the strains with no R-M system compared to the ones with single R-M systems (p=0.00001). Overall, the strains with multiple systems, displayed a lower content of resistance genes. The R-M negative strains displayed the second highest prevalence of resistance genes. Lastly, the strains with single R-M systems had the highest percentage of resistance, especially MDR and ESBL.

In terms of plasmids, 13% of the R-M negative group, 10% of the single R-M systems and 8% of the group with multiple systems had no plasmids. The plasmid load ≥4 plasmids were significant different between the R-M negative group and multiple R-M group (p=0.00001).

Both the single and multiple R-M systems displayed the same plasmid content higher than for the R-M negative strains. Overall the R-M negative group seemed to display the least number of plasmids increasing for the number of R-M systems.

5.6 Co-occurrence of CRISPR-Cas- and R-M systems: population comparisons, virulence score, AMR classification and plasmid content

Figure 24 displays the PCA plot grouped by absent (neither CRISPR-Cas- or R-M system), only CRIPSR-Cas systems, only R-M systems or both CRISPR-Cas and R-M systems.

In total, the group carrying both CRISPR-Cas and R-M systems consisted of 71 carrier strains and 45 NORM strains (non-ESBL (n=33) and ESBL (n=12)). Only R-M (n=342) was seen for 143 carrier strains, 108 NORM strains (non-ESBL (n=82) and ESBL (n=26)) and 91 ST307. The group only containing CRISPR-Cas systems (n=140) consisted of 73 carrier strains and 67 NORM strains (non-ESBL (n=38) and ESBL (n=29). The group that did not contain any systems (n=401) was comprised of a total 197 carrier strains, 194 NORM strains (non-ESBL (n=117) and ESBL (n=77)) and ST307 (n=10) strains. ST307 was only represented in the

group harbouring R-M systems and no other systems.

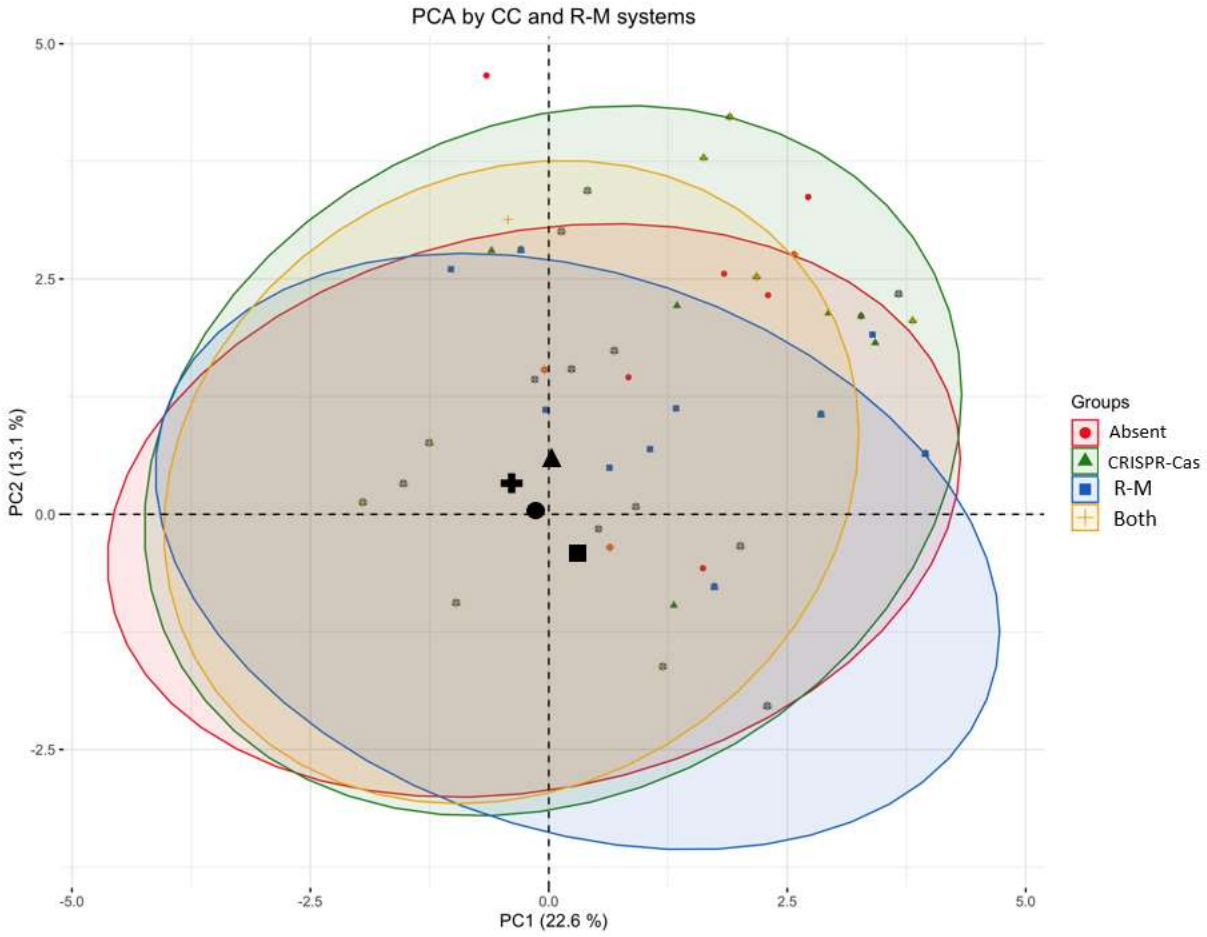


Figure 24: PCA plot displaying the distribution of the selected features (virulence score, AMR classification and plasmid content) sorted by the presence of only CRISPR-Cas systems (n=140), only R-M systems (n=342), both CRISPR-Cas- and R-M systems (n=116) and absence of both systems (n=401). Including their designated concentration ellipses and average points (highlighted by added black symbols), displaying the average point of the group. This PCA scatter plot was generated using R studio and FactorMineR

By looking at the concentration ellipses, there were small differences in the features between the groups. The maximum percentage difference was 22,6%. All groups had overlaps, but it seemed as the largest similarity was between strains harbouring only CRISPR-Cas systems and those with both systems. Secondary, the groups without any system also showed resemblance with the strains not harbouring any systems. This was seen for both the concentration ellipses and the average concentration points. The group with the largest difference was the ones only harbouring R-M systems. They revealed the largest difference at PC2.

Closer analysis of the selected features (virulence score, AMR classification and plasmid content) grouped by the presence of only CRISPR-Cas-, only R-M-, negative for all systems

and both CRISPR-Cas- and R-M systems, displayed small variances between the systems (Figure 25).

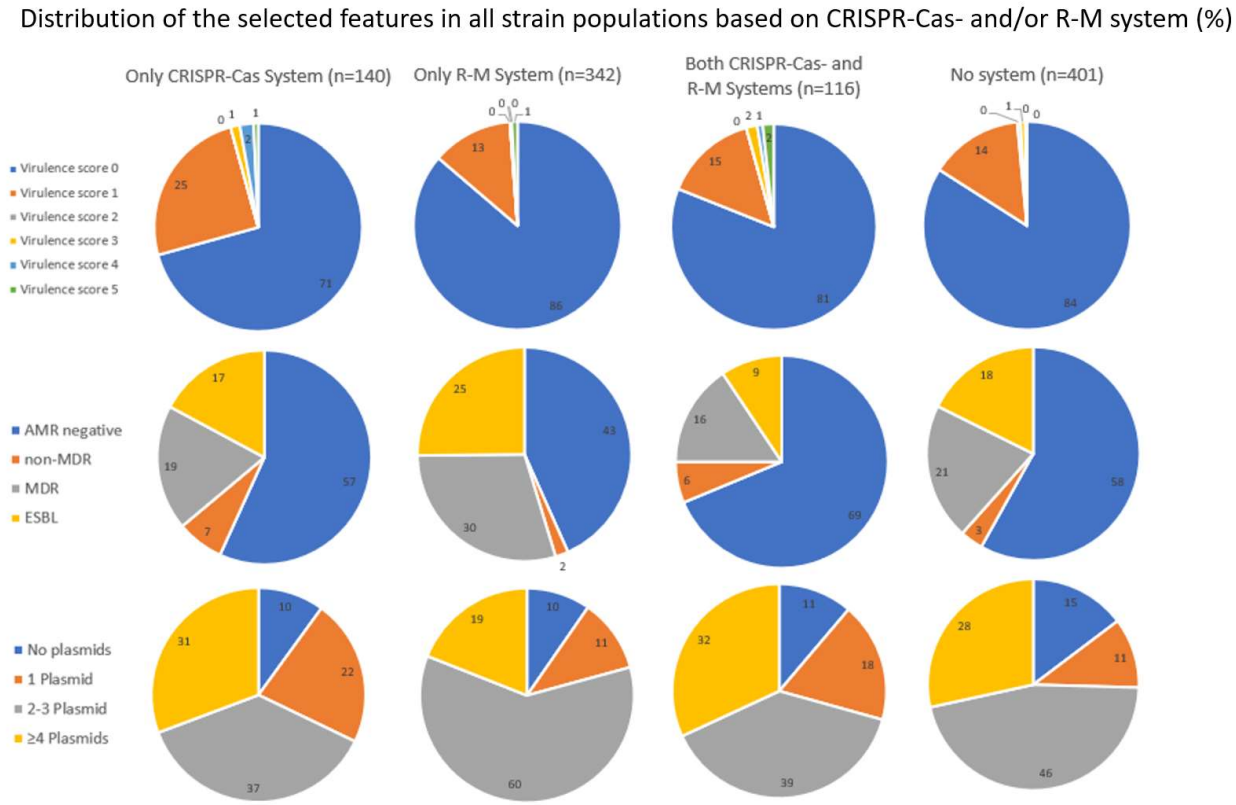


Figure 25: Percentage distribution of the selected features (virulence score, AMR classification and plasmid content) for all the strain collections grouped by the presence of only CRISPR-Cas systems (n=140), only presence of R-M systems (n=342), both CRISPR-Cas- and R-M systems (n=116) and the strains without CRISPR-Cas- and R-M systems (n=401).

An overall low **virulence score** was observed within each group. Virulence score 0 was lowest in the CRISPR-Cas system only group (71%) compared to the R-M system only group (86%), CRISPR-CAS and R-M system group (81%), and the CRISPR-Cas and R-M negative group (84%). The proportion of virulence positive strains with only CRISPR-Cas were significantly different from the virulence positive strains with both systems (p=0.0118). In addition, there was a significance with virulence score 0 between harbouring only CRISPR-Cas systems and no systems (p=0.00001).

Resistance profiles displayed some variations between the groups. The highest proportion of AMR negative strains was observed in the CRISPR-Cas and R-M group (69%) compared to the CRISPR-Cas and R-M negative group (58%), the CRISPR-Cas only group (57%), and the R-M only group (43%). There was no significant difference in the proportion of AMR negative strains between the group with only CRISPR-Cas systems and the group with both systems

(p=0.5783). There was however a significant difference in AMR negative strains between both systems and the group with no system (p=0.00001). In total, the group with only R-M systems had the highest number of resistance genes, followed by the group without systems and the similar group with only CRISR-Cas systems, leaving the group of both systems with the lowest percentage of resistance mechanisms.

The **plasmid profile** in the CRISPR-Cas only group showed that a total of 10% were without plasmids, which was similar to the R-M only group (10%), the CRISPR-Cas and R-M group (11%) and higher in the CRISPR-Cas and R-M negative group (15%). The proportion of strains with ≥ 2 plasmid was approximately the same within each group: CRISPR-Cas only group (68%), R-M only group (79%), the CRISPR-Cas and R-M group (71%), and the CRISPR-Cas and R-M negative group (74%). There was a significant difference in carrying ≥2 plasmids between the group with only R-M systems and the group with only CRISPR-Cas systems (p=0.00001). There was a significant difference in harbouring ≥4 plasmids between the group with only R-M systems and the group with no systems (p=0.0018). There was a significant difference carrying ≥4 plasmids between the group with both systems and the strains with only R-M systems (p=0.0027). There was not a significance in harbouring ≥4 plasmids between the group with both systems and the group with only CRISPR-Cas systems (p=0.5434).

5.6.1 Plasmid content in relation to CRISPR-Cas- and R-M systems

Moving into analysing the plasmid content of all the strains, they all had similarities between the strain populations (Figure 26). Lack of threshold in the bioinformatic plasmid detection, resulted in a total of 17% false positive nucleotide matches. These strains are marked with red in the Supplementary Table 2.

On average of 2.3 plasmids per carrier strain, 2.7 plasmids per NORM strain and 2.5 plasmids per ST307 strain were observed. In all strain populations the maximum number of plasmids was seven.

The most frequently found plasmid replicons were the ColRNAI_1 (n=250), INFIB(K)_1_Kpn3 (n=725), IncFII_1_pKp91 (n=548), IncFIA(HI1)_1_HI1 (n=135), IncFIB(Mar)-1_pNDM-Mar (n=41), IncFIB(pKPHS1)_1_pKPHS1 (n=76), and IncN_1 (n=60) and IncR_1 (n=130). In addition to these groups, some were seen more exclusively for certain groups. For example, the Col(MGD2)_1 (n=35) had a higher presence in the carrier and the NORM non-ESBL strain collection, decreasing or absent in the NORM-ESBL and ST307 population. The IncFiB(pQil)-1_pQil (n=24) was predominantly seen in the NORM ESBL strains and in the ST307 population.

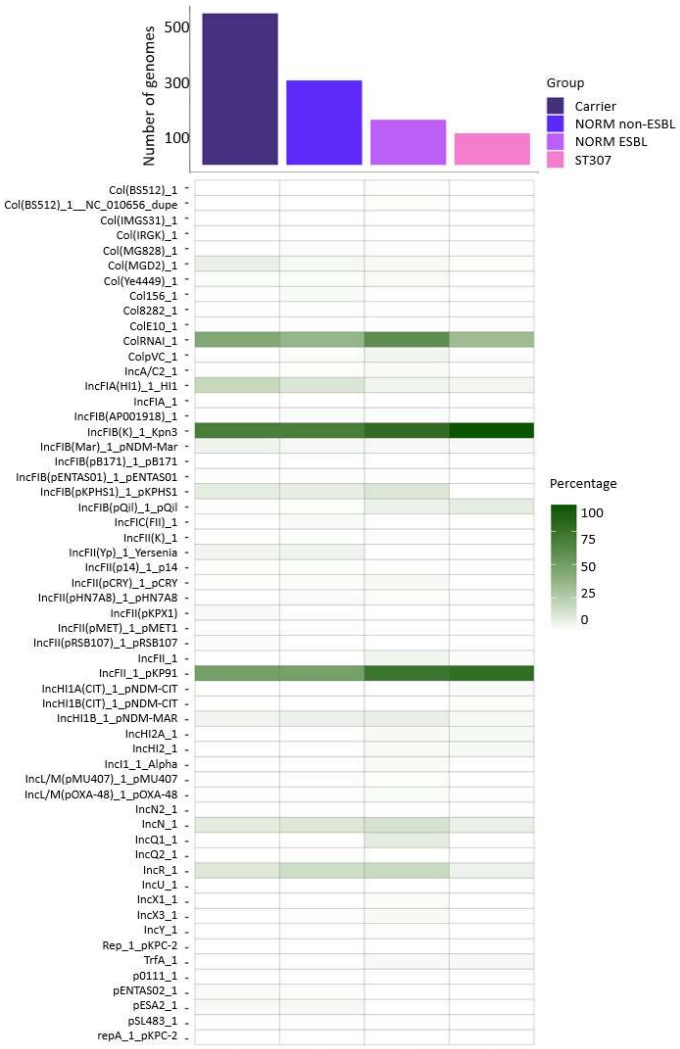


Figure 26: Percentage distribution of the total plasmid profile for all the strain populations. This heatmap was made using R -studio FactorMineR. Note that ~17% might display false positives.

5.6.1.1 Plasmid distribution in the strain populations

Looking into the total of plasmids sorted into the selected groups, displayed similarities between the carrier- and the NORM non-ESBL strains, but a higher prevalence and different profile in the NORM ESBL- and ST307 strains (Figure 27).

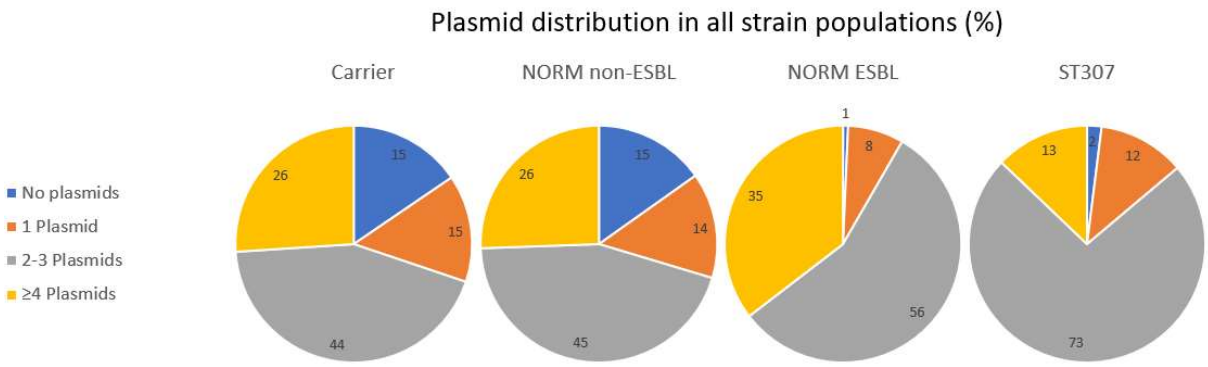


Figure 27: Percentage distribution of the plasmid categories within the different strain populations.

The total **plasmid content** for the carrier- and the NORM non-ESBL strains was only differing by 1%. The proportions of strains in the carrier (15%) and the NORM non-ESBL (15%) were significantly lower than in the NORM-ESBL (1%) and the ST307 (2%) populations. Correspondingly, the proportions of strains with a plasmid load ≥ 2 were significantly higher in the NORM-ESBL (91%) and ST307 (86%) populations compared to the carrier (70%) and NORM-non-ESBL (71%) populations.

To better understand the effects of CRISPR-Cas- and R-M systems effects on plasmid content in the different strain populations, one must look at the individual strain populations. Starting by the carrier strains (Figure 28). The total of the strain population is summarised in the black frame and the rest is classified by the presence and/or absence of CRISPR-Cas- and R-M systems.

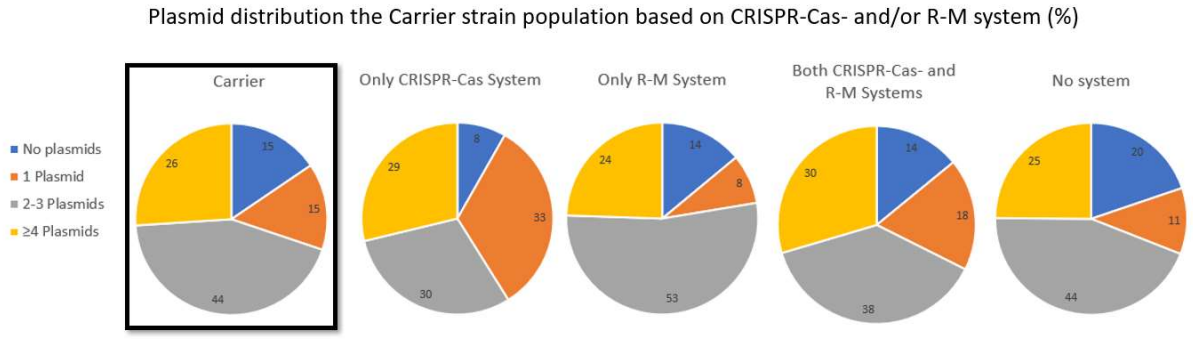


Figure 28: Percentage distribution of the plasmid categorisation for the carrier strain population (n=484), based on the total strain population (marked by frame), presence of only CRISPR-Cas systems (n=73), only R-M systems (n=143), both CRISPR-Cas- and R-M systems (n=71) and the strains without CRISPR-Cas- and R-M systems (n=197).

The proportion of carrier strains with no plasmids varied between groups: CRISPR-Cas only (8%), R-M only (14%), CRISPR-Cas and R-M (14%), and finally CRISPR-Cas and R-M negative group (20%). There was a significant difference in no plasmids between the group

with only CRISPR-Cas systems and the group with only R-M systems (p=0.0044). There was no significant difference in no plasmids between only CRISPR-Cas systems and the group with both systems (p=0.4284). The proportion of strains with a plasmid content ≥ 2 was highest in the R-M only group (77%), followed by the CRISPR-Cas and R-M negative group (69%), the CRISPR-Cas and R-M positive group (68%), and the CRISPR-Cas only group (59%). In total, the lowest plasmid load was found in the strains with only CRISPR-Cas systems, judging by the total percentage of the classes over two plasmids. There was a significant difference between the plasmid content ≥4 between the group with only CRISPR-Cas- and the group with only R-M systems (p=0.0483).

For the NORM non-ESBL strain population, a total of 38 strains had only CRISPR-Cas systems, 82 with only R-M systems, 33 with both CRISPR-Cas- and R-M systems and 117 without any of these systems (Figure 29).

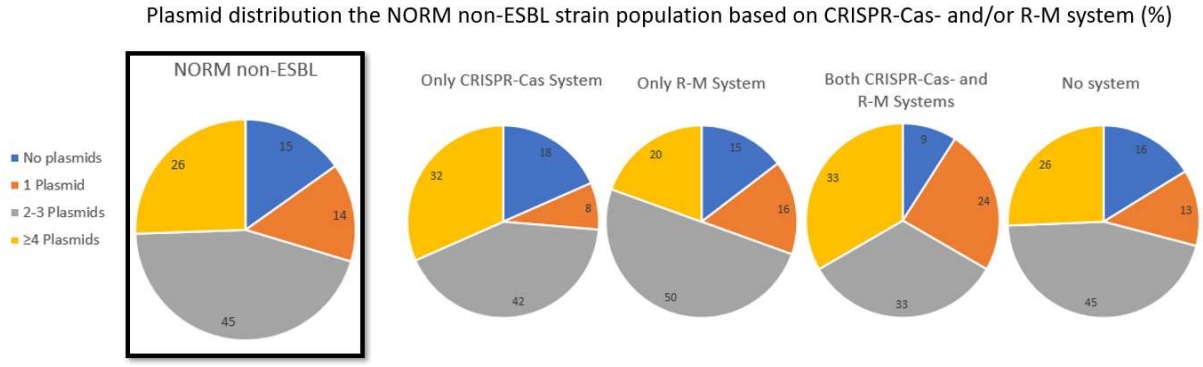


Figure 29: Percentage distribution of the plasmid categorisation for the NORM non- ESBL strain population (n=270), based on the total strain population (marked by the frame), presence of only CRISPR-Cas systems (n=38), only R-M systems (n=82), both CRISPR-Cas- and R-M systems (n=33) and the strains without CRISPR-Cas- and R-M systems (n=117).

The proportion of NORM non-ESBL strains with no plasmids varied between groups: CRISPR-Cas only (18%), R-M only (15%), CRISPR-Cas and R-M (9%), and finally CRISPR-Cas and R-M negative group (16%). The proportion of strains with a plasmid content ≥ 2 was highest in the CRISPR-Cas only group (74%), followed by CRISPR-Cas and R-M negative group (71%), the R-M only group (70%), and the CRISPR-Cas and R-M positive group (66%). The observed differences were not considered significant.

The NORM ESBLs displayed a total of 29 strains with Only CRISPR-Cas systems, 26 with only R-M systems, 12 with both CRISPR-Cas- and R-M systems and 77 without any systems (Figure 30).

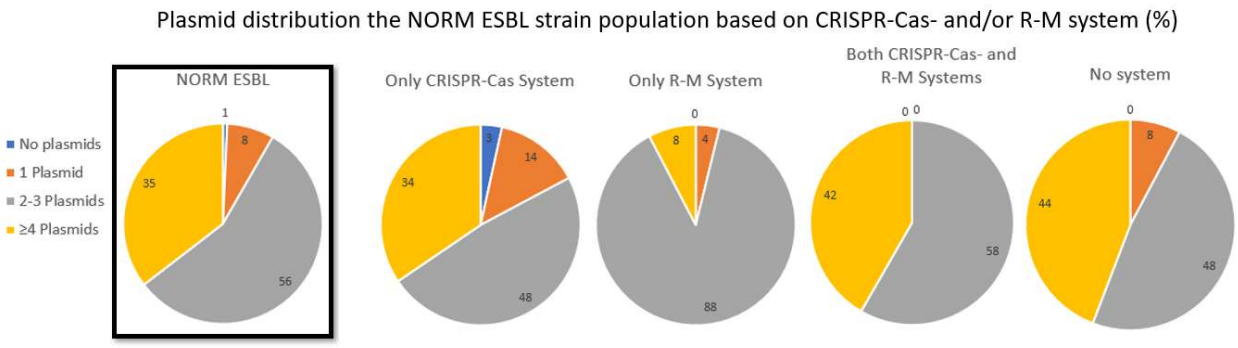


Figure 30: Percentage distribution of the plasmid categorisation for the NORM ESBL strain population (n=144), based on the total population (marked by the frame), the presence of only CRISPR-Cas systems (n=29), only R-M systems (n=26), both CRISPR-Cas- and R-M systems (n=12) and the strains without CRISPR-Cas- and R-M systems (n=77).

Almost all NORM-ESBL strains contained plasmids. The proportion of strains with a plasmid content ≥ 2 was highest in the CRISPR-Cas and R-M positive group (100%), followed by the R-M only group (96%), the CRISPR-Cas and R-M negative group (92%), and the CRISPR-Cas only group (82%). The difference in ≥4 plasmids was significant between the group with no systems and both systems (p=0000.1). Moreover, there was a significant difference in ≥4 plasmids for the only CRISPR-Cas group compared to the only R-M group (p=0.0279).

The ST307 strain population was the only group without CRISPR-Cas systems, and therefore the groups were divided into only R-M systems (n=91) and the ones without R-M systems (n=10) (Figure 31).

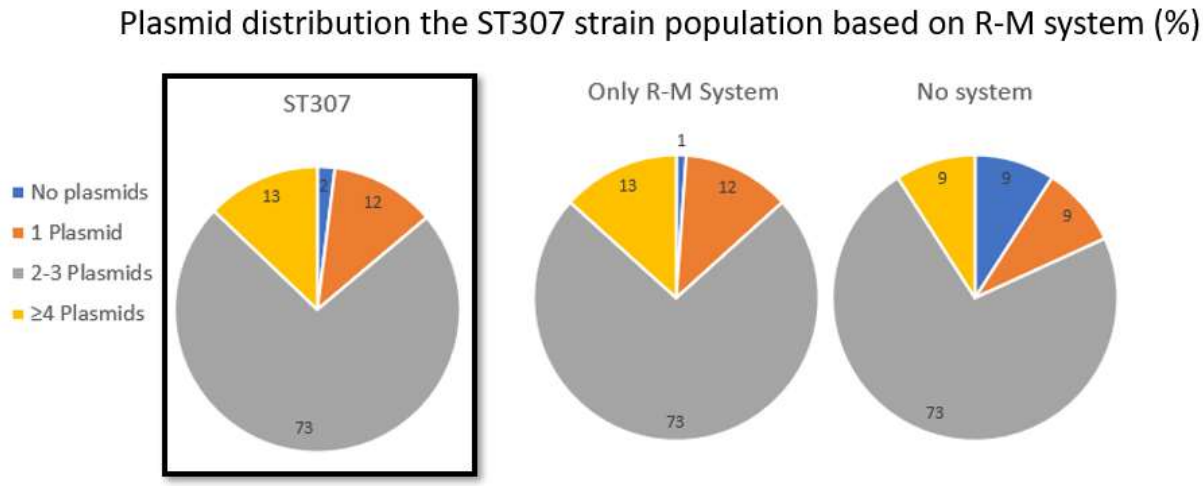


Figure 31: Percentage distribution of the plasmid categorisation for the ST307 strain population (n=101) based on the total population (marked by the frame), presence of only R-M systems (n=91) and the absence of R-M systems (No system) (n=10). All the ST307 strains were CRISPR-Cas negative.

The group with only R-M systems had a total of 1% without any plasmids in contrast to the no-system group (9%). The proportion of strains with ≥ 2 plasmids was similar in the R-M (86%) and the non-system group (82%). The number of non-system strains (n=10) was too low to consider any statistical analysis.

5.7 Sub analysis of dominant STs and high-risk STs

Features of the dominant STs, high-risk hypervirulent- and MDR STs found in the carrier strain population, are listed in Table 8. Briefly, the most common ST20 (n=15) is considered a global high-risk MDR clone. However, all carrier ST20 strains were either AMR negative or non-MDR. Only two strains had a virulence score 1. Moreover, sT26 (n=13) and ST35 (n=10), had an AMR negative profile and a total of two ST35 strains had acquired yersiniabactin. The plasmid content of all the dominant ST types was mostly from 2-3 plasmids and up, only one strain had only one plasmid (ST20). In terms of CRISPR-Cas systems, all ST20 strains were negative, all ST26 strains harboured Class 1 Type I-E and all ST35 contained Class 1 Type I-E*. R-M systems were seen for the ST20 and ST26, in 15 of 28 strains with Type I (n=9) and Type IIG (n=6).

Table 8: Summary of features seen for the dominant-, globally associated high-risk hypervirulent- and MDR strains found in the carrier strain population.

	ST type	CRISPR-Cas system	R-M system	Virulence score	AMR genes	MDR and/or ESBL	Number of plasmids
Dominant ST types (n=38)	**ST20*** (n=15)	Negative	Type I (n=2), Type IIG (n=6) and negative (n=7)	Negative (n=13) score 1 (n=2)	Negative (n=14) non-MDR (n=1)	Negative	One plasmid (n=1), 2-3 plasmids (n=3) and ≥4 plasmids (n=10)
	ST26 (n=13)	Class 1 Type I-E	Type I (n=7), negative (n=6)	Negative	Negative	Negative	2-3 plasmids (n=3) and ≥4 plasmids (n=10)
	ST35 (n=10)	Class 1 Type I-E*	Negative	Negative (n=8) and score 1 (n=2)	Negative	Negative	One plasmid (n=9) and ≥4 plasmids (n=1)
Globally associated high-risk hypervirulent strains (n=7)	**ST23** (n=1)	Class 1 Type I-E*	Negative	Score 5	Non-MDR	Negative	One plasmid
	ST375 (n=1)	Negative	Type IV	Score 3	Negative	Negative	No plasmids
	ST25 (n=5)	Negative	Type IV (n=4) Type IIG (n=1)	Score 1	Negative (n=1)	MDR (n=4)	No plasmid (n=1) and 2-3 plasmids (n=4)
Globally associated high-risk MDR strains (n=18)	**ST70** (n=3)	Negative	Type I (n=2) and negative (n=1)	Negative	Negative (n=2)	MDR (n=1)	≥4 plasmids
	ST17 (n=3)	Negative	Type I (n=1), Type III (n=1) and negative (n=1)	Negative (n=2) and score 1 (n=1)	Negative (n=2)	MDR (n=1)	2-3 plasmids
	ST29 (n=2)	Negative	Type I (n=1) and Type I+IIG (n=1)	Negative	Negative	Negative	2-3 or ≥4 plasmids
	ST11 (n=1)	Class 1 Type I-E	Type IIG	Score 1	Non-MDR	Negative	2-3 plasmids
	ST37 (n=9)	Negative	Negative (n=7) Type I (n=1), Type IV (n=1)	Negative (n=7) and score 1 (n=2)	Negative	Negative	No plasmids (n=2), one plasmid (n=2), 2-3 plasmids (n=3) and ≥4 plasmids (n=2)

*Also associated with globally spread high-risk MDR clones

High-risk MDR associated ST types were seen in 33 carrier strains; ST20 (n=15) (one of the dominant ST types), ST70 (n=3), ST17 (n=3), ST29 (n=2), ST11 (n=1) and ST37 (n=9). However, only a total of two strains were MDR, the rest were AMR negative (n=29) or non-

MDR (n=2). In terms of virulence factors, only five strains had acquired yersiniabactin, whereas the rest had no virulence factors. The plasmid content for these strains was diverse. The presence of CRISPR-Cas systems was only seen in ST11 (Class 1 Type I-E system). The presence of R-M systems was seen in 17 of the 33 strains.

High-risk hypervirulent ST types (n=7) were seen in seven carrier strains: ST23 (n=1), ST375 (n=1) and ST25 (n=5). However, virulence score 5 was only observed for the single ST23 strain. Moreover, four of the ST25 were also classified MDR. In terms of CRISPR-Cas systems, only ST23 were Class 1 Type I-E* positive and the rest were negative. However, R-M systems were found in six strains.

Actual MDR genotypes was observed in 12 carrier strains, all non-ESBL; ST70 (n=1), ST25 (n=4), ST17 (n=1), ST697 (n=1), ST3008 (n=1), (ST1693 (n=1) only 2 locus variants (LV) recognised), (ST499 (n=1) only one LV recognised), ST405 (n=1), and ST48 (n=1). Eleven strains were CRISPR-Cas negative, whereas one strain harboured CRISPR-Cas Class 1 Type I-E*. The distribution of R-M systems was: R-M negative (n=6), R-M Type IV (n=5, including co-occurrence), R-M Type I (n=2, including co-occurrence). A low virulence profile was observed in these strains: virulence score 0 (n= 6), virulence score 1 (n=6), plasmid content was diverse.

Features of the dominant STs, high-risk hypervirulent- and MDR strains STs found in the **NORM strain population**, are listed in Table 9. **Among the dominant ST types** (n=90), ST20 (n=17), ST37 (n=15), ST70 (n=14) and ST15 (n=13) are all ST types associated with high-risk MDR clones (n=59). Moreover, a total of 36 strains were classified as MDR and 34 were ESBL -producing strains. ST14 and ST45 in the dominant NORM STs displayed MDR (n=22) and ESBL (n=19). A total of 50 strains were virulence positive. 37 with virulence score 1, one strain with virulence score 3 and two strains with virulence score 4. A total 50 strains were CRISPR-Cas negative, 31 were CRISPR-Cas Class 1 Type I-E* positive and 8 CRISPR-Cas Class 1 Type I-E positive. R-M systems were found in most of the strains, 40 out of 59. See table 9.

Table 9: Summary of features seen for the dominant-, globally associated high-risk hypervirulent- and MDR strains found in the NORM strain population.

	ST type (number)	CRISPR-Cas systems	R-M systems	Virulence score	AMR genes	MDR and/or ESBL	Number of plasmids
Dominant ST types (n=90)	**ST14** (n=18)	Class 1 Type I-E*	Negative (n=12), Type I (n=5), Type III (n=1)	Negative (n=16), score 1(n=1) and score 3 (n=1)	Negative (n=1), non MDR (n=2)	MDR (n=15) ESBL (n=11)	No plasmid (n=1), one plasmid (n=6), 2-3 plasmids (n=4) or ≥4 plasmids (n=7)
	ST20* (n=17)	Negative	Type I (n=1), Type III (n=1), Type IIG (n=10) or Negative (n=6)	Negative (n=13) and score 1 (n=4)	Negative (n=14) and non-MDR (n=1)	MDR and ESBL (n=2)	1 plasmid (n=7), 2-3 plasmids (n=8) or ≥4 plasmids (n=2)
	ST37* (n=15)	Negative	Type III (n=1), Type IV (n=1) and Negative (n=13)	Negative (n=13) or score 1 (n=2)	Negative (n=6) and non-MDR (n=1)	MDR (n=8) and ESBL (n=7)	No plasmid (n=2), one plasmid (n=2), 2-3 plasmids (n=6) or ≥4 plasmids (n=5)
	ST70* (n=14)	Negative	Type I (n=11) or negative (n=3)	Negative (n=1) or score 1 (n=13)	-	MDR and ESBL (n=13)	2-3 plasmids (n=13) or ≥4 plasmids (n=1)
	ST15* (n=13)	Class 1 Type I-E*	Type I (n=6), Type IIG (n=2) or negative (n=7)	Negative (n=6), score 1 (n=5) and score 4 (n=2)	-	MDR (n=13) and ESBL (n=12)	2-3 plasmids (n=5) or ≥4 plasmids (n=8)
	ST45 (n=13)	Class 1 Type I-E (n=8) or negative (n=5)	Type III (n=1) or Negative (n=12)	Negative (n=1) or score 1 (n=12)	Negative (n=2) and non-MDR (n=4)	MDR (n=7) and ESBL (n=8)	No plasmid (n=1), one plasmid (n=1), 2-3 plasmids (n=4) or ≥4 plasmids (n=7)
Globally associated high-risk hypervirulent strains (n=8)	**ST23** (n=2)	Class 1 Type I-E*	Type III	Score 5	Negative	-	2-3 plasmids (n=1) or ≥4 plasmids (n=1)
	ST25 (n=6)	Negative	Type IV (n=5) or negative (n=1)	Negative (n=1) or score 1 (n=5)	-	MDR	2-3 plasmids
Globally associated high-risk MDR strains (n=13)	**ST11** (n=4)	Class 1 Type I-E (n=1) and negative (n=3)	Negative	Negative (n=2) or score 1 (n=2)	AMR negative (n=1)	MDR (n=3) and ESBL (n=2)	2-3 plasmids (n=3) and ≥4 plasmids (n=1)
	ST17 (n=6)	Negative	Negative	Negative (n=3) or score 1 (n=3)	Negative (n=2)	MDR (n=4) and ESBL (n=2)	No plasmid (n=1), 2-3 plasmids (n=3) and ≥4 plasmids (n=2)
	ST29 (n=3)	Negative	Type I (n=1) or negative (n=2)	Negative	Negative (n=1)	MDR (n=2)	One plasmid (n=1), ≥4 plasmids (n=2)

*Also associated with globally spread high-risk MDR clones

Additionally, globally associated **MDR high-risk ST strains**, ST11 (n=4), ST17 (n=6) and ST29 (n=3) were also seen. These strains were classified as MDR (n=9) with co-occurring ESBL -production (n=4). However, four strains were AMR negative. The virulence profile displayed virulence score 1 (n=5) and score 0 (n=8). All strains were CRISPR-Cas negative, except for one ST11 with CRISPR-Cas Class 1 Type I-E. Only one strain (ST29) had a R-M system (Type I). The plasmid content was diverse.

The global **high-risk hypervirulent** ST25 (n=6) and ST23 (n=2) were seen in the NORM strain collection. The ST25 strains had R-M system Type IV in five strains and no CRISPR-Cas systems. They displayed a low virulence profile with a virulence score 1 (n=5). The ST23 were CRISPR-Cas Class 1 Type I-E* positive and R-M Type III positive. Both ST23 strains had a virulence score 5 and were AMR negative.

Some STs were commonly found in both carrier- and NORM strain populations. In total, the NORM strain collection represented the same global **high-risk MDR ST types** found in the carrier strains: ST70, ST17, ST29 and ST37. In terms of the rest of the profiles, they had many similarities like the absence of CRISPR-Cas systems.

In addition to the globally MDR associated STs, there was also commonly associated **hypervirulent high-risk ST types** seen in both strain collections. The NORM strain collection harboured all the same known global high-risk hypervirulent strains, except for the ST375 which were only seen in the carrier strains. The ST25 (n=6 NORM) displayed the same profiles being CRISPR-Cas negative and R-M positive between the strain populations. However, they were classified MDR in the NORM strain collection. For all ST23 in both collections, they were CRISPR-Cas Class 1 Type I-E*, but the carrier strain was R-M negative, while the NORM strains were R-M Type III positive, they were all classified with a virulence score 5.

Other shared prevalent STs between carrier- and NORM strains have some general features in common. **ST14** strains (n=25) were also found in both the carrier (n=7) and NORM (n=18) populations. All ST14 (n=25) were CRISPR-Cas Class 1 Type I-E* positive. The R-M system profile was diverse; R-M type I (n=9), R-M Type III (n=1, NORM), R-M Type IIG (n=1, carrier) or R-M negative (n=14). ST 14 had a low virulence profile; virulence negative (n=23). All carrier strains were AMR negative (n=7) whereas fifteen NORM strains were classified as MDR and eleven as ESBL -producing. **ST35** (n=19) displayed an almost identical and profile for both the carrier (n=10) and NORM (n=9) strains. They had CRISPR-Cas systems

(n=18) and were R-M system negative. The differences amongst these strains were the NORM strains having 6 strains with virulence score 1, versus two in the carrier strains and one classified MDR in the NORM strain collection. **ST26** (n=19) were found in 13 carrier strains and 6 NORM strains. The ST26 had almost the same features, with CRISPR-Cas Class 1 Type I-E system and about half had presence of R-M systems (both populations). All had virulence score 0, AMR negative (n=18), one non -MDR, all MDR negative with varying plasmid content.

6 Discussion

The aims in this study were to investigate the distribution of CRISPR-Cas- and R-M systems in different *K. pneumoniae* strain populations and their association to AMR -classification, virulence score and plasmid content. To achieve this, a large collection of human KpSC strains, previously characterized by WGS, was selected and examined thoroughly by various bioinformatic tools and statistical analyses.

The strain collection consisted of faecal carrier strains from adults in the Tromsø municipality, clinical isolates (blood and urine) collected through the national NORM system and NORKAB study, as well as an international collection of ST307 strains. The ST307 collection was selected as an interesting emerging high-risk MDR clone in Norway and abroad (1)(79)(75). The strain collection is quite large compared to previous studies of CRISPR-Cas/R-M systems distributions in *K. pneumoniae* (11)(51)(90)(91). The strain collection is also relevant and representative for the genetic diversity of *K. pneumoniae* strains circulating in the Norwegian human population. Population structure analysis of the carrier, NORM-ESBL and non-ESBL as well as ST307 revealed a total of 503 different ST types. A total of nine different ST types associated with high-risk clones were seen in the carrier and NORM collections. The ST307 also displayed diversity within the population and the strains had a variety of nationality. The plasmid content indicated a strong presence of INFIB(K)_1_Kpn3 (n=725), IncFII_1_pKp91 (n=548) and ColRNAI_1 (n=250) in all strain populations. Additionally, some plasmids, like Col(MGD2)_1 (n=35) was seen more frequently in the carrier- and NORM strain collections compared to the ST307. Whereas IncFiB(pQil)-1_pQil (n=24) was predominantly seen for the NORM ESBL- and ST307 strains. However, there was a presence of 17% potential false positive nucleotide matches for the plasmid detection because of a mistake in the bioinformatic analysis.

Bioinformatic identification of structurally complete CRISPR-Cas- and R-M systems was performed by screening the strains for CRISPR-Cas systems using CRISPRCasFinder and

manual assessment of the SimpleSynteny results. R-M system detection was performed using HMMer profiles. Only structural complete systems could possibly preform a biologic function, and therefore a lot of effort was put into this step. In addition to determining the presence, the CRISPR-Cas systems were classified using in-house profiles to determine their subtypes within the Class 1 Type I systems. Correspondingly, the R-M systems was also assessed by their types and subfamilies of MTases and REases to make sure they had potential to perform biological functions. It is fair to say that more programs were tested then what was found to be useful in producing results. In addition, many of the programs needed additional in-house profiles, adjustments and manual evaluation. The manual assessment led to discoveries in terms of potential mechanisms for CRISPR-Cas systems riddance. The results from assessing CRISPR-Cas- and R-M systems will be further elaborated.

Previous studies have documented a strong link between defined STs and their presence/absence of CRISP-Cas systems (11)(59)(90)(92). A similar strong association between defined STs and the presence or absence of CRISPR-Cas- and/or R-M systems was observed in this study for ST14, ST15, ST23 and ST307. The consistent relationship between defined STs and their CRISPR-Cas-/R-M systems content supports the quality of the bioinformatic work in detection of structural complete systems performed in this study.

6.1 CRISPR-Cas systems: distribution and correlation to virulence score, AMR -classification and plasmid content

Structurally complete CRISPR-Cas systems was found in 26% of the strains: carrier strains (30%), NORM non-ESBL strains (26%) and NORM ESBL strains (29%). Thus, the CRISPR-Cas system seems to be equally distributed between faecal carrier and clinical strains. In contrast, a total absence of both structurally complete systems and fragmented CRISPR-Cas elements was observed in the ST307 population.

A study in 2018, displayed a prevalence of 31% of CRISPR-Cas systems in clinical *K. pneumoniae* strains (n=176)(11). In the same study, they also documented the prevalence in publicly available *K. pneumoniae* genomes to be 41% (11). One other study analysing the prevalence in publicly available databases has documented the prevalence to be 54% in 68 *Klebsiella* genomes (51). Variations in the observed prevalence of CRISPR-Cas systems could be due to differences in strains collections and methods or criteria for systems identification.

Class 1 systems are the most commonly observed in *K. pneumoniae* (11)(51)(59). The subtype IV, I-E and the recent Type I-E* has been frequently seen in *K. pneumoniae* (11)(51)(59)(60). The study looking into both clinical and publicly available *K. pneumoniae* strains in 2018, displayed a total of 28% CRISPR-Cas Class 1 Type I-E (n=15) and 72% presence of Type I-E* (n=39) in the clinical strains, respectively (11). The distribution of Type I-E and I-E* systems were different in this study. The reason for this could be connected to the overrepresentation of ST23 (n=15) and ST15 (n=9), associated with Type I-E* in their study (11)(90). However, for the publicly available *K. pneumoniae* genomes (n=97) a total of 58% were classified Type I-E and 43% Type I-E* (11). Similar distributions of CRISPR-Cas subsystems were seen in this study. In addition, a total of 1% (n=7) strains displaying co-occurrence of CRISPR-Cas Class 1 Type I-E and I-E*. Co-occurrence of I-E and I-E* systems have to our knowledge not been previously reported and will be commented later.

Some CRISPR-Cas subtypes have been assumed to be associated with certain **MLST types** (11)(90). A study of carbapenem-resistant *K. pneumoniae* revealed that ST11, ST45 and ST147 were associated with Type I-E and ST23, ST15, ST11, ST65 and ST685 with the Type I-E* (90). The results in this study also connected ST23 and ST15 to Type I-E* and ST11 and ST45 to Type I-E. Observations in this study are in line with previous reports showing ST-specific patterns in the distribution of CRISPR-Cas systems including subtypes.

6.1.1 Differences in virulence score, AMR-classification and plasmid content based on the presence and absence of CRISPR-Cas systems

Comparison of all strains based on their CRISPR-Cas profile grouped by the absence of CRISPR-Cas (n=743), Type I-E positive (n=140), Type I-E* positive (n=109) and CRISPR-Cas Class 1 Type I-E and I-E* positive (n=7), revealed groups specific features (11)(90). Overall, the strains without CRISPR-Cas systems seemed to have a lot in common with primarily the single subtypes I-E and I-E* groups. The strains with both subtypes displayed the most distant relationship to the strains with absence of CRISPR-Cas in the PCA analysis.

CRISPR-Cas Class 1 Type I-E* strains have been associated with a lower susceptibility to ampicillin-sulbactam, cefazolin, cefuroxime and gentamicin, lower phage- an plasmid content compared to the I-E positive strains of *K. pneumoniae* (11). Resistance to the four classes of antibiotics would result in MDR classification in this study. Another study looking into 16 virulence positive carbapenem-resistant *K. pneumoniae*, found the Type I-E* system to be associated with a higher virulence score compared to the strains with Type I-E or absence of

CRISPR-Cas (90). In this study, Type I-E* positive strains did not display significant higher prevalence of acquired virulence factors. Moreover, the prevalence of MDR was significant different between the Type I-E*compared to the absence of CRISPR-Cas systems. This was not consistent with previous studies (11)(90).

There was a significant difference in the number of AMR negative strains between the strains without CRISPR-Cas and the presence of Type I-E*. **CRISPR-Cas Type I-E strains** displayed a higher susceptibility towards antimicrobial agents and a slightly increased virulence score possibly associated with the presence of ST11 and ST45. A study found ST45 to be associated with CRISPR-Cas Class 1 Type I-E systems (11).

A total of seven strains were positive for **both CRISPR-Cas Class 1 Type I-E and I-E*** subtypes. This has previously not been documented for *K. pneumoniae*, possibly because of limitations in the strain populations investigated. The biological function of harbouring both subtypes is today unknown and further studies are needed. The results in this study indicated the presence of features separating them from the other subtype-groups in the PCA analysis. The most interesting feature was all being AMR negative.

Almost all **CRISPR-Cas positive systems** displayed a higher presence of acquired virulence factors. The combination of both systems displayed about the same profile compared to the CRISPR-Cas negative strains. The difference between virulence positive strains in Type I-E and I-E* was not significant. All the ST types in this study associated with high-risk hypervirulence harboured the I-E* subtype. And 77% of the strains with presence of virulence factors and CRISPR-Cas systems displayed presence of the I-E* subtype.

6.2 Restriction- Modification systems: distribution and correlation to virulence score, AMR -classification and plasmid content

R-M systems were found in 48% of the strains varying across populations; carrier (43%), NORM (44%) and ST307 (90%). Thus, the presence of R-M ysstems seems to be equally distributed between faecal carrier and clinical strains. There was significant difference between the prevalence of R-M systems and subtypes in the NORM ESBL- and ST307 collections. This may well be due to ST- specific distributions of R-M systems and subtypes. Previous large studies have indicated an average of 2.1-2.6 R-M systems per prokaryote genome (63)(66). However, none of the studies focused on R-M systems in *K. pneumoniae* and both stated that the distribution could vary amongst bacterial genera and species (63)(66).

To our knowledge, there are no studies looking directly into the prevalence and function of R-M systems in *K. pneumoniae.* However, the occurrence of R-M systems are found to be affected by the presence of MGEs, CRISPR-Cas systems, integrons and natural transformation (66). Previous studies have displayed rapid acquisition and loss of R-M primarily through HGT (63). In addition, some solitary R-M associated genes are transferred autonomously in small MEGs (66).

The previous study by Oliveira et al. comprised 2261 prokaryotic genomes and stated the **prevalence of R-M Types** were the following: Type II (42%), Type I (30%), Type IV (29%) and Type III (8%) (66). The Type IIG is considered a subtype of Type III and is often not included in studies as they usually look into the most common types (63)(66). One interesting observation is that the Type I and Type IIG (total of 80% of R-M) are assumed to be almost impossible to avoid (61). Mechanisms like removal of recognition sites or expressions of R-M system inhibiting proteins, like Orc proteins, is rarely seen as a mechanism to avoid Type I systems (61). Type I is also connected to affect expression of certain genes which could give the host advantages, however these mechanisms are poorly understood (66). These features could be potentially useful for bacteria in environments with high selection pressures and in the presence of MGEs, but in need of limiting further acquisition of MGE affecting bacterial fitness, like the ST307. The high prevalence of Type I could potentially be explained by R-M systems ability to propagate themselves selfishly through HGT and their ability to select MGEs containing fragments of the present R-M Type (66). However, the specific functions of the different R-M Types in *K. pneumoniae* remains to be thoroughly investigated.

Co-occurrence of R-M systems was observed in 12% for the R-M positive strains (n=488). The most common co-occurrence was Type I and Type IV (n=34). This co-occurrence was also significant in the Oliveira study (66). In addition to R-M system co- occurrence, there was also a significant association between R-M- and prevalence of CRISPR-Cas systems (66). CRISPR-Cas related spacers do seldom target R-M systems (66). In this study, a total of 24% of the R-M positive strains displayed a co-occurring CRISPR-Cas system.

6.2.1 Differences in virulence score, AMR-classification and plasmid content based on the presence and absence of R-M systems

Group analysis of strains with or without R-M systems indicated a large overlap between the two groups. Because of the large complexity of co-occurring systems, it was not considered beneficial to represent the individual systems and co-occurrences in a PCA plot. The R-M

positive strains were grouped into single- (one R-M system) (n=398) and multiple systems (more than one type of R-M system) (n=60). The PCA plot indicated overall differences between absence, single and multiple R-M systems.

The group with **absence of R-M systems** displayed the highest **virulence score** and the lowest plasmid content. There was a significant difference in virulence negative strains between the R-M negative strains and the single R-M systems. A closer look at the virulence positive strains harbouring R-M systems revealed a prevalence of 55% R-M Type I and 26% R-M Type IV. These R-M types were also seen for some of the global MDR STs and the Type IV at the hypervirulent ST types. Presence of **AMR -genes** was found to be highest in the strains with single R-M systems. There was a significant difference in AMR negative strains between R-M negative strains and single R-M systems. One thing to note is that this population includes 90% of the ST307 MDR strains. Moreover, the R-M Type I was seen for 81% of the R-M positive strains with MDR. R-M Type I system was also seen in some of the global high-risk MDR STs.

The **plasmid profiles** displayed the highest plasmid content in the R-M positive strains. Interestingly, the strains with multiple R-M systems displayed the highest plasmid loads. They also displayed a significant difference in plasmid content ≥2 plasmids. This could potentially be because of the R-M systems stabilising the MGEs in the DNA with a high potentially because of R-M Type specific fragments in the plasmid (61)(63)(66).

6.3 Co-occurrence of CRISPR-Cas- and R-M systems

In total, 40% of the strains displayed absence of both systems, 12% contained both CRISPR-Cas- and R-M systems, 14% showed only presence of CRISPR-Cas- and 34% R-M systems only. PCA analysis of the groups indicated similarities between the strains harbouring both CRISPR-Cas- and R-M systems compared to the strains with only CRISPR-Cas systems. These two groups had the most distant relationships to the strains with only R-M systems. The strains with complete absence of both CRISPR-Cas- and R-M systems had the average point of the population in between these systems.

Absence of CRISPR-Cas- and R-M systems represented the situation unaffected by CRISPR-Cas- and R-M systems in the different strains. They had the second lowest virulence score by 1%. Looking at the resistance profile, this group had the second highest presence of MDR and ESBLs. There was no significant difference in the plasmid content ≥2 for this

group compared to the strains with only R-M systems. None of the dominant or high-risk clonal strains were included in this category.

The strains only harbouring R-M systems previously displayed a significant difference in virulence negative strains compared to the R-M negative group. In this context, they also had the highest prevalence of virulence negative strains. This indicated that R-M systems had the highest association with virulence negative strains. Previous we also saw that there was a significant difference in AMR negative strains between the strains with no R-M systems and the ones with single R-M systems, indicating a potential association between AMR genes and R-M systems. Moreover, the analysis of only R-M systems also displayed a significant difference in harbouring ≥4 plasmids between the strains with absence of R-M systems and presence of multiple R-M systems. Here there was also a significant difference in carrying ≥4 plasmids between the strains with R-M systems only and the strains with no system. In addition, there was a significant difference in carrying ≥4 plasmids between the strains only harbouring R-M systems and the strains with only CRISPR-Cas systems in the NORM ESBL strains. These observations suggests a significant association between presence of R-M system and higher plasmid load.

The strains only harbouring CRISPR-Cas systems features displayed the highest percentage of strains with acquired virulence. Previously there was not proven a significance between the CRISPR-Cas Type I-E and I-E* with the highest numbers of virulence factors. However, the difference in virulence negative strains between strains with only CRISPR-Cas systems and the group without any system was significant. This observation suggests an association between the presence of CRISPR-Cas systems and a higher virulence score. The associations between a lower MDR profile for the group with absence of CRISPR-Cas compared to the CRISPR-Cas Type I-E* (second highest MDR profile) were significant. Here, they displayed the second lowest MDR profile, but there was no significance in AMR negative strains between the strains with only CRISPR-Cas systems and both CRIPSR-Cas- and R-M systems. In terms of plasmid content there was a significant difference in carrying ≥2 plasmids between the group with only CRISPR-Cas- and only R-M systems. Moreover, there was a significant difference between plasmid content ≥4 plasmids between the CRISPR-Cas systems and the R-M systems for the NORM ESBL strains. These observations underline a restrictive effect of CRISPR-Cas systems in terms of AMR genes and plasmid acquisition, but a possible association with higher virulence.

Analysis of the group with **both CRISPR-Cas- and R-M systems** indicated a significant difference between in virulence positive strains between the strains with only CRISPR-Cas- and the strains with both CRISPR-Cas- and R-M systems. The previously displayed significance in R-M systems restricting virulence factors seems to be in interplay with the CRISPR-Cas systems significant association with higher acquisition of virulence factors. There was also a significant difference in the presence of virulence factors between the strains only harbouring CRISPR-Cas systems and the strains with both CRISPR-Cas- and R-M systems. The AMR profile for these strains had the highest presence of AMR negative strains amongst all groups. There was no significant difference in the prevalence of AMR negative strains between the group with only CRISPR-Cas systems and the strains with both systems. There was a significant difference in AMR negative strains between the group with both systems and no systems. This indicated a restricting effect of CRISPR-Cas systems affecting the association of R-M systems to acquire AMR. Previously we saw the association of R-M systems and their higher plasmid content with a significant difference compared to the strains with only CRISPR-Cas systems. There was a significant lower prevalence in carriage of ≥4 plasmids in the strains with both systems compared to the strains with only R-M systems, indicating the higher effect of CRISPR-Cas systems. This was supported by the absence of significant differences in carrying ≥4 plasmids between the group with both systems and the group with only CRISPR-Cas systems.

6.4 Strengths and limitations

This study has an overall strength because of the unique and large strain collection. The strain collections display diversity both within and between themselves. The carrier strains displayed common ST types as found in the NORM strain collection, representing the opportunistic pathogen side of *K. pneumoniae.* Moreover, there high-risk associated ST307 strains originated from Norway and additional countries with significant genetic diversity.

The thorough assessment of the CRISPR-Cas- and R-M systems, allowed a certainty of the best prediction of potential functional systems. However, the complete systems can still be inactive because of self-targeting sequences (90).

The weaknesses in this study are primarily represented the time limitations of a master's degree. Learning bioinformatics and analysing a total of 999 strains sadly sets limits for how comprehensive the study can be. To make good conclusions and elaborate the potential relationships indicated in this study, it would have been useful to look into the spacers

specificity in the CRISPR-Cas systems, the phage content and do more advanced statistical analyses. Additionally, the PCA plots could have provided more information about the relationships of the results if they were not classified, but took in consideration the actual typing of plasmids, not just the categorised amounts. A potential mistake in plasmid detection procedure was detected late in this study. The homology cut off in the analysis was lacking because of word converting two double "-" into one long, resulting in exclusion in the command line and no call back on the flag. This resulted in a total of ~17% of the plasmids hits in the study being below 80% homology and potential false positive results. The distribution of this mistake is divided 9.2% in the carrier-, 7.2% in the NORM- and 0.4% in the ST307 strain collections.

7 Conclusions and future perspectives

- The presence of CRISPR-Cas- and R-M systems seems to be prevalent and equally distributed in faecal carrier and clinical strains.
- The content and distribution of CRISPR-Cas- and R-M systems, including subtypes, seems to have ST specific associations.
- There was also some significant correlations between the presence and absence of CRISPR-Cas- and R-M systems in terms of MGE acquisition, reflected in virulence factors, AMR genes and plasmid content. These associations were seen across the strain collections.
- CRISPR-Cas systems were associated with a higher virulence profile, a lower AMR and plasmid load.
- The R-M systems were associated with lower virulence score, a higher AMR and plasmid load.
- Future studies should include analysis of CRISPR spacer specificity, overall phage content and utilise more advanced comparisons and statistical analyses.

8 Citations

1. Wyres KL, Lam MMC, Holt KE. Population genomics of Klebsiella pneumoniae. Nat Rev Microbiol [Internet]. 2020; Available from: http://dx.doi.org/10.1038/s41579-019-0315-1

2. Lee CR, Lee JH, Park KS, Jeon JH, Kim YB, Cha CJ, et al. Antimicrobial resistance of hypervirulent Klebsiella pneumoniae: Epidemiology, hypervirulence-associated determinants, and resistance mechanisms. Front Cell Infect Microbiol. 2017;7(NOV).

3. Wyres KL, Holt KE. Klebsiella pneumoniae as a key trafficker of drug resistance genes from environmental to clinically important bacteria. Curr Opin Microbiol [Internet]. 2018;45:131–9. Available from: https://doi.org/10.1016/j.mib.2018.04.004

4. Rettedal S, Löhr IH, Natås O, Giske CG, Sundsfjord A, Øymar K. First outbreak of extended-spectrum β-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):612–21.

5. Dalgleish T, Williams JMG., Golden A-MJ, Perkins N, Barrett LF, Barnard PJ, et al. Klebsiella pneumoniae- En nasjonal studie ac sykdomsbyrde, populasjonsstruktur, resistensutvikling og virulens hos en stadig viktigere humanpatogen. 2007. p. 1–10.

6. Brolund A. Overview of ESBL-producing Enterobacteriaceae from a Nordic perspective. Infect Ecol Epidemiol. 2014;4(1).

7. Tacconelli E et al. Global Priority List of Antibiotic-Resistant Bacteria to Guide Research, Discovery, and Development of New Antibiotics. World Health Organisation Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics. 2017.

8. Wyres KL, Wick RR, Judd LM, Froumine R, Tokolyi A, Gorrie CL, et al. Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of Klebsiella pneumoniae. PLoS Genet. 2019;15(4):e1008114.

9. Koonin E V., Makarova KS, Zhang F. Diversity, classification and evolution of CRISPR-Cas systems. Curr Opin Microbiol [Internet]. 2017;37:67–78. Available from: http://dx.doi.org/10.1016/j.mib.2017.05.008

10. Shen J, Li L, Wang X, Xiu Z, Chen G. Comparative analysis of CRISPR-Cas systems in Klebsiella genomes. J Basic Microbiol. 2017;57:325–36.

11. Li HY, Kao CY, Lin WH, Zheng PX, Yan JJ, Wang MC, et al. Characterization of CRISPR-Cas systems in clinical Klebsiella pneumoniae isolates uncovers its potential association with antibiotic susceptibility. Front Microbiol. 2018;9(JUL):1–9.

12. Adeolu M, Alnajar S, Naushad S, Gupta RS. Genome-based phylogeny and taxonomy of the 'Enterobacteriales': Proposal for enterobacterales ord. nov. divided into the families Enterobacteriaceae, Erwiniaceae fam. nov., Pectobacteriaceae fam. nov., Yersiniaceae fam. nov., Hafniaceae fam. nov., Morgane. Int J Syst Evol Microbiol. 2016;66(12):5575–99.

13. Barbier E, Rodrigues C, Depret G, Passet V, Gal L, Piveteau P, et al. The ZKIR Assay, a Real-Time PCR Method for the Detection of Klebsiella pneumoniae and Closely Related Species in Environmental Samples. Appl Environ Microbiol. 2020;86(7):1–15.

14. Ludden C, Moradigaravand D, Jamrozy D, Gouliouris T, Blane B, Naydenova P, et al. A One Health Study of the Genetic Relatedness of Klebsiella pneumoniae and Their Mobile Elements in the East of England. Clin Infect Dis. 2019;(Xx):1–8.

15. Laxminarayan R, Duse A, Wattal C, Zaidi AKM, Wertheim HFL, Sumpradit N, et al. Antibiotic resistance-the need for global solutions. Lancet Infect Dis. 2013;13(12):1057–98.

16. Gorrie CL, Mirc Eta M, Wick RR, Edwards DJ, Thomson NR, Strugnell RA, et al. Gastrointestinal Carriage Is a Major Reservoir of Klebsiella pneumoniae Infection in Intensive Care Patients. Clin Infect Dis. 2017;65(2):208–15.

17. Lin YT, Siu LK, Lin JC, Chen TL, Tseng CP, Yeh KM, et al. Seroepidemiology of Klebsiella pneumoniae colonizing the intestinal tract of healthy chinese and overseas chinese adults in Asian countries. BMC Microbiol. 2012;12.

18. Raffelsberger N, Andrea M, Hetland K, Andreassen L, Gravningen K, Löhr H, et al. P1772 Human gut carriage of Klebsiella pneumoniae in an adult community population-ECCMID poster. 2019.

19. Blin C, Passet V, Touchon M, Rocha EPC. Metabolic diversity of the emerging pathogenic lineages of Klebsiella pneumoniae. Environ Microbiol. 2017;19(5):1881–98.

20. Leadbetter ER. Microbiology, an Evolving Science [Internet]. 4th editio. Twitchell B, editor. Vol. 4, Microbe Magazine. London, England: Norton & Company; 2009. 236–354 p. Available from: hhps://lccn.loc.gov/2016051604

21. Diancourt L, Passet V, Verhoef J, Grimont PAD, Brisse S. Multilocus sequence typing of Klebsiella pneumoniae nosocomial isolates. J Clin Microbiol. 2005;43(8):4178–82.

22. Bialek-Davenet S, Criscuolo A, Ailloud F, Passet V, Jones L, Delannoy-Vieillard AS, et al. Genomic definition of hypervirulent and multidrug-resistant klebsiella pneumoniae clonal groups. Emerg Infect Dis. 2014;20(11):1812–20.

23. Dupuis MÈ, Villion M, Magadán AH, Moineau S. CRISPR-Cas and restriction-modification systems are compatible and increase phage resistance. Nat Commun. 2013;4(May):1–7.

24. Giske CG, Sundsfjord AS, Kahlmeter G, Woodford N, Nordmann P, Paterson DL, et al. Redefining extended-spectrum β-lactamases: Balancing science and clinical need. J Antimicrob Chemother. 2009;63(1):1–4.

25. Fostervold A, Raffelsberger N, Andrea M, Hetland K, Bernhoff E, Sundsfjord A, et al. Klebsiella pneumoniae ST107 : an emerging invasive clone in Norway -ECCMID poster. Vol. 29 th ECCM. 2019.

26. Villa L, Feudi C, Fortini D, Brisse S, Passet V, Bonura C, et al. Diversity, virulence, and antimicrobial resistance of the KPCproducing klebsiella pneumoniae ST307 clone. Microb Genomics. 2017;3(4).

27. García-Martínez J, Maldonado RD, Guzmán NM, Mojica FJM. The CRISPR conundrum: Evolve and maybe die, or survive and risk stagnation. Microb Cell. 2018;5(6):262–8.

28. Rossolini GM, D'Andrea MM, Mugnaioli C. The spread of CTX-M-type extended-spectrum β-lactamases. Clinical Microbiology and Infection. 2008.

29. Rezazadeh M, Baghchesaraei H, Peymani A. Plasmid-Mediated Quinolone-Resistance (qnr) Genes in Clinical Isolates of Escherichia coli Collected from Several Hospitals of Qazvin and Zanjan Provinces, Iran. Osong Public Heal Res Perspect [Internet]. 2016;7(5):307–12. Available from: http://dx.doi.org/10.1016/j.phrp.2016.08.003

30. Peirano G, Pitout JDD. Extended-Spectrum β-Lactamase-Producing Enterobacteriaceae: Update on Molecular Epidemiology and Treatment Options. Drugs. 2019;79(14):1529–41.

31. Samuelsen Ø. ESBL-A ESBL-M -Høstkonferansen. Nasjonal kompetansetjeneste for påvisning av antibiotikaresistens (K-res); 2018.

32. Hansen DS, Schumacher H, Hansen F, Stegger M, Hertz FB, Schønning K, et al. Extended-spectrum β-lactamase (ESBL) in Danish clinical isolates of Escherichia coli and Klebsiella pneumoniae: Prevalence, β-lactamase distribution, phylogroups, and co-resistance. Scand J Infect Dis. 2012;44(3):174–81.

33. Magiorakos AP, Srinivasan A, Carey RB, Carmeli Y, Falagas ME, Giske CG, et al. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: An international expert proposal for interim standard definitions for acquired resistance. Clin Microbiol Infect [Internet]. 2012;18(3):268–81. Available from: http://dx.doi.org/10.1111/j.1469-0691.2011.03570.x

34. Naseer U, Sundsfjord A. The CTX-M Conundrum: Dissemination of Plasmids and *Escherichia coli* Clones. Microb Drug Resist [Internet]. 2011;17(1):83–97. Available from: http://www.liebertonline.com/doi/abs/10.1089/mdr.2010.0132

35. Shon AS, Bajwa RPS, Russo TA. Hypervirulent (hypermucoviscous) Klebsiella Pneumoniae: A new and dangerous breed. Virulence. 2013;4(2):107–18.

36. Bush K, Fisher JF. Epidemiological Expansion, Structural Studies, and Clinical Challenges of New β-Lactamases from Gram-Negative Bacteria. Annu Rev Microbiol [Internet]. 2011;65(1):455–78. Available from: http://www.annualreviews.org/doi/10.1146/annurev-micro-090110-102911

37. Gu D, Dong N, Zheng Z, Lin D, Huang M, Wang L, et al. A fatal outbreak of ST11 carbapenem-resistant hypervirulent Klebsiella pneumoniae in a Chinese hospital: a

molecular epidemiological study. Lancet Infect Dis [Internet]. 2018;18(1):37–46. Available from: http://dx.doi.org/10.1016/S1473-3099(17)30489-9

38. Paczosa MK, Mescas J. Klebsiella pneumoniae : Going on the Offense with a Strong Defense. Microbiol Mol Biol Rev. 2016;80(3):629–61.

39. Follador R, Heinz E, Wyres KL, Ellington MJ, Kowarik M, Holt KE, et al. The diversity of Klebsiella pneumoniae surface polysaccharides. Microb genomics. 2016;2(8):e000073.

40. Gomez-Simmonds A, Uhlemann AC. Clinical implications of genomic adaptation and evolution of carbapenem-resistant klebsiella pneumoniae. J Infect Dis. 2017;215(Suppl 1):S18–27.

41. Lam MMC, Wick RR, Wyres KL, Gorrie CL, Judd LM, Jenney AWJ, et al. Genetic diversity, mobilisation and spread of the yersiniabactin-encoding mobile element ICEKp in Klebsiella pneumoniae populations. Microb genomics. 2018;4(9).

42. Löhr IH, Rettedal S, Natås OB, Naseer U, Øymar K, Sundsfjord A. Long-term faecal carriage in infants and intra-household transmission of CTX-M-15-producing Klebsiella pneumoniae following a nosocomial outbreak. J Antimicrob Chemother. 2013;68(5):1043–8.

43. Martin RM, Cao J, Brisse S, Passet V, Wu W, Zhao L, et al. Molecular Epidemiology of Colonizing and Infecting Isolates of Klebsiella. Clin Sci Epidemiol. 2016;1(5):1–12.

44. Thomas CM, Nielsen KM. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol. 2005;3(9):711–21.

45. Kamruzzaman M, Iredell JR. CRISPR-Cas System in Antibiotic Resistance Plasmids in Klebsiella pneumoniae. Front Microbiol. 2020;10(January).

46. De la Cruz F, Davies J. Horizontal gene transfer and the origin of species: Lessons from bacteria. Trends in Microbiology. 2000.

47. McDonald ND, Regmi A, Morreale DP, Borowski JD, Fidelma Boyd E. CRISPR-Cas systems are present predominantly on mobile genetic elements in Vibrio species. BMC Genomics. 2019;20(1):1–23.

48. Ramirez MS, Iriarte A, Reyes-Lamothe R, Sherratt DJ, Tolmasky ME. Small Klebsiella pneumoniae Plasmids: Neglected Contributors to Antibiotic Resistance. Front Microbiol. 2019;10(September):1–14.

49. Carattoli A, Zankari E, Garciá-Fernández A, Larsen MV, Lund O, Villa L, et al. In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother. 2014;58(7):3895–903.

50. Marraffini LA, Sontheimer EJ. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat Rev Genet. 2010;11(March):181–90.

51. Shen J, Lv L, Wang X, Xiu Z, Chen G. Comparative analysis of CRISPR-Cas systems in Klebsiella genomes. J Basic Microbiol. 2017;57(4):325–36.

52. Faure G, Shmakov SA, Yan WX, Cheng DR, Scott DA, Peters JE, et al. CRISPR–Cas in mobile genetic elements: counter-defence and beyond. Nat Rev Microbiol [Internet]. 2019;17(8):513–25. Available from: http://dx.doi.org/10.1038/s41579-019-0204-7

53. Krebs , Jocelyn E Goldstein, Elliott S Kilpatrick ST. Lewin`s Genes XII. Kane, Matthew Schwinn, Audrey Hoffman NHN, editor. Burlington, Massachusetts; 2018. 769–780 p.

54. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, et al. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol. 2015;13(11):722–36.

55. Gholizadeh P, Aghazadeh M, Asgharzadeh M, Kafil HS. Suppressing the CRISPR/Cas adaptive immune system in bacterial infections. Eur J Clin Microbiol Infect Dis. 2017;36(11):2043–51.

56. Makarova KS, Wolf YI, Iranzo J, Shmakov SA, Alkhnbashi OS, Brouns SJJ, et al. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. Nat Rev Microbiol. 2019;

57. Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P, et al. Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol. 2011;9(6):467–77.

58. Makarova KS, Wolf YI, Koonin E V. Classification and Nomenclature of CRISPR-Cas Systems: Where from Here? Cris J. 2018;1(5):325–36.

59. Mackow NA, Shen J, Adnan M, Khan AS, Fries BC, Diago-Navarro E. CRISPR-Cas influences the acquisition of antibiotic resistance in Klebsiella pneumoniae. PLoS ONE2 [Internet]. 2019;14(11):13. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0225131

60. Newire E, Aydin A, Juma S, Enne VI, Roberts AP, Free R, et al. Identification of a Type IV CRISPR-Cas system located exclusively on IncHI1B/IncFIB plasmids in Enterobacteriaceae. 2019.

61. Rusinov IS, Ershova AS, Karyagina AS, Spirin SA, Alexeevski A V. Avoidance of recognition sites of restriction-modification systems is a widespread but not universal anti-restriction strategy of prokaryotic viruses. BMC Genomics. 2018;19(1):1–11.

62. Wilson GG, Murray NE. Restriction and modification systems. Annu Rev Genet. 1991;25:585–627.

63. Ershova AS, Rusinov IS, Spirin SA, Karyagina AS, Alexeevski A V. Role of restriction-modification systems in prokaryotic evolution and ecology. Biochem. 2015;80(10):1373–86.

64. Pleška M, Guet CC. Effects of mutations in phage restriction sites during escape from restriction–modification. Biol Lett. 2017;13(12):10–3.

65. Shen BW, Xu D, Chan SH, Zheng Y, Zhu Z, Xu SY, et al. Characterization and crystal structure of the type IIG restriction endonuclease RM.BpuSI. Nucleic Acids Res. 2011;39(18):8223–36.

66. Oliveira PH, Touchon M, Rocha EPC. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. 2014;42(16):10618–32.

67. Wang X, Liotta L. Clinical bioinformatics: A new emerging science. J Clin Bioinforma. 2011;1(1):2–4.

68. illumina. A high-resolution view of the genome [Internet]. Available from:

https://www.illumina.com/techniques/sequencing/dna-sequencing/whole-genome-sequencing.html

69. Inouye M, Dashnow H, Raven LA, Schultz MB, Pope BJ, Tomita T, et al. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Med. 2014;6(11):1–16.

70. Sczyrba A. Critical Assessment of Metagenome Interpretation- a benchmark of computational metagenomics software. Nat Methods. 2017;14(11):163–78.

71. UiT NAU. Den sjuende Tromsø-undersøkelsen (Tromsø 7). [Internet]. Available from: https://uit.no/forskning/forskningsgrupper/sub?p_document_id=367276&sub_id=503778

72. Rafaelsberger N et. al. unpublished.

73. Matuschek E, Brown DFJ, Kahlmeter G. Development of the EUCAST disk diffusion antimicrobial susceptibility testing method and its implementation in routine microbiology laboratories. Clin Microbiol Infect [Internet]. 2014;20(4):O255–66. Available from: http://dx.doi.org/10.1111/1469-0691.12373

74. EUCAST. EUCAST- The European Committee on Antimicrobial Susceptibility Testing. Breakpoint tables fon interpreation of MICs and zone diameters. [Internet]. Available from: http://www.eucast.org

75. Hetland M et al. unpublished.

76. NORM - Norsk overvåkingssystem for antibiotikaresistens hos mikrober [Internet]. Available from: https://unn.no/fag-og-forskning/norm-norsk-overvakingssystem-for-antibiotikaresistens-hos-mikrober

77. Fostervold A et al. unpublished.

78. Fostervold A. The Norwegian Klebsiella pneumonia bacteremia study. Reg Kom Med og helsefaglig forskningsetikk [Internet]. 2016; Available from: https://helseforskning.etikkom.no/prosjekterirek/prosjektregister/prosjekt?p_document_id=729278&p_parent_id=742566&_ikbLanguageCode=n

79. Fostervold A. The Norwegian Klebsiella pneumoniae bacteremiae study. 2016;(12):1–12.

80. ENA. ENA-European Nucleotide Archive [Internet]. Available from: https://www.ebi.ac.uk/ena

81. Wick R, Whyres K, Holt KE. Kleborate github [Internet]. Australia; 2018. Available from: https://github.com/katholt/Kleborate#klebsiella-species

82. Institut Pasteur MLST and whole genome MLST databases. Primers used for MLST og Klebsiella pneumoniae [Internet]. Available from: https://bigsdb.pasteur.fr/klebsiella/primers_used.html

83. Martin C J, Maiden MJ, Rensburg J van, Bray JE, Earle SG, Ford SA, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol. 2013;11(10):728–36.

84. Carattoli A, Zankari E, Garciá-Fernández A, Larsen MV, Lund O, Villa L, et al. PlasmidFinder and pMLST: In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother. 2014;58(7):3895–903.

85. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691–3.

86. Alkhnbashi OS, Meier T, Mitrofanov A, Backofen R, Voß B. CRISPR-Cas bioinformatics. Methods [Internet]. 2019;(July):1–9. Available from: https://doi.org/10.1016/j.ymeth.2019.07.013

87. CRISPRCasFinder / CRISPRCasViewer manual [Internet]. 2019. p. 1–9. Available from: https://github.com/dcouvin/CRISPRCasFinder

88. Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, Néron B, et al. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res. 2018;46(W1):W246–51.

89. Kassambara A. A practical Guide To Principal Component Methods in R [Internet]. STHDA-Statistical tools for high-throughput data analysi. 2017. Available from: http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/

90. Liu Y, Chen C, Li J, Du F, Long D, Zhang W, et al. Distribution of CRISPR-Cas Systems in Clinical Carbapenem-Resistant Klebsiella pneumoniae Strains in a Chinese Tertiary Hospital and Its Potential Relationship with Virulence. Microb Drug Resist. 2019;00(00):1–7.

91. Lin TL, Pan YJ, Hsieh PF, Hsu CR, Wu MC, Wang JT. Imipenem represses CRISPR-Cas interference of DNA acquisition through H-NS stimulation in Klebsiella pneumoniae. Sci Rep [Internet]. 2016;6(July):1–10. Available from: http://dx.doi.org/10.1038/srep31644

92. Tang Y, Fu P, Zhou Y, Xie Y, Jin J, Wang B, et al. Absence of the type I-E CRISPR-Cas system in Klebsiella pneumoniae clonal complex 258 is associated with dissemination of IncF epidemic resistance plasmids in this clonal complex. J Antimicrob Chemother. 2019;1–6.

Appendix 1: Bioinformatic command lines

Downloading strains from ENA using https://github.com/stevenjdunn/enaget

$ python path/to/enaget -l path/to/list.txt -o downloadedENA $

##Annotation and removal of contigs below 200nt by Prokka using https://github.com/tseemann/prokka

$ for F in *.fasta; do N=$(basename $F .fasta) ; prokka --prefix "$F" --locustag "$F" --cpus 70 --usegenus --compliant --mincontiglen 200 --genus Klebsiella --species pneumoniae --outdir "$F"_prokka --force --addgenes $F; done $

##Assembly

#Assembly was done using https://github.com/marithetland/Asmbl, the commands and parameters for trimming and assmbly are:

#TrimGalore

$ trim_galore --paired -trim1 --retain_unpaired Sequence_?.fastq.gz # Outputs Sequence_1_val_1.fq.gz and Sequence_2_val_2.fq $

#Unicycler:

$ unicycler -1 Sequence_1_val_1.fq.gz -2 Sequence_2_val_2.fq.gz -o Sequence_assembly --verbosity 2 --keep 2 $

##Kleborate using https://github.com/katholt/Kleborate.git

$ kleborate -r --all -a *.fasta $

##Plasmidfinder through abricate (https://github.com/tseemann/abricate)

$ for file in *.fasta ; do abricate --db plasmidfinder -- mincov 80 -- minid 80 $file > ${file}_abricate.tsv ; done $

$ abricate --summary *.tsv >summary_Abricate.tsv $

##Phylogenetic tree by Roary

$roary -n -e --mafft -p 64 Gff_Files_ST307/*.gff

create_pan_genome_plots.R *.Rtab

roary_plots.py --labels accessory_binary_genes.fa.newick gene_presence_absence.csv $

##Running CRISPRCasFinder

$ for file in $(ls *fasta); do
~/Klebs_project/Programs/CRISPRCasFinder/CRISPRCasFinder.pl -html -copyCSS -cas -minDR 19 -minSP 20 -def SubTyping -getSummaryCasfinder -so ~/Klebs_project/Programs/CRISPRCasFinder/sel392v2.so -ccvr -in $file ; done $

##BLAST using https://anaconda.org/bioconda/blast

$; for x in *.fasta; do blastn -query $x -db Merged_Cas1_Ref.fa -outfmt 6 -out $x.BCas1.tsv; done $

Structural evaluation and subtyping by SimpleSynteny using https://www.dveltri.com/simplesynteny/about.html

$ for F in *.fa do N=$(basename $F *.fa); ~/SimpleSyntheny/SyntenyMapper.rb I-ES_Merged.fasta $F $F.out -e 0.001 -cov 80 ; done $

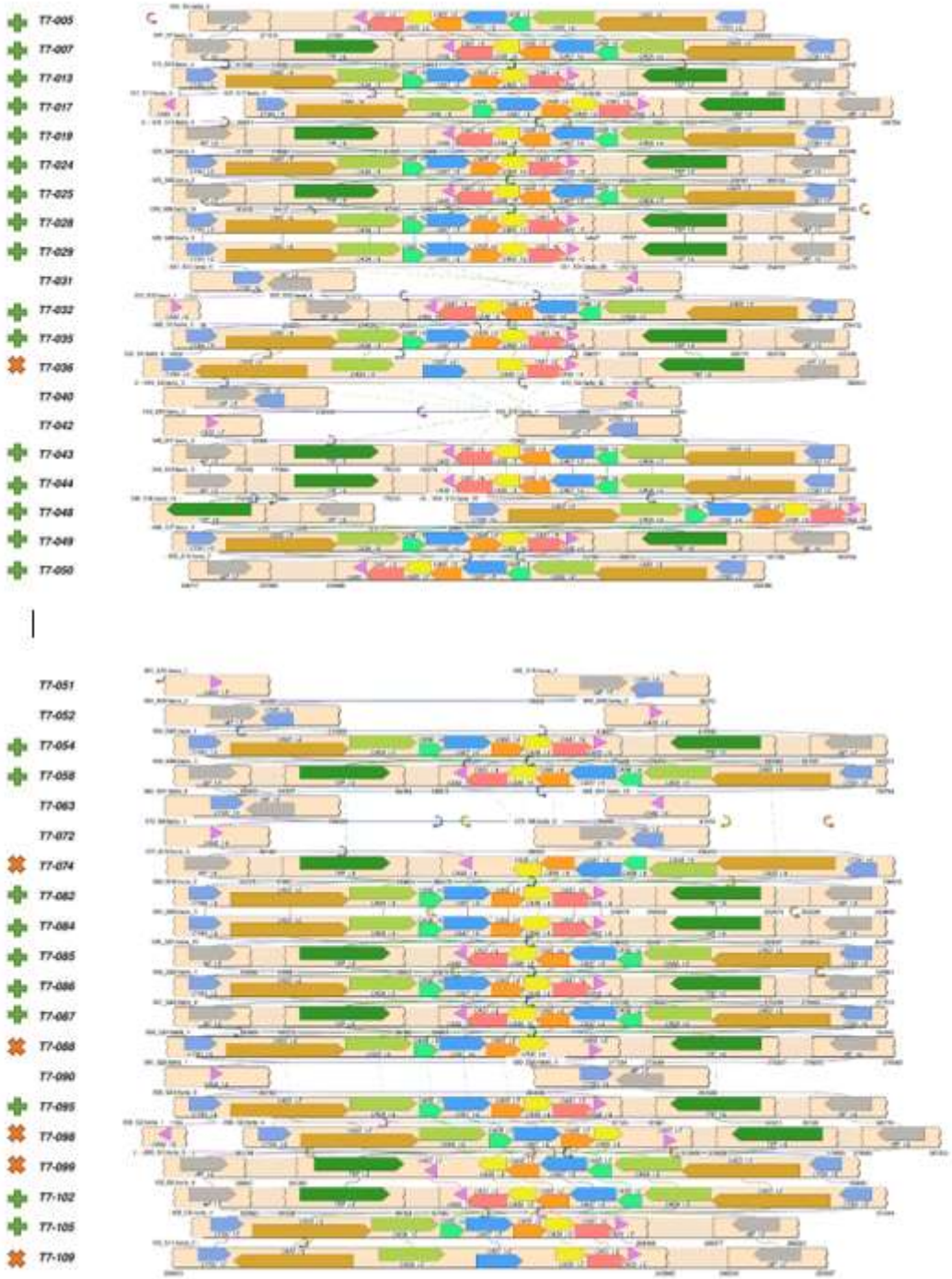
##R-M system profile by HMMER using https://github.com/EddyRivasLab/hmmer

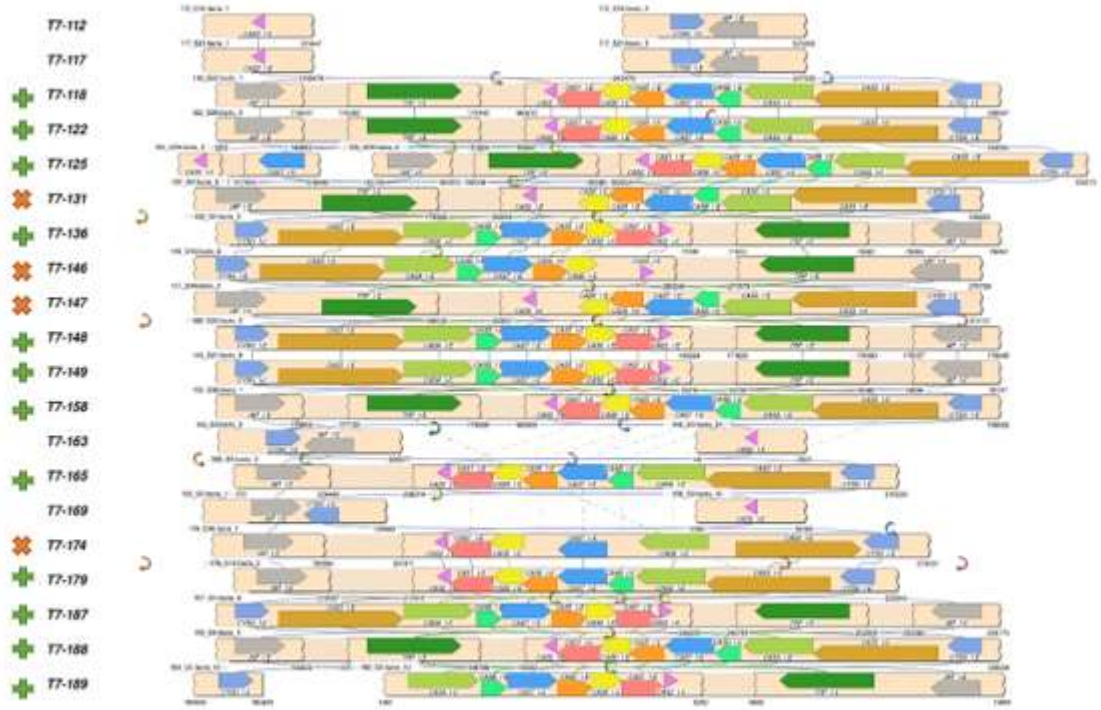
$ for h in *.hmm; do for f in *.faa ; do csh Go_HMM.sh $f $h –xfile 200 ; done ; done $

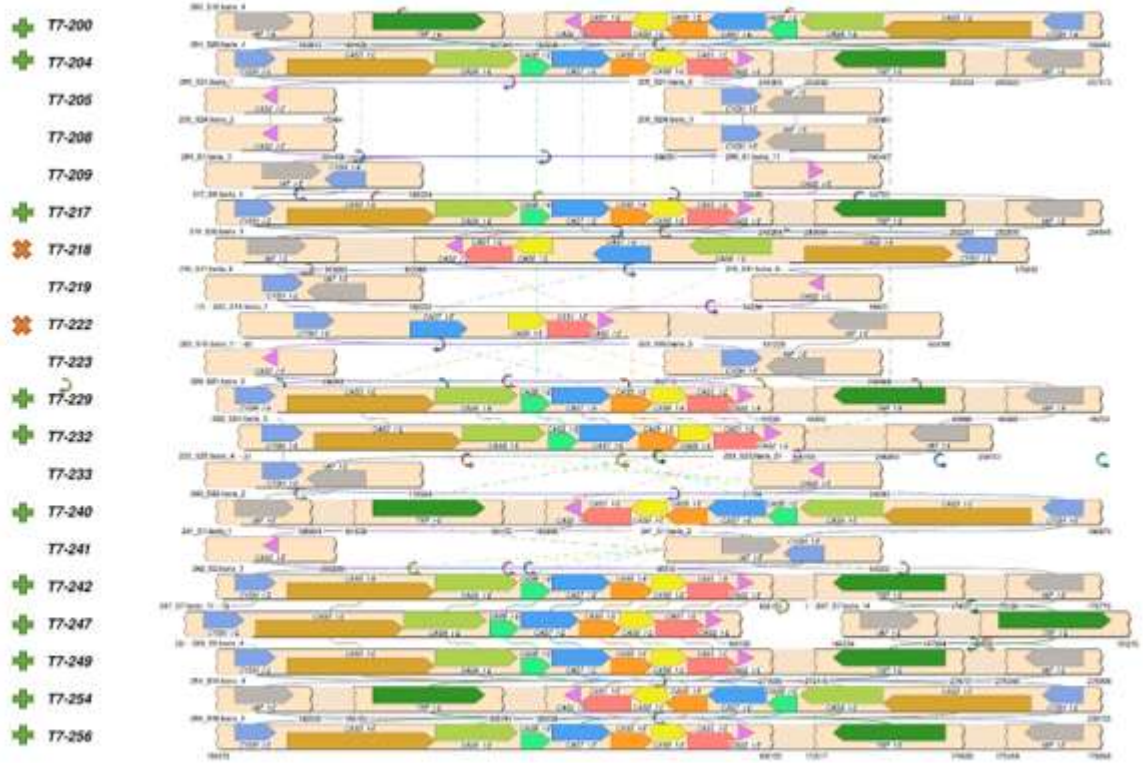
Appendix 2: SimpleSynteny evaluation

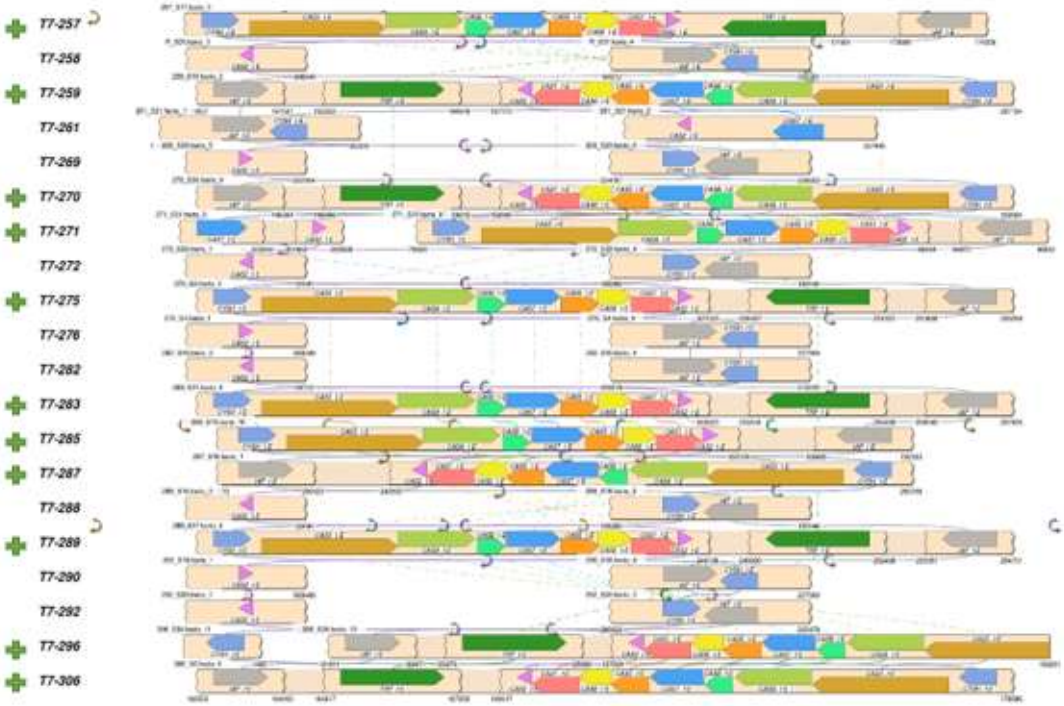
Carrier strains

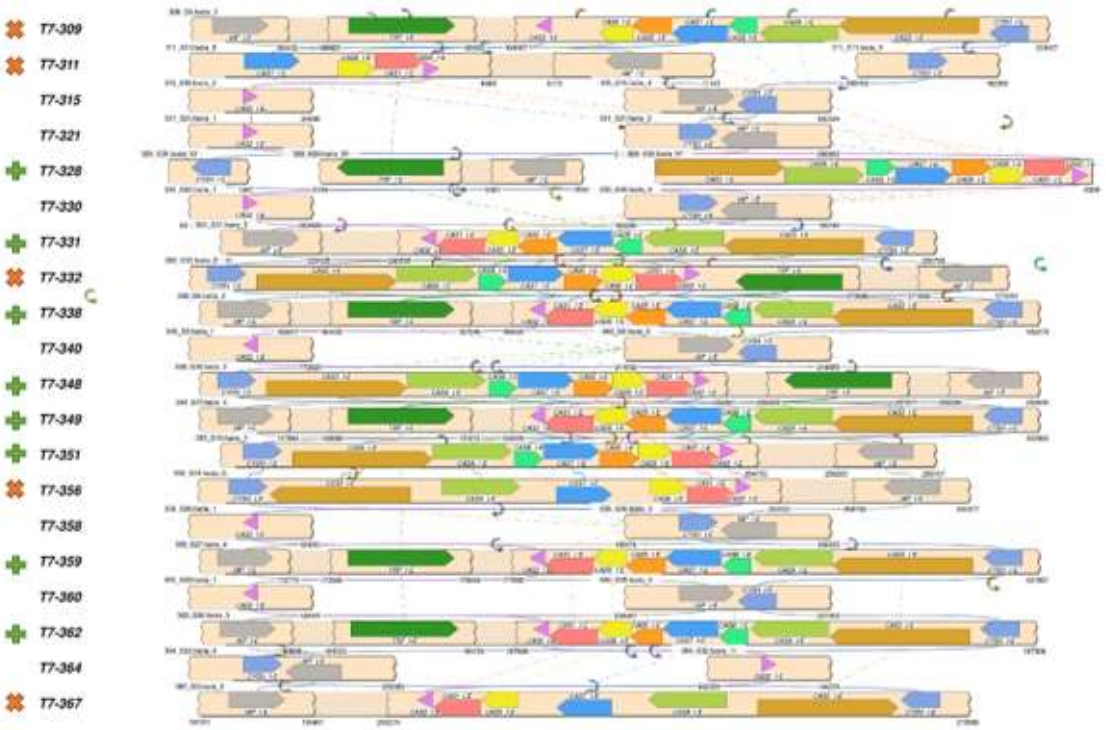
Type I-E 80% coverage

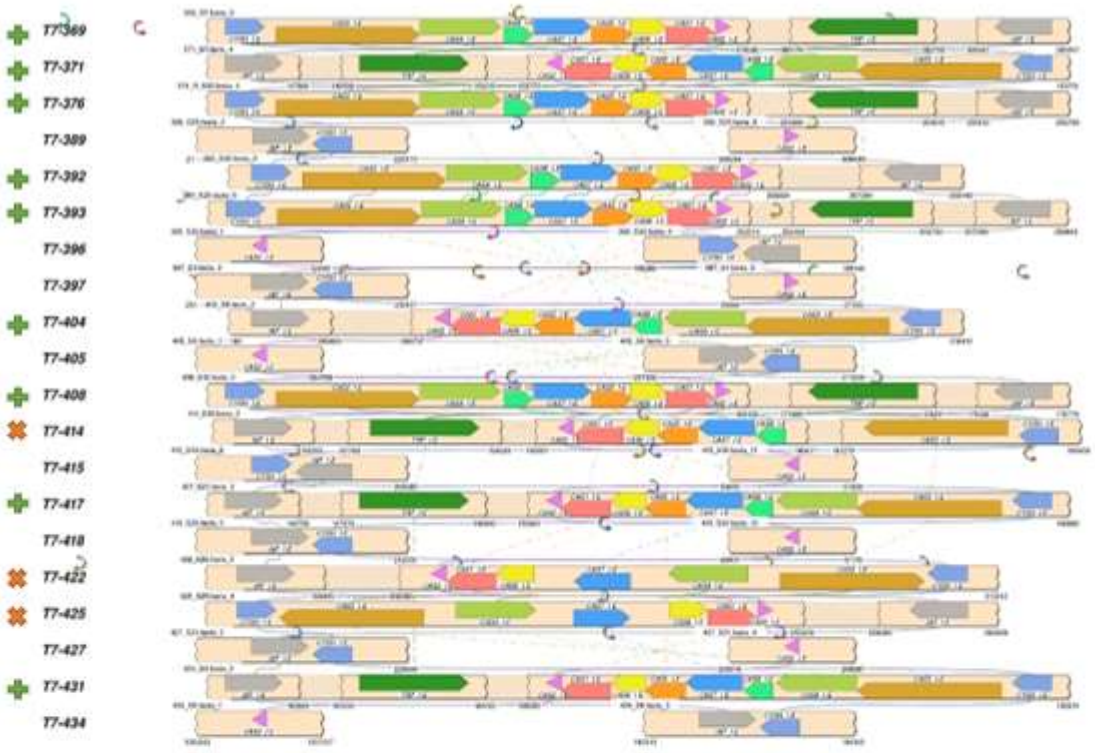


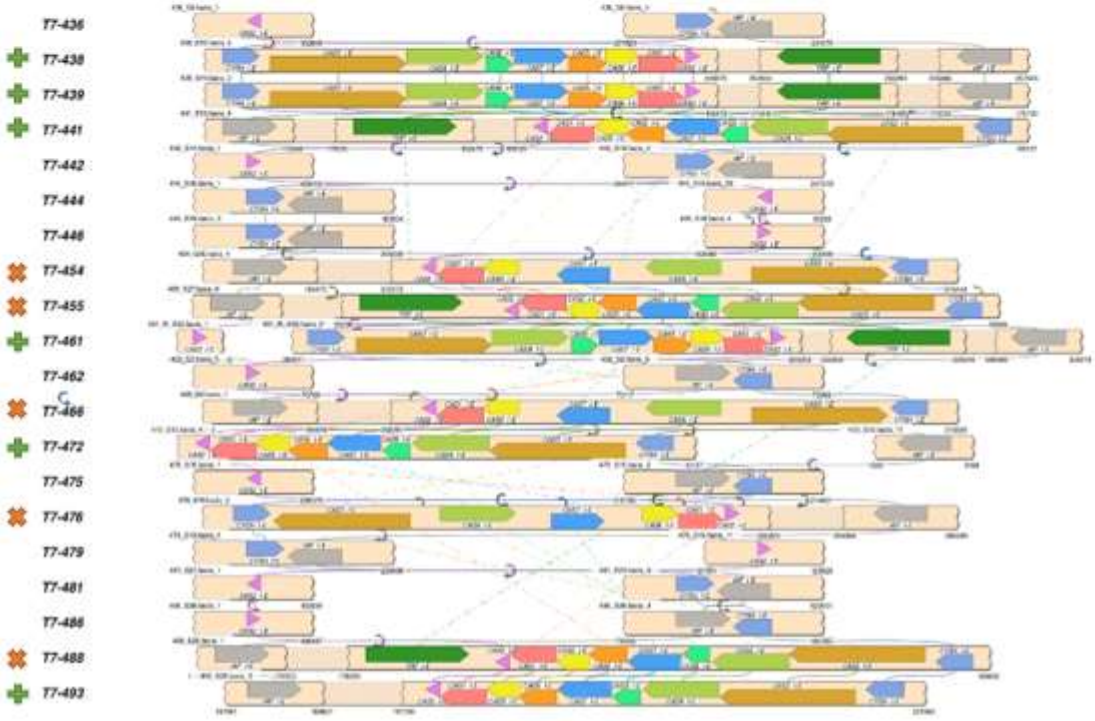

T7-112
T7-117
T7-118
T7-122
T7-125
T7-131
T7-136
T7-146
T7-147
T7-148
T7-149
T7-158
T7-163
T7-165
T7-169
T7-174
T7-179
T7-187
T7-188
T7-189


T7-200
T7-204
T7-205
T7-208
T7-209
T7-217
T7-218
T7-219
T7-222
T7-223
T7-229
T7-232
T7-233
T7-240
T7-241
T7-242
T7-247
T7-249
T7-254
T7-256

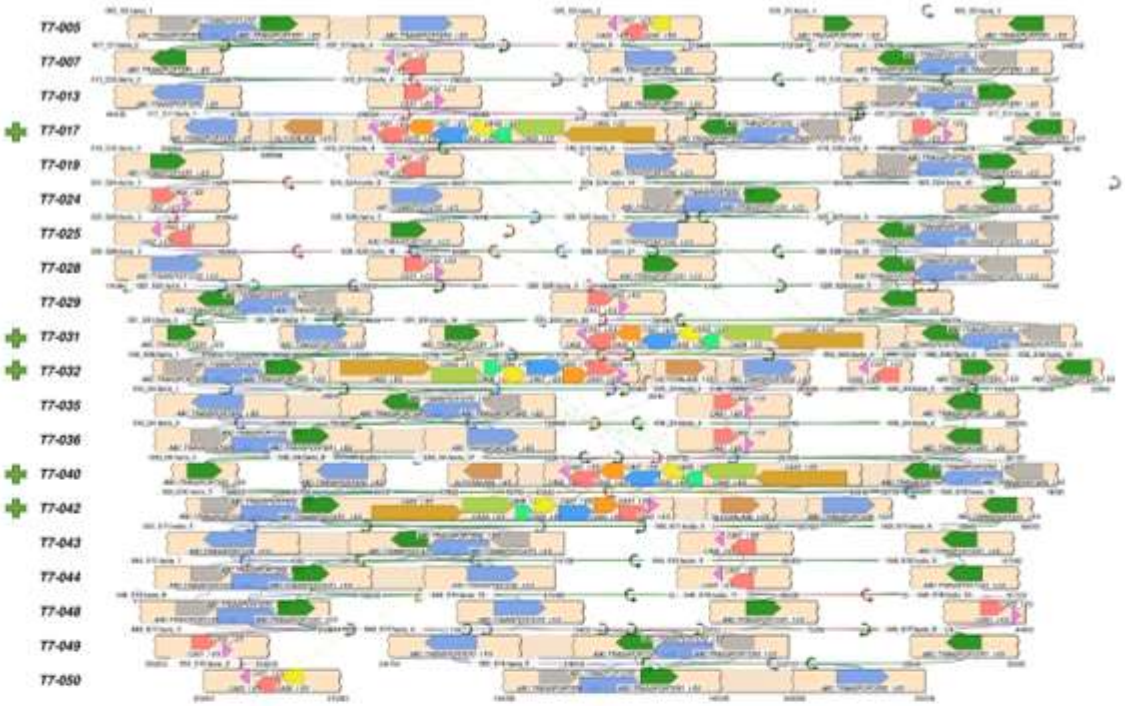

T7-257
T7-258
T7-259
T7-261
T7-269
T7-270
T7-271
T7-272
T7-275
T7-276
T7-282
T7-283
T7-285
T7-287
T7-288
T7-289
T7-290
T7-292
T7-296
T7-306

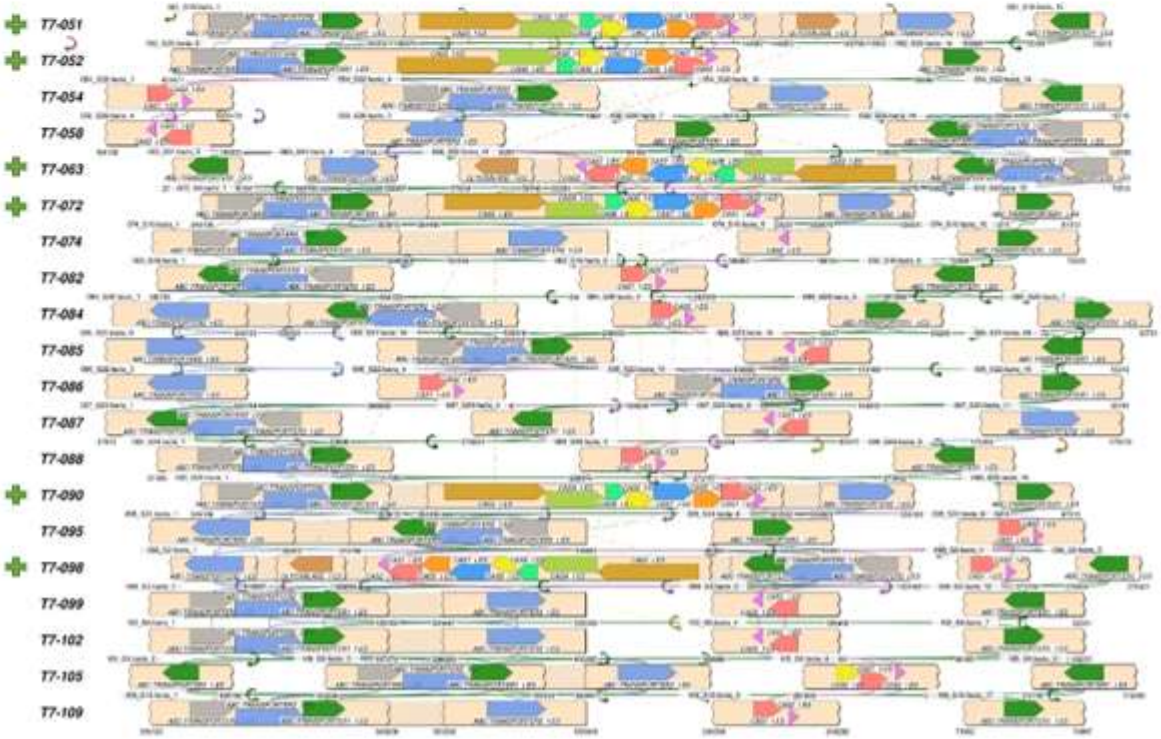

T7-309
T7-311
T7-315
T7-321
T7-328
T7-330
T7-331
T7-332
T7-338
T7-340
T7-348
T7-349
T7-351
T7-356
T7-358
T7-359
T7-360
T7-362
T7-364
T7-367

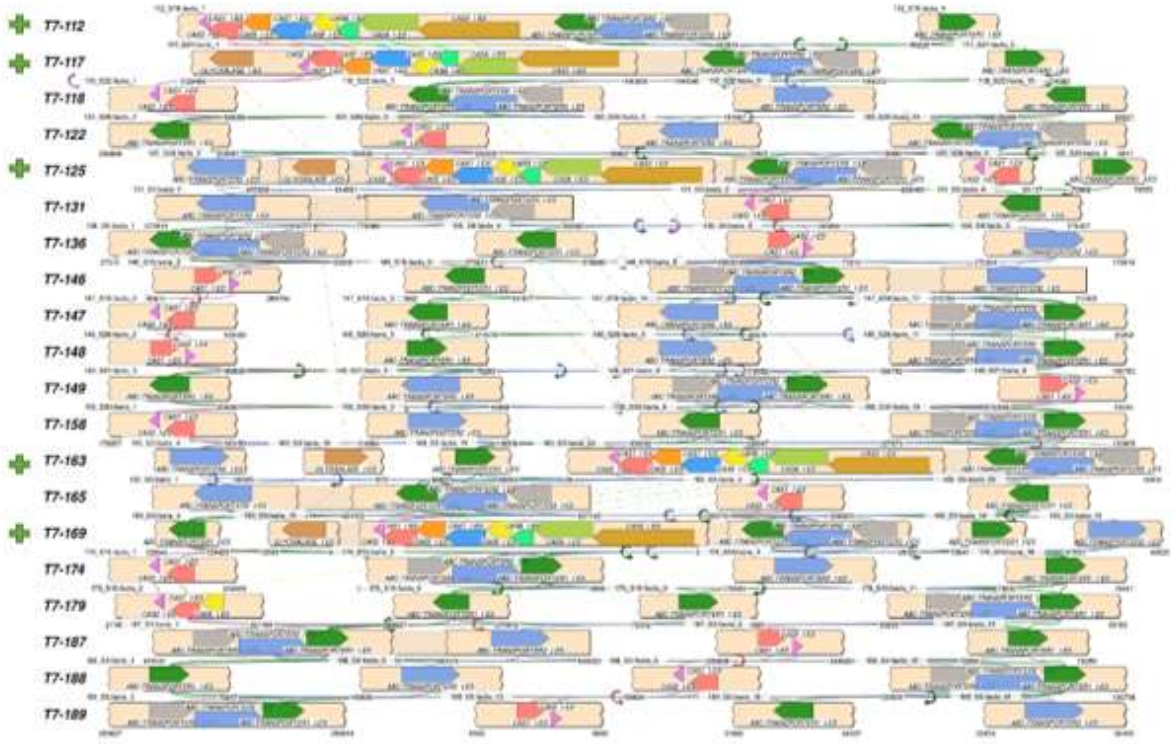

T7-369
T7-371
T7-376
T7-389
T7-392
T7-393
T7-396
T7-397
T7-404
T7-405
T7-408
T7-414
T7-415
T7-417
T7-418
T7-422
T7-425
T7-427
T7-431
T7-434

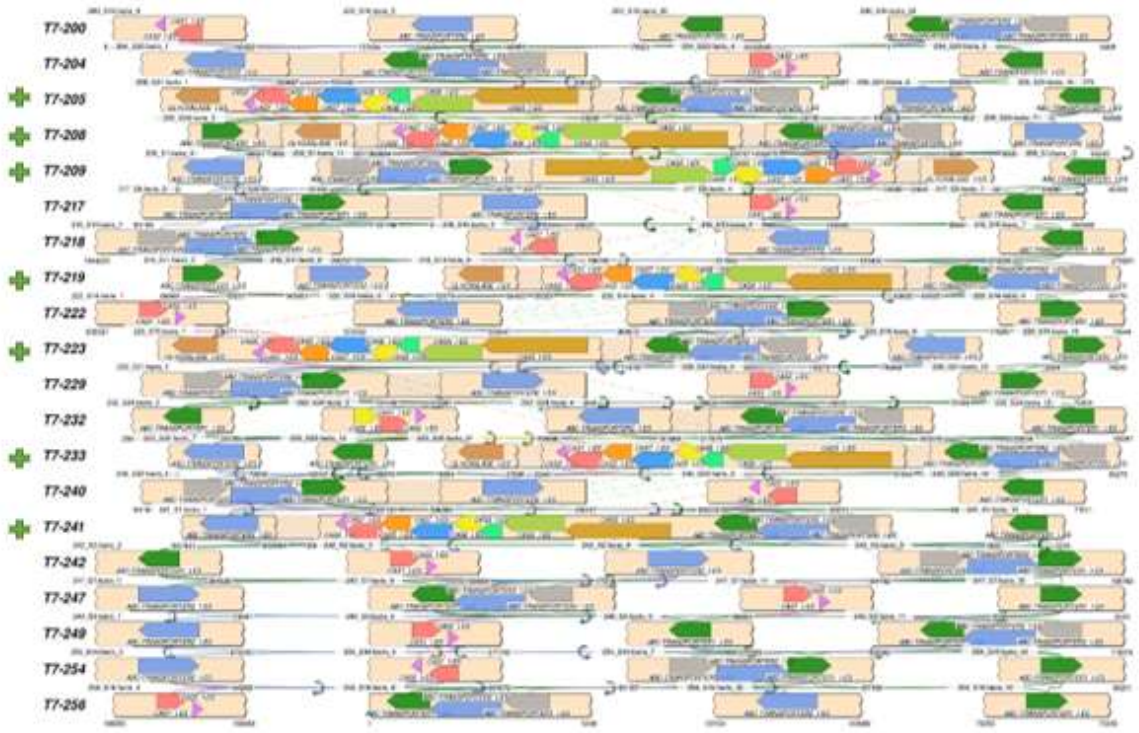

T7-436
T7-438
T7-439
T7-441
T7-442
T7-444
T7-446
T7-454
T7-455
T7-461
T7-462
T7-466
T7-472
T7-475
T7-476
T7-479
T7-481
T7-486
T7-488
T7-493

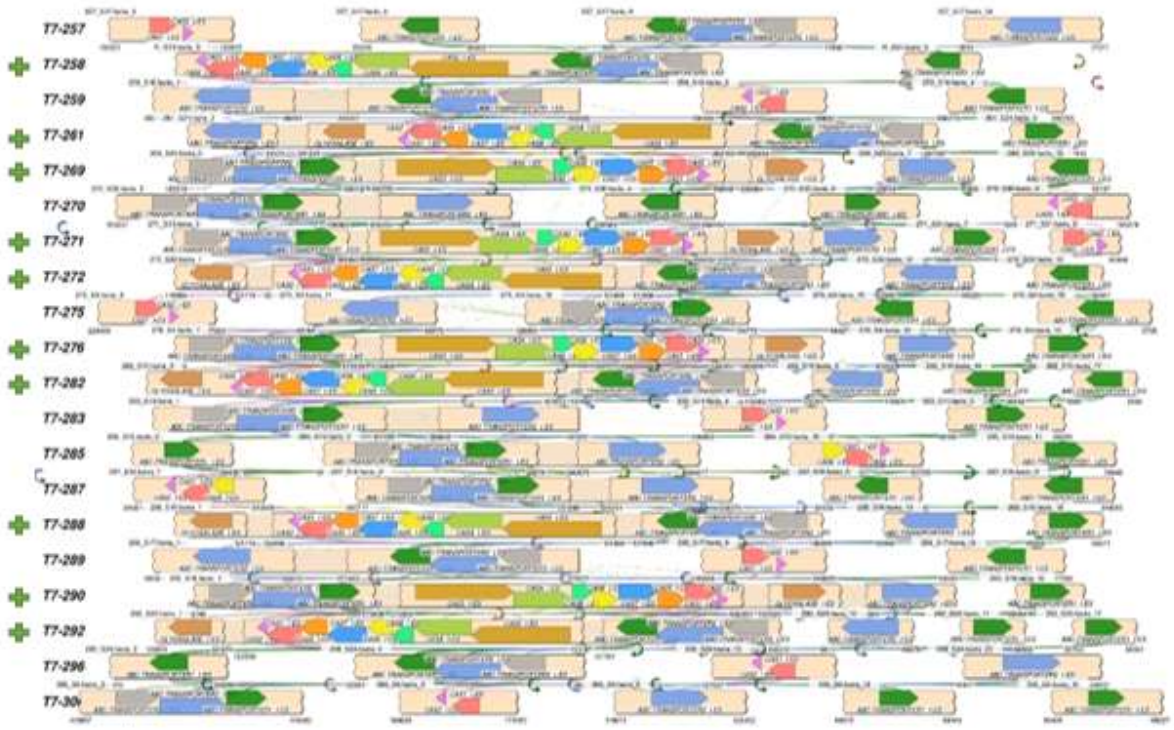
CRISPR-Cas Type I-E*

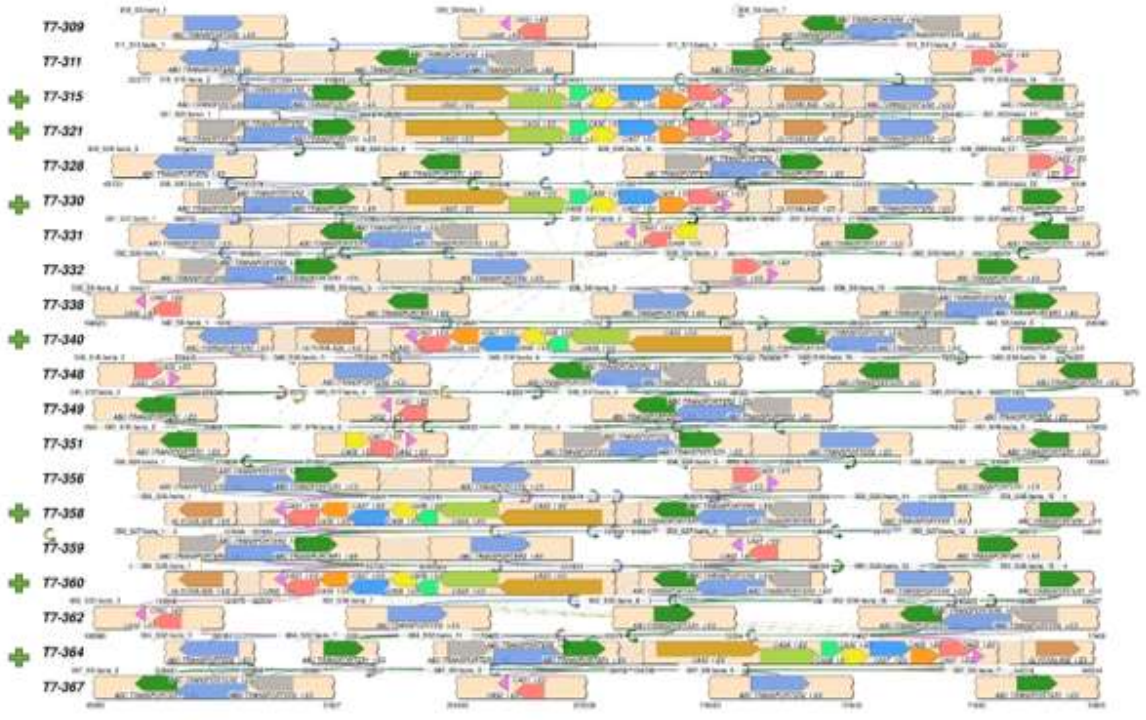


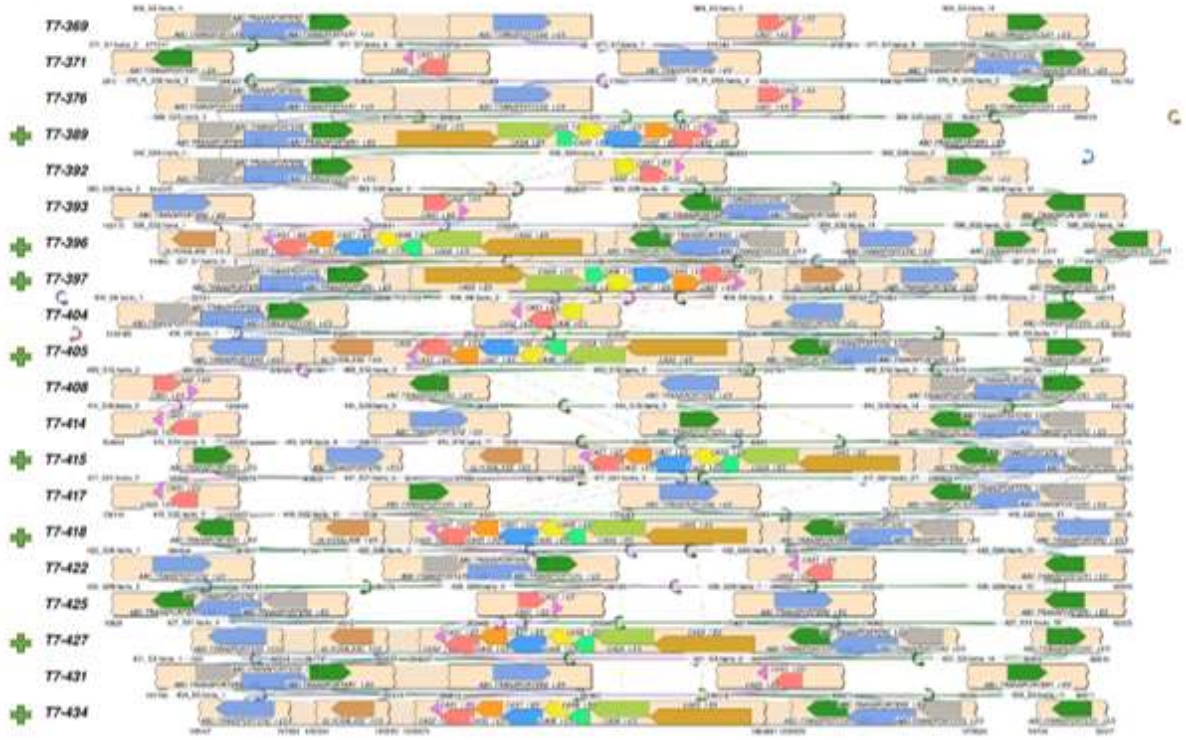


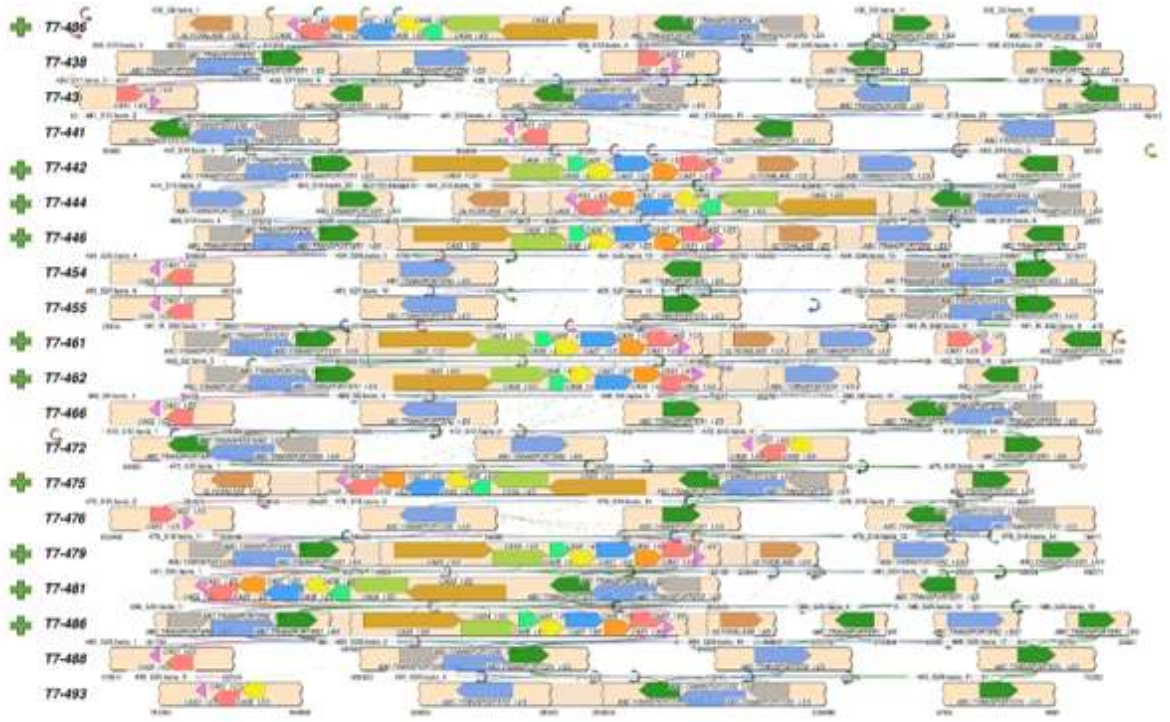

T7-112
T7-117
T7-118
T7-122
T7-125
T7-131
T7-136
T7-146
T7-147
T7-148
T7-149
T7-158
T7-163
T7-165
T7-169
T7-174
T7-179
T7-187
T7-188
T7-189


T7-200
T7-204
T7-205
T7-208
T7-209
T7-217
T7-218
T7-219
T7-222
T7-223
T7-229
T7-232
T7-233
T7-240
T7-241
T7-242
T7-247
T7-249
T7-254
T7-256


T7-257
T7-258
T7-259
T7-261
T7-269
T7-270
T7-271
T7-272
T7-275
T7-276
T7-282
T7-283
T7-285
T7-287
T7-288
T7-289
T7-290
T7-292
T7-296
T7-30

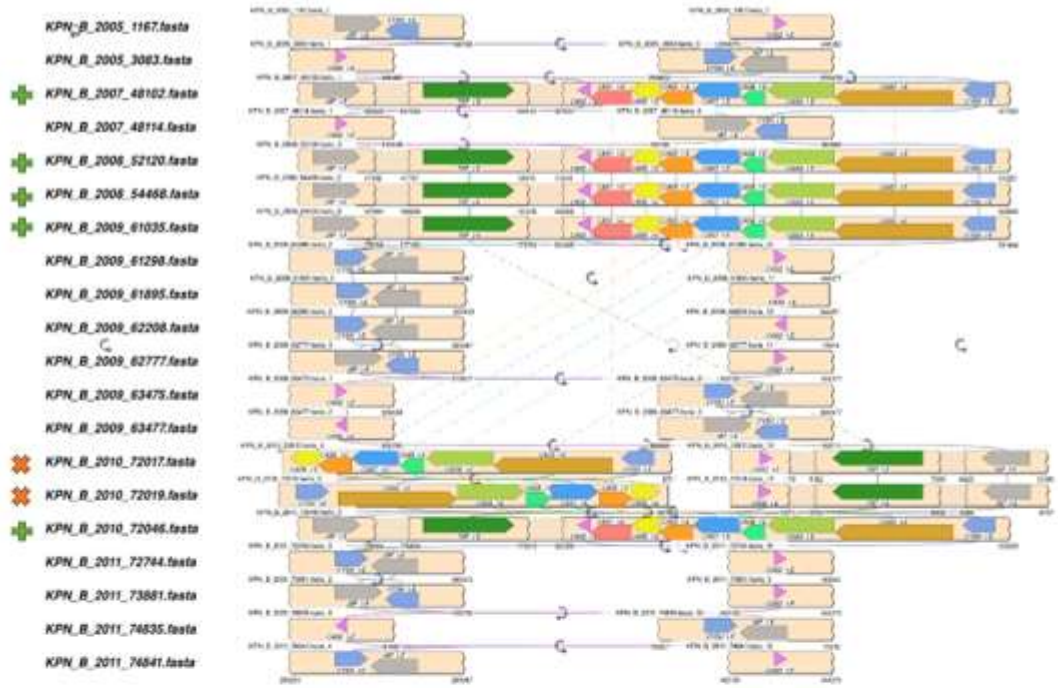

T7-309
T7-311
T7-315
T7-321
T7-328
T7-330
T7-331
T7-332
T7-338
T7-340
T7-348
T7-349
T7-351
T7-356
T7-358
T7-359
T7-360
T7-362
T7-364
T7-367

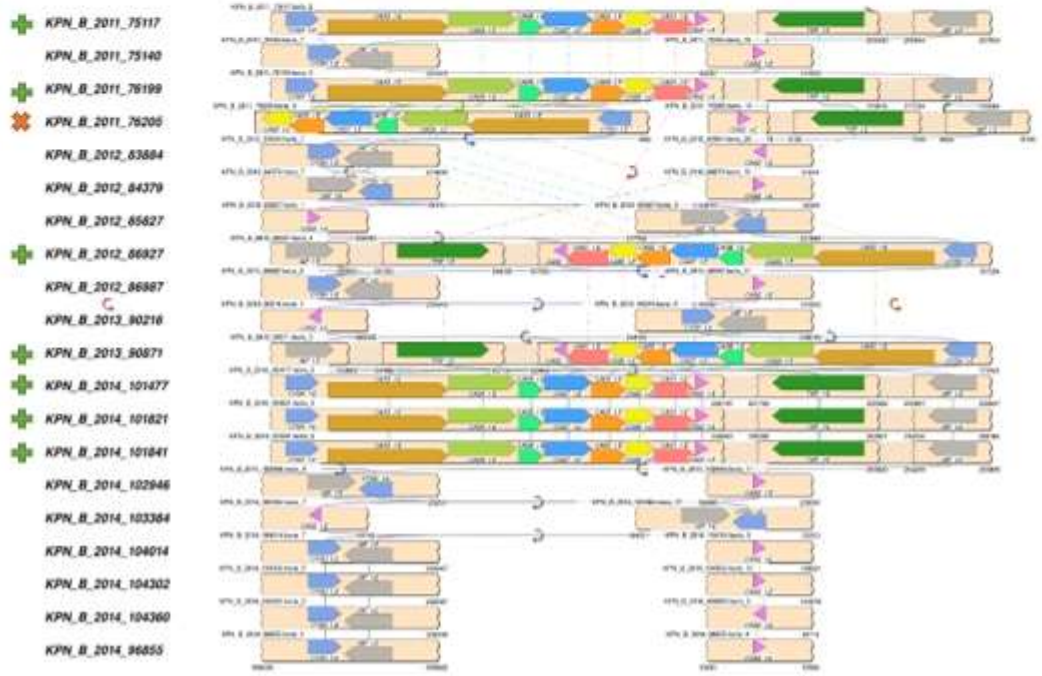


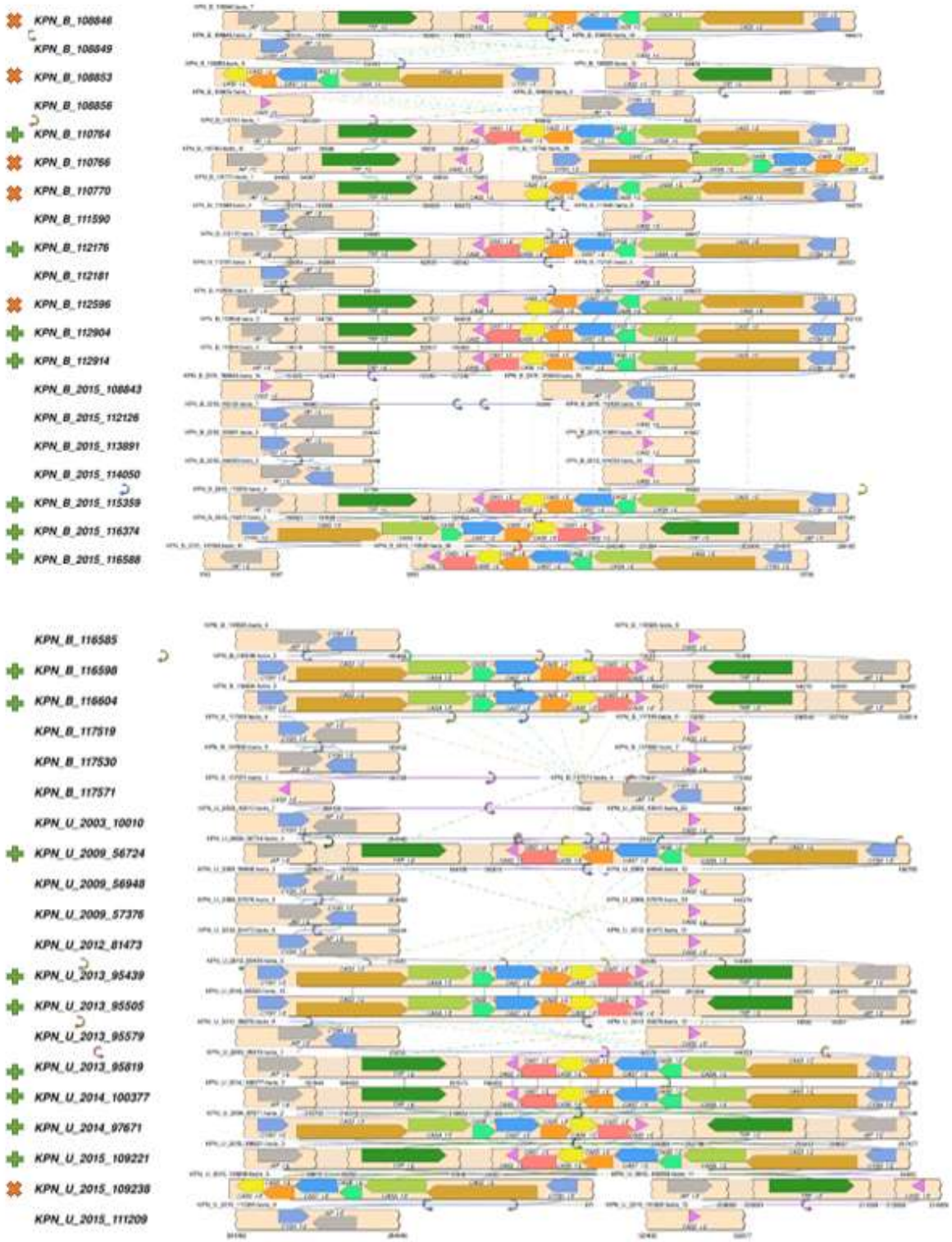


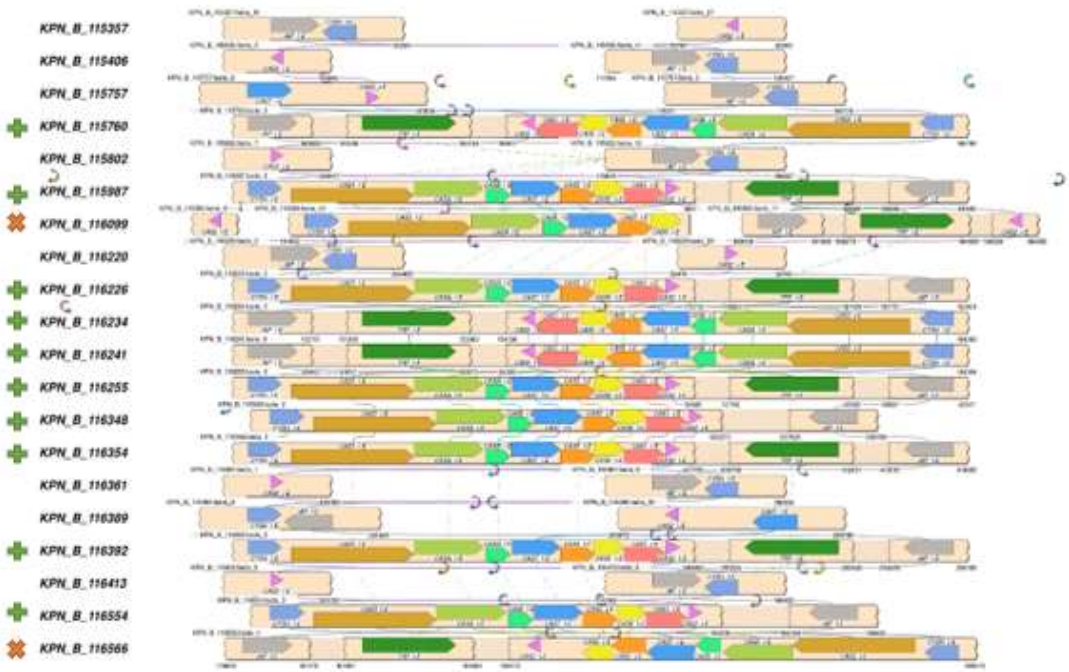
NORM strain collection

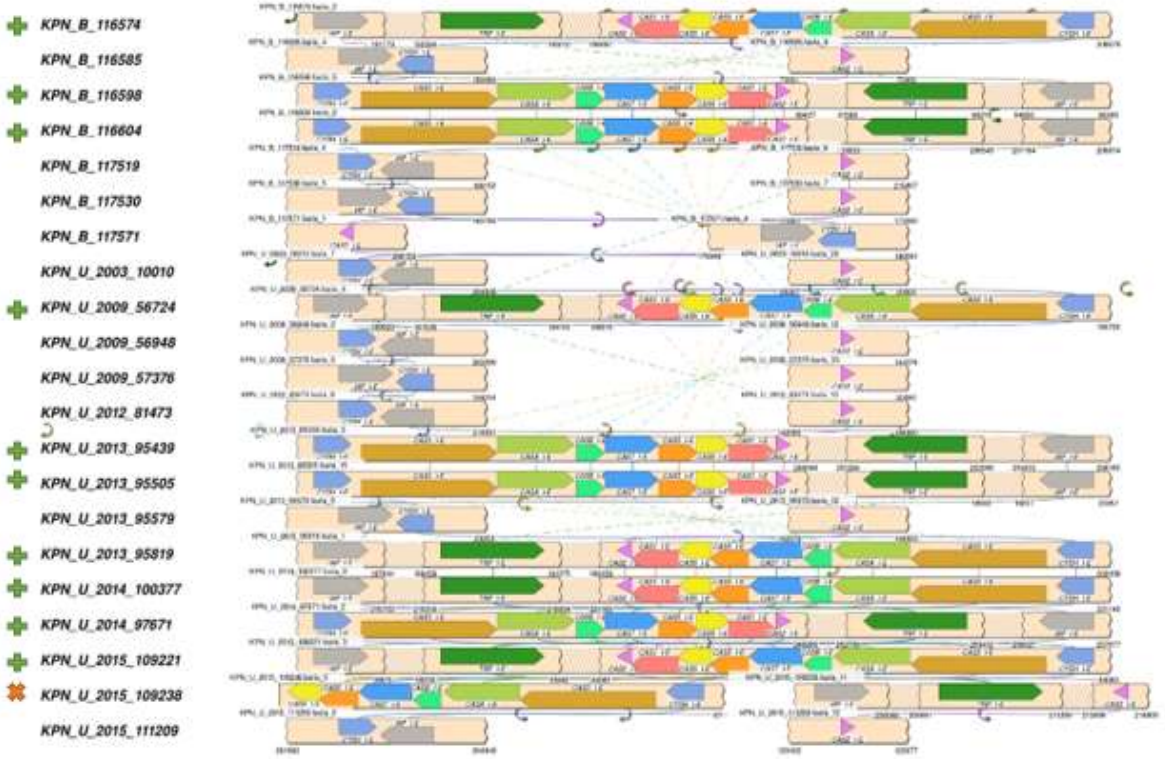
CRISPR-Cas Class 1Type I-E

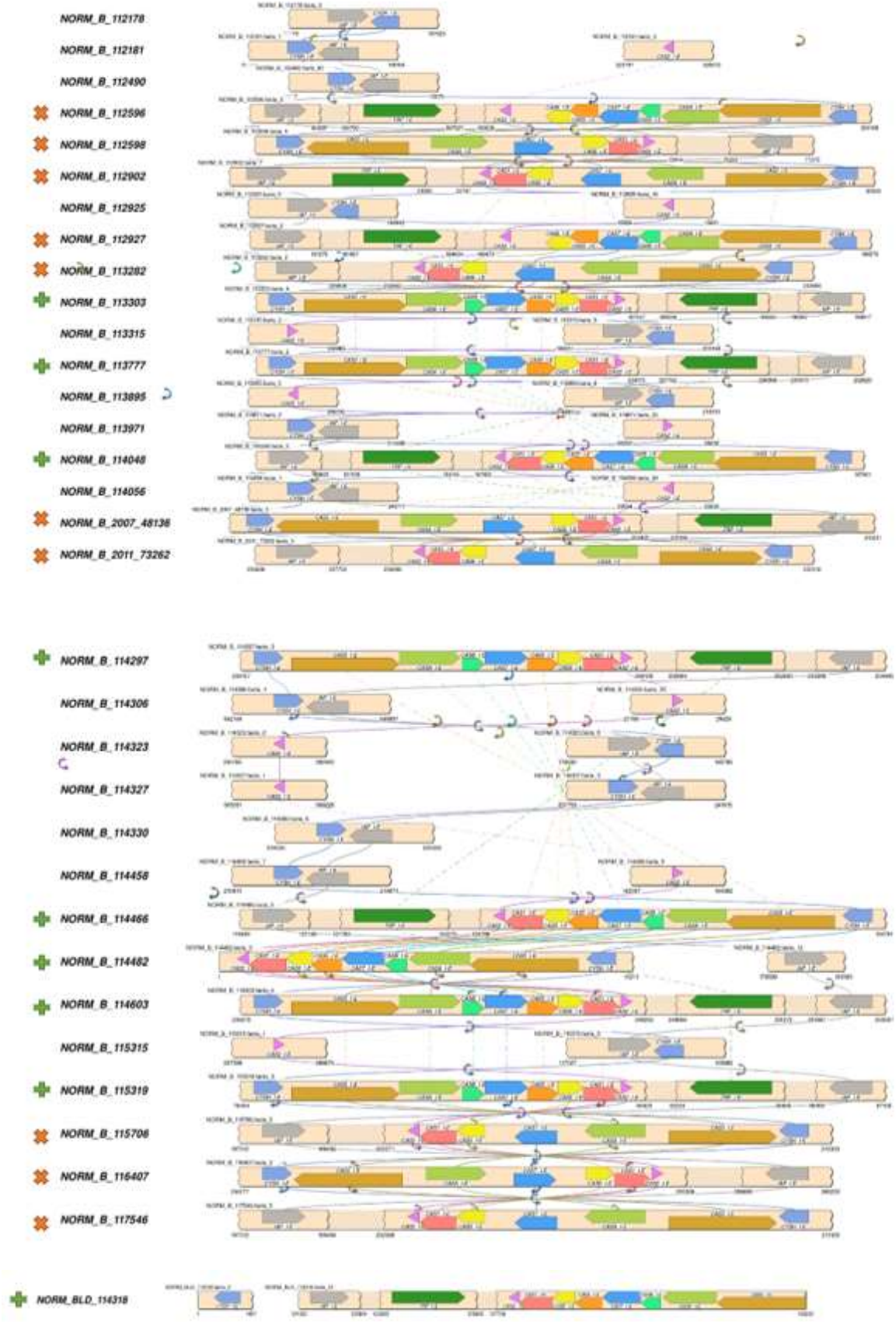




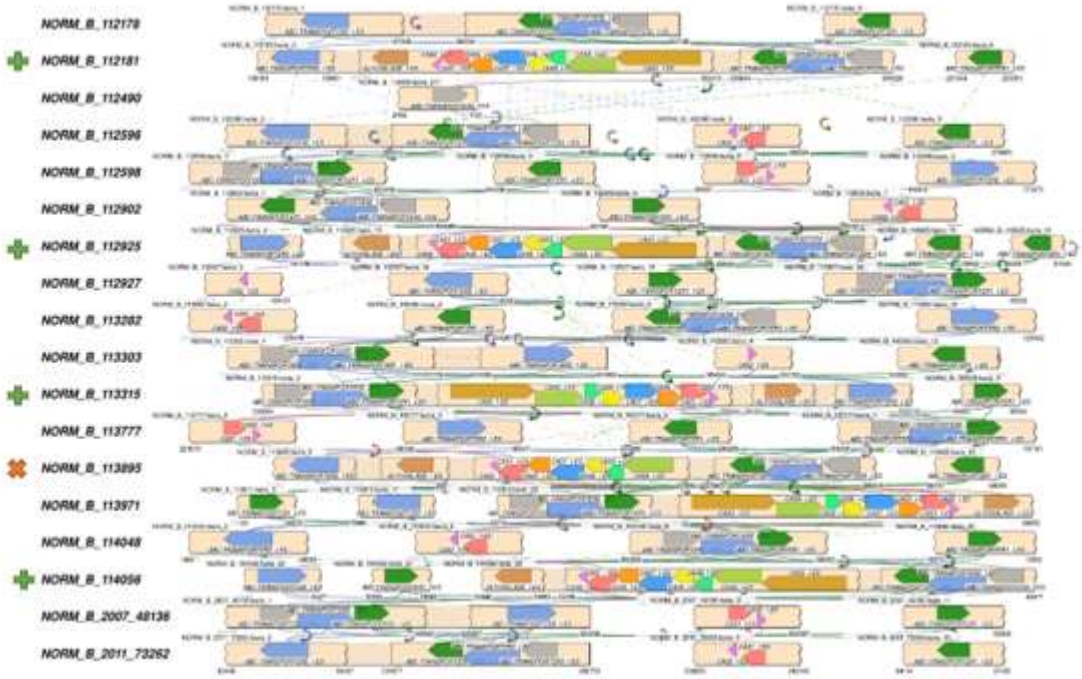

KPN_B_108846
KPN_B_108849
KPN_B_108853
KPN_B_108856
KPN_B_110764
KPN_B_110766
KPN_B_110770
KPN_B_111590
KPN_B_112176
KPN_B_112181
KPN_B_112596
KPN_B_112904
KPN_B_112914
KPN_B_2015_108843
KPN_B_2015_112126
KPN_B_2015_113891
KPN_B_2015_114050
KPN_B_2015_115359
KPN_B_2015_116374
KPN_B_2015_116588
KPN_B_116585
KPN_B_116598
KPN_B_116604
KPN_B_117519
KPN_B_117530
KPN_B_117571
KPN_U_2003_10010
KPN_U_2009_56724
KPN_U_2009_56948
KPN_U_2009_57376
KPN_U_2012_81473
KPN_U_2013_95439
KPN_U_2013_95505
KPN_U_2013_95579
KPN_U_2013_95819
KPN_U_2014_100377
KPN_U_2014_97671
KPN_U_2015_109221
KPN_U_2015_109238
KPN_U_2015_111209

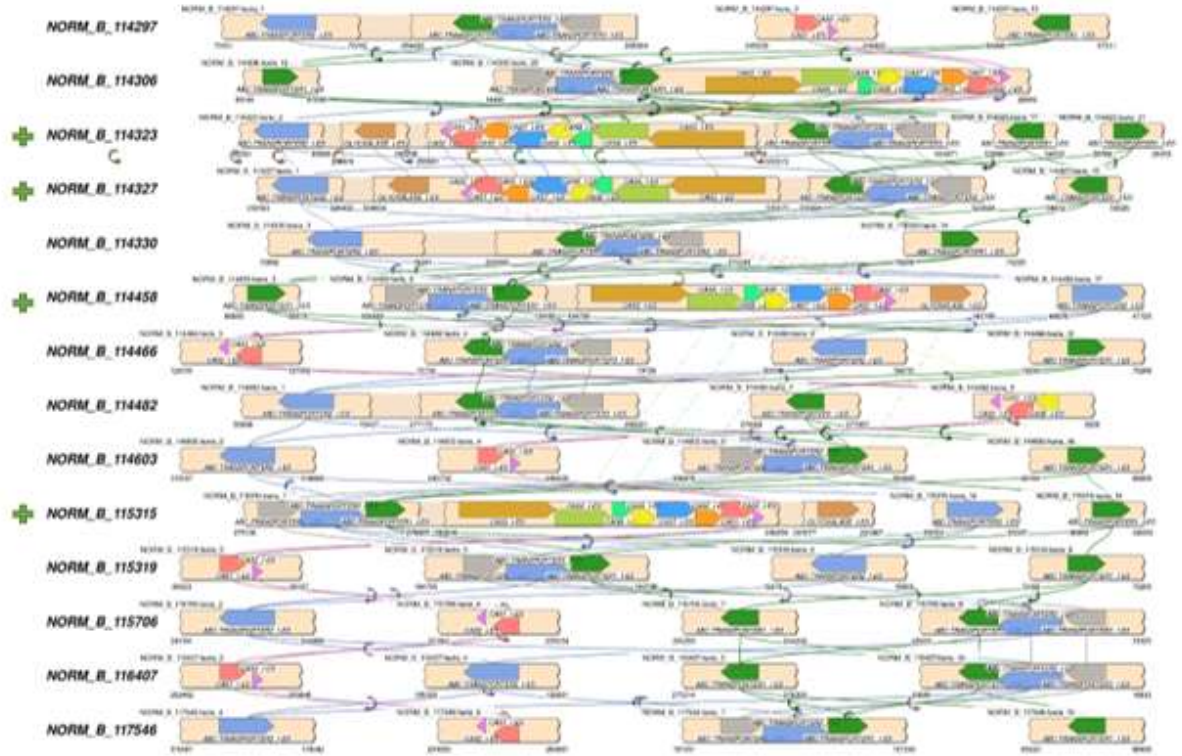

KPN_B_115357
KPN_B_115406
KPN_B_115757
KPN_B_115760
KPN_B_115802
KPN_B_115987
KPN_B_116099
KPN_B_116220
KPN_B_116226
KPN_B_116234
KPN_B_116241
KPN_B_116255
KPN_B_116348
KPN_B_116354
KPN_B_116361
KPN_B_116389
KPN_B_116392
KPN_B_116413
KPN_B_116554
KPN_B_116566

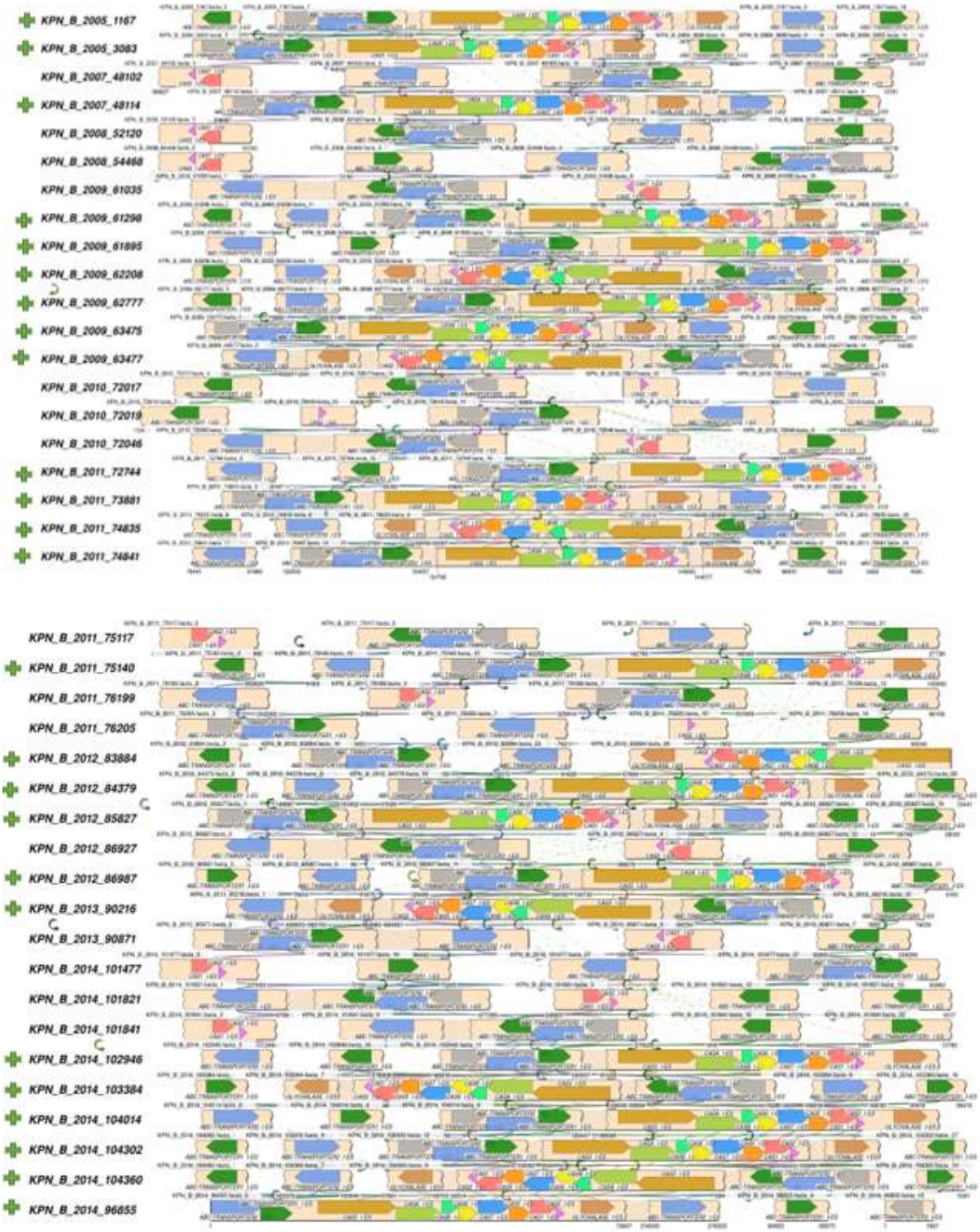

KPN_B_116574
KPN_B_116585
KPN_B_116598
KPN_B_116604
KPN_B_117519
KPN_B_117530
KPN_B_117571
KPN_U_2003_10010
KPN_U_2009_56724
KPN_U_2009_56948
KPN_U_2009_57376
KPN_U_2012_81473
KPN_U_2013_95439
KPN_U_2013_95505
KPN_U_2013_95579
KPN_U_2013_95819
KPN_U_2014_100377
KPN_U_2014_97671
KPN_U_2015_109221
KPN_U_2015_109238
KPN_U_2015_111209

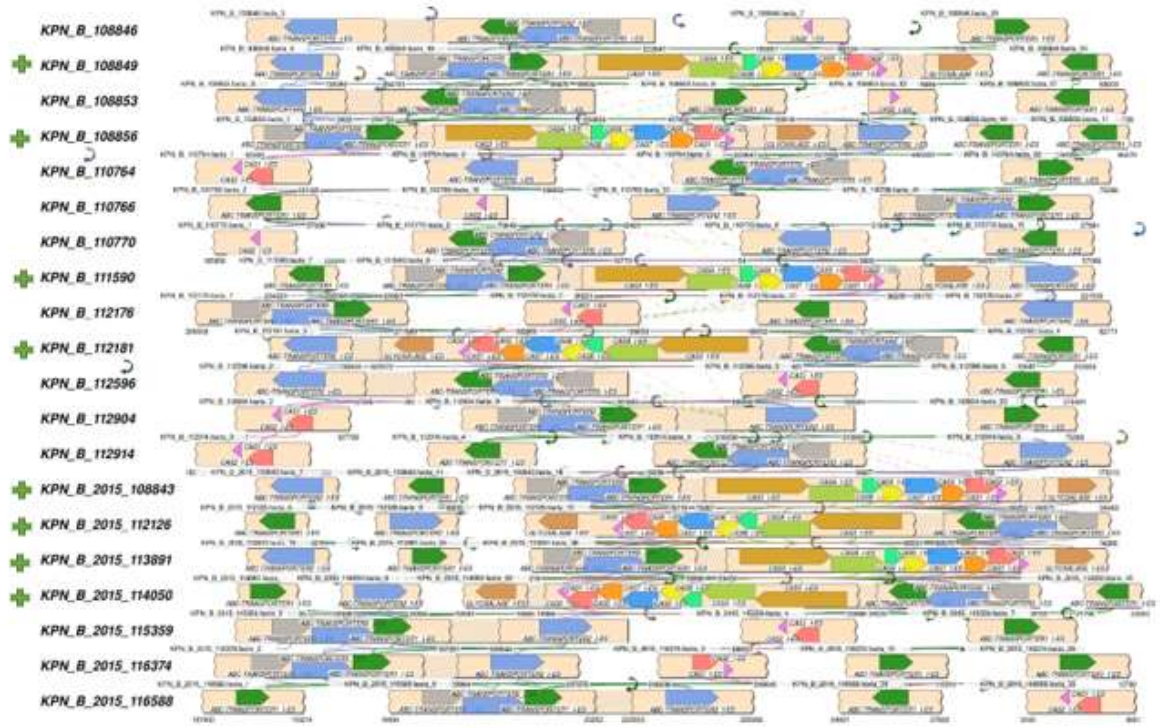


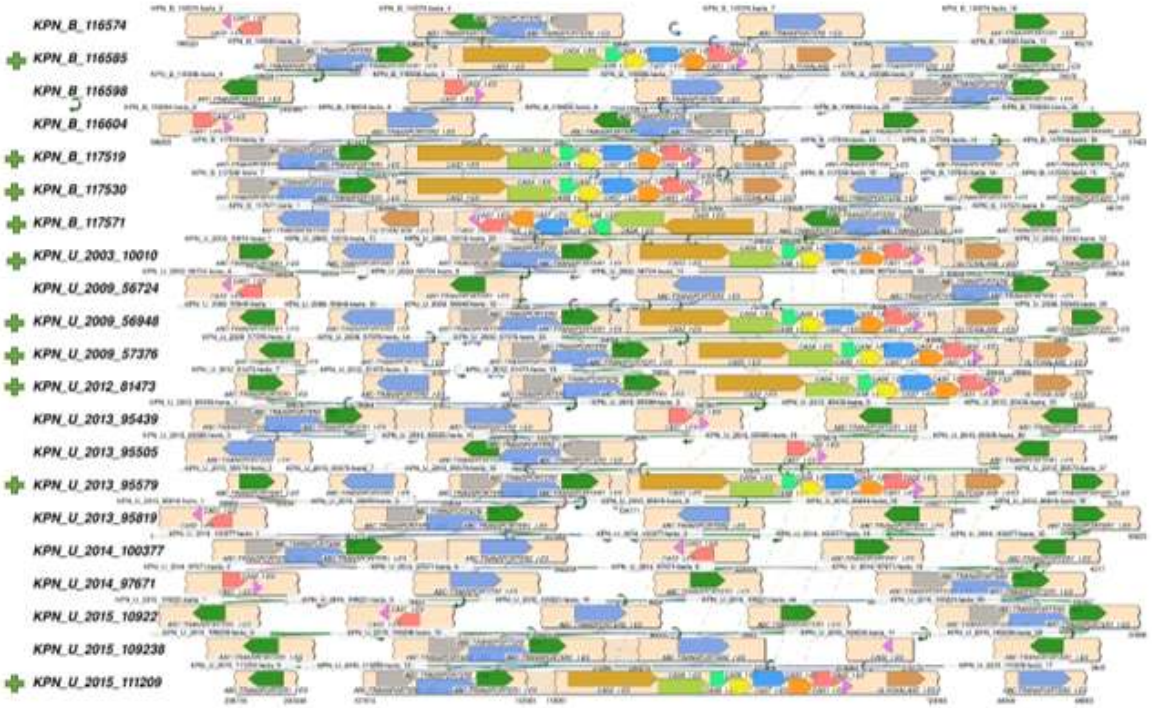
CRISPR-Cas Class 1Type I-E*

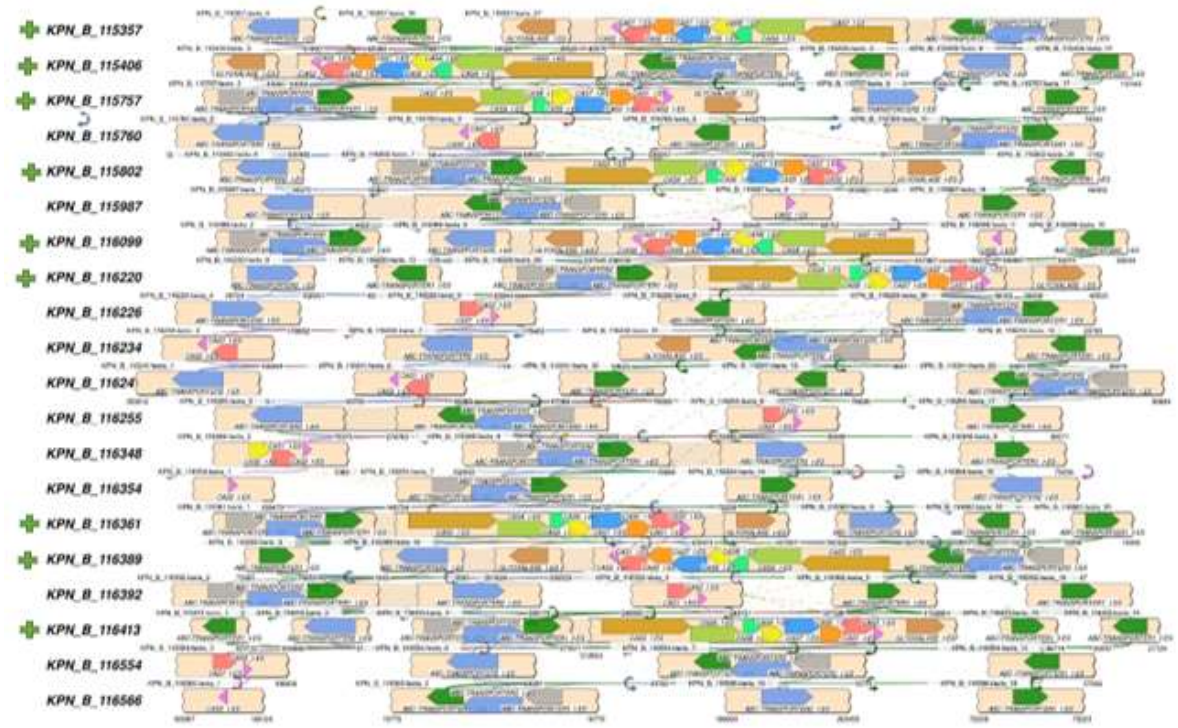

NORM_B_112178
NORM_B_112181
NORM_B_112490
NORM_B_112596
NORM_B_112598
NORM_B_112902
NORM_B_112925
NORM_B_112927
NORM_B_113282
NORM_B_113303
NORM_B_113315
NORM_B_113777
NORM_B_113895
NORM_B_113971
NORM_B_114048
NORM_B_114056
NORM_B_2007_48136
NORM_B_2011_73262

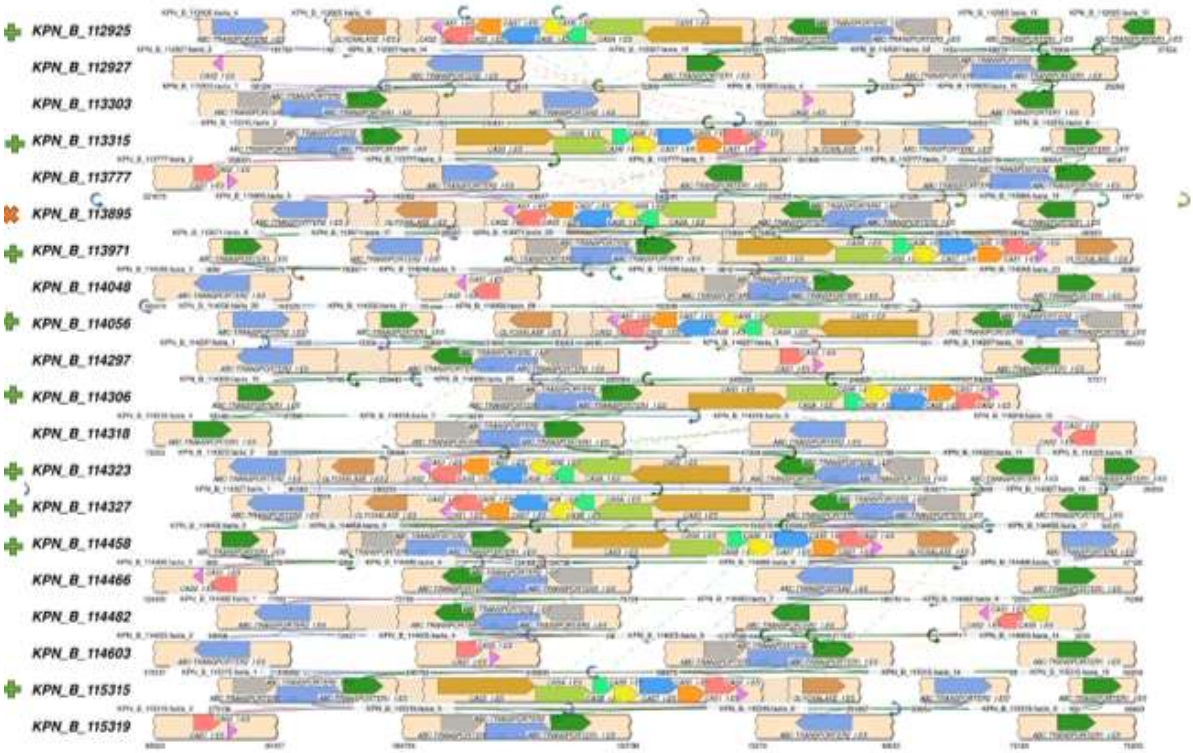

NORM_B_114297
NORM_B_114306
NORM_B_114323
NORM_B_114327
NORM_B_114330
NORM_B_114458
NORM_B_114466
NORM_B_114482
NORM_B_114603
NORM_B_115315
NORM_B_115319
NORM_B_115706
NORM_B_116407
NORM_B_117546


KPN_B_2005_1167
KPN_B_2005_3083
KPN_B_2007_48102
KPN_B_2007_48114
KPN_B_2008_52120
KPN_B_2008_54468
KPN_B_2009_61035
KPN_B_2009_61298
KPN_B_2009_61895
KPN_B_2009_62208
KPN_B_2009_62777
KPN_B_2009_63475
KPN_B_2009_63477
KPN_B_2010_72017
KPN_B_2010_72019
KPN_B_2010_72046
KPN_B_2011_72744
KPN_B_2011_73881
KPN_B_2011_74835
KPN_B_2011_74841
KPN_B_2011_75117
KPN_B_2011_75140
KPN_B_2011_76199
KPN_B_2011_76205
KPN_B_2012_83884
KPN_B_2012_84379
KPN_B_2012_85827
KPN_B_2012_86927
KPN_B_2012_86987
KPN_B_2013_90216
KPN_B_2013_90871
KPN_B_2014_101477
KPN_B_2014_101821
KPN_B_2014_101841
KPN_B_2014_102946
KPN_B_2014_103384
KPN_B_2014_104014
KPN_B_2014_104302
KPN_B_2014_104360
KPN_B_2014_96855


KPN_B_108846
KPN_B_108849
KPN_B_108853
KPN_B_108856
KPN_B_110764
KPN_B_110766
KPN_B_110770
KPN_B_111590
KPN_B_112176
KPN_B_112181
KPN_B_112596
KPN_B_112904
KPN_B_112914
KPN_B_2015_108843
KPN_B_2015_112126
KPN_B_2015_113891
KPN_B_2015_114050
KPN_B_2015_115359
KPN_B_2015_116374
KPN_B_2015_116588


KPN_B_116574
KPN_B_116585
KPN_B_116598
KPN_B_116604
KPN_B_117519
KPN_B_117530
KPN_B_117571
KPN_U_2003_10010
KPN_U_2009_56724
KPN_U_2009_56948
KPN_U_2009_57376
KPN_U_2012_81473
KPN_U_2013_95439
KPN_U_2013_95505
KPN_U_2013_95579
KPN_U_2013_95819
KPN_U_2014_100377
KPN_U_2014_97671
KPN_U_2015_10922
KPN_U_2015_109238
KPN_U_2015_111209


KPN_B_115357
KPN_B_115406
KPN_B_115757
KPN_B_115760
KPN_B_115802
KPN_B_115987
KPN_B_116099
KPN_B_116220
KPN_B_116226
KPN_B_116234
KPN_B_116241
KPN_B_116255
KPN_B_116348
KPN_B_116354
KPN_B_116361
KPN_B_116389
KPN_B_116392
KPN_B_116413
KPN_B_116554
KPN_B_116566


KPN_B_112925
KPN_B_112927
KPN_B_113303
KPN_B_113315
KPN_B_113777
KPN_B_113895
KPN_B_113971
KPN_B_114048
KPN_B_114056
KPN_B_114297
KPN_B_114306
KPN_B_114318
KPN_B_114323
KPN_B_114327
KPN_B_114458
KPN_B_114466
KPN_B_114482
KPN_B_114603
KPN_B_115315
KPN_B_115319