# MECANISMO DE DISEMINACIÓN Y TRAZADO DE LA HISTORIA EVOLUTIVA DE LA ISLA GENÓMICA GIE492 EN *KLEBSIELLA PNEUMONIAE*

**Tesis**

**Entregada A La Universidad De Chile**

**En Cumplimiento Parcial De Los Requisitos**

**Para Optar Al Grado De**

**Magíster en Ciencias Biológicas**

**Facultad De Ciencias**

**Por**

**Patricio Andrés Arros Muñoz**

**Marzo, 2022**

**Director de Tesis: Dr. Andrés Marcoleta Caldera**

**Co - Director de Tesis: Dra. Rosalba Lagos Mónaco**

**FACULTAD DE CIENCIAS**

**UNIVERSIDAD DE CHILE**

**I N F O R M E  D E  A P R O B A C I O N  T E S I S  D E  M A G Í S T E R**

Se informa a la Escuela de Postgrado de la Facultad de Ciencias que la Tesis de Magister presentada por el candidato.

**Patricio Andrés Arros Muñoz**

Ha sido aprobada por la comisión de Evaluación de la tesis como requisito para optar al grado de Magíster en Ciencias Biológicas, en el examen de Defensa Privada de Tesis rendido el día …….

**Director de Tesis:**

Dr. Andrés Marcoleta Caldera ----------

**Co – Director de Tesis:**

Dra. Rosalba Lagos Mónaco ----------

**Comisión de Evaluación de la Tesis**

Dr. Francisco Chávez ----------

Dr. Miguel Allende ----------

For Juan Fernández (RIP) and Alexis Pérez

## BIOGRAPHICAL SUMMARY

My name is Patricio Andrés Arros Muñoz, 25 years old and born in Talagante, Chile. I live with both my parents and an older brother. Our life has been humble but pleasant. During 14 years I attended the same subsided private school where I became interested in maths, language learning and science. I never hesitated to expecialize in biology in college, so after finishing school, I applied to the Bachelor's Degree in Biology program in Universidad de Chile. Currently, as a Bachelor of Science, and six years studying Biology in University, I think I have made the best decision.

## ACKNOWLEDGMENTS

I want to express my gratitude to the Integrative Microbiology Group (GMI) and to the Structural and Molecular Biology Laboratory (BEM) for their constant support, constructive criticism and for giving me a place where I can express my interests and ideas.

We would like to thank Dr. Francisco Chávez for his support and for giving us access to his laboratory equipment. Also, we want to thank Dr. Miguel Allende for his support with the materials required for Nanopore sequencing.

This project was funded by the grant FONDECYT 11181135 (Andrés Marcoleta).

**TABLE OF CONTENTS**

## LIST OF TABLES

## LIST OF FIGURES

## ABBREVIATIONS

| | |
|---|---|
| **aa** | Amino acid |
| **ABC** | ATP-binding cassette |
| **ACD** | Average Coverage Depth |
| **AMR** | Antimicrobial Resistance |
| **asn** | Asparagine |
| **ATOP** | asn-tDNAs Occupation Pattern |
| **ATP** | Adenosine triphosphate |
| **bp** | Base Pairs |
| **BSR** | BLAST score Ratio |
| **CG** | Clonal Group |
| **cgMLST** | core-genome MLST |
| **CRKp** | Carbapenem Resistant Kpn |
| **DNA** | Deoxyribonucleic acid |
| **DOR** | Diagnostic Odds Ratio |
| **Ent** | Enterobactin also known as Enterochelin |
| **ESBL** | Extended Spectrum β-Lactamase |
| **GC Content** | Guanine-Cytosine content |
| **gDNA** | Genomic DNA |
| **GI** | Genomic Island |
| **Gv1** | Genome Database version 1 |
| **Gv2** | Genome Database version 2 |
| **Gv2_GIE492** | Genomes from Gv2 carrying GIE492 |
| **HGT** | Horizontal Gene Transfer |
| **HTH-type** | Helix-Turn-Helix-type |
| **hvKp** | Hypervirulent Kpn |
| **ICE** | Integrative Conjugative Element |
| **Kae** | *K. aerogenes* |
| **Kgr** | *K. grimontii* |
| **Khu** | *K. huaxiensis* |
| **Kmi** | *K. michiganensis* |

| | |
|---|---|
| **Kor** | *K. ornithinolytica* |
| **Kox** | *K. oxytoca* |
| **Kp1** | *K. pneumoniae sensu stricto* |
| **Kp2** | *K. quasipneumoniae subsp quasipneumoniae* |
| **Kp3** | *K. variicola subsp variicola* |
| **Kp4** | *K. quasipneumoniae subsp similipneumoniae* |
| **Kp5** | *K. variicola subsp tropica* |
| **Kp6** | *K. quasivariicola* |
| **Kpa** | *K. pasteurii* |
| **Kpl** | *K. planticola* |
| **Kpn** | *Klebsiella pneumoniae* |
| **KpSC** | *Klebsiella pneumoniae* Species Complex |
| **KpVP** | Kpn Virulence Plasmid |
| **Kte** | *K. terrigena* |
| **LB** | Luria-Bertani |
| **LPS** | Lipolysaccharide |
| **MAPQ** | Mapping Quality |
| **maxcc-MSA** | Maximum column confidence MSA |
| **Mcc** | Microcin |
| **MDR** | Multidrug Resistance |
| **MGE** | Mobile Genetic Element |
| **MLST** | Multi Locus Sequence Typing |
| **MSA** | Multiple Sequence Alignment |
| **MST** | Minimum Spanning Tree |
| **NA** | Not Available |
| **NCBI** | National Center for Biotechnology Information |
| **ND** | Not Determined |
| **NGS** | Next Generation Sequencing |
| **nLV** | Number of Loci Variants |
| **NTP** | Nucleotide triphosphate |
| **ONT** | Oxford Nanopore Technologies |
| **oriT** | Transfer Origin |
| **PBS** | Phosphate Buffered Saline |
| **PCR** | Polymerase Chain Reaction |
| **QC** | Quality Check |
| **qPCR** | Quantitative PCR |
| **qRT-PCR** | Quantitative reverse transcription PCR |
| **RNA** | Ribonucleic acid |
| **rRNA** | Ribosomal RNA |

| | |
|---|---|
| **RS** | Resistance Score |
| **ST** | Sequence Type |
| **SV** | Structural variant |
| **T4SS** | Type IV Secretion System |
| **TAE** | Tris-acetate-ethylenediaminetetraacetic acid |
| **tDNA** | tRNA coding sequence |
| **tmDNA** | tmRNA coding sequence |
| **tmRNA** | Transfer-messenger RNA |
| **tRNA** | Transfer RNA |
| **UDP-glucose** | Uridine-5-diphospoglucose |
| **VS** | Virulence Score |

## ABSTRACT

*Klebsiella pneumoniae* (Kpn) is considered an urgent health threat due to the emergence of multidrug-resistant and hypervirulent strains attributed to the acquisition of mobile genetic elements (MGEs). One MGE possibly associated with hypervirulent strains is the genomic island GIE492, allowing the production of both, the antimicrobial peptide microcin E492 and salmochelin siderophore. However, GIE492 distribution and its dissemination mechanism remain unknown. Here, we determined the prevalence of GIE492 in different lineages of *Klebsiella*, the existence of structural variants, and the loci where it is integrated. In addition, we searched for evidence of its dissemination through horizontal gene transfer and possible co-mobilization with the ICE*Kp* conjugative element family.

A total of 3878 *Klebsiella* genomes were searched for GIE492, from which 6.8% had the island (all *K. pneumoniae* except for one *K. michiganensis*). Analysis of GIE492 indicated it has up to 23 coding sequences, which in total showed 53 alleles that were used to create a classification scheme for this MGE, identifying 4 structural variants. Of the total GIE492+ strains, 71% of them showed co-occurrence with ICE*Kp*. GIE492 and ICE*Kp* integrated into one of the four asparagine tDNAs of the chromosome (*asn1A*-*D*), the former with a strong bias towards *asn1C*. GIE492+ strains belonged to 17 distinct

clonal groups. This allows inferring that the structural variant has been mainly vertically inherited. However, we found evidence that during the evolution of *Klebsiella*, multiple events of acquisition by horizontal gene transfer has occurred: the sporadic presence of GIE492 in different branches of the *Klebsiella* phylogeny and the finding of at least two cases where the island is integrated into a tDNA other than *asn1C*.

We tested the presence of GIE492 SV I and ICE*Kp*10 as a predictor for the hypervirulent phenotype in *K. pneumoniae* finding a strong relationship. We also developed a new strategy based on next-generation sequencing to detect genomic islands excision events and then used it to detect the excision of both GIE492 and ICE*Kp*10 in *K. pneumoniae* SGH10.

## RESUMEN

*Klebsiella pneumoniae* (Kpn) se considera una amenaza urgente para la salud debido a la aparición de cepas multirresistentes e hipervirulentas atribuidas a la adquisición de elementos genéticos móviles (MGE). Un MGE posiblemente asociado a las cepas hipervirulentas es la isla genómica GIE492, que permite la producción del péptido antimicrobiano microcina E492 y del sideróforo salmochelina. Sin embargo, la distribución de GIE492 y su mecanismo de diseminación siguen siendo desconocidos. Aquí, determinamos la prevalencia de GIE492 en diferentes linajes de *Klebsiella*, la existencia de variantes estructurales y los loci donde ocurre la integracion. Además, buscamos evidencias de su diseminación a través de transferencia genética horizontal y la posible co-movilización con la familia de elementos conjugativos ICE*Kp*.

Se buscó GIE492 en un total de 3878 genomas de *Klebsiella*, de los cuales el 6,8% tenía la isla (todos *K. pneumoniae* excepto uno *K. michiganensis*). El análisis de GIE492 indicó que tiene hasta 23 secuencias codificantes, que en total presentaron 53 alelos, los que se utilizaron para crear un esquema de clasificación para esta GI, identificando 4 variantes estructurales. Del total de cepas GIE492+, el 71% mostró co-ocurrencia con ICE*Kp*. GIE492 e ICE*Kp* se integraron en uno de los cuatro tDNAs de asparagina del cromosoma (*asn1A-D*), la primera con un fuerte sesgo hacia *asn1C*.

Las cepas GIE492+ pertenecen a 17 grupos clonales distintos. Esto permite inferir que la variante estructural se ha heredado principalmente de forma vertical. Sin embargo, encontramos evidencias de que durante la evolución de *Klebsiella* se han producido múltiples eventos de adquisición por transferencia genética horizontal: la presencia esporádica de GIE492 en diferentes ramas de la filogenia de *Klebsiella* y el hallazgo de al menos dos casos en los que la isla está integrada en un tDNA distinto de *asn1C*.

Probamos la presencia de GIE492 SV I e ICE*Kp*10 como predictor del fenotipo hipervirulento en *K. pneumoniae* encontrando una fuerte relación. También desarrollamos una nueva estrategia basada *next generation sequencing* para detectar eventos de escisión de islas genómicas, y luego la utilizamos para detectar la escisión tanto de GIE492 como de ICE*Kp*10 en *K. pneumoniae* SGH10.

# INTRODUCTION

## 1. The genomics era and the increasing concern of infectious diseases and antimicrobial resistance

The current era in which we live is a complex association of nature's biological characteristics and processes correlated to the environmental impact of social development. Thus, it is crucial to consider holistic approaches when investigating biological phenomenons. In this context, one process that has gained critical importance in microbiology during the twentieth century and recent years corresponds to the emergence or re-emergence of pathogens with new or exacerbated phenotypic traits, leading to new infectious diseases (Jones *et al.,* 2008). This phenomenon would be favored by our sustained intervention of the ecosystems where microorganisms develop due to routinary modern society practices, which commonly leads to a decrease in biodiversity, favoring the growth of only specific taxa (Skovgaard, 2007).

For example, during the mid-1960s, *Yersinia enterocolitica* emerged as a new pathogen that gained critical importance during the following decades. One of the main features of *Yersinia* is its ability to multiply at temperatures close to 0 ºC, which was greatly favored by the developing practice of keeping the *chilling chain* of food, using vacuum and refrigeration for food preservation. These technologies and practices

favored the growth of this bacterial species and the corresponding increase of human yersiniosis due to contaminated food ingestion (Skovgaard, 2007).

The complex microbial population dynamics determined by anthropogenic and natural factors make pathogen emergence and dissemination a fructiferous and dynamic research topic in microbiology. In particular, intensive research focuses on knowing pathogen diversity and understanding molecular pathogenesis to guide epidemiological and therapeutical actions. Moreover, significant gaps remain in understanding how genomic changes contribute to the arising of clinically relevant traits among pathogens and how medical procedures to treat bacterial infections impact microbial genotype and phenotype.

The standard therapy for bacterial infections is administering a specific antibiotic cocktail to the patient. However, the effectiveness of this strategy is rapidly decreasing, as it also acts as a stimulus for the development, activation, and selection of antibiotic resistance mechanisms in the bacterial population. Moreover, the diversified and massive use of antibiotics has favored the development of multidrug-resistant (MDR) bacteria, which can resist multiple structurally unrelated antimicrobials with different molecular targets (Méndez-Vilas, 2013). In this context, the spreading of such MDR bacteria had led to the so-called global antimicrobial resistance crisis, causing high mortality rates and high medical costs, posing a serious worldwide menace to public health (Tanwar *et al.,* 2014).

In the presence of antibiotics or other toxic substances, bacterial cells face a trade-off between the potential benefits of an increased mutation rate and diminishing the harmful effects of deleterious mutations by activating dedicated molecular systems

under stress conditions such as the presence of antibiotics in the environment (McKenzie & Rosenberg, 2001). Moreover, antimicrobial therapy could promote the acquisition or loss of distinct mobile genetic elements, potentially impacting antibiotic resistance, virulence, and cell fitness (Anderson, 2003).

The biological cost is a crucial parameter to consider when the dynamics of an antibiotic resistance mechanism is modeled, in terms that this concept regulates the possibility of antibiotic resistance reversibility (Levin, 2002). For example, if a resistance plasmid or a mutation reduces the fitness of resistant strains, they will be outcompeted by susceptible cells in the absence of antibiotic selection. Thus, the clearance of the antibiotic, the differences in fitness between both types of bacterial populations, and the possibility that the resistance mechanism is transferred between populations are the key factors that regulate the emergence and persistence of antibiotic-resistant bacterial populations in a community, as previously modeled (Austin *et al.,* 1999, Cohen *et al.*, 2003). All this evidence points out that there is a close relationship between the arising of pathogens with increased virulence and resistance and the evolution of their genomes due to intrinsic changes or horizontal gene transfer. For this reason, the development of massive or “next-generation” DNA sequencing (NGS) technologies and genomic approaches have revolutionized the study of the epidemiology and evolution of pathogen populations.

On the other hand, classical culture-dependent pathogen detection and surveillance approaches have significant limitations. Most bacteria do not grow in conventional culture broths, and some of the species that can be cultured may present a viable but not a cultivable state. For example, *Campylobacter* can retain its virulence

and stop multiplying, acquiring a particular coccoid form under unfavorable environmental conditions. This viable but not cultivable state hampers detecting these organisms, especially in highly treated foods (Oliver, 2000). In this aspect, NGS and (meta)genomic approaches can provide significantly more detailed and less biased information regarding pathogen presence and molecular features.

NGS has many uses, and the most important ones are to be able to detect relevant pathogens, to be able to discover new microorganisms without the need to cultivate them, and to be able to characterize these microbes in a detailed, comparative, and in a high throughput way. Currently, NGS has become the gold standard for many procedures in bacteriology research (MacCannell, 2016), having the potential to determine the causative agents of infectious diseases and the epidemiology and evolution of various infecting pathogens inside healthcare settings and the community (Sintchenko & Holmes, 2015).

NGS technologies can complement different steps of a classical clinical diagnosis cycle (i.e., patient examination and sample collection, pathogen isolation and culturing, biochemical identification, serological tests, directed molecular analysis, antibiotic susceptibility testing, and interpretation of results); but can also reduce the diagnosis time itself by doing all the wet analyses in tandem, and providing the researcher of more detailed data, including subtyping, resistome, and virulome typing, phenotypic inference, detection of genetic variants, mobile genetic elements, toxins, among others (Didelot *et al.*, 2012).

A typical NGS workflow in a clinical environment consists of the following: sample collection and preparation, nucleic acid extraction, NGS library preparation,

sequencing, data analysis, and data storage (Gilchrist *et al.,* 2015). The rapid advances in NGS technologies and their capabilities have proven and further promise to be a game-changer in diagnostic microbiology, significantly reducing the time from diagnosis to clinical treatment while also broadening the toolbox of clinical and public health microbiology laboratories, physicians, and medical decision-makers (Motro & Moran-Gilad, 2017).

## 2. *Klebsiella pneumoniae*, an urgent threat to global health

### 2.1 *K. pneumoniae* general features

*Klebsiella,* originally denominated Friedlander's bacillus, is a bacterial genus from the *Enterobacteriaceae* family, which initially grouped four main species: *K. pneumoniae, K. ozaenae, K. rhinoscleromatis,* and *K. oxytoca* (Buchanan & Gibbons, 1974). All *Klebsiella* species are normally found in soil, water, and in the mouth, nose, and intestines of humans and animals. Moreover, it corresponds to the second most abundant bacterial genus in the human gastrointestinal tract (Ristuccia & Cunha, 1984).

Gram-negative and typically non-motile, the shorter and thicker rods of this genus are easily distinguishable from the rest of the members of *Enterobacteriaceae*. The species of this genus are aerobic and facultative anaerobic that grow in conventional laboratory media such as blood agar, where grayish-white convex mucoid colonies can be seen due to a capsule made of polysaccharides, the K antigen (Wilson, 1975). The K antigen is the main component that defines *Klebsiella* capsule serotyping. *Klebsiella* bacteriocin production is also a useful way to typify strains as it was originally described by Hamon and Peron (1960).

*Klebsiella* species are considered opportunistic pathogens colonising mucosal surfaces without causing pathology; however, from mucosae *Klebsiella* may disseminate to other tissues causing life-threatening infections including pneumonia, urinary tract infections, bloodstream infections and sepsis (Paczosa & Mecsas, 2016).

All *K. pneumoniae* (Kpn) strains harbor a subset of chromosomally encoded virulence factors required for the establishment of opportunistic infections in mammalian hosts, including the siderophore enterobactin (Ent), the genes *fim* and *mrk* encoding type 1 and type 3 fimbriae, as well as the capsule and LPS (O antigen) biosynthesis loci (Follador *et al.,*2016). Ent is also required for organisms growing in most niches (Bachman *et al.,* 2012). Also, as part of the Kpn accessory genome, the *ybt* locus encoding the siderophore yersiniabactin is statistically associated with infections, being present in 30-40% of all human clinical Kpn isolates (Lam *et al.,* 2018a). This locus is usually mobilized by an integrative conjugative element (ICE) called ICE*Kp* that integrates into an asparagine (asn) tRNA gene, but also occasionally in plasmids (Lin *et al.,* 2008).

*K. pneumoniae* has been declared an urgent threat due to the frequent detection of MDR strains causing outbreaks globally (Holt *et al.,* 2015). Moreover, the emergence of MDR Kpn strains and the intricacies in treating their infections made the World Health Organization give Kpn a priority status as a target for new therapies and antibiotics discovery research (World Health Organization, 2017). In this sense, especially problematic strains correspond to those producing extended-spectrum β-lactamases, carbapenemases, and determinants for colistin resistance, prompting an increased reliance on colistin and β-lactam/β-lactamase inhibitors combinations

(Petrosillo *et al.*, 2019, Tooke *et al.,* 2019). In addition, population genomics studies have revealed that the rising of MDR Kpn has correlated with both horizontally acquired antimicrobial resistance (AMR) genes and mutations in core genes (Wyres & Holt, 2018).

Even though Kpn has been traditionally associated with opportunistic healthcare infections, this organism has currently gained relevance as the causative agent of severe community-acquired infections, including pyogenic liver accesses and endophthalmitis (Siu *et al.,* 2012). These invasive infections have been mainly associated with the so-called hypervirulent Kpn (hvKp) strains commonly reported in East or Southeast Asia and individuals with southeast Asian ancestry (Shon *et al.,* 2013).

## 2.2 Factors, mobile genetic elements, and lineages associated with *K. pneumoniae* virulence and antimicrobial resistance

The hvKp strains accumulate a distinctive set of characteristics and virulence factors, including a type K1, K2, or K5 capsular polysaccharide, mobile genetic elements encoding determinants for the production of aerobactin (*iuc* genes) and salmochelin (*iro* genes) siderophores, the genotoxin colibactin (*clb* genes) and the hypermucoviscous phenotype conferred by the *rmpADC* locus (Brisse *et al.*, 2019, Walker *et al.*, 2019, Wyres *et al.*, 2019, Walker *et al.,* 2020, Wyres *et al.,* 2020).

The *iro* and *iuc* genes are mainly mobilized in large (>100 kbp) FIBK plasmids known as *K. pneumoniae* virulence plasmids (KpVPs). The two main KpVPs associated with hypervirulent strains (KpVP-I and KpVP-II) have been extensively characterized

due to their association with these types of clones (Lam *et al.,* 2018c). KpVPs are non-conjugative, but they are mobilized between bacterial cells by other conjugative elements or through mosaic conjugative plasmids containing KpVP coding sequences (Nassif *et al.,* 1989, Yang *et al.,* 2019).

Colibactin corresponds to a genotoxic polyketide that induces DNA damage in eukaryotic cells, mainly encoded in a specific variant of the ICE*Kp* family, ICE*Kp*10, by the *clb* locus (Nougayrède *et al.,* 2006, Lam *et al.,* 2018a). It has been proposed that colibactin promotes gut and mucosal colonization and dissemination through the blood and other organs, although in Kpn, it is mainly associated with liver abscess infections (Lam *et al.,* 2018a) and some types of colorectal cancer (Lai *et al.,* 2014).

In contrast with previously described MDR clones, hvKp strains are rarely MDR and remain susceptible to most drugs, except for ampicillin, antibiotic to which Kpn is intrinsically resistant due to a chromosomally encoded SHV β-lactamase (Holt *et al.,* 2015) and to a lesser degree to quinolones and fosfomycin, due to the *oqxAB* and *fosA* genes respectively, also chromosomally encoded (Ito *et al.,* 2017, Li *et al., 2019*). However, hvKp carrying MDR encoding plasmids and co-occurrence of AMR and virulence factors in non-hvKp strains have all been reported (Gu *et al.,* 2018, Zhang *et al.,* 2016). This convergence of MDR and hypervirulence potentiates the emergence of both invasive and hard-to-treat infections, an alarming and urgent problem for public health, especially when at least one outbreak of carbapenemases-producing hvKp in China has already occurred, and patients with this type of infections become more common by the day (Xu *et al.,* 2019).

Recent Kpn population structure studies (summarized in Wyres *et al.,* 2020) have revealed clear and deep-branching phylogenetic lineages, delineated mainly by a seven gene Multi Locus Sequence Typing (MLST) scheme (Diancourt *et al.,* 2005) comprising sequence types (STs), and whole-genome phylogenetic analyses defining clonal groups (CGs). Some of these lineages have been deeply characterized and associated with specific types of Kpn, such as CG258, CG15, and ST307, which host a high number of AMR genes (MDR clones) and are mainly linked to healthcare-associated infections and hospital outbreaks worldwide (Wyres & Holt, 2016). In contrast, lineages such as CG23, CG65, and CG86 carry most of the virulence factors previously described and have been designated as hvKp clones. The phylogenetic distinction between MDR and hvKp lineages is also correlated with the capsular and LPS antigens (K and O): while MDR clones are highly diverse due to homologous recombination, hvKp rarely deviate from K1, K2, and K5, as previously mentioned.

It is relevant to point out that, as revealed by genomic analyses, strains classified by biochemical or mass spectroscopy methods as *K. pneumoniae* are actually part of a closely related group of species and subspecies, denominated *Klebsiella pneumoniae* Species Complex (KpSC). The members of KpSC differ by 3-4% of nucleotide divergence across the core chromosomal genes but share a set of AMR and virulence genes.

However, non-Kpn members of KpSC have been reported to have a significantly lower disease burden than *K. pneumoniae sensu stricto* (10-20%) (Long *et at.,* 2017, Potter *et al.,* 2018). Furthermore, genomic analyses also revealed that *K. ozaenae* and

*K. rhinoscleromatis*, classically defined as distinct species, are specific lineages of Kpn, CG90, and CG3, respectively (Brisse *et al.,* 2009).

In terms of genomic structure, typical Kpn genomes are 5-6 Mbp in size and encode ~5000-6000 genes, from which approximately 1700 are part of the core genome while the rest are part of the accessory genome of the strain (Holt *et al.,* 2015).

Considering the vast number of lineages of Kpn and how each contributes to the total gene pool, it is estimated that the complete Kpn pan-genome is likely to exceed 100000 protein-coding sequences.

Kpn clones can be distinguished based on the accessory genome content, as it results from the clone-specific niche-adaptation through horizontal gene transfer (HGT) (Mclnerney *et al.,* 2017). However, the HGT phenomenon can also happen between clones of the same lineage, mainly driven by chromosomal recombination, plasmid conjugation, phage transduction, and the action of Integrative Conjugative Elements (ICEs) (Wyres *et al.,* 2019, Wyres *et al.,* 2015).

A diversified pool of plasmids has been sequenced and characterized in Kpn, but incompatibility types IncFIIK and IncFIBK tend to be the most prevalent (Navon-Venezia *et al.,* 2017). In contrast with other members of the *Enterobacteriaceae* family, Kpn seems to be particularly permissive in terms of plasmid uptake, being not uncommon to detect strains carrying up to 10 plasmids, a number significantly higher than in the rest of the ESKAPE pathogens (Wyres & Holt, 2018), a faction of AMR bacteria including *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacter* spp., highlighted by the Infectious Diseases Society of America for representing new

paradigms in pathogenesis, transmission and resistance (Rice, 2008, Cerit, 2020). Furthermore, Kpn genomes tend to carry multiple prophages, many of which have been sequenced and characterized (Lam *et al.,* 2018b).

Because of how diverse, intricate, and clinically relevant the Kpn population is, some of the widely distributed clones that contribute disproportionately to the global disease burden have been extensively researched and denominated “global problem clones” (Wyres *et al.,* 2020). As an example, CG258, CG15, CG20, CG29, CG37, CG147, CG101, and CG307, even though they lack a phylogenetic relationship, all of them form part of the “global MDR problem clones”, which include the vast majority of carbapenem-resistant Kpn (CRKp) and third-generation-cephalosporin resistant Kpn (David *et al.,* 2019).

In terms of hvKp, CG23, CG65, CG25, CG66, and CG380, to a lesser extent, dominate hypervirulent infections worldwide. CG23 has been identified as one of the two topmost prevalent Kpn associated with blood infections in China, Vietnam, and Laos, accounting for >10% of all isolates (Siu *et al.,* 2011, Lee *et al.,* 2016). CG23 has several sublineages, where CG23 sublineage I (CG23-I) have become globally distributed, and it includes approximately 82% of all liver abscess strains (Lam *et al.,* 2018b). All CG23 isolates are associated with the K1 capsule phenotype, *ybt* lineage 1 and *clb* lineage 2, both encoded in the ICE*Kp*10 element, and most of the isolates also present a KpVP encoding *iuc, iro, rmpA,* and *rmpA2.* Other characteristics of CG23 are the presence of a genomic island encoding the virulence factor microcin E492 (named GIE492) (Marcoleta et al., 2016; Lam et al., 2018b), and the chromosomal allantoinase locus (Bialek-Davenet *et al.,* 2014, Struve *et al.,* 2015, Lam *et al.,* 2018b). Due to the

clinical importance of CG23-I, the Kpn SGH10 strain has been selected as a reference genome for CG23 hvKp, and its genomic sequence has been sequenced and assembled (Lam *et al.,* 2018b).

## 3. Genomic islands and their role in *K. pneumoniae* pathogenesis and genome evolution

### 3.1 General features of Genomic Islands from Gram-negative bacteria

The current era of bacteriology and the increasing number of prokaryotic gene and genome sequences has revealed that the exchange of genetic information through HGT and homologous recombination have been greatly underestimated in both quantity and quality (Dobrindt *et al.,* 2004). This revelation implies a need to change the classical focus of bacterial evolution, making it broader and not only relying on clonal divergence and periodic selection. So, gene acquisition, loss, and other genomic alterations have an essential role in adaptative prokaryotic evolution.

In the context of evolutive genomic alteration and variability, unique insights into the role and contribution of accessory and mobile genetic elements (MGEs) have been developed. MGEs in bacteria are part of the so-called accessory genome of a species representing a flexible gene pool encoding additional traits that can be favorable in specific environmental conditions, such as AMRs and virulence factors, and the acquisition of them by HGT allows the inheritance of complex disease-related characteristics in a single step (Ochman *et al.,* 2000). MGEs come in different flavors distinguished by their genetic structure, including bacteriophages, plasmids, genomic islands, insertion sequences, transposons, and integrons.

A specific type of MGE, formerly known as pathogenicity islands (Hacker *et al.,* 1990, Lee, 1996) has gained much attention in recent years due to their key role in chromosomal evolution in all types of bacteria, including commensal, environmental, and symbiotic organisms, not only pathogens as initially described, receiving the more general name of “Genomic Islands” (GIs) (Dobrindt *et al.,* 2004).

GIs are chromosomal segments variably present among related strains, which typically show a G+C content that differs from the rest of the chromosome, frequently integrate into tRNA and tmRNA genes, and are often flanked by direct repeat sequences (Dobrindt *et al.,* 2004, Mao *et al.,* 2009). For convenience, genes encoding tRNAs and the tmRNA will be referred to as “t(m)DNAs”. In addition, GIs tend to be unstable and often contain mobility genes, such as integrases or transposases (Dobrindt *et al.,* 2002). The integration and excision of these elements occur by site-specific recombination catalyzed by the GI-encoded integrase, which usually recognizes part of the t(m)DNA sequence as a recombination site. The direct repeats flanking the GIs arise from the duplication of the recombination site due to the integration process.

GIs can mediate the transfer of up to hundreds of genes encoding one or more metabolic routes or functional capabilities, potentially allowing for a successful adaptation and increased fitness in a particular ecological niche. In Bacteria, the acquisition of foreign genetic elements is frequently counterbalanced by the loss of native genes, which can result in adaptative advantage (Maurelli *et al.,* 1998). In the case of symbionts or obligate intracellular bacteria, gene loss occurs with higher frequency, empathizing that genomic reduction and HGT mediate genome optimization,

which relies on the lifestyle of the microorganism itself (Moran, 2002). The presence of virulence factors inside GIs is not necessary, but the original description as pathogenicity islands implied their presence. Virulence factors that have been encountered in GIs can be divided in several groups, including adherence factors, siderophores, capsular antigens, the LPS endotoxin in gram negative bacteria, exotoxins, invasins, and type III and IV secretion systems.

The origin of GIs is uncertain, though it was proposed that they would have evolved from prophages or plasmids that lost genes required for phage particle assembly, replication, and self-transfer, in exchange for a more stable association with the bacterial chromosome and a higher inheritance probability (Dobrindt *et al.,* 2004). Additionally, some GIs encode conjugative machinery and a transfer origin (oriT) that allow the element transfer by conjugation, commonly referred to as Integrative Conjugative Elements (ICEs), which can be included in the same group as conjugative transposons and plasmids. These elements normally include genes required for chromosomal excision, the formation of a circular intermediary, and the posterior conjugation and site-specific recombination (Burrus *et al.,* 2002).

According to the GI features described above, the hypothetical life cycle of these mobile elements could be described as follows. A lysogenic prophage or integrative plasmid is acquired by HGT, followed by chromosomal integration by site-specific recombination, and then inherited and maintained due to positive selection. Posterior genetic rearrangements can lead to deleting or inactivating genes related to phage particle assembly, mobilization, and replication, although keeping the integrase-coding genes. Afterwards, successive rearrangements and genetic modifications leading to

gene loss and gain could continue shaping the GI, which sometimes, upon excision, could be transferred to a new bacterial cell, starting the cycle once again.

### 3.2 The impact of genomic islands in *K. pneumoniae* evolution and virulence

Previous evidence indicates that GIs have a relevant role in *K. pneumoniae* genome evolution and are an important source of accessory genes that define specific characteristics of many lineages of this species complex (Marcoleta *et al.,* 2016, Lam et al., 2018b). According to previous work from our research group, Kpn chromosomes tend to have 86 tDNAs and one tmDNA that in theory could harbor a putative GI, although a reduced subset of them (roughly twenty) are largely preferred for GI integration (Berríos-Pastén *et al.*, 2020). This study also showed that the genomic context of these t(m)DNAs is highly conserved and would influence their usage as integration sites. Moreover, given this usage bias and the relevance of t(m)DNAs as hotspots for mobile elements insertion, we developed a consistent nomenclature for all the tDNAs in a Kpn genome which allows distinguishing identical tDNAs located in different conserved genomic contexts to facilitate further analyses in this topic (Berríos-Pastén *et al.*, 2020). According to this nomenclature, the tDNA name includes the three-letter code (in lowercase) of the amino acid transferred by the encoded tRNA, an Arabic numeral that is different for every anticodon specifying the same amino acid, and a capital letter that conveys the specific genomic context where the gene is located.

As introduced before, we demonstrated that the upstream genomic context of t(m)DNAs, defined as the 4-kbp DNA sequence upstream of these genes, and to a lesser extent the downstream context, tends to be highly conserved across *K.*

*pneumoniae* genomes. We leveraged this finding to develop a strategy to identify all the t(m)DNAs-associated GIs in a given *K. pneumoniae* genome, based on the structure of the genomic region where each t(m)DNA is normally located and detecting possible interruptions of the locus continuity due to GI integration (Berríos-Pastén et al., 2020). Using this approach, we studied the GI content in a set of 66 Kpn genomes from different lineages, discovering that a mean of six GIs exist in every Kpn genome, with sizes between 3.5 kbp to 133 kbp, and accounting for roughly 25% of the chromosomally encoded accessory genome. All this evidence supports a major role of GIs in Kpn evolution.

Although most of the proteins codified in these GIs have an unknown function, thus being considered hypothetical, many of the known Kpn virulence factors and AMR genes are encoded in them, including fimbriae assembly, type III and IV secretion systems, colibactin toxin, yersiniabactin, and salmochelin production, and drug efflux systems.

As previously described, among the most relevant virulence-related mobile elements in the Kpn population are the GIs from the ICE*Kp* family. ICE*Kp* elements are self-transmissible through conjugation, requiring the proteins encoded by *virB1* and *mob* genes and an oriT sequence. Also, they encode an integrase protein which catalyze the integration of the element in one of the four asparagine tDNAs present in the Kpn chromosome (Marcoleta *et al.,* 2016; Lam et al. 2018a). One of the most relevant characteristics of ICE*Kp* is that it carries the *ybt* locus, encoding proteins for the biosynthesis of yersiniabactin and its receptor (Lin *et al.,* 2008). The importance of yersiniabactin for the Kpn metabolism is that even if all Kpn produce enterobactin, this

siderophore and its iron scavenging functions are inhibited by the human protein lipocalin-2 inducing an inflammatory response (Goetz *et al.,* 2002, Bachman *et al.,* 2009), while yersiniabactin can avoid this effect enhancing bacterial growth and its dissemination in the spleen (Bachman *et al.,* 2011). In addition, some varieties of ICE*Kp* have an increased virulence factors load, also carrying *iro* (salmochelin siderophores) and *clb* (colibactin genotoxin) genes.

ICE*Kp* structural variants in Kpn share several features: a P4-like integrase gene, the *ybt* locus, a *xis* excisionase, a *virB* T4SS, a transfer origin (oriT), and *mobBC* proteins for mobilization. Additionally, a variable selection of cargo genes can be found. These genes were used to identify 14 distinct structural variants of ICE*Kp,* which correlate with only one of the clusters of the *ybt* locus phylogeny (Lam *et al.,* 2018a). The only exception to this rule is ICE*Kp*10, associated with 3 *ybt* lineages and carries the *clb* locus in its cargo region. ICE*Kp* elements circulate dynamically within the Kpn population, which is suggested by the high number of distinct ICE*Kp* acquisitions within individual Kpn CGs, indicating that the MGE is highly transmissible within the bacterial population (Lam *et al.,* 2018a).

Besides ICE*Kp*, a second GI that carries a Kpn virulence factor is GIE492 encoding determinants for producing the antimicrobial peptide microcin E492 (MccE492) and salmochelin siderophores (Marcoleta *et al.,* 2016).

## 4. The antibacterial peptide microcin E492 encoded in the GIE492 island, and its possible role in *K. pneumoniae* virulence

### 4.1 Microcin E492: an antibacterial peptide produced by *K. pneumoniae*

MccE492 is a pore-forming bacteriocin with activity against *Escherichia, Klebsiella, Salmonella, Citrobacter* and *Enterobacter* (de Lorenzo, 1984; Lagos *et al*., 1993). Like other microcins, MccE492 is a hydrophobic protein with high stability against extreme temperature or pH conditions and possesses strong antibacterial activity. These characteristics are believed to play an important role in shaping the community structure of the intestinal microbiome (Lagos *et al.,* 2009). MccE492 can be found in two different forms: unmodified or covalently linked to salmochelin molecules at serine 84 as a form of post-translational modification (Thomas *et al.*, 2004). This process requires the synthesis of an enterochelin molecule as a precursor of salmochelin, its glycosylation catalyzed by the MceC protein from the MccE492 cluster, and its subsequent linkage to the end of a specific ~12 amino acid region at the C-terminus of MccE492 (Nolan *et al.,* 2007).

MccE492 exerts its antibacterial effect by causing depolarization of the cytoplasmic membrane of the target cells in a process dependent on the presence of the protein TonB (De Lorenzo & Pugsley, 1985). Accordingly, MccE492 does not present antibacterial activity against *E. coli* strains carrying mutations in *tonB, exbB, manY, and manZ,* and against the triple mutant *fepA-fiu-cir* (De Lorenzo & Pugsley, 1985, Bieler *et al.,* 2006, Patzer *et al.,* 2003, Pugsley *et al.,* 1986, Pugsley, 1985). The outer membrane proteins FepA, Fiu, and Cir are the receptors necessary for MccE492 recognition (Hantke *et al.,* 2003, Patzer *et al.,* 2003, Strahsburger *et al.,* 2005). These

three receptors are used to transport iron-catecholate complexes in a TonB-dependent manner (Braun *et al.,* 2002, Hantke *et al.,* 2003). TonB, ExbB, and ExbD form the TonB system that works by transducing the proton motive force from the cytoplasmic membrane to the outer membrane (Postle & Kadner, 2003, Koebnik, 2005, Wiener, 2005). The inner membrane proteins ManYZ are the inner membrane component of the mannose permease system, and they physically interact both with MccE492 and its immunity protein (Bieler *et al.,* 2010).

MccE492, as other microcins, has a modular structure where each module has a distinct function (Azpiroz & Laviña, 2007). The N-terminal module is required for its export to the extracellular milieu, the central module determines the antibacterial activity, while the C-terminal module is necessary for the receptor recognition, which needs the salmochelin linkage mentioned above (Mercado *et al*., 2008). Moreover, the C-terminal region of MccE492 can be used to design molecules that exploit the Fe-siderophore uptake system as a cellular entry way. This strategy is called the “Trojan horse” mechanism of antibacterial activity, because the MccE492 structure mimics the essential siderophore element at the C-terminal, which can be recognized by the outer membrane receptors and translocated across the membrane to exert its action (Lagos *et al.,* 2009).

Among other salient features, MccE492 can induce apoptosis on human tumor cells lines, and has a demonstrated antitumoral effect in human colorectal cancer xenografts made in zebrafish, giving this peptide an interesting cancer treatment suitability because cell death by apoptosis induction does not involve an inflammatory response (Hetz *et al.,* 2002, Varas et al. 2021). Additionally, MccE492 can form amyloid

fibers, both *in vivo* and *in vitro,* which was proposed as a mechanism for toxin inactivation and storage, due to the high stability of this type of fibrils (Bieler *et al.,* 2005, Marcoleta *et al*., 2013b).

### 4.2 Genetic determinants for MccE492 production

In terms of the genetic structure of the determinants involved in MccE492 production, all these genes form part of a ~13-kbp cluster present in the chromosome of *K. pneumoniae* RYC492, a clinical strain from which this microcin was originally isolated (De Lorenzo *et al.,* 1984). The genes involved in this process are named *mce* (microcin E) followed by a capital letter to distinguish them (Lagos *et al.,* 2001). Other genetic elements present in the cluster that could regulate the expression of MccE492 but without a well-established function are simply named *orf* (open reading frame).

The *mceA* gene encodes the MccE492 precursor that includes an N-terminal leader peptide that is cleaved once the microcin is secreted to the extracellular space (Lagos *et al.,* 1999). *mceB* encodes a protein located in the inner membrane of producer cells that confers immunity to MccE492 (Lagos *et al.,* 2001). The way MceB neutralizes the action of MccE492 involves interaction with TonB (Lagos *et al.,* 2009).

*mceC* encodes an amino acid glycosyl transferase homologous to IroB of *Salmonella enterica.* This protein transfers the glucose moiety of uridine-5'-diphosphoglucose (UDP-glucose) to enterochelin, forming a molecule of salmochelin. *mceI,* encoding an amino acid acyltransferase, and *mceJ* without homology to any protein known, both are responsible of covalently binding the salmochelin molecule to the last amino acid of MccE492 (Nolan *et al.,* 2007). *mceD* encodes an esterase that

hydrolyzes the apo and ferric stages of salmochelin (Nolan *et al.,* 2007). Additionally, *mceE* encodes a protein with unclear function with a C-terminal region similar to MchS4, a protein of the MccH47 system that promotes enterochelin production (Azpiroz *et al.,* 2004).

On the other hand, MccE492 export depends on the ABC exporter encoded by *mceG* and the accessory protein encoded by *mceH,* where both act together with the outer membrane protein TolC (Lagos et al., 2001). Finally, the protein encoded by *mceF* has an unknown function, however, it possesses 75% similarity with its equivalent McmM from microcin H47 and M systems. There is no experimentally determined function to these proteins, but homologs are frequently found in pathogenicity islands (Dobrindt *et al.,* 2002, Lagos *et al.,* 2009).

Bacteriocins such as colicin G, colicin H or PAI III(536) have a similar genetic context to the MccE492 system (Braun *et al.,* 2002). These systems have open reading frames analogous to sequences also found in the MccE492 cluster, such as *orfK*, a gene encoding a protein where the last 56 amino acids are like McmA, the structural gene of MccM, and its first 75 amino acids have similarity to McmI, its immunity protein. However, even if OrfK is transcribed, its product does not confer immunity to MccM nor possess any antibacterial activity (Lagos *et al.,* 2009).

On the other hand, the N-terminal region of the peptide codified by *mceL* is similar to MccE492, MccH47, and MccM with a double-glycine motif compatible with a precursor that will be cleaved and exported through a dedicated ABC exporter. Similarly, the C-terminal region ends with the glycine-serine motif, also found in MccE492, MccH47 and MccM. The MceL peptide is also a target for salmochelin post-

translational attachment (Lobos, Hurtado, Marcoleta, unpublished results). This second microcin shares the exportation mechanism with MccE492 (unpublished results). Additionally, *orfS2* and *orfS3* are probably inactive equivalents of proteins present in the MccH47 system. *orfS2* and *orfS3* might encode truncated proteins with significant identity to MchS2 and MchS3, equivalent to MccI47 and its immunity protein, respectively. *mceX* is similar to *mchX* (Lagos *et al.,* 2009; Marcoleta *et al*., 2018).

### 4.3 Regulation of the MccE492 production and activity

Studies performed in Kpn RYC492 demonstrated that MccE492 is mainly produced in the exponential phase of growth and is positively affected by lactate or citrate in the culture medium (De Lorenzo, 1985). Even if *mceA* is translated both in exponential and stationary phases of bacterial growth, the inactivation of MccE492 during the latter phase is associated with the absence of transcription of the maturation genes *mceI* and *mceJ,* which are transcribed as a bicistronic unit (Corsini *et al.,* 2002). The loss of antibacterial activity during the stationary phase is also associated with the formation of MccE492 fibrils with amyloid properties, which lack antibacterial activity (Bieler *et al.,* 2005; Marcoleta et al., 2013b). *K. pneumoniae* RYC492 cultures also produce an antagonist to MccE492 when the nutrients of the medium have been depleted (De Lorenzo *et al.,* 1984), which corresponds to the non-ferric form of enterochelin (Orellana & Lagos, 1996).

MccE492 production is closely linked to siderophore production and iron metabolism through the action of the Ferric Uptake Regulator (Fur). In *K. pneumoniae*, *E. coli,* and other *Enterobacteriaceae,* Fur is a protein that acts as a transcriptional

repressor sensing iron availability: at high levels of $Fe^{2+}$ it acts as a co-repressor forming a Fur-$Fe^{2+}$ complex that binds to a 19 bp region denominated Fur box, usually located in the promoter region of iron-regulated genes (Bagg & Neilands, 1987, Escolar *et al.,* 1998, Chen *et al.,* 2007). In contrast, at low levels of $Fe^{2+}$, the Fur-$Fe^{2+}$ complex disassembles, allowing gene expression.

The MccE492 production cluster possesses putative Fur Boxes upstream of *mceC, mceD, mceE,* and *mceX* (Marcoleta *et al.,* 2018). However*,* Fur only binds in *vivo* to the region upstream of *mceX*, suggesting a role in controlling *mceX* and possibly *mceJI* expression, which was later confirmed, showing that *mceX* and *mceJI* are co-transcribed, sharing a unique transcription starting site located upstream of the Fur box (Marcoleta *et al.,* 2018). Due to the direct regulation of Fur over the maturation genes of MccE492, high $Fe^{2+}$ concentrations indirectly down-regulate the activity of MccE492. Furthermore, Fur also represses the expression of genes required for enterochelin synthesis, limiting its availability for salmochelin synthesis, acting as a double negative regulator over MccE492 maturation and activity (Marcoleta *et al.,* 2018). Moreover, the expressed MceX protein acts as a partial repressor of the expression of the bi-cistronic unit *mceBA* (Marcoleta *et al.,* 2018).

### 4.4 The MccE492 cluster is located in a genomic island named GIE492

Even if the most important genetic factors required for MccE492 synthesis, maturation, immunity, and exportation were already known by 2009 (Lagos *et al.,* 2009) and were enough to understand the antibacterial activity from Kpn RYC492, the sequencing and analysis of the RYC492 genome in 2013 (Marcoleta *et al.,* 2013a)

revealed that the MccE492 production cluster might be part of a genomic island named GIE492. First, roughly 300 bp downstream of the *mceA* gene, there is an integrase-coding gene from the tyrosine recombinase family, which normally catalyzes the excision and integration of elements such as GIs (Boyd *et al.,* 2009). Second, immediately after the integrase gene, an asparagine tDNA can be found; and third, the ~21 kbp chromosomal region that includes all the Mcc cluster genes presents a GC content lower than the chromosome average and a different codon bias in terms of usage for protein-coding sequences (Marcoleta *et al.,* 2016). The putative GIE492 element was concluded to be a 22.3 kbp genomic island located between coordinates 1705122 and 1727413 of the Kpn RYC492 genome, which is flanked by 17 and 20 bp direct repeats (Marcoleta *et al.,* 2016).

Besides the MccE492 ~13kbp production cluster encompassing genes *mceA to mceF*, GIE492 also carries the previously mentioned integrase gene and at least six putative genes of unknown function, provisionally designated *u1-u6* (Marcoleta *et al.,* 2016). These putative genes outside of the main cluster were compared with public databases in search of a probable protein function, revealing that *u1* encodes a putative protein with significant identity to type-11 methyltransferases, which transfer methyl groups from S-adenosylmethionine to DNA, RNA, proteins, or small molecules such as catechols (Schubert *et al.,* 2003, Nelson *et al.,* 2007). Also, *u5* encodes a putative protein with tetratricopeptide repeats, which have been proposed to participate in protein-protein interactions and scaffolding of higher-order macromolecular complexes. Further, the putative protein encoded by *u6* possesses an NTPase motif and a topoisomerase-primase nucleotidyl transferase/hydrolase domain, related to

endonuclease proteins of the old-like family. Finally, the rest of the hypothetical proteins, u2, u3, and u4, did not have an identified or proposed function (Marcoleta *et al.,* 2016).

On the other hand, GIE492 presence was assessed in a collection of Kpn genomes previously reported to have the *mceA* gene (Struve *et al.,* 2015), finding positive hits to the GI in 35 out of 71 genomes tested. Moreover, all the identified GIE492 shared 99% sequence identity, were inserted in the same asparagine tDNA (*asn1C*), and had identical direct repeats (Marcoleta et al., 2016). Strikingly, 32 out of the 35 positive genomes belonged to CG23 or a closely related group, and 24 of them were isolated from liver abscess infections, strongly suggesting a relationship between GIE492 and hvKp strains, and the possible role of MccE492 and salmochelin in the development of these types of aggressive infections (Struve *et al.,* 2015, Marcoleta *et al.,* 2016).

As mentioned previously, the presence of an integrase inside a GI sequence strongly suggests that it can be excised from the genome under certain conditions, implying that the GI is unstable. In this regard, our research group used a nested PCR assay strategy to successfully detect the theoretical scar left if GIE492 was excised from the genome (the region formed by joining both the upstream and downstream chromosomal regions immediately adjacent to the island borders). However, this same strategy failed to detect the circular GI that should form upon excision (Marcoleta *et al.,* 2016). Furthermore, a qPCR-based strategy was used to determine the excision frequency in different growth phases, being around $6x10^{-7}$ (Marcoleta *et al.,* 2016). This frequency increased upon adding the DNA-damaging agent mitomycin C in a dose-

dependent way without affecting cell growth and upon overexpressing the integrase protein. Further, the overexpression of *int* in presence of mitomycin C results in a synergistic excision frequency increase up to 1000-fold, mainly noticed in the early exponential phase of growth (Marcoleta *et al.,* 2016). These results suggested that the *int* gene promotes GIE492 excision and that mitomycin C-mediated induction involves a different molecular mechanism than the one mediated by *int* (Marcoleta *et al.,* 2016).

Sequence comparison between GIE492 and several putative asn-related GIs, revealed that most of them share a 300-500 bp region, that according to a previous study describing the ICE*Kp*1 element that also integrates into asparagine tDNA, comprises the ~250 bp transfer origin (oriT), located next to the *virB* and *mobB* genes, all of which mediate the conjugal transfer of the ICE. Moreover, cloning this oriT sequence in a plasmid vector was sufficient to allow its conjugal mobilization from a host encoding conjugation-related proteins (Lin *et al.,* 2008).

Among the GIs that shared this region, variations in the content of conjugation-related genes and the genetic structure were detected, but the cryptic oriT sequence was always conserved (Marcoleta et al., 2016). In the case of GIE492, only the oriT sequence was found. Thus, GIE492 and other asn-related GIs would be mobilized by conjugative GIs such as ICE*Kp* that shares a similar oriT (Marcoleta *et al.,* 2016). In this regard, it is important to consider that hvKp CG23 strains are characterized for carrying both an ICE*Kp* element and GIE492, supporting that these two GIs could be co-mobilized and co-selected.

Besides the known general features of GIE492 described above, highly relevant aspects of this mobile element remain to be addressed, including:

- What function do the still unknown genes of GIE492 fulfill? Are these functions related to the MccE492 system?
- Do structural variants of this element exist? If so, which part of the element show variations?
- Is GIE492 only present in Kpn strains? What about the rest of the *Klebsiella* genus?
- How is this MGE distributed in the *Klebsiella* population? Which Kpn lineages carry this element? Do they form a monophyletic clade?
- How is GIE492 being disseminated in the *Klebsiella* population?
- Is there genomic evidence of possible associations between GIE492 and the ICE*Kp* family?
- Is the cryptic oriT conserved across GIE492 from different *K. pneumoniae* lineages?
- Is GIE492 associated with hypervirulent *K*. *pneumoniae* strains*?*
- Is it possible to apply NGS technologies to experimentally detect the excision of GIE492 and other genomic islands present in hypervirulent strains?

**HYPOTHESES**

Based on the background discussed above, we hypothesize that:

1). GIE492 has multiple structural variants distributed among specific Clonal Groups and is mainly vertically inherited, although it was acquired multipe times by horizontal transfer events during *K. pneumoniae* evolution.

2). GIE492 is strongly associated with ICE*Kp* elements and hypervirulent *K. pneumoniae* strains.

To test these hypotheses, the following main objectives are proposed:

***GENERAL GOAL***

To investigate the evolutionary history, phylogenomic distribution, and dissemination route of GIE492 in *Klebsiella* and its possible association with the ICE*Kp* family and hypervirulent strains.

**SPECIFIC GOALS**

- To identify GIE492 structural variants, integration sites, associated Clonal Groups, and coexistence with ICE*Kp* elements among a set of thousands of *Klebsiella* genomes.
- To determine the possible ways of GIE492 dissemination in the *Klebsiella* population.
- To test the association of GIE492 with the hypervirulent *K. pneumoniae* strains.
- To develop and test a methodology based on massive sequencing to detect the excision of GIE492 and other genomic islands in a population of *K. pneumoniae* cells.

# MATERIALS AND METHODS

## 1. *K. pneumoniae* strains studied in this work

*K. pneumoniae* RYC492 (available from our laboratory collection) and ten *K. pneumoniae* strains isolated during 2018 and 2019 as part of the Chilean Program for Surveillance of carbapenem-resistant *Enterobacteriaceae* performed by the National Public Health Institute were subjected to complete genome sequencing using Illumina and Nanopore technologies as part of the 1000 Genomas project led by Dr. Miguel Allende, member of the Center for Genome Regulation. For this, they were grown in LB broth overnight and then subjected to DNA extraction.

Additionally, the hypervirulent reference strain *K. pneumoniae* SGH10 (kindly provided by Prof. Yunn Hwen Gan from the National University of Singapore) was used for GI excision detection experiments. For this, Kpn SGH10 was grown in 100 mL of LB broth until the exponential phase and then divided into two samples of 50 mL each. Then, 625 µL of a Mitomycin C (Sigma) stock solution prepared in sterile PBS (0.4 mg/mL) was added to one of the SGH10 samples (5 µg/mL final concentration), while the other remained untreated. Both SGH10 samples were incubated for 2 hours at 37° C and then processed for genomic DNA extraction.

For safety reasons, work with SGH10 strains was performed in a biosafety cabinet belonging to SysMicroLab led by Dr. Francisco Chávez.

## 2. DNA purification and quality assessment

High molecular weight genomic DNA was extracted using the GeneJet Genomic DNA Purification Kit (Thermo Scientific) following the manufacturer's recommendations.

Upon extraction, the gDNA was quantified using a Qubit fluorimeter (Invitrogen), and the integrity was assessed by 1% TAE-agarose gel electrophoresis.

## 3. Genome sequencing, quality checking, and assembly

The whole genome of *K. pneumoniae* RYC492 and the 10 Chilean isolates were sequenced using Illumina and Oxford Nanopore technologies. Illumina sequencing services were hired to Macrogen Inc. (Korea). For this, upon quality checking, sequencing libraries were prepared using the TruSeq Nano DNA kit, and the size distribution of PCR enriched fragments was checked by running on an Agilent Technologies 2100 Bioanalyzer using a DNA 1000 chip. Sequencing was performed in a Hiseq4000 machine to obtain 101-bp paired-end reads.

Nanopore sequencing was performed at our laboratory (Laboratorio de Biología Estructural y Molecular BEM, Facultad de Ciencias, Universidad de Chile) using a MinION Mk1C device and FLO-MIN106 (R9.4) flow cells, according to the manufacturer's instructions. Barcoded sequencing libraries were prepared using the Native Barcoding Expansion (EXP-NBD104) and the 1D Ligation sequencing kit (SQK-

LSK109). The basecalling of Nanopore reads was performed using Guppy (Nanopore, Oxford, UK).

After sequencing, the Illumina reads were quality checked using FastQC v0.11.9 (Andrews, 2017) followed by quality trimming using Trimmomatic v0.36 (Bolger *et al.,* 2014) and Fastp v0.22.0 (Chen *et al.,* 2018). Raw Nanopore reads were corrected, trimmed, and posteriorly assembled using Canu v2.3 (Koren *et al.,* 2016).

Hybrid genome assemblies were performed with Unicycler v0.4.9b (Wick *et al.,* 2017), using as input the trimmed short and long reads and the long read assembly made with Canu. Additionally, to obtain an improved assembly of VA833 and RYC492, we conducted the following strategy for each of them. We first generated a total of 15 assemblies following three pipelines (five assemblies each) using the following tools: 1) Flye v2.9-b1768 (Kolmogorov *et al.,* 2019), 2) Minimap2 v2.22-r1105-dirty (Li, 2018) + Miniasm v0.3-r179 (Li, 2016) + Minipolish v0.1.3 (Wick & Holt, 2020), and 3) Raven v1.5.3 (Vaser & Šikić, 2021). Next, the long-read assemblies were used as input for the Trycycler v0.5.0 pipeline (Wick *et al.,* 2021) generating a consensus assembly. Afterward, a first long-reads-based polishing step was performed using Medaka v1.4.3 (Nanopore, Oxford, UK) and the model r941_min_fast_g507. A final polishing based on short reads was performed using Pilon v1.24 (Walker *et al.,* 2014). Assembly quality evaluation was performed using QUAST v5.0.2 (Gurevich *et al.,* 2013). General genome annotation was performed using the RAST annotation pipeline v2.0 (Aziz *et al.,* 2008).

For GI excision experiments, two Illumina-only datasets were obtained, corresponding to the following conditions: untreated Kpn SGH10 and Kpn SGH10

treated with mitomycin C. The raw reads generated were used for mapping-based downstream bioinformatic analyses.

### 4. *Klebsiella* genome databases construction

To construct a first curated database of thousands of *Klebsiella* genomes, the National Center for Biotechnology Information website was accessed in December 2020, searching for genome assemblies annotated as *Klebsiella* in both the Genbank and RefSeq databases. This search yielded 4006 unique genomes. Additionally, we added the *K. pneumoniae* RYC492 genome to this set and the genomes of the *K. pneumoniae* Chilean strains previously described. In sum, this collection process yielded a total of 4017 assembled *Klebsiella* genomes in FASTA format (Genome Database v1, Gv1).

Then, different filters were applied to Gv1, based on general assembly features and taxonomic classification using Kleborate v2.0.1 (Lam *et al.,* 2021). After applying these, the final curated database for downstream analyses was created, containing 3878 *Klebsiella* genome assemblies, Gv2.

### 5. GIE492 screening among *Klebsiella* genome sequences

All the assemblies in Gv2 were screened for GIE492 proteins using DIAMOND (Buchfink *et al.,* 2015) in blastx mode for local alignments, requiring at least 70% of sequence identity. GIE492+ candidates were defined based on the concomitant presence of *int, mceA* and *mceB.* This way, we identified 265 GIE492+ genomes (Gv2_GIE492 set). Afterwards, the gene sequences of GIE492 found in the

Gv2_GIE492 genomes were extracted and aligned “all vs all” using the megablast algorithm of BLASTn v2.9.0+ (Altschul *et al.,* 1990), considering as positive hits those with 70% or more identity. These alignments were used as input for the clusterization process described below.

## 6. Clusterization of the observed nucleotide sequences for each GIE492-encoded gene

All nucleotide sequences obtained from the local BLAST alignments were classified by gene. Later, following the method described by Rasko *et al*. (2005), we calculated the Blast Score Ratio (BSR) for all alignments, according to the following formula:

$$BSR = \frac{Query\ Blast\ Score}{Reference\ Blast\ Score} \quad (1)$$

In equation 1, “Score” corresponds to the “Raw Score” parameter determined for the alignment, “Query Blast Score” is the value of the raw score of an alignment against any reference sequence, and “Reference Blast Score” is the value of a reference sequence aligned against itself. Thus, the BSR ranges between 0 and 1. The BSR values calculated for each gene bundle were tabulated in a square matrix (all against all) and then used as input for the ComplexHeatmap R package (Gu *et al,* 2016), utilizing K-means as the default clusterization parameter.

Each sequence bundle was iterated until the maximum K-means value was reached, corresponding to the highest number of possible logical groups according to the data distribution. Then, each group determined for each gene received a distinct number. Using this information, we assigned to each genome assembly in Gv2_GIE492

a unique text string comprising the contig name(s) where BLAST hits were detected, and for each gene, the start and end coordinates, the gene name, and the clustering group number.

**7. Non-redundant clusterization of gene sequences, allele definition, and structural variant typing**

Each of the previously described text strings was analyzed exhaustively, and in cases where two or more hits for the same gene and genome were found, one of them was chosen based on the following criteria:

- The evaluated gene should have "close proximity" to the rest of the GIE492-encoded genes. Here, close proximity is defined as a combination of two factors: Distance less than the median length of the whole GI (23 kbp) and gene position following the canonical GIE492 order (*int, mceA, mceB, mceC, mceD, mceS2, mceS3, mceM, mceL, mceE, mceX, mceJ, mceI, mceH, mceG, mceK, mceF, u1, u2, u3, u4, u5, u6).*
- If the assembly shows alignments for the same gene in multiple contigs, the one in the same contig as the other genes should be preferred.
- If none of the above criteria is enough to select a confident hit, give the gene a "*" tag, indicating that the gene is present in the assembly, but it is impossible to determine the distinctive group to which this gene belongs.

Subsequently, the nucleotide sequence of each GIE492-encoded gene previously selected was extracted as a FASTA file using the faidx module from the SAMTools package v1.10 (Li *et al.,* 2009). Next, all the extracted gene sequences were

concatenated into a single MULTIFASTA file, which was used as input for CD-HIT-EST v4.8.1 with the following parameters: -n 10 -s 0.9 -A 0.9 -g 1 -c 0.95 (Li & Godzik, 2006, Fu *et al.,* 2012). CD-HIT-EST yielded non-redundant clusters of nucleotide sequences, which were considered alleles for each of the GIE492 genes.

Using a Multi Locus Sequence Typing (MLST) (Larsen *et al,* 2012) approach, specific and distinct combinations of allelic sequences were determined. Based on these combinations, we defined structural variants of GIE492, and each one of them received a unique roman number identifier. Two caveats were considered:

- If any of the genes got the "*" assignation previously, no structural variant was assigned in these cases. Instead, the NA designation was given.
- If the determined Structural Variant was utterly unique, the "u" identifier was used instead of a roman number.

**8. Determination of consensus alleles for the GIE492-encoded genes**

Upon determining the GIE492 structural variants, a consensus sequence was defined for each allele, storing it in a FASTA file for future use. For this, all the sequences belonging to the same non-redundant cluster were used as input for MUSCLE v5.0.1428 (Edgar, 2004) to perform a Multiple Sequence Alignment (MSA) in diversified mode. The diversified alignment ensemble was then used to infer the alignment with the maximum column confidence available (maxcc-MSA), using MUSCLE maxcc mode. Next, both the maxcc-MSA and the diversified ensemble were used to calculate the letter confidence of the alignment using the MUSCLE letterconf mode. Once all MSAs were done, the corresponding files were analyzed using Jalview

v2.11.1.7 (Waterhouse *et al.,* 2009) to get the final consensus sequence of each allele. All the allele sequences were then stored in a properly formatted MULTIFASTA file as a database for SRST2 (Inouye *et al.,* 2014), which can then be used to infer both the presence and structural variant of GIE492 in assembled genomes or raw Illumina reads.

### 9. OriT identification

The cryptic oriT sequence of GIE492 was searched in all Gv2_GIE492 genomes using BLASTn v2.9.0+ as described previously. Genomic coordinates and BSR values were stored for all the positive hits.

### 10. GIE492 Structural Variants graphical representation

After each genome in Gv2_GIE492 received a proper identification of its GIE492 structural variant and its oriT was identified, one representative version of each distinct GIE492 structural variant was annotated in genbank format and these files were then aligned using Easyfig v2.2.5 (Sullivan *et al.,* 2011), to obtain a graphical representation.

### 11. GIE492 integration site determination

All the Gv2_GIE492 assemblies were functionally annotated using PROKKA v1.14.16 (Seemann, 2014) to have a consistent nomenclature for the predicted coding sequences across all the genomes. Upon annotation, the tDNA where GIE492 was integrated was identified by visually inspecting the genomic context of the island using SnapGene Viewer v5.3 (Biotech). Additionally, the tDNA integration site for ICE*Kp* was determined in every case where Kleborate indicated that this element was also present.

To easily understand and discriminate the tDNAs used as integration sites by both GIs, a 4-character string named asn-tDNAs Occupation Pattern (ATOP), was developed:

- 0000 by default, indicating that all the four asparagine tDNAs are not interrupted by any sequence (virgin context).
- Each 0 is a unique asparagine tDNA, in the following order: *asn*1A, *asn*1B, *asn*1C and *asn*1D.
- G indicates the tDNA in which GIE492 is integrated.
- I indicates the tDNA in which ICE*Kp* is integrated.
- 1 indicates the tDNA in which a putative GI different from both GIE492 and ICE*Kp* is integrated.
- If one or more of the asparagine tDNAs could not be identified, the assembly received the “ND” value.

For example, 10GI, represents a genome harboring a putative GI inserted in *asn*1A, no GI integrated at *asn*1B, GIE492 integrated at *asn*1C, and ICE*Kp* integrated at *asn*1D.

## 12. Clonal Group assignation

To identify proper Clonal Groups in *Klebsiella*, we followed the strategy initially developed by Bialek-Davenet *et al.* (2014), consisting of eight main steps:

1. To determine all the possible coding sequences in a closely related genome database.

2. To identify which genes are present in “all” genomes under a certain presence threshold. This set corresponds to the *core-genome.*

3. To define an identity threshold that determines when a sequence corresponds to an allele of a gene and not to another coding sequence.

4. To assign a unique identifier to every allele of each gene.

5. To determine the combination of alleles for each assembly. This is called the cgMLST (*core-genome* Multi Locus Sequence Typing) approach.

6. To calculate the Genomic Distance between all assemblies, defined as the number of allelic mismatches between two assemblies, expressed as a percent of all genes of the *core-genome*.

7. To analyze the Genomic Distances distribution to empirically determine two distance thresholds: short distances between members inside the same CG and large distances between members of different CGs.

8. To assign a unique CG to each assembly.

We used chewBBACA v2.7.0 (Silva *et al.,* 2018) to assign a CG to each Gv2_GIE492 assembly following this strategy. First, to make chewBBACA more efficient, a reference genome from Gv2_GIE492 needed to be selected. It was opted to use the SGH10 genome, because this assembly is known to be complete, it was previously designated as the reference hypervirulent CG23 *K. pneumoniae,* and it is also a GIE492+ ICE*Kp*+ genome, so it is ideal for this role.

The SGH10 genome was then converted to a binary training file using Prodigal v2.6.3 (Hyatt *et al.,* 2010) in training mode. Next, considering that the *core-genome* needs to be properly designated, only those genomes that were “complete enough”

were used to create the schema. To define these representative genomes, the following two criteria were established:

- No QC warnings according to Kleborate, this means that the assembly does not have ambiguous bases or a non-standard *Klebsiella* genome length (less than 5Mbp or more than 6Mbp).
- N50 > 5Mbp.

The representative set and the reference training file were used as input for the full chewBBACA pipeline to obtain the representative cgMLST schema (Create Schema, Allele Call, Remove Paralogs, Test Genome Quality, and Extract cgMLST) with default parameters. Once this schema was defined, it was used to determine the cgMLST for the rest of Gv2_GIE492. Next, all cgMLST profiles were joined in a single matrix which was used as input for PHYLOViZ Online (Ribeiro-Gonçalves *et al.,* 2018) to determine a Minimum Spanning Tree (MST) between all assemblies. This MST was used to calculate genomic distances and distance thresholds. Next, CGs were identified following the Bialek-Davenet criteria, and graphically represented through a nLV-graph using PHYLOViZ online.

A MST corresponds to a map of points joint by lines, where each point corresponds to an n-dimentional element in a set and each line represents the distance between two elements, corresponding to the amount of dimensions the points differ. In this case, every point corresponds to a genome with an n-dimentional (number of genes in the *core*-genome) cgMLST profile, and every line to the amount of loci with different alleles between the two genomes. The number of loci variants (nLVs) defined as the “inside CGs” distance threshold, can then be used to transform the MST into a nLV-

graph, where all points inside a group appear clustered as if they were a single element and “between CGs” distances are exacerbated.

## 13. Phylogenetic trees

A phylogenetic tree of all genomes in Gv2 based on MASH distances (Ondov *et al.,* 2016), was inferred using mashtree v1.2.0 (Katz *et al.,* 2019). Node support was calculated using 1000 bootstrap trees. This tree was graphically represented using FigTree v1.4.4 (Rambaut, 2009).

In a similar way, a phylogenetic tree of all Kpn assemblies in in Gv2 and one for all genomes in Gv2_GIE492 (and selected outgroups) were also inferred. Graphical representations of both these trees were obtained using the iTOL online tool (Letunic & Bork, 2021).

## 14. Development of a tool to detect genomic regions with differential sequencing depth of coverage

Multiple scripts written in Perl programming language (Wall, 1994), were developed with the objective of detecting GI excision events based on discriminating genomic regions with differential coverage depth. These scripts integrated the use of Bioperl (Stajich *et al*., 2002), Bowtie2 v2.4.4 (Langmead & Salzberg, 2012), SAMtools v1.13 (Li *et al*, 2009) and R (Team, 2000).

The developed programs were used for the following pipeline: 1) To divide the assembled chromosome into small regions, 2) To map genomic reads to these regions, 3) To determine a coverage depth threshold representing the baseline amount of reads

that map to any region, 4) To determine “excision probabilities” for each region according to the the coverage depth threshold, and 5) To identify which regions are above the coverage depth threshold.

Testing of these tools was performed using the already available assembled genome of Kpn SGH10 and the Illumina reads obtained after sequencing both treatments of Kpn SGH10 described previously.

Identification of relevant MGEs in the SGH10 chromosome was done using Kintun-VLI v1.0 (Berríos-Pastén *et al.,* 2020) for GIs, PHASTER (Arndt *et al.,* 2016) for prophages and ISEScan v1.7.2.3 (Xie & Tang, 2017) for insertion sequences.

# RESULTS

## 1. GIE492 structure and presence in the *Klebsiella* population

### 1.1 Genome sequencing of selected *K. pneumoniae* strains

At the time of this work, no complete genome sequences of *K. pneumoniae* isolated in Chile were available in public databases. Therefore, to contribute in overcoming this significant limitation and to include Chilean Kpn genomes in our analyses, we determined the complete genome sequence of ten isolates collected during 2018 and 2019 as part of the National Surveillance of carbapenem-resistant *Enterobacteriaceae*, in collaboration with the Chilean Health Institute (Instituto de Salud Pública (ISP)). Additionally, we determined the complete genome sequence of *K. pneumoniae* RYC492, the strain where MccE492 and the GIE492 island was first described, for which only a draft genome assembly was previously available.

We performed Illumina and Nanopore genome sequencing for all of these strains obtaining 1.15 – 3.45 and 0.12 – 9.84 Gbp of data per strain, respectively. Next, we combined the short (Illumina) and long (Nanopore) reads to generate hybrid assemblies. For all the strains we obtained a closed circular chromosome and a variable number of closed plasmids. The general assembly features of these genomes

are specified in Table 1. Due to the novelty, neither Refseq nor Genbank assembly numbers of these genomes are yet available.

**Table 1.** General Assembly features of genome assemblies made for this work.

| Strain | Genome Length (Mbp) | Chromosome Length (Mbp) | Plasmid Number | ACD[1] | ST | Assembly Protocol[2] |
|---|---|---|---|---|---|---|
| **VA04** | 5.76 | 5.46 | 4 | 176X | ST25 | A |
| **VA32** | 5.73 | 5.29 | 5 | 202X | ST11 | A |
| **VA126** | 5.76 | 5.12 | 5 | 181X | ST25 | A |
| **VA172** | 5.40 | 5.25 | 2 | 198X | ST25 | A |
| **VA564** | 5.72 | 5.41 | 4 | 182X | ST25 | A |
| **VA569** | 5.71 | 5.58 | 4 | 164X | ST11 | A |
| **VA591** | 5.65 | 5.43 | 4 | 180X | ST45 | A |
| **VA681** | 5.55 | 5.25 | 5 | 204X | ST11 | A |
| **VA684** | 5.78 | 5.18 | 6 | 195X | ST505 | A |
| **VA833** | 5.76 | 5.22 | 6 | 220X | ST25 | B |
| **RYC492** | 5.37 | 5.31 | 2 | 1108X | ST35 | B |

1. Average Coverage Depth, expressed as X times the genome length.
2. A. Hybrid, Canu+Unicycler; B. Hybrid, Trycycler Pipeline.

### 1.2 General features of the *Klebsiella pneumoniae* RYC492 complete genome

Due to the importance of this genome, as it corresponds to the strain where MccE492 was originally described, it was necessary to have a high-quality assembly to be able to get the most proper nucleotidic sequence of GIE492 and its genomic context. Moreover, we took advantage of the complete genome sequence determined to get relevant information of this strain. For this purpose, the complete assembly was annotated using both PROKKA and RAST.

As shown in Table 1, the RYC492 strain has three replicons corresponding to the chromosome and two plasmids: a 53-kbp circular plasmid containing T4SS related genes and a TEM beta-lactamase, and a 4-kbp plasmid encoding a regulatory protein Rop and two hypothetical proteins. Annotated genomic features across all replicons include 5163 coding sequences, 25 rRNA genes and 86 tDNAs. According to Kleborate,

the phylogenetic lineage of the strain corresponds to ST35, its genetic capsular type is K22 (*wzi*37) and the O-locus is O1v1. Also, Kpn RYC492 lacks important virulence factors, being negative for *ybt, clb, iuc* and plasmid encoded *iro genes*, ICE*Kp* and KpVPs.

Regarding antibiotic resistance, a relatively small number (five) of acquired resistance genes were found in this genome: APH(3')-IIa, *sph, strA* and *strB,* conferring resistance to aminoglycosides, and the β-lactamase TEM-1. Additionally, GIE492 presence was confirmed through DIAMOND alignments to the island sequence presented by Marcoleta *et al. (*2016), then it was extracted and used for the following analyses.

### 1.3 Construction of a curated *Klebsiella* genomes database

The assembled genomes of the ten chilean *K. pneumoniae* isolates and the RYC492 strain were complemented with 4006 assemblies downloaded from the NCBI Refseq and Genbank databases, which were identified as *Klebsiella*. This initial database was named Genomic Database v1 (Gv1), and was screened against *Klebsiella* housekeeping genes, virulence factors, resistance genes, and evaluated for general assembly characteristics using the software Kleborate.

To ensure better results in the downstream analyses for the objectives of this study, Gv1 needed to be filtered out of low-quality genomes trying to keep most of the assemblies, which are mainly draft genomes, excluding the extreme outliers. This was made based on the following criteria:

- Only keep assemblies designated as *Klebsiella* by Kleborate with a strong confidence value. This filter was needed due to possible taxonomic classification errors in public databases.
- Only keep those assemblies where at least one contig has a length 10 times the length of GIE492 (223 kbp). This filter was created to exclude very poorly assembled genomes and maximize the probability of finding GIE492 as a complete sequence in a single contig.

The application of the second filter was made using a custom-made Perl script. After both filters were applied, the curated Gv2 database containing 3878 genomes was created. For easier usage of Gv2, every assembly received a unique number, from 0000 to 3877.

These databases included genomes of 15 *Klebsiella* species, comprising 6 species of the KpSC (*Kp1 – Kp6)* and 9 more *Klebsiella* species*.* The species distribution of Gv2 is specified in Table 2.

**Table 2.** Species distribution in Genome database v2 (Gv2).

| **Species**[1] | **Number of Strains** |
|---|---|
| *K. pneumoniae sensu stricto (Kp1)* | 3453 |
| *K. quasipneumoniae* subsp *quasipneumoniae (Kp2)* | 32 |
| *K. variicola* subsp *variicola (Kp3)* | 123 |
| *K. quasipneumoniae* subsp *similipneumoniae (Kp4)* | 40 |
| *K. variicola* subsp *tropica (Kp5)* | 2 |
| *K. quasivariicola (Kp6)* | 8 |
| *K. aerogenes (Kae)* | 99 |
| *K. michiganensis (Kmi)* | 60 |
| *K. oxytoca (Kox)* | 18 |
| *K. grimontii (Kgr)* | 18 |
| *K. ornithinolytica (Kor)* | 7 |
| *K. terrigena (Kte)* | 7 |
| *K. planticola (Kpl)* | 5 |
| *K. huaxiensis (Khu)* | 3 |
| *K. pasteurii (Kpa)* | 3 |
| **Total** | **3878** |

1. All species names are followed by a three-letter code in parenthesis. These codes are based on the three-letter codes of the KpSC species (Rodrigues *et al.,* 2019).

All genomes in Gv2, including assembly name, unique identifier and Kleborate screening are available as a supplementary spreadsheet file (Appendix 1).

Due to high amount of Kp1 genomes present in Gv2 (~89%), Table 3 describes the STs distribution of this species in the database.

**Table 3.** Kp1 sequence type (ST) distribution in Gv2.

| **ST** | **fi**[1] | **Fi%**[2] | **ST** | **fi** | **Fi%** | **ST** | **fi** | **Fi%** |
|---|---|---|---|---|---|---|---|---|
| ST11 | 635 | 18.39 | ST35 | 54 | 58.44 | ST340 | 23 | 69.82 |
| ST258 | 477 | 32.20 | ST307 | 53 | 59.98 | ST152 | 22 | 70.46 |
| ST101 | 189 | 37.68 | ST48 | 47 | 61.34 | ST395 | 22 | 71.10 |
| ST15 | 170 | 42.60 | ST17 | 45 | 62.64 | ST111 | 19 | 71.65 |
| ST45 | 126 | 46.25 | ST37 | 44 | 63.92 | ST39 | 19 | 72.20 |
| ST14 | 88 | 48.80 | ST231 | 43 | 65.16 | ST86 | 19 | 72.75 |
| ST147 | 83 | 51.20 | ST29 | 40 | 66.32 | ST20 | 17 | 73.24 |
| ST23 | 79 | 53.49 | ST25 | 39 | 67.45 | ST628 | 17 | 73.73 |
| ST405 | 62 | 55.29 | ST336 | 32 | 68.38 | Other | 907 | 100 |
| ST16 | 55 | 56.88 | ST268 | 27 | 69.16 | Total | 3453 | 100 |

1. Absolute frequency.
2. Cummulative relative frequency.

### 1.4 Presence of GIE492 in the *Klebsiella* population

Once obtained the curated database, we searched the 3878 *Klebsiella* genomes from Gv2 for the presence of the GIE492 island. For this, the protein-coding genes present in the GIE492 sequence extracted from the RYC492 chromosome were translated and stored in a MULTIFASTA file. The 23 ordered coding sequences are the following: *int, mceA, mceB, mceC, mceD, mceS2, mceS3, mceM, mceL, mceE, mceX, mceJ, mceI, mceH, mceG, mceK, mceF, u1, u2, u3, u4, u5* and *u6.* Then, the protein sequences were locally aligned to the assemblies in Gv2 using DIAMOND blastx, with an identity cuttoff of at least 70%. Because this work was focused in finding possible GIE492 structural variants, not all 26 coding sequences needed to be present in the assemblies for them to be considered positive hits.

A hit was considered positive when the “Fundamental GIE492” paradigm was met, this defined as the first three genes of the island:

- *int,* previously demonstrated to encode a P4-like integrase that mediates GIE492 excision (Marcoleta et al., 2016).
- *mceA,* for being the structural gene of MccE492, encoding one of the key functions of GIE492.
- *mceB,* due to coding the immunity protein of MccE492 and being fundamental for the survival of a MccE492 producing strain.

265 genomes from Gv2 were positive hits for GIE492 presence, which were included in a database named Gv2_GIE492. 264 of these genomes were taxonomically classified as *K*. *pneumoniae sensu stricto* (Kp1), and one of them as *K. michiganensis.*

## 1.5 GIE492 structural variants

We next aimed to compare the sequence and structure of the GIE492 island found in the *Klebsiella* strains. For this purpose, DIAMOND alignments were used to obtain the nucleotidic sequences of the GIE492 genes in all Gv2_GIE492 genomes. A total of 6040 sequences were stored this way, and then aligned “all vs all” using the megablast algorithm of BLASTn. 3992 pair-wise alignments were obtained.

The alignments obtained for each GIE492-encoded protein were used to calculate BSR square matrices following the BSR definitions presented in the Methods section, using custom Perl scripts.

The fundamental reason for using calculated BSR values instead of E-values provided by BLAST relies on the intrinsic match length skewness of E-values. Small regions of high similarity can generate artificially low E-values and negate the global level of similarity exhibited by the sequence. This bias is eliminated when the raw BLAST scores are used, but these types of values are query length dependent, so BSR calculations can supersede both problems (Rasko *et al.,* 2005).

The BSR-based matrices were used as input for the ComplexHeatmap R package to obtain 254 distinct sequence groups, using the maximum number of K-means as the clusterization algorithm.

This first clusterization strategy distinguished between 27 unique combinations of GIE492 proteins, a number way too restrictive to be useful, due to making ~1 nucleotide differences in coding sequences enough to be classified as different. Also, this method of classification made impossible to identify the GIE492 structural variant in 95 genomes (35.8%).

To solve this problem, a non-redundant clusterization approach was considered. To do this, all the already available nucleotidic sequences were used as input for the CD-HIT-EST clustering software. The clusterization process was restricted under the following parameters: 90% of sequence coverage and 95% of identity between all sequences in a same group.

This process only yielded 53 unique groups in contrast to the previous 254. Next, repeating the steps in the first method, every genome in Gv2_GIE492 received a GIE492 structural variant based on the combination of protein groups identified.

These 53 unique groups were considered alleles to the corresponding coding sequences of GIE492 in RYC492. The number of alleles (N) for each gene is specified in table 4. If a gene only has 1 allele, it means that the coding sequence never varies.

**Table 4** Number of alleles for each gene in GIE492.

| **Gene** | **N** | **Gene** | **N** | **Gene** | **N** | **Gene** | **N** |
|---|---|---|---|---|---|---|---|
| *int* | 9 | *mceS3* | 2 | *mceI* | 2 | *u2* | 4 |
| *mceA* | 1 | *mceM* | 2 | *mceH* | 2 | *u3* | 2 |
| *mceB* | 3 | *mceL* | 1 | *mceG* | 1 | *u4* | 1 |
| *mceC* | 2 | *mceE* | 2 | *mceK* | 6 | *u5* | 1 |
| *mceD* | 1 | *mceX* | 1 | *mceF* | 3 | *u6* | 2 |
| *mceS2* | 2 | *mceJ* | 1 | *u1* | 2 | **Total** | 53 |

However, 11 of these alleles are present only in one genome each, indicating that they are probable products of assembly errors.

With this approach, the combination of alleles for all the GIE492 encoded genes accounted for a total of four structural variants, and only in 38 genomes the variant could not be identified (14.3%, mainly due to the GI being divided in multiple contigs). Each of these structural variants received a roman number as unique identifier. Moreover, the four variants fell into two larger groups that are easily distinguishable: big

islands (B) and small islands (S). This distinction relies on the consistent absence of 5 consecutive coding sequences in small islands: *u1, u2, u3, u4* and *u5,* suggesting that these genes could be dispensable for MccE492 production. If one of the coding sequences was previously classified as "*" or a different set of genes was absent rather than *u1-u5,* the GI received the NA distinction. Structural variants I and III correspond to large islands, while structural variants II and IV to small islands.

In contrast, a selection of genes is completely constant in all 4 structural variants, lacking any alternative allele sequences: *mceA, mceB, mceC, mceD, mceS2, mceS3, mceM, mceL, mceE, mceX, mceJ, mceI, mceH, mceG* and *mceF*. Additionally, the other three genes were fundamental for distinguishing between the main structural variants:

- *int* showed two relevant alleles, the originally described one encoding a 424 aa protein present in the structural variants I, III and IV, and a second one coding a truncated ~419 aa version present in II.
- *mceK* showed two relevant alleles, the originally described one encoding a 136 aa protein in the structural variant III, and an alternative sequence coding a smaller 77 aa peptide present in the variants I and IV, due to a frameshift in the region where the fusion of the putative microcin and its immunity protein was produced
- In the structural variant II, *mceK* is apparently absent, but upon further inspection, a frameshift is present in codon 26 generating an early nonsense mutation.
- *u6* showed two relevant alleles, the originally described one encoding a 626 aa protein present in the structural variants I and III, and an alternative sequence coding a smaller ~348 aa protein present in the structural variants II and IV.

To get a schematic visualization and comparative sequence alignment of the structural variants, a genome comprising one of the four variants was selected arbitrarily (1370, 1791, 3340, and 3292 for structural variants I, II, III, and IV, respectively), the GIE492 sequence was extracted using SAMtools faidx, and then the gene positions were manually annotated in genbank format. The obtained genbank files were aligned using Easyfig to obtain the graphical representation shown in Figure 1.

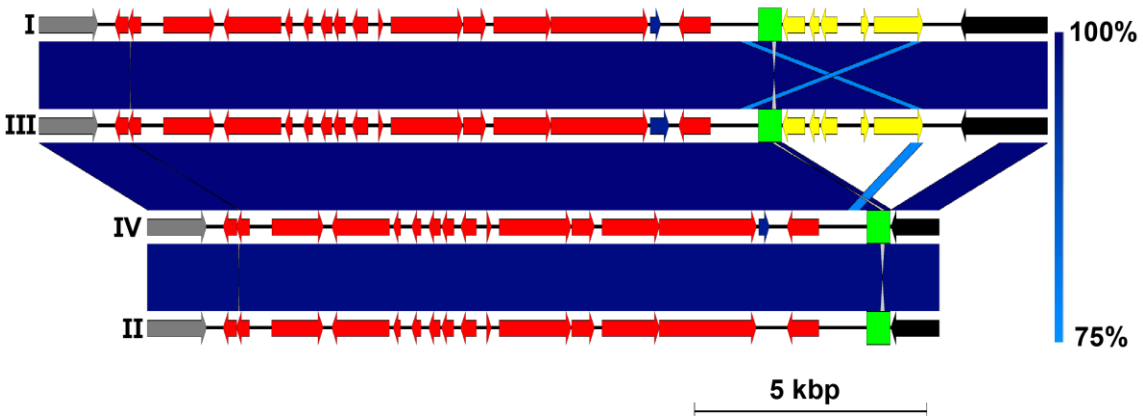


**Figure 1.** Graphical alignment of the four structural variants of GIE492. Each roman number identifies the respective structural variant. Red arrows represent constant genes (*mceA, mceB, mceC, mceD, mceS2, mceS3, mceM, mceL, mceE, mceX, mceJ, mceI, mceH, mceG* and *mceF).* Yellow arrows correspond to expendable genes (*u1, u2, u3, u4 and u5).* Gray arrows represent *int* genes. Blue arrows correspond to *mceK* sequences. Black arrows correspond to *u6* genes. Green rectangles show the position of the cryptic oriT sequence present in the GI. Blue bands illustrate the nucleotidic sequence conservation (% of identity) following the gradient shown left.

Even though four easily distinguishable structural variants of GIE492 do exist, the high level of sequence conservation of this GI is the most relevant observable result, as it is shown by figure 1, impliying that the genes required for MccE492 and salmochelin production are always conserved and theoretically active in all the assemblies of Gv2_GIE492.

## 2. GIE492 distribution across different *K. pneumoniae* lineages

### 2.1 GIE492 is found in different *K. pneumoniae* lineages

All genomes in Gv2 were used to construct a phylogenetic tree based on MASH distances using the software Mashtree. This tree was iterated 1000 times for node bootstrap support, and then annotated using the FigTree software. The species distribution of the database, described previously on Table 2, and their phylogenetic relationships are shown in Figure 2.

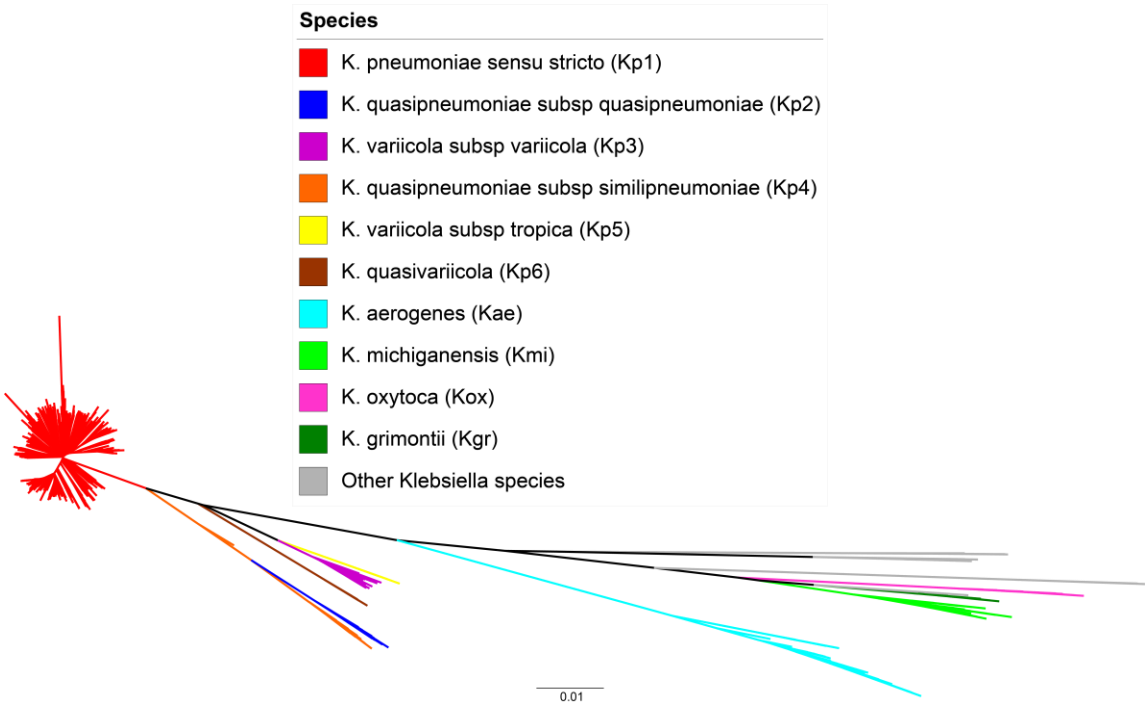


**Figure 2.** Phylogenetic tree based on MASH distances between all members in Genome Database v2. Clades are colored according to species. Other *Klesiella* species refers to: *K. terrigena, K. planticola, K. ornithinolytica* and *K. huaxiensis.*

Because 264 out of 265 genomes in Gv2_GIE492 belong to the *K. pneumoniae sensu stricto* species, a second phylogenetic tree was constructed following the same procedure as before, however, this tree used as input only the assemblies classified as

Kp1 in Gv2. This representation, made using the iTOL online tool, allows to better appreciate the different *K. pneumoniae* lineages and sequence types, as seen in Figure 3.

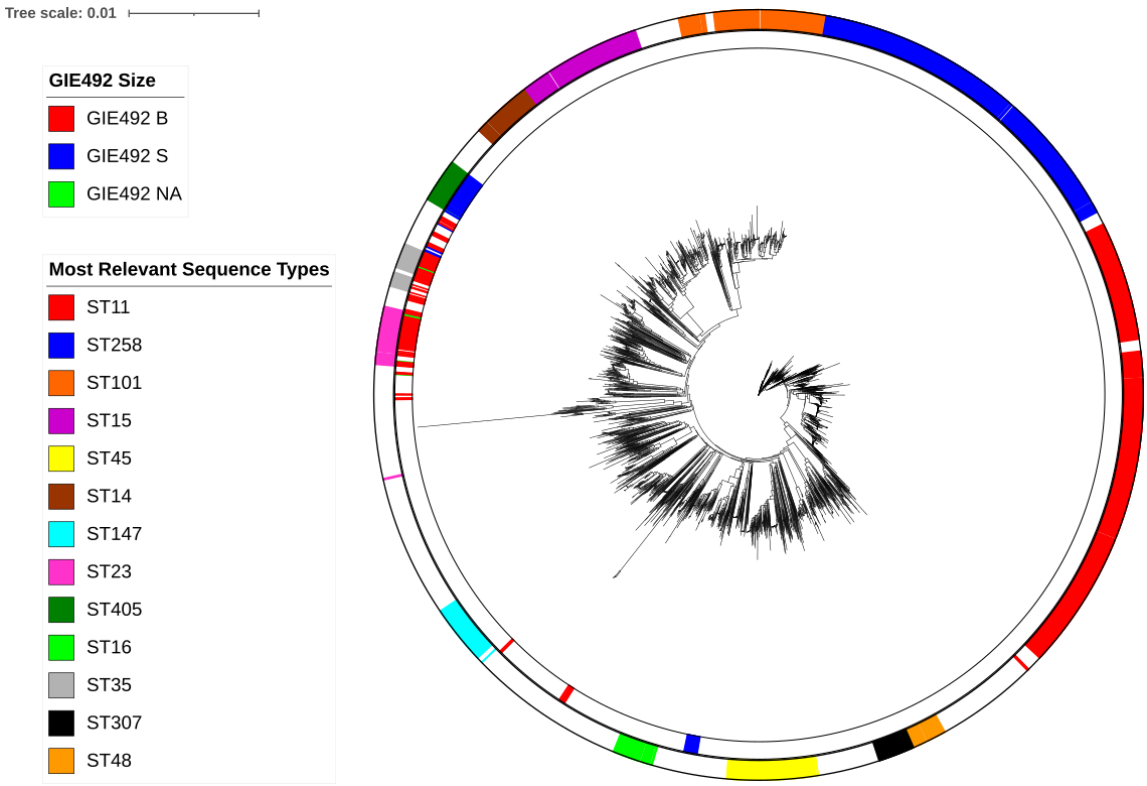


**Figure 3.** Phylogenetic tree based on MASH distances between all members in Genome Database v2 taxonomically classified as *K. pneumoniae sensu stricto.* From the inside out, the first track indicates if GIE492 is present in that genome and if the GI is B, S or NA. The second track shows the most relevant Sequence Types (ST) in the database. The cumulated relative sum of the most relevant STs corresponds to ~60% of all Kp1 genomes.

In terms of GIE492 distribution among the *K. pneumoniae* population, we observed that it is present sporadically in defined Kpn lineages, some of them phylogenetically distant. Out of all the 13 most relevant Kp1 STs represented in figure 3, GIE492 is absent in 10 of them, including ST258, ST101, ST15, ST147 and ST307, all of which were mentioned in the Introduction section as MDR global problem clones.

GIE492 is also absent in ST11, the CRKp most prevalent in China (Qi *et al.,* 2011). GIE492 is also absent from ST25, ST45 and ST505, making all Chilean isolates presented in this study negative for this GI. However, GIE492 is present in ST23, hvKp by excellence as explained in the Introduction section.

As it was mentioned before, GIE492 is present in one genome taxonomically classified as *K. michiganensis,* however it was not possible to determine the structural variant this genome carried, due to it lacking hits to the genes *u3*-*u6*, and presenting unique alleles in many of the core genes.

The highly sparse distribution of GIE492 in the *K. pneumoniae* population hampered to infer which variant would be the most ancestral. Therefore, it remains unclear if the missing region in the small variants is due to a deletion occurred in a large variant, or if this region arises from an insertion occurred in the former.

### 2.2 GIE492 variants associate with specific *K. pneumoniae* clonal groups

Due to limitations of the MLST approach to illustrate deep phylogenetic relationships between groups, the gold standard in Kpn research is to follow a cgMLST strategy, being able to identify proper Clonal Groups (CGs). For this, we followed the approach developed by Bialek-Davenet *et al.* (2014), as explained in the Methods section. This task was done using the chewBBACA pipeline. The first step was creating the cgMLST schema using high-quality assemblies from Gv2_GIE492, which showed a standard *Klebsiella* genome length, no ambiguous bases, and a N50 of at least 5 Mbp. 27 genomes were selected this way, including the assemblies from SGH10 and

RYC492 strains. A reference genome was also needed to be used as a training file, and SGH10 was selected.

The created schema included 3751 unique coding sequences as the core-genome of GIE492+ Kpn strains, under a 95% of presence threshold. This schema was applied to all genomes in Gv2_GIE492, to find the corresponding cgMLST profile. All profiles were then joint to form a matrix, and this matrix was used as input for PHYLOViZ online to obtain a Minimum Spanning Tree (MST).

PHYLOViZ online allows users to calculate genomic distances between all elements in the MST. For N genomes, $N^2$ distances are calculated. As explained in the Methods section, genomic distance is defined as the number of allelic mismatches between two cgMLST profiles, expressed as a percent of the total number of genes in the core-genome; in this case 3751.

Upon determining genomic distances, PHYLOViZ detected six genomes sharing an identical cgMLST profile with other strain, hence only 259 genomes were kept for the following analyses.

All 67081 genomic distances were compiled in an interval frequency distribution table as shown in Table 5.

**Table 5.** Frequency Distribution table of Genomic Distances between Gv2_GIE492 assemblies.

| Genomic Distance Interval | Class Mark | Absolute Frequency | Relative Frequency | Cumulative Relative Frequency |
|---|---|---|---|---|
| **[0 – 5[** | 2.5 | 4435 | 6.61 | 6.61 |
| **[5 – 10[** | 7.5 | 5620 | 8.38 | 14.99 |
| **[10 – 15[** | 12.5 | 910 | 1.36 | 16.35 |
| **[15 – 20[** | 17.5 | 178 | 0.27 | 16.61 |
| **[20 – 25[** | 22.5 | 318 | 0.47 | 17.09 |
| **[25 – 30[** | 27.5 | 84 | 0.13 | 17.21 |
| **[30 - 35[** | 32.5 | 8 | 0.01 | 17.22 |
| **[35 - 40[** | 37.5 | 2 | 0.00 | 17.23 |
| **[40 – 45[** | 42.5 | 126 | 0.19 | 17.41 |
| **[45 – 50[** | 47.5 | 6 | 0.01 | 17.42 |
| **[50 – 55[** | 52.5 | 100 | 0.15 | 17.57 |
| **[55 – 60[** | 57.5 | 12 | 0.02 | 17.59 |
| **[60 – 65[** | 62.5 | 6 | 0.01 | 17.60 |
| **[65 – 70[** | 67.5 | 0 | 0.00 | 17.60 |
| **[70 -75[** | 72.5 | 184 | 0.27 | 17.87 |
| **[75 – 80[** | 77.5 | 204 | 0.30 | 18.18 |
| **[80 – 85[** | 82.5 | 1242 | 1.85 | 20.03 |
| **[85 – 90[** | 87.5 | 50478 | 75.25 | 95.28 |
| **[90 – 95[** | 92.5 | 2652 | 3.95 | 99.23 |
| **[95 – 100]** | 97.5 | 516 | 0.77 | 100.00 |
| **Total** | | 67081 | 100.00 | |

If the Cumulative Relative Frequency of each interval is plotted against the class mark, representing the mean genomic distance of each interval, we can empirically guess both thresholds required to establish Clonal Groups, according to specifications described in the Methods section (Figure 4).

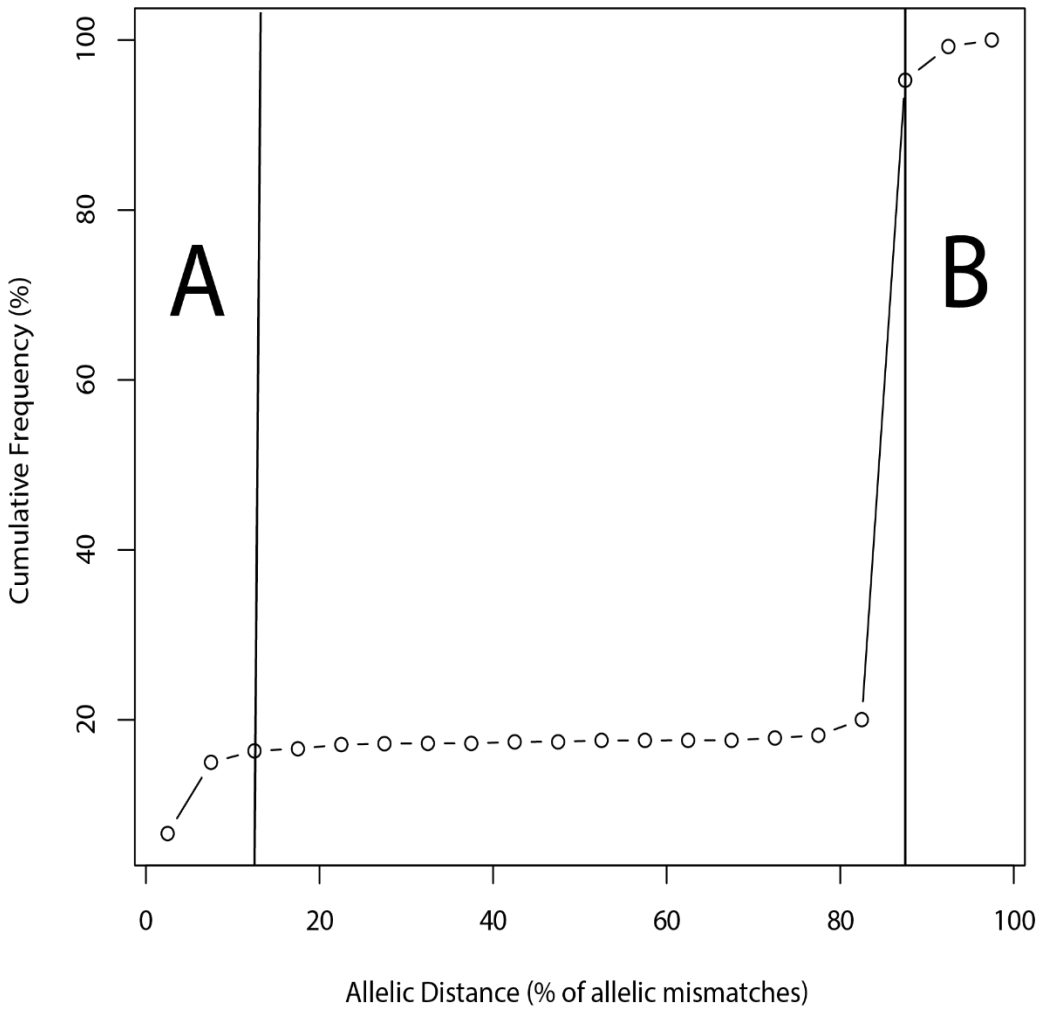


**Figure 4.** Distribution plot of Genomic Distances between Gv2_GIE492 assemblies. Regions A and B correspond to distances “inside groups” and “between groups” respectively.

As Figure 4 shows, the two inflection points corresponding to the distance thresholds can be easily determined. In this case, the threshold for “inside groups” genomic distances corresponded to 12.5% of allelic mismatches or 469 Loci Variants (nLVs). nLVs are defined as the number of loci in which two cgMLST profiles differ. Knowing the nLV threshold for “inside groups” genomic distances, we can transform the previously computed MST into a nLV graph, which clearly shows the Clonal Groups that

are determined by the distance thresholds. The MST and the nLV graphs are shown in Figure 5A and 5B, respectively.

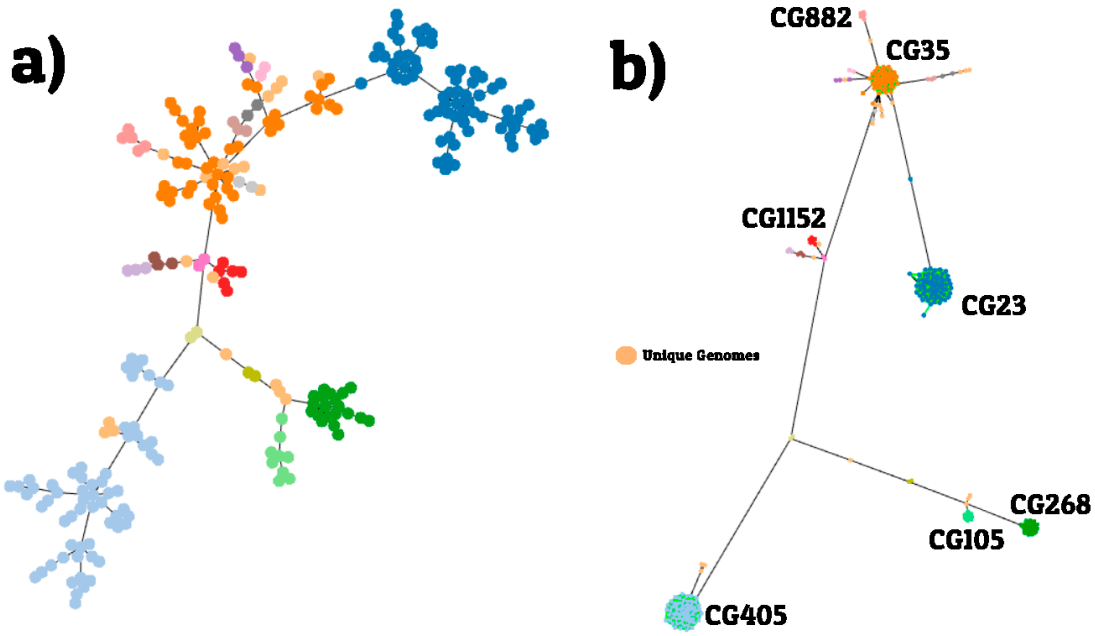


**Figure 5.** A) Minimum Spanning Tree of genomic distances between Gv2_GIE492 genomes. Each point corresponds to one assembly, and the connecting lines between them represent the genomic distance needed to convert one cgMLST profile into another. The longer the distance between two points in the graph, the more different they are to each other. B) nLV-graph computed from the previous MST, using 469 LVs as the nLV clustering threshold. The nLV-graph properties are equivalent to the MST, but points clustered into a Clonal Group act as single point in the graph. Both plots are color coded according to the determined CGs, which are also named in the nLV-graph. Unique genomes that do not match any CG are shown in light orange. An interactive version of these graphs can be found here https://bit.ly/MSTGIE492

Each genome in Gv2_GIE492 was assigned to a CG, except if the assembly was completely unique and dissimilar to all other assemblies, as shown in Figure 5B. Seventeen CGs were determined to carry GIE492, and each of them was named according to the most relevant ST present in the CG. CGs and the carried GIE492 structural variant are detailed in Table 6.

**Table 6.** *Klebsiella pneumoniae* GIE492+ Clonal Groups and the GIE492 Structural Variant (SV) present in them.

| CG | GIE492 SV | CG | GIE492 SV | CG | GIE492 SV |
|---|---|---|---|---|---|
| CG23 | **I** | CG380 | **I** | CG1552 | **I** |
| CG35 | **III** | CG405 | **II** | CG1630 | **I** |
| CG105 | **I** | CG542 | **I** | CG1799 | **I** |
| CG151 | **I** | CG875 | **I** | CG2990 | **I** |
| CG268 | **II** | CG882 | **I** | CG3661 | **I** |
| CG321 | **I** | CG1145 | **IV** | | |

The results from Table 6 clearly show that every GIE492+ CG associates with one specific structural variant. However, comparing results shown in figure 5B and table 6, the GIE492 structural variant carried by CGs is not related to proximity between them. The structural variant I was the most common among CGs, while structural variants III and IV were restricted to specific CGs. In this regard, the GIE492 variant present in Kpn RYC492 (III) seems to be quite rare and only associated to CG35.

### 2.3 Association of GIE492 with hypervirulent and carbapenem-resistant clones

We aimed to test the possible association of GIE492 with clinically relevant clones, namely, hypervirulent and carbapenem-resistant strains. For this, we used the tool Kleborate to calculate virulence and antibiotic resistance scores for the Gv2_GIE492 genome set. The Virulence Score (VS) is an integer number between 0 and 5, that specifies if a genome carries the more relevant virulence factors associated to Kpn. Each number has a specific meaning explained below:

- 0 = negative for all yersiniabactin (*ybt*), colibactin (*clb*) and aerobactin (*iuc*) loci.
- 1 = yersiniabactin only.
- 2 = yersiniabactin and colibactin (or colibactin only).
- 3 = aerobactin (without yersiniabactin or colibactin).

- 4 = aerobactin with yersiniabactin (without colibactin).
- 5 = yersiniabactin, colibactin and aerobactin.

Note that neither salmochelin (*iro)* locus nor *rmpADC* are explicitly considered in VS. The *iro* and *rmpADC* loci typically appear alongside the aerobactin *iuc* locus on KpVPs, so a score from 3 – 5 generally implies the presence of both *iro* and *rmpADC*. Moreover, aerobactin is specifically associated with growth in blood and is the most accurate predictor of the hypervirulence phenotype (Marr & Russo, 2019). For this reason, VS 3 – 5 will be used to predict if a strain is hypervirulent or not.

The Resistance Score (RS) works the same way as the VS, but ranges from 0 to 3. RS meanings are specified below:

- 0 = no ESBL (Extended-Spectrum β-lactamases), no carbapenemase.
- 1 = ESBL, no carbapenemase.
- 2 = Carbapenemase without colistin resistance.
- 3 = Carbapenemase with colistin resistance.

We next integrated for all the GIE492+ genomes the information of VS and RS, the structural variants of GIE492 identified, and the CG assigned, building a phylogenic tree based on MASH distances following the same cgMLST methodology described before. This tree included the 259 Gv2_GIE492 assemblies plus 31 outgroup genomes from Gv2, covering different *Klebsiella* lineages. These outgroup genomes do not carry GIE492 and were used to maintain the general structure of the full tree presented in Figure 2. The 31 outgroup genomes are listed in Table 7.

**Table 7.** Outgroup Gv2 assemblies used to construct a reduced phylogenetic tree of Gv2_GIE492 assemblies.

| Genome | Species | Genome | Species | Genome | Species |
|---|---|---|---|---|---|
| 0008 | *Kmi* | 1351 | *Kae* | 2376 | *Kp1* |
| 0075 | *Kox* | 1831 | *Kp2* | 2745 | *Kox* |
| 0083 | *Kp3* | 1928 | *Khu* | 2778 | *Kp4* |
| 0091 | *Kor* | 2012 | *Kp2* | 3800 | *Kor* |
| 0604 | *Kp1* | 2124 | *Kp1* | 3813 | *Kte* |
| 0799 | *Kgr* | 2169 | *Kpa* | 3826 | *Kpl* |
| 0924 | *Kp6* | 2243 | *Kmi* | 3827 | *Kpl* |
| 0982 | *Kp1* | 2267 | *Kae* | 3831 | *Kp3* |
| 1031 | *Kp4* | 2314 | *Kgr* | 3850 | *Kte* |
| 1200 | *Kp6* | 2315 | *Kp1* | | |
| 1304 | *Kp1* | 2331 | *Kp5* | | |

Figure 6 shows the phylogenetic tree representing the 290 selected genomes and the relationship between CGs, GIE492 variants, VS, and RS. As expected, all the Kp1 GIE492+ genomes clustered together, close to the KpSC outgroup genomes and distant to the rest of *Klebsiella* species, including the *K. michiganensis* GIE492+ genome. Also, hypervirulence (VS=3-5) and carbapenem resistance (RS=2-3) occurred in GIE492+ strains from defined CGs. More details regarding the CGs, GIE492 variants, hypervirulence, and carbapenem resistance among the GIE492+ *Klebsiella* genomes are summarized in Table 8. Only CG23 and CG380 harboring GIE492 are hypervirulent strains. Also, carbapenem-resistant GIE492+ strains belonged mostly to CG405 and to a lesser extent CG23 and CG35. Moreover, CG151 and CG1799 showed CRKp GIE492+ clones, although in these cases it is hard to generalize due to the very limited number of published genomes from these lineages. It is important to mention that according to our results, 21.7% of the CG23 hypervirulent GIE492+ strains were also carbapenem-resistant, proving the existence of GI492+ genomes where hypervirulence and carbapenem resistance converges.

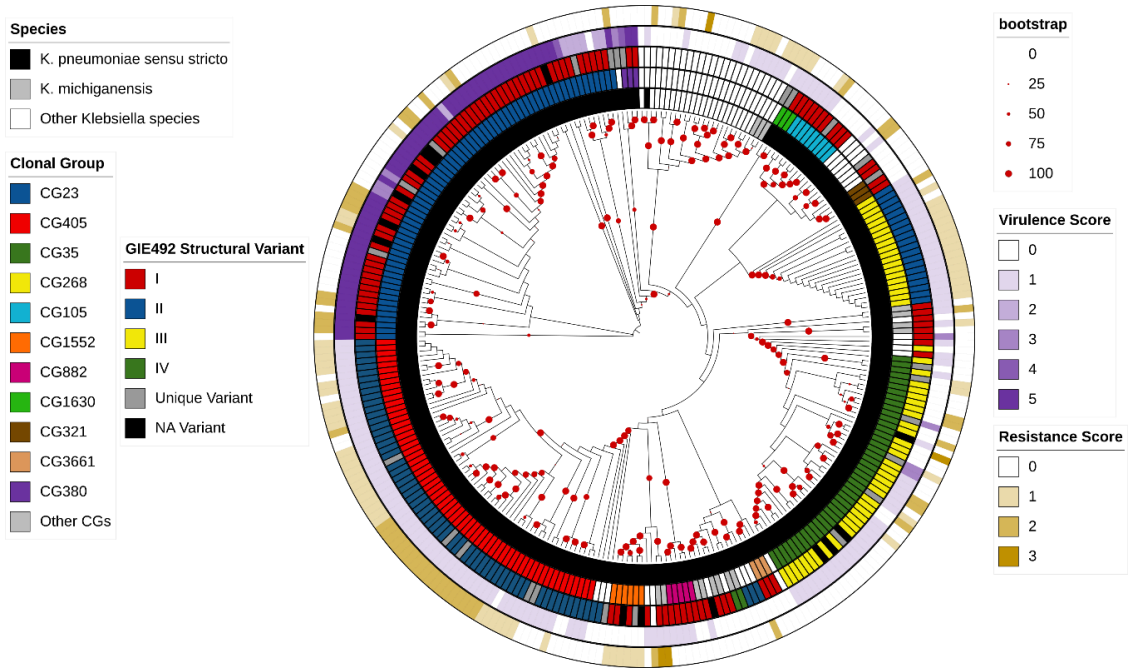


**Figure 6.** Reduced Phylogenetic Tree based in MASH distances of selected genomes. In this tree, branch lengths were ignored to make phylogenetic relationships clearer, due to including outgroups of 14 different species. Bootstrap values are represented as red dots. From the inside out, the first track makes the distinction between species. The second track identifies the clonal group. The third track shows the GIE492 Structural Variant carried by the genome. The fourth and fifth tracks identify the Virulence and Resistance Scores associated with the assemblies, respectively.

**Table 8.** Relationship between clonal groups, GIE492 structural variants, hypervirulence and carbapenem resistance.

| Clonal Group | Number of Strains | GIE492 SV | Percentage of hvKp strains | Percentage of CRKp strains |
|---|---|---|---|---|
| CG23 | 69 | I | 88.4 | 21.7 |
| CG405 | 63 | II | 0.0 | 31.7 |
| CG35 | 46 | III | 6.5 | 15.2 |
| CG268 | 21 | II | 0.0 | 4.8 |
| CG105 | 9 | I | 0.0 | 11.1 |
| CG882 | 6 | I | 0.0 | 0.0 |
| CG1552 | 6 | I | 0.0 | 0.0 |
| CG321 | 3 | I | 0.0 | 0.0 |
| CG380 | 3 | I | 100.0 | 0.0 |
| CG1630 | 3 | I | 0.0 | 0.0 |
| CG3661 | 3 | I | 0.0 | 0.0 |
| CG151 | 2 | I | 0.0 | 50.0 |
| CG542 | 2 | I | 0.0 | 0.0 |
| CG875 | 2 | I | 0.0 | 0.0 |
| CG1145 | 2 | IV | 0.0 | 0.0 |
| CG1799 | 2 | I | 0.0 | 100.0 |
| CG2990 | 2 | I | 0.0 | 0.0 |

To get a broader perspective, the 30 most relevant STs of Gv2, considering the relevant data presented in the Introduction section and the results from table 3 were also screened for hypervirulence and carbapenem resistance, results shown in table 9.

**Table 9.** Distribution of GIE492+, hvKp and CRKp strains in the most relevant Sequence Types of Gv2.

| ST | Number of strains | GIE492 presence (%) | Percentage of hvKp | Percentage of CRKp |
|---|---|---|---|---|
| ST11 | 635 | 0.00 | 14.17 | 84.25 |
| ST258 | 477 | 0.00 | 0.21 | 94.76 |
| ST101 | 189 | 0.00 | 0.53 | 66.14 |
| ST15 | 170 | 0.00 | 3.53 | 45.29 |
| ST45 | 126 | 0.00 | 1.59 | 50.79 |
| ST14 | 88 | 0.00 | 0.00 | 50.00 |
| ST147 | 83 | 0.00 | 1.20 | 81.93 |
| ST23 | 79 | 81.01 | 91.14 | 21.52 |
| ST405 | 62 | 98.39 | 0.00 | 32.26 |
| ST16 | 55 | 0.00 | 0.00 | 65.45 |
| ST35 | 54 | 72.22 | 14.81 | 20.37 |
| ST307 | 53 | 0.00 | 0.00 | 50.94 |
| ST48 | 47 | 0.00 | 14.89 | 12.77 |
| ST17 | 45 | 0.00 | 0.00 | 22.22 |
| ST37 | 44 | 0.00 | 2.27 | 38.64 |
| ST231 | 43 | 0.00 | 74.42 | 74.42 |
| ST29 | 40 | 0.00 | 5.00 | 7.50 |
| ST25 | 39 | 0.00 | 10.26 | 15.38 |
| ST336 | 32 | 0.00 | 0.00 | 21.88 |
| ST268 | 27 | 77.78 | 18.52 | 14.81 |
| ST340 | 23 | 0.00 | 0.00 | 60.87 |
| ST152 | 22 | 0.00 | 0.00 | 40.91 |
| ST395 | 22 | 0.00 | 18.18 | 68.18 |
| ST111 | 19 | 0.00 | 0.00 | 15.79 |
| ST39 | 19 | 0.00 | 0.00 | 5.26 |
| ST86 | 19 | 0.00 | 57.89 | 26.32 |
| ST20 | 17 | 0.00 | 0.00 | 5.88 |
| ST65 | 12 | 0.00 | 75.00 | 33.33 |
| ST380 | 3 | 100 | 100 | 0.00 |
| ST66 | 2 | 0.00 | 100 | 0.00 |
| **Total** | 2546 | 7.38 | 10.25 | 63.35 |

As it is shown in table 9, a clear bias towards CRKp strains is present in Gv2, a expected feature in assemblies obtained from a public database due to their clinical

relevance. Moreover, this characteristic is also useful to reveal the prevalence of GIE492+ genomes compared to other types of assemblies in public databases, being far less that CRKp and even less than hvKp strains. Table 9 also reveals that carbapenem resistance is broadly distributed in the Kpn population, being a relevant characteristic of many lineages, and that hypervirulence is a rare phenotype, even inside STs that present these types of strains.

Therefore, GIE492 was found in strains from different CGs with widely ranging virulence and resistance genetic potential. In this scenario, no direct association between the presence or the structural variant of GIE492 and hypervirulence can be inferred, since many GIE492+ genomes have a low predicted virulence. However, we found that GIE492, is highly prevalent among the hypervirulent clones CG23 and CG380, and thus it may contribute, in combination with virulence factors encoded elsewhere in the genome to the pathogenicity of these strains. Moreover, the sporadic presence of GIE492 inside some clonal groups including CG23 supports the instability of this element, which could be gained or lost.

## 3. GIE492 instability and dissemination

### 3.1 GIE492 integration site usage and evidence for horizontal transference.

To further explore possible evidence supporting GIE492 instability and dissemination through horizontal gene transfer, we first analyzed the integration sites used by this element across the different GIE492+ *Klebsiella* genomes. For this, all the assemblies in Gv2_GIE492 were annotated using PROKKA. This process involved establishing the coordinates (start and end positions) of all possible genomic features

that exist in the genome, such as protein-coding genes, tDNAs, and rRNAs, among others. Next, these features were aligned to an internal database and identified according to a best-hit approach. Those features where only the open reading frame can be deduced, but do not align to any known sequence, are labelled as hypothetical features. PROKKA annotation process is standardized, meaning that under the same configuration parameters, any genome will always be annotated in the same way, according to the internal database. This trait allows comparison between PROKKA annotated genomes to know which genomic features are shared, and the genomic context where they are positioned, meaning which other genomic features are next to them.

Because the genomic coordinates of GIE492 features were already stablished, the genome annotation can be used to determine the genomic context of this GI in each genome. According to what is already known about Kpn genomes and GIs, any element that interrupts the virgin context of any of the 86 core tDNAs, constitutes a putative GI. Moreover, as mentioned in the Introduction section, Marcoleta *et al.* (2016) did a proper description of the Kpn tDNAs most frequently used as integration sites for GIs, and proposed a nomenclature to distinguish between them, according to the anticodon codified, the nucleotidic sequence and the genomic context. Furthermore, in this previous study, GIE492 was found found associated to one specific asparagine tDNA (*asn1C*). We considered all this evidence to analyze the GIE492 integration site.

GIE492 was visually inspected using SnapGene Viewer in annotated genomes, according to its genomic coordinates, and the tDNA where GIE492 is inserted was identified for each of the genomes in Gv2_GIE492 and named according to the

proposed tDNA nomenclature (Berríos-Pastén *et al*., 2020). Following this strategy, the tDNA used for GIE492 integration was properly identified in 237 out of 259 assemblies, while in the rest the genomic context was indistinguishable. In 235 genomes (99.2%) the GIE492 integration site was *asn1C*, with only two exceptions:

- GIE492 was inserted into *asn1A* in genome GCF_001031475.1 (CG405, SV II).
- GIE492 was inserted into *asn1D* in genome GCF_900508825.1 (CG105, SV I).

These results indicate that the two alelles identified for the *int* gene would encode a non-promiscuous integrase directing GIE492 integration mainly into *asn1C*, although other asn tDNAs could be used. Additionally, the two exceptions are evidence supporting that GIE492 is an unstable GI and that it is being disseminated in the Kpn population by horizontal gene transfer since, if GIE492 was only vertically inherited, the integration site should be the same in all cases. Besides, if we consider the previously shown evidence that GIE492 is present in multiple unrelated Kpn lineages and even in *K. michiganensis,* these two observations are also evidence that this GI is being disseminated in the *Klebsiella* population through HGT.

These observations and the results shown in table 6, will then suggest that GIE492 has been disseminated in the *Klebsiella* population through HGT events, and then been fixated in specific clonal groups, showing no structural variation through vertical inheritance.

### 3.2 Possible co-evolution of GIE492 and the ICE*Kp* family: association with hvKp clones

As mentioned in the Introduction section, previous evidence indicate that ICE*Kp* and GIE492 biology could be related, including a highly similar integrase targeting asn tDNAs and the presence in GIE492 of a cryptic transfer origin that would allow parasitizing the conjugative machinery of ICE*Kp*. Thus, we aimed to search for genomic evidence of the possible co-evolution of these two GIs and its potential association with hvKp. To this end, we used Kleborate to determine the presence and structural variant of the ICE*Kp* element, along with the *ybt* gene lineage, as previously defined (Lam et al., 2018b). This way, the ICE*Kp* presence in Gv2_GIE492 genomes was evaluated, determining the co-occurrence of both genomic islands.

188 out of 265 GIE492+ strains (~71%) also carried an ICE*Kp* GI. On the other hand, 188 out of 2955 (~6%) Gv2 genomes carrying an ICE*Kp* element also harbor GIE492. The ICE*Kp* presence and structural variants distribution among different GIE492+ CGs followed a mixed pattern, although some findings can be highlighted (Table 10). Among the most commonly ICE*Kp* variants co-occurring with GIE492 in different CGs are ICE*Kp*10, 3, 4, 2, 12 and an unknown ICE*Kp* element present in CG405. Also, the co-occurrence of these two GIs is specially frequent in CG23, CG405, and CG268. In general, from these strains only CG23 showed high VS, while the two last CGs showed reduced virulence potential, suggesting that the co-occurrence of both GIs can be highly frequent outside hypervirulent clones. However, an interesting finding was that the co-occurrence of GIE492 with the specific variant ICE*Kp*10 is highly restricted to CG23 and CG380 and occur with an outstanding frequency in these *K.*

*pneumoniae* lineages. Moreover, these two CGs carrying ICE*Kp*10 differ in the *ybt* locus present in the GI, which also makes them differ in the *clb* locus (Lam *et al.,* 2018a).

**Table 10.** Prescence of ICE*Kp* and structural variants seen for the element in GIE492+ strains.

| **Clonal Group** | **Number of Strains** | **ICE*Kp*+ strains (%)[1]** | **Seen ICE*Kp* SVs[2]** |
|---|---|---|---|
| CG23 | 69 | 91.3 | ICE*Kp*10 (91.3)[3] |
| CG405 | 63 | 93.65 | ICE*Kp*2 (7.94)<br>ICE*Kp*5 (1.59)<br>ICE*Kp*12 (1.59)<br>Unknown ICE*Kp* (80.95)<br>Unknown *ybt* (1.59) |
| CG35 | 47 | 60.87 | ICE*Kp*2 (4.35)<br>ICE*Kp*3 (19.57)<br>ICE*Kp*4 (15.22)<br>ICE*Kp*5 (2.17)<br>ICE*Kp*6 (6.52)<br>ICE*Kp*9 (10.87)<br>ICE*Kp*12 (2.17) |
| CG268 | 21 | 100 | ICE*Kp*3 (100) |
| CG105 | 9 | 44.44 | ICE*Kp*6 (11.11)<br>ICE*Kp*11 (11.11)<br>Unknown *ybt* (22.22) |
| CG1552 | 6 | 0.0 | - |
| CG882 | 6 | 16.67 | ICE*Kp*4 (16.67) |
| CG1630 | 3 | 100.0 | ICE*Kp*12 (100) |
| CG321 | 3 | 0.0 | - |
| CG3661 | 3 | 0.0 | - |
| CG380 | 3 | 100.0 | ICE*Kp*10 (66.67)[4]<br>ICE*Kp*12 (33.33) |
| CG1145 | 2 | 0.0 | - |
| CG1799 | 2 | 0.0 | - |
| CG151 | 2 | 100.0 | ICE*Kp*4 (50)<br>ICE*Kp*11 (50) |
| CG2990 | 2 | 0.0 | - |
| CG542 | 2 | 50.0 | ICE*Kp*6 (50) |
| CG875 | 2 | 50.0 | ICE*Kp*5 (50) |
| Unique Genomes | 21 | 9.52 | ICE*Kp*12 (4.76)<br>Unknown *ybt* (4.76) |

1. Because all genomes considered for this table are part of Gv2_GIE492, this percentage is also the co-occurrence of both GIs.
2. The prevalence percentage of the ICE*Kp* structural variant inside the CG is stated in parenthesis.
3. ICE*Kp*10 carrying *ybt* 1.
4. ICE*Kp*10 carrying *ybt* 12.

To easily visualize the results described in table 10, a slight modification of the phylogenetic tree presented in figure 6 was done to show the presence of ICE*Kp* in GIE492+ genomes and the ICE*Kp* structural variant associated with them (Figure 7).

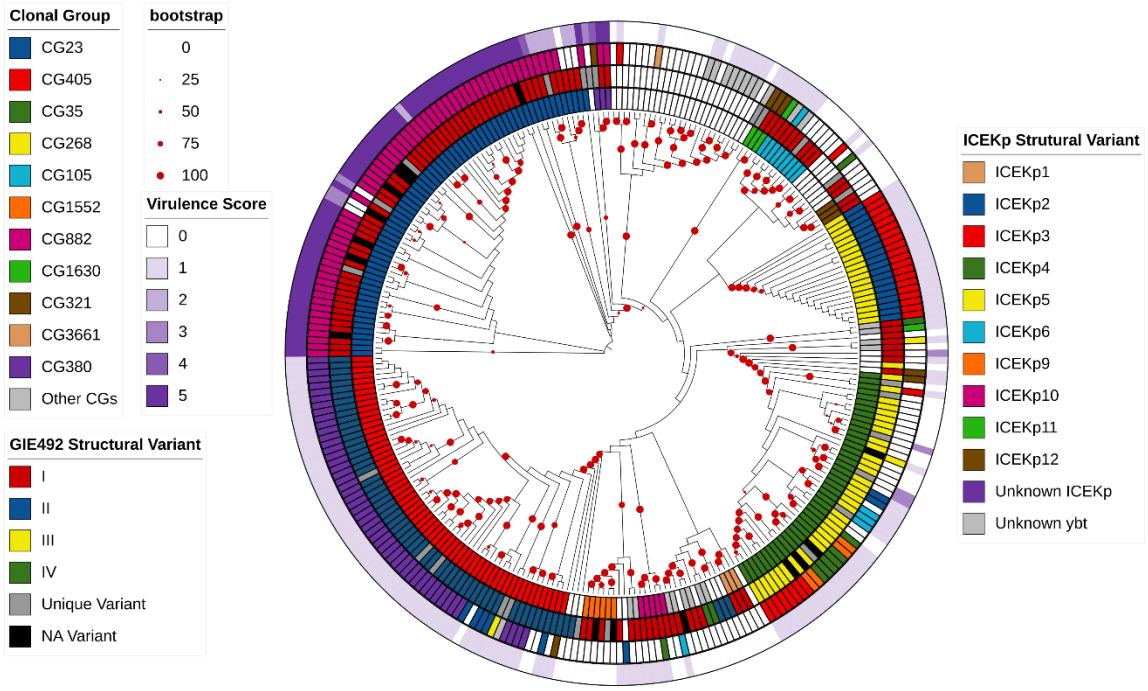


**Figure 7.** Reduced Phylogenetic Tree based in MASH distances of selected genomes. In this tree, branch lengths were ignored to make phylogenetic relationships clearer, due to including outgroups of 14 different species. Bootstrap values are represented as red dots. From the inside out, the first track shows the clonal group of the genome. The second and third tracks correspond to the type of structural variant found for both GIE492 and ICE*K*p respectively. The fourth track represents the virulence score associated.

Analyzing Figure 7, we can infer that the combination of carrying GIE492 I and possessing ICE*Kp*10 is an strong indicator of hypervirulence. To validate this statement, performance metrics were calculated considering the following: 222 genomes in Gv2 do not have a GIE492 SV, ICE*Kp* SV or both properly identified, so only 3656 genomes were considered. The calculated values are presented in Table 11 below:

**Table 11.** Performance metrics for the simultaneous presence of GIE492 SV I and ICE*Kp*10 integrated in the chromosome as a predictor of the hypervirulence phenotype in *Klebsiella pneumoniae.*

| Metric | Value | Formula |
|---|---|---|
| **Positives (P)** | 363 | - |
| **Negatives (N)** | 3293 | - |
| **True Positives (TP)** | 50 | - |
| **True Negatives (TN)** | 3289 | - |
| **False Positives (FP)** | 4 | - |
| **False Negatives (FN)** | 313 | - |
| **Sensitivity** | 0.138 | TP/P |
| **Specificity** | 0.999 | TN/N |
| **Precision** | 0.926 | TP/(TP + FP) |
| **Miss Rate** | 0.862 | FN/P |
| **Fall-out** | 0.001 | FP/N |
| **Diagnostic Odds Ratio** | 131.350 | (TP/FP)/(FN/TN) |

As stated in the original work where DOR (Diagnostic Odds Ratio) was defined (Glas *et al.,* 2003), the value of this metric for a diagnostic test, ranges from 0 to infinity, where useful tests have a value greater than 1, improving the performance as the value increases. According to this definition, the simultaneous presence of GIE492 I and ICE*Kp*10 in a *Klebsiella* genome is a reliable predictor for hypervirulence. However, this extremely reductionist approach strongly affects the sensitivity of the test, due to ignoring the virulence factors present in extrachromosomal elements, such as the KpVPs presented in the Introduction section.

Next, we investigated the integration site usage of ICE*Kp* in GIE492+ strains, using the same strategy described before for identifying the integration site of GIE492. Because it is known that ICE*Kp* can be integrated in any of the four asn-tDNAs in the Kpn genome (Lam *et al.,* 2018a) and also that these tDNAs are hotspots for GI integration (Marcoleta *et al.,* 2016), the way these tDNAs are used in GIE492+ genomes, was an interesting topic to consider.

Following the strategy presented in the Methods section, the ATOP for all genomes in Gv2_GIE492 was determined and is presented in figure 8.

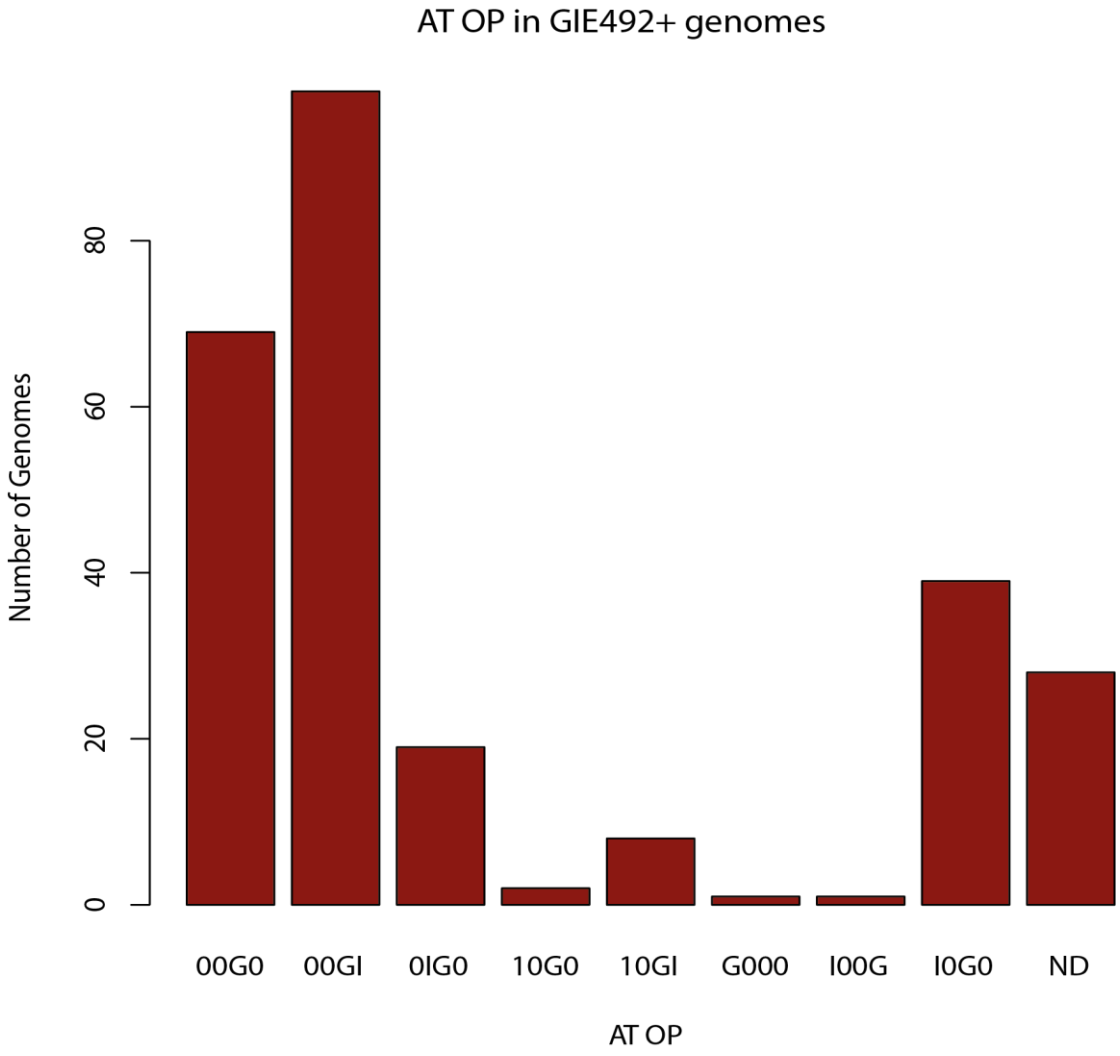


**Figure 8.** Barplot representing the frequencies of all asn-tDNAs Ocuppation Patterns found in Gv2_GIE492 genomes.

According to figure 8, the most common insertion site for ICE*Kp* in GIE492+ genomes corresponds to *asn1D* followed by *asn1A* and *asn1B.* According to what was previously stated by Lam *et al.* (2018a), ICE*Kp* can be found inserted into any of the asn tDNAs, so multiple types of ATOP were found as expected. However, the

prevalence of GIE492 for *asn1C* gives an apparent impossibility for ICE*Kp* to be inserted into this tDNA when both GIs are present in a genome. It is important to mention that in 10 different genomes a third putative GI inserted into *asn1A* was found.

On the other hand, as mentioned before, GIE492 carries a cryptic transfer origin (oriT) sequence that is shared with ICE*Kp* elements and other putative GIs inserted in asn-tDNAs. According to *in vitro* experiments, the presence of this oriT is enough for a plasmid to be transferred by conjugation in strains that carry a conjugative system such as the one encoded in ICE*Kp* (Lin *et al.,* 2008). This could be potentially true for other circular DNA elements such as excised GIs. Thus, it was proposed that upon excision, the circular GIE492 intermediate could be mobilized by conjugation in genomes where this GI coexists with ICE*Kp* (Marcoleta *et al.*, 2016). Given the importance of the cryptic oriT, we aimed to evaluate the presence and conservation of this sequence element among the GIE492+ genomes.

According to this idea, local alignments using BLASTn were done in Gv2_GIE492 genomes against the cryptic oriT sequence, the same way as described before. The 508 bp oriT sequence was found in all genomes, with a BSR value of at least 0.97. Non redundant clusterization of oriT sequences using CD-HIT-EST, yielded a single group. The oriT sequence was invariably located after *mceF* and before the group of expendable genes in GIE492, as shown in Figure 1.

Considering all this evidence, all four GIE492 SVs share characteristics that would classify them as unstable GIs: the presence of an active integrase gene that targets *asn1C;* an sporadic and unrelated prevalence in distinct clonal groups in the *Klebsiella* phylogeny; a strong evidence of active events of HGT; a clear relationship

with elements of the ICE*Kp* family, specially with ICE*Kp*10 in hvKp strains, suggesting a possible co-evolution of both GIs; and a conserved oriT sequence that also gives this co-evolution an apparent co-mobilization approach.

## 4. Tracking the excision of GIE492 and other mobile genetic elements in hypervirulent *K. pneumoniae* cultures.

### 4.1 Setting up a strategy based on next-generation sequencing for tracking GI dynamics

Considering the importance of GIE492 and other GIs in *K. pneumoniae* evolution, the complex dynamics of these elements, and the scarcity of tools leveraging whole-genome sequencing data to evaluate genome-wide GI instability, we aimed to develop a bioinformatics pipeline to experimentally detect the excision of GIE492 and other mobile elements in cultures of a model hypervirulent *K. pneumoniae* strain. Although a previous work described a tool that would serve for these purposes (Schoeniger *et al.,* 2016), upon several attempts, we were unable to successfully run it with our datasets due to software errors of unknown nature. Unfortunately, no other published works reported succesfull application of this tool, besides the work in where it was originally described. Thus, there is a significant gap in the available toolbox to study GI instability that we aimed to address using a reads mapping-based approach. Before describing the pipeline development, some concepts must be introduced.

Next-generation sequencing of genomic isolates using Illumina technologies is a process where genomic DNA is first fragmented in millions of small sequences, amplified by PCR and then subjected to a fast and accurate sequencing by synthesis

strategy using fluorescent modified nucleotide bases that emit a unique fluorescence signal that is detected by the sequencing machine. In microbial genomics applications, the sequencing reads generated are then assembled trying to reconstruct the complete genomic sequence from which the reads were obtained. Most protocols consider a final step of aligning the original reads to the contigs obtained, called polishing, with the objective of solving the multiple assembly errors that occur during the process. The strategy of aligning reads to contigs is called “reads mapping”.

Due to the importance of reads mapping in genomic procedures, multiple software have been developed for this purpose, such as Bowtie (Langmead, 2010). The mapping process to assemblies introduces the concept of “depth of coverage”, which corresponds to the amount of reads that align to any region R of length L in the assembly. In a complete assembly, a consistent number of reads should map to all regions of the genome. This is sign of robustness and is the reason why a high average depth of coverage is associated with high quality genomes.

A depth of coverage map is a graph that shows how many reads are aligned to every nucleotide in the genomic sequence. However, due to the small length of Illumina reads (100-150 bp), part of them tend to map to multiple genomic regions that share a similar sequence. This consideration introduces three key concepts: 1) Mapping quality (MAPQ): is a measure of the probability that an alignment is wrong and is expressed as the negative logarithm of this probability, so the bigger the MAPQ value, the smaller the probability of a mapping error. 2) When a read is being aligned, all regions to where the read maps receive a MAPQ value. The region with the highest MAPQ value is called

“primary alignment” and is unique. 3) Any other region where the read aligns is called a “secondary alignment”.

Even if reads can map to multiple regions, the baseline in mapping counts per region tends to be consistent, as explained before. However, it is possible that some specific genomic regions can concentrate more mappings than others. Likely, these regions should belong to any of these two types: 1) Genomic regions with multiple copies in a genome, such as rRNA clusters; or 2) Unique genomic regions that can exist in two different states in a bacterial population, integrated and exscinded, such as unstable GIs. Therefore, reads mapping approaches measuring differences in coverage depth across the chromosome should allow to detect both kinds of regions, being the latter the main focus of our approach. With these concepts in mind, we developed the following strategy to attempt detecting GI excisions using next-generation sequencing data:

1. To divide an assembled chromosome in multiple small regions “R” of length “L”. This length should be about the size of a single gene (500 bp).

2. To map the reads that were used for assembling the chromosome to these regions and count the number of total alignments and primary alignments.

3. To determine the following probabilities for every region R based on the number of alignments:

- P(A) = Probability of primary alignments in R.
- P(B) = Probability of R having more mappings that the rest.
- P(B|A) = Probability of R having more mappings than the rest given that these mappings correspond to primary alignments.

- P(A|B) = Probability of primary alignments in R given that R has more mappings than the rest. This probability was named “excision probability”.

4. To tabulate all the excision probabilities obtained and get the median value of P(A|B) considering all regions.

5. To normalize P(A|B) dividing by the median in every region to obtain the coverage depth fold-change with respect to the baseline.

6. To graph the coverage depth fold-change vs genomic coordinate and identify all the regions where the fold change is greater than a threshold value “T”. For the purpose of this work, the value of T was set to the 95th percentile of the excision probability distribution.

7. To inspect the genes associated to these regions in the annotated chromosome searching for evidence of being part of a mobile genetic element.

To put this strategy in practice, custom scripts were developed in Perl programming language. During the mapping step, the software Bowtie2 was used, and the following considerations were implemented: 1) Only consider global alignments, that is, map the whole read without any trimming; and 2) Report all alignments for every read instead of the default mode that only reports the best one.

The counting step was performed using SAMtools, only considering alignments with a MAPQ value 30 or higher. The P(A) probability was calculated dividing the number of primary alignments in a region by the total amount of primary alignments in all regions. The probabilities P(B) and P(B|A) were calculated using the following strategy:

- To use R default libraries to make a histogram of the number of alignments observed in every region.
- To calculate the cumulative relative frequency of every interval in the histogram.
- To assign to R the cumulative relative frequency value of the interval as probability, if the number of alignments in the region falls into that interval.
- If a region has 0 alignments the probability is always 0.
- P(B) uses the number of total alignments for calculations, while P(B|A) uses the number of primary alignments.

The excision probability was calculated using the standard Bayes Theorem. The simplified pipeline followed by the developed Perl script is shown in Figure 9.

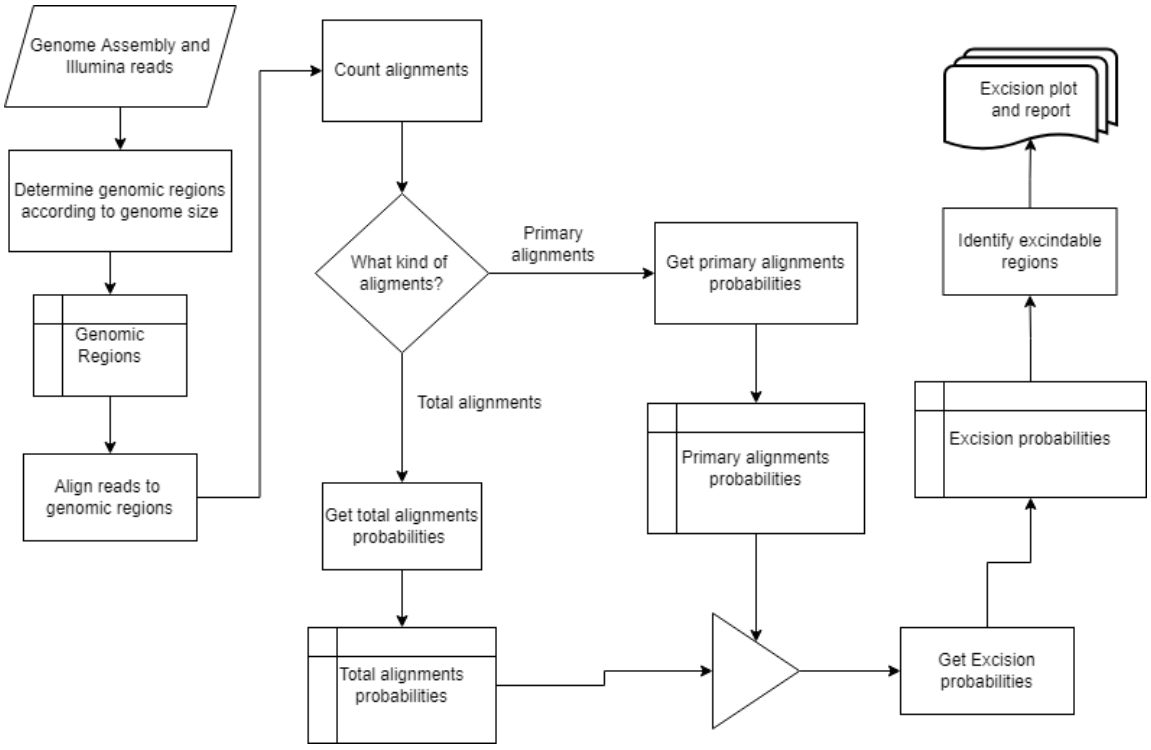


**Figure 9** Flowchart representing the developed strategy to identify genomic regions which can be exscinded from the chromosome using Illumina reads obtained by next-generation sequencing.

As it was explained before, the chromosomal regions used for determination of excision probabilities in this pipeline are very short, 500 bp to be exact, an approach based on the length of a single gene instead of the length of an arbitrarily long putative MGE. This is due to the necessity of being able to identify if the candidate regions with normalized excision probabilities above the baseline threshold belong to elements inside excisdanble MGEs or correspond to genes that exist in multiple copies in the bacterial chromosome. However, once the uniqueness of the region has been identified, this is not enough evidence that the region belongs to a putative excisdanble MGE which can be thousands of bp long.

To solve this problem, the concept of “blocks” needs to be introduced: a block corresponds to a group of candidate regions that posses close proximity to each other, where the close proximity parameter “C” corresponds to a arbitraly chosen distance length threshold that groups all candidate regions and non-candidate neighbors into a single block, if the distance between them is below this threshold. For the purposes of this work, the value of C was arbitrarily set to 5000 bp.

## 4.2 Known GIs and EGM sequences can be predicted in their excised states using the developed strategy

The previously described strategy and associated developed Perl script, was tested using the Kpn SGH10 assembled chromosome FASTA sequence retrieved from the NCBI genome database, and Illumina reads generated in the context of the present study (more details can be found in the Methods section). Two readsets were generated from genomic DNA extracted from exponential phase cultures of the model

hypervirulent strain *K. pneumoniae* SGH10, treated or not with the DNA damaging agent mitomycin C, previously demonstrated to induce the excision of lysogenic phages and GIs including GIE492 (Marcoleta *et al.,* 2016).

After following the complete pipeline using both read sets and comparing to the PROKKA annotated chromosome (complemented with MGE annotations obtained using Kintun-VLI, PHASTER and ISEScan), several regions which concentrate significantly higher mapped reads were identified and then were grouped in blocks as described previously. For testing purposes and due to possessing two types of read sets, an special consideration was applied when constructing blocks: candidate regions could belong to any ot the two sets, whether or not they were above the baseline threshold in only one set or both. According to this, 3 types of blocks were created: “A blocks” when candidate regions were above the baseline only in the untreated condition set (SG), “B blocks” when the candidate regions were above the baseline only in the Mitomycin C treatment set (SGM), or “C blocks” when candidate regions where above the baseline in both conditions.

As expected, candidate regions belonged to both predicted groups: genes that have multiple copies in the chromosome, such as rRNAs, and genes related to EGMs expected to be excised from the chromosome in some proportion such as GIs. The most relevant blocks obtained by grouping these regions and genes contained in them are specified in table 12.

**Table 12.** Selected blocks identified by grouping candidate regions above the baseline depth of coverage threshold in the SGH10 chromosome by applying the complete developed pipeline to detect excidanble chromosomal regions.

| Block | Start | End | Block Type | Relevant Features |
|---|---|---|---|---|
| **1** | 42751 | 87750 | C | Putative GI *sec1A*<br>*lamB*<br>7 HTH-type transciptional regulators |
| **2** | 221751 | 231250 | C | 2 IS3 transposases |
| **3** | 321251 | 324251 | C | *livJ, livH* |
| **4** | 453751 | 459250 | C | 16S rRNA, 5s rRNA |
| **5** | 729251 | 737750 | C | Putative GI *phe1A* |
| **6** | 1136751 | 1138250 | C | IS5 transposase |
| **7** | 1209251 | 1214750 | B | 16S rRNA, 5s rRNA |
| **8** | 1834751 | 1840250 | C | ICE*Kp*10<br>Mobilization module genes |
| **9** | 1922251 | 1924750 | C | GIE492<br>*int, mceA, mceB* |
| **10** | 1943751 | 1944250 | A | GIE492<br>*u6* |
| **11** | 2012251 | 2225750 | C | Intact Phage<br>*livJ, livH*<br>4 HTH-type transcriptional regulators |
| **12** | 2232751 | 2251250 | C | 1 HTH-type transcriptional regulator |
| **13** | 2375251 | 2379750 | A | CRISPR locus |
| **14** | 2400751 | 2404750 | A | 2 IS3 transposases |
| **15** | 2718251 | 2726250 | A | 2 HTH-type transcriptional regulators |
| **16** | 3105251 | 3111250 | A | 1 HTH-type transcriptional regulator |
| **17** | 3318251 | 3332250 | C | Intact Phage |
| **18** | 3687251 | 3687750 | A | 1 HTH-type transcriptional regulator |
| **19** | 3970751 | 3972750 | B | 2 IS3 transposases |
| **20** | 4392251 | 4398250 | B | 16S rRNA, 5s rRNA |
| **21** | 4747751 | 4753250 | B | 16S rRNA, 5s rRNA |
| **22** | 4758251 | 4774750 | C | 1 HTH-type transcriptional regulator<br>*lamB* |
| **23** | 4867751 | 4903750 | C | 1 HTH-type transcriptional regulator |
| **24** | 5187751 | 5193750 | B | 16S rRNA, 5s rRNA |
| **25** | 5227251 | 5238750 | B | 16S rRNA, 5s rRNA |
| **26** | 5324751 | 5335250 | B | 16S rRNA, 5s rRNA |
| **27** | 5423751 | 5447250 | C | 16S rRNA, 5s rRNA<br>1 HTH-type transcriptional regulator |

According with table 12, the genes that are detected due to being in multiple copies correspond to: *livJ* and *livH,* both coding proteins involved in amino acids transport (Nazos *et al.,* 1986, Trakhanov *et al.,* 2005), *lamB,* involved in transport of maltose and maltodextrins (Schirmer *et al.,* 1995), and also the integration site for prophages such as lambda (Randall-Hazelbauer & Schwartz, 1973), Helix-Turn-Helix-type transcriptional regulators (Müller, 2001), IS3 and IS5 family transposases, and the 16S and 5S rRNA genes. The CRISPR locus also probably falls into this group due to its inherent repeated sequences. As expected, genes specific to unique excindable elements are also detected, including genes in the mobilization module of ICE*Kp*10 (Lam *et al.,* 2018a) and *mceA, mceB* and *u6* of GIE492, multiple genes (many of them coding hypothetical proteins) belonging to two putative GIs and to intact lysogenic bacteriophages. Interestingly, many of the rRNA clusters are only detected in B blocks, containing regions that increase their depth of coverage only under Mitomycin C treatment.

To compare both read sets, the normalized fold-change for SGM was divided by the equivalent measured value in SG for every region (Excision probabilities for all regions are included in Appendix 2). Figure 10 shows the calculated coverage depth fold change across the chromosome coordinate of the untreated read set, and of the quotient between the treated and untreated read sets (SGM/SG).

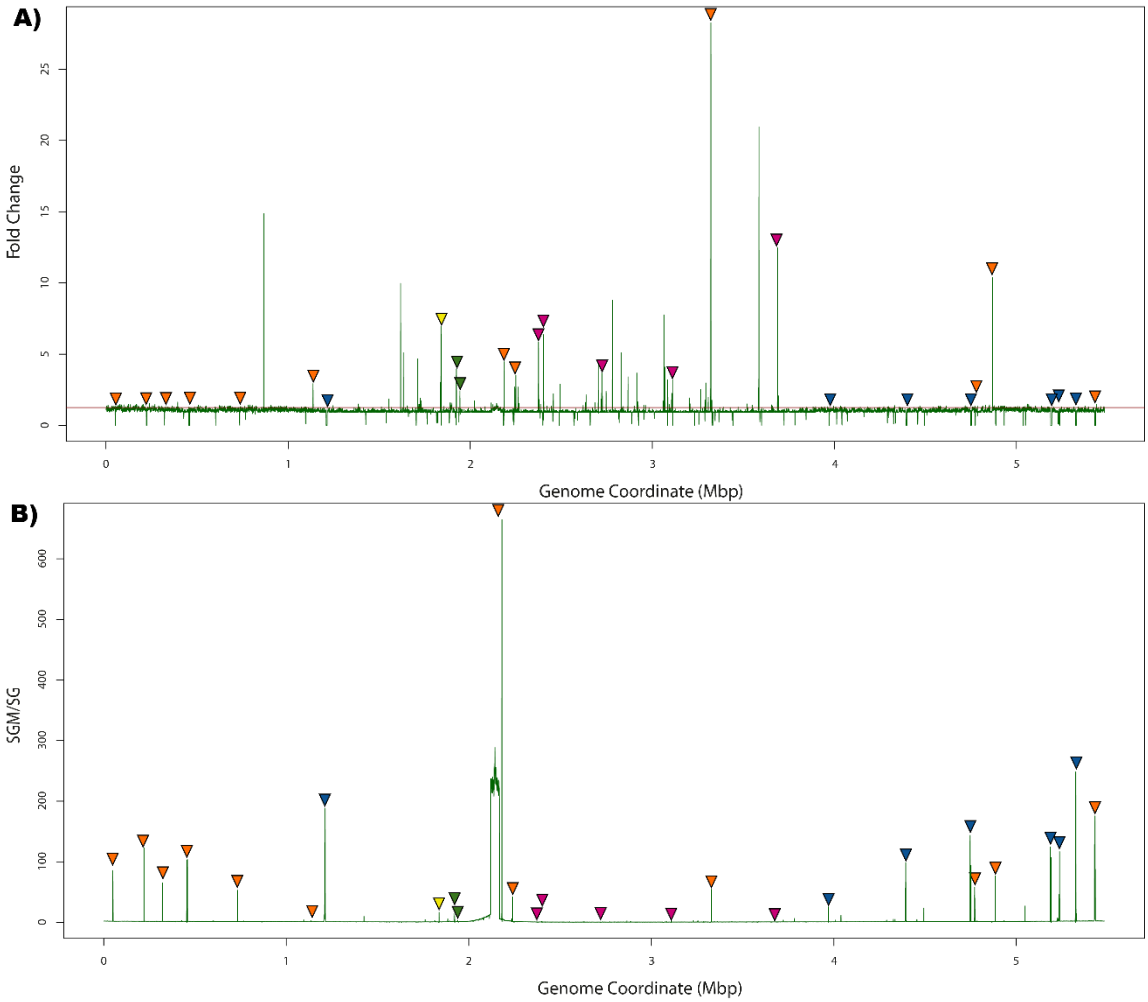


**Figure 10.** Comparison between the depth of coverage fold change graph obtained from the untreated read set (A) and the plotted quotient SGM/SG (B). The red horizontal line in figure 10A corresponds to the threshold value 1.26. Triangles point to every block described in table 12 in the same order. Magenta triangles correspond to A Blocks, blue triangles to B Blocks, orange triangles to C Blocks. The yellow and green triangles give emphasis to ICE*Kp* and GIE492, respectively.

The measured fold change in SG for all 27 relevant blocks shown in table 12 and the related calculated SGM/SG quotient, both of which were plotted in figure 10, are detailed below in table 13.

**Table 13.** Measured fold change values in SG and SGM/SG quotients for all selected blocks presented in table 12.

| Block | Highest SG fold change | Highest SGM/SG | Block Type |
|---|---|---|---|
| **1** | 1.48 | 85.65 | C |
| **2** | 1.31 | 122.41 | C |
| **3** | 1.31 | 64.90 | C |
| **4** | 1.27 | 103.66 | C |
| **5** | 1.32 | 52.81 | C |
| **6** | 2.94 | 2.69 | C |
| **7** | 0.00 | 188.39 | B |
| **8** | 6.97 | 16.61 | C |
| **9** | 3.97 | 14.97 | C |
| **10** | 2.43 | 0.38 | A |
| **11** | 4.46 | 664.78 | C |
| **12** | 3.53 | 42.29 | C |
| **13** | 5.85 | 0.28 | A |
| **14** | 6.37 | 0.31 | A |
| **15** | 3.67 | 0.46 | A |
| **16** | 3.20 | 0.37 | A |
| **17** | 28.26 | 54.48 | C |
| **18** | 12.45 | 0.04 | A |
| **19** | 0.21 | 25.32 | B |
| **20** | 0.41 | 99.03 | B |
| **21** | 0.50 | 143.64 | B |
| **22** | 1.36 | 58.16 | C |
| **23** | 10.37 | 76.53 | C |
| **24** | 0.10 | 124.66 | B |
| **25** | 0.66 | 117.11 | B |
| **26** | 1.05 | 248.24 | B |
| **27** | 1.50 | 175.66 | C |

The results reported in table 13 clearly show that many blocks that are naturally above the depth of coverage threshold (1.26) massively increase this value when treated with Mitomycin C, especially apparent in C Blocks that include all 6 excindable elements: two intact lysogenic prophages, two putative GIs, ICE*Kp* and GIE492.

The types of genes identified in the selected blocks and the increased fold change of the same blocks when treated with Mitomycin C, are evidence that the developed strategy can detect both types of regions that accumulate read mappings: genes in multiples copies in the chromosome, and genes related to EGMs that can be excised from the chromosome.

It is important to mention that particular peaks in both graphs shown in figure 10 that were not explained in this work correspond to blocks where genes mainly codify for hypothetical proteins, so a clear explanation for them was impossible to infer.

Now that the proof of concept for this strategy resulted successfully, interesting projections to consider include:

- To develop a strategy to automate the classification between blocks containing genes in multiple copies and genes contained in excindable elements.
- To make the tool work with draft genomes.
- To test the tool with different species and strains.
- To define a function to automatically detect which threshold is best suited to discriminate regions above the baseline.
- To define a function to automatically detect the close proximity threshold between regions inside the same block.
- To make the tool work with reads coming from different NGS technologies, such as ONT, which produces low quality long reads.
- To give the user the alternative to select between different software to map reads.

- To make the tool automatically annotate the genomic sequence with the regions above the baseline.

# DISCUSSION

## 1. Evidence of GIE492 dissemination by horizontal gene transfer and conservation during vertical inheritance

The genomic island GIE492 is a genomic feature present in specific Kpn lineages which are not phylogenetically related, trait that gives an apparent sporadic distribution in the *Klebsiella* phylogeny, as can be seen in the multiple phylogenetic trees presented in this work. If GIE492+ lineages were phylogenetically related, they should all have been part of a single monophyletic clade. However, we provided strong evidence against this possibility. Moreover, not even the main groups of structural variants, B and S, formed a monophyletic clade. This GIE492 variant distribution strongly suggest that this element was adquired multiple times in different *Klebsiella* lineages by horizontal gene transfer.

However, every clonal group that carries GIE492, only carries one specific structural variant of the element which remain highly conserved. This contrasting observation indicates that once a GIE492 structural variant is acquired by a Kpn clone through horizontal gene transfer, the element is fixated in the lineage through vertical inheritance. It is also important to remember that one of the GIE492+ genomes was taxonomically classified as *K. michiganensis*. This rare case can only be explained by a

horizontal gene transfer event. In this line, important questions arise regarding how rare is the presence of GIE492 in *K. michiganensis*, and why GIE492 is not found in other *Klebsiella* species which likely share the asparagine tDNAs used as integration site for this element. The GIE492 structural variant present in *K. michiganensis* is completely unique and it cannot even be classified as B or S, so it might be its own lineage developed independenly once the GI was adquired by HGT.

Another piece of evidence supporting horizontal gene transfer of GIE492 that cannot be missed, is the that in two completely unrelated genomes, GIE492 was inserted in an asparagine tDNA different from *asn1C*. If GIE492 was only vertically inherited, the GI should be found integrated into the same tDNA in all the genomes. According to these observations, GIE492 is disseminated through the *Klebsiella* phylogeny by horizontal gene transfer events, but very rarely, as only 17 known clonal groups carry the element. Complementarily, conservation of GIE492 in the population is product of vertical inheritance and likely strong selective pressure favoring the maintenance and invariability of the functions encoded in the island, mainly MccE492 production determinants.

Although our evidence indicates that GIE492 has been transferred horizontally, some barriers would prevent its dissemination to other *Klebsiella* species besides *K. pneumoniae* and other *Enterobacteriaceae*. The search for GIE492 in *Enterobacteriaceae* genomes excluding *K. pneumoniae* indicated that it can only be found in three other assemblies, the *K. michiganenis* strain presented in this work and two possible clones of *Escherichia coli* ST410, according to the Atchman MLST scheme (Wirth *et al.,* 2006), GCF_009913765.1 and GCF_009913755.1, having an

almost perfect match to the GIE492 SV III. Both of them carry ICE*Kp*11 and the aerobactin locus *iuc*, features that would classify them as hypervirulent clones. *E. coli* ST410 is know for being a ESBL+ clone present in southeast Asia (Nadimpalli *et al.,* 2019), feature that was confirmed, as both GIE492+ *E. coli* strains carry both a CMY and a CTX-M ESBLs.

Even though a concrete scheme for distinguishing asn-tDNAs in species different from Kpn has not been developed, GIE492 in Kmi and *E. coli* is inserted into the asn-tDNA closest to the genes *yeeO* and *amn,* key features of *asn1C* in Kpn (Marcoleta *et al.,* 2016), showing that *asn1C* tDNA targeting activity of the GIE492 integrase protein is highly conserved and basically an intrinsic feature of this GI. Moreover, ICE*Kp*11 in *E. coli* GIE492+ strains is inserted in a asn tDNA similar to *asn1D* (Marcoleta *et al.,* 2016), making a pseudo 00GI ATOP, which is the most frequent integration pattern observed in this work.

Even though GIE492 presence outside Kpn is very rare, it was expected to happen due to 1) the MccE492 production cluster is functional when expressed in other *Enterobacteriaceae* including *E. coli* and *Salmonella enterica* (Lagos et al., 2009, Marcoleta et al., 2018), thus potentially conferring the same adaptive advantage than when produced in *K. pneumoniae*; and 2) *E. coli* and other *Enterobacteriaceae* share with *K. pneumoniae* 100% conserved asparagine tDNAs that could potentially act as integration sites for GIE492 (Berríos-Pastén et al., 2020). Further studies are required to shed light on these barriers preventing a higher inter-species GI dissemination.

## 2. GIE492 Structural Variants are restricted in an evolutionary sense

Comparison between the four main structural variants of GIE492 gives an important result: 15 of the 23 encoded sequences are completely conserved at the protein level, including all genes required for MccE492 production, immunity, maturation, exportation, and regulation (Lagos *et al.,* 2009). On the other side of the spectrum, those genes that vary between SVs, all follow patterns that may explain variability. Expendable genes (*u1 – u5),* all encode proteins of unknown function, and even though all of them are transcribed (Marcoleta *et al.,* 2016), they do not seem to be involved in MccE492 production. Considering the evolutionary mechanisms shaping GIs and bacterial genomes where unnecessary genes tend to be lost (Dobrindt *et al.,* 2004), it is plausible that small GIE492 structural variants originate from a deletion event involving the *u1-u5* genes unrelated to MccE492 production. However, as this small variant was found in the unrelated CG268, CG405 and CG1145, multiple transfer events would have occurred after the deletion event in an originary CG thus reaching the other CGs. Alternatively, the deletion would have occurred multiple times in different CGs upon acquisition of the large, more ancestral variant. Fine analysis of the *u1-u5* region and a multiple sequence alignment of all the small GIE492 variants found in different CGs, indicated that the sequence downstream of *mceF* and upstream of *u6* is completely conserved among all genomes carrying small variants, with the exception of both CG1145 strains carrying the small structural variant IV and the three closest related genomes carrying the SV II, all of which are unique genomes without a CG assigned to them. This difference relies in a frameshift in position 15354 of the MSA, prior to the position of the conserved oriT sequence. This discovery might imply that at least two

different deletion events of the expendable cluster of genes have occurred. This evidence and the prescense of a large GIE492 SV in *E.coli* used as an outgroup organism may also imply that the B variants of GIE492 are the more ancestral types.

Interestingly, no insertion sequences or mobile elements remnants were found that could explain why and by which mechanisms the deletion would occur in this specific GIE492 region. If instead of a *u1-u5* deletion event, this cluster was inserted into a small variant to get a large one, the *u1-u5* cluster should exist in a different genomic context. Extensive research of this region in publicly available genomes revealed that small fragments of *u1, u2* and *u3* align to genes coding hypothetical proteins in ICE*Kp*5 and ICE*Kp*12. However, the *u1-u5* sequence was only found in large variants of GIE492, so an insertion theory of the expendable cluster into small variants lacks severe support.

*mceK* has been considered as an strange gene, since one half of the encoded peptide is similar to the structural gene of Microcin M, and the other half is similar to the MccM immunity protein. However, *mceK* fails to confer any of these phenotypes when expressed in *E. coli* (Lagos *et al.,* 2009). In view of our results, this strange structure of *mceK* identified originally in the MccE492 production cluster from Kpn RYC492, would result from a deletion affecting two originally separated genes (microcin M-like and immunity) leading to a vestigial fused gene. Indeed, several variations were concentrated in this GIE492 region, where *mceK* was shorter in SVs I and IV due to a frameshift in the fusion region and absent at the protein level in II due to an early nonsense mutation. Therefore, the clonal groups that carry SVs with reduced or absent

*mceK* have lost or are losing this vestigial gene, probably through accumulation of small mutations.

Additionally, the big and small alleles of *u6* are directly linked to the size of GIE492 and the coordinates of the cryptic oriT sequence, as shown in Figure 1. It has been shown previously in plasmids, that the relaxosome machinery starts formation in the oriT, prior to DNA nicking in the nic site and posterior genetic transfer (Pansegrau *et al.,* 1990). Thus, at least one of the small GIE492 structural variants could be a product of a deleterious mutation occurred during the initiation of conjugative transfer, where the expendable gene cluster and part of *u6* were lost. Supporting the previous statement, CG268 carries both the small SV II of GIE492 and ICE*Kp*3, a possible combination of both islands in single CG that could be the product of deleterious mutation caused during co-mobilization by conjugation.

Another possibility to explain *u6* alleles, might be related to the way the protein is structured, due to the shorter allele aligning in the UniProt public database, to an AAA domain containing protein, part of a putative Type IV Toxin-Antitoxin system, and the longer allele aligning to the same domains of that protein. Therefore, it is possible that both alleles codify the active domain of the protein, and the hypothetical function of this protein is not lost, even if the shorter protein is clearly truncated.

The alleles of the *int* gene of GIE492, as described in the results section, seem to retain the integration site specificity, meaning that all variants may conserve all functional domains. This hypothesis is supported by the multiple sequence alignment of both alleles, showing that the translation of the alternative allele originates a truncated protein lacking only the five last amino acids from its C-terminus.

As it was shown in the results section, every GIE492+ CG only carries one structural variant of the element, evidencing the vertical inheritance conservation mechanism. However, only four known SVs exist, but 17 CGs carry them. This implies that the GIE492 structure is extremely conserved and hard to mutate through bacterial evolution. If the previous statement was not true, the number of SVs should be similar to the number of Clonal groups related to them. This is also supported by what was previously established, 15 out 23 coding sequences in the GI never vary. Therefore, GIE492 Structural Variants are restricted in an evolutionary sense, both in a functional and structural perspective.

### 3. GIE492 SV I and ICE*Kp*10 are strongly related in hypervirulent *K. pneumoniae* clones

Through this work, multiple relationships between GIE492 and ICE*Kp* elements, have been established. First, there is a high (~71%) prevalence of ICE*Kp* in GIE492+ genomes, although this cannot be said the other way around: only ~6% of ICE*Kp*+ genomes also carry GIE492. This evidence shows an inherent difference in the prevalence of these two GIs in the *Klebsiella* population and is compatible with the idea that GIE492 would leverage ICE*Kp* for its mobilization but not reciprocally.

Therefore, there is a high probability that relationships between these two GIs are clonal group specific and were developed inside these groups due to the limited prevalence of GIE492 in the Kpn population and its low levels of HGT. If GIE492 was as prevalent as ICE*Kp* is (Lam *et al.,* 2018a), then the relationship between both GIs would have developed in ancestral clones of Kpn and be as evident as the relationship

between genes belonging to the *core-genome.* This implies that trying to find a general relationship between both GIs might be impossible. The previous statement is supported by what was shown in Table 10, where a mixed pattern between ICE*Kp* SVs and CGs carrying GIE492 was observed, because if both GIs were related deeply in a evolutionary way, all genomes carrying one GI should also carry the other, and both GIs should vary together presenting a single SV for each of them in every CG.

However, as it was demonstrated by the GIE492 SV I and ICE*Kp*10 presence test developed in this work, there is a clear relationship between both islands in hvKp strains. It is important to mention that all true positives predicted by this test belonged to CG23 or CG380. This relationship is quite relevant, due to ICE*Kp*10 being the only structural variant of this element that carries the genotoxin colibactin locus *clb* (Lam *et al.,* 2018a), which is among the most defining features of CG23 (Lam *et al.,* 2018b) and CG380 being well known as a hvKp clone (Bialek-Davenet *et al.,* 2014). CG23 isolates also carry KpVP-1 and as was stated in the results, sometimes even genes conferring carbapenem resistance. A synergistic effect of virulence factors encoded in GIE492, ICE*Kp*10 and KpVP-1 may render CG23 as the most dangerous Kpn strains known nowadays. These observations may help establish GIE492 SV I as a potential virulence factor, but instead of directly giving potential virulence by itself, acts as a virulence potentiator in specific types of strains. In this regard, the increased virulence of strains showing the concomitant presence of determinants for the production of the yersiniabactin and salmochelin siderophores, microcins and colibactin was previously reported and extensively studied in uropathogenic *E. coli* strains. Moreover, the

production of all these virulence factors was proven to share some connections at the molecular level (Massip *et al*., 2020).

## 4. Theoretical evidence supporting that GIE492 and ICE*Kp* are actively being excised in *K. pneumoniae* SGH10 and co-mobilized

It is important to remember that all GIE492+ genomes carry the full cryptic oriT sequence observed in asn-tDNAs related GIs, and as it was established before, ICE*Kp* is present in ~71% of these genomes. The ICE*Kp* family of elements carry all the genes needed for horizontal gene transfer by conjugation (Lam *et al.,* 2018a), so it is plausible that these two islands are co-mobilized. Marcoleta *et al.* (2016) established that GIE492 can be exscinded from the Kpn RYC492 chromosome, and in this work, we theoretically demonstrated, using an innovative NGS based approach, that both ICE*Kp* and GIE492 are actively being exscinded in an exponential culture of Kpn SGH10. The results of this work strongly suggest that the co-mobilization hypothesis of GIE492 and ICE*Kp* might be true, and that these two islands are strongly related in CG23 hvKp strains. However, further studies are needed to get experimental evidence of this co-mobilization.

## CONCLUSIONS

- GIE492 is a genomic island highly restricted to *K. pneumoniae* sensu stricto, which harbors up to 23 protein-coding genes and shows at least four structural variants carried by seventeen different clonal groups.
- GIE492 is disseminated in the *Klebsiella* population mainly by vertical inheritance, remaining highly conserved across related genomes. However, the sporadic presence of this element in several unrelated CGs supports multiple GIE492 acquisition events during *Klebsiella* evolution.
- GIE492 structural variation was restricted to only 8 genes, the rest were always conserved.
- GIE492 has a marked preference for the *asn1C* integration site, although it can integrate into other asparagine tDNAs.
- All GIE492+ strains carry a conserved cryptic transfer origin sequence, that might be used for co-mobilization by conjugation with ICE*Kp.*
- GIE492 strongly associate with ICE*Kp*10 in hypervirulent CG23 and CG380. The presence of both GIE492 SV I and ICE*Kp*10 is a reliable predictor for hypervirulence in *K. pneumoniae*.

- The excision of genomic islands in *K. pneumoniae* and possibly other *Enterobacteriaceae* can be detected using our new developed strategy based on next-generation sequencing.

## REFERENCES


Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of molecular biology, 215(3), 403-410.

Andersson, D. I. (2003). Persistence of antibiotic resistant bacteria. Current opinion in microbiology, 6(5), 452-456.

Andrews, S. (2017). FastQC: a quality control tool for high throughput sequence data. 2010.

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., & Wishart, D. S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. Nucleic acids research, 44(W1), W16-W21.

Austin, D. J., Kristinsson, K. G., & Anderson, R. M. (1999). The relationship between the volume of antimicrobial consumption in human communities and the frequency of resistance. Proceedings of the National Academy of Sciences, 96(3), 1152-1156.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., ... & Zagnitko, O. (2008). The RAST Server: rapid annotations using subsystems technology. BMC genomics, 9(1), 1-15.

Azpiroz, M. F., & Laviña, M. (2004). Involvement of enterobactin synthesis pathway in production of microcin H47. Antimicrobial agents and chemotherapy, 48(4), 1235-1241.

Azpiroz, M. F., & Laviña, M. (2007). Modular structure of microcin H47 and colicin V. Antimicrobial agents and chemotherapy, 51(7), 2412-2419.

Bachman, M. A., Lenio, S., Schmidt, L., Oyler, J. E., & Weiser, J. N. (2012). Interaction of lipocalin 2, transferrin, and siderophores determines the replicative niche of *Klebsiella pneumoniae* during pneumonia. MBio, 3(6), e00224-11.

Bachman, M. A., Miller, V. L., & Weiser, J. N. (2009). Mucosal lipocalin 2 has pro-inflammatory and iron-sequestering effects in response to bacterial enterobactin. PLoS pathogens, 5(10), e1000622.

Bachman, M. A., Oyler, J. E., Burns, S. H., Caza, M., Lépine, F., Dozois, C. M., & Weiser, J. N. (2011). *Klebsiella pneumoniae* yersiniabactin promotes respiratory tract infection through evasion of lipocalin 2. Infection and immunity, 79(8), 3309-3316.

Bagg, A., & Neilands, J. B. (1987). Ferric uptake regulation protein acts as a repressor, employing iron (II) as a cofactor to bind the operator of an iron transport operon in *Escherichia coli*. Biochemistry, 26(17), 5471-5477.

Bellanger, X., Payot, S., Leblond-Bourget, N., & Guédon, G. (2014). Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. FEMS Microbiology Reviews, 38(4), 720-760.

Berríos-Pastén, C., Acevedo, R., Arros, P., Varas, M. A., Wyres, K. L., Lam, M. M., ... & Marcoleta, A. E. (2020). Properties of genes encoding transfer RNAs as integration sites for genomic islands and prophages in *Klebsiella pneumoniae*. bioRxiv.

Bialek-Davenet, S., Criscuolo, A., Ailloud, F., Passet, V., Jones, L., Delannoy-Vieillard, A. S., ... & Brisse, S. (2014). Genomic definition of hypervirulent and multidrug-resistant *Klebsiella pneumoniae* clonal groups. Emerging infectious diseases, 20(11), 1812.

Bieler, S., Estrada, L., Lagos, R., Baeza, M., Castilla, J., & Soto, C. (2005). Amyloid formation modulates the biological activity of a bacterial protein. Journal of Biological Chemistry, 280(29), 26880-26885.

Bieler, S., Silva, F., & Belin, D. (2010). The polypeptide core of Microcin E492 stably associates with the mannose permease and interferes with mannose metabolism. Research in microbiology, 161(8), 706-710.

Bieler, S., Silva, F., Soto, C., & Belin, D. (2006). Bactericidal activity of both secreted and nonsecreted microcin E492 requires the mannose permease. Journal of bacteriology, 188(20), 7049-7061.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114-2120.

Boyd, E. F., Almagro-Moreno, S., & Parent, M. A. (2009). Genomic islands are dynamic, ancient integrative elements in bacterial evolution. Trends in microbiology, 17(2), 47-53.

Braun, V., Patzer, S. I., & Hantke, K. (2002). Ton-dependent colicins and microcins: modular design and evolution. Biochimie, 84(5-6), 365-380.

Brisse, S., Fevre, C., Passet, V., Issenhuth-Jeanjean, S., Tournebize, R., Diancourt, L., & Grimont, P. (2009). Virulent clones of *Klebsiella pneumoniae*: identification and evolutionary scenario based on genomic and phenotypic characterization. PloS one, 4(3), e4982.

Buchanan, R. E., & Gibbons, N. E. (1949). Bergey's manual of determinative bacteriology. Williams & Wilkins.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. Nature methods, 12(1), 59-60.

Burrus, V., Pavlovic, G., Decaris, B., & Guédon, G. (2002). Conjugative transposons: the tip of the iceberg. Molecular microbiology, 46(3), 601-610.

Cerit, E. E. ESKAPE (*Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, Enterobacter* spp.) lies at the heart of the AMR crisis. Global Health Journal 2020, 46.

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, 34(17), i884-i890.

Chen, Z., Lewis, K. A., Shultzaberger, R. K., Lyakhov, I. G., Zheng, M., Doan, B., ... & Schneider, T. D. (2007). Discovery of Fur binding site clusters in *Escherichia coli* by information theory models. Nucleic acids research, 35(20), 6762-6777.

Cohen, T., Sommers, B., & Murray, M. (2003). The effect of drug resistance on the fitness of *Mycobacterium tuberculosis*. The Lancet infectious diseases, 3(1), 13-21.

Corsini, G., Baeza, M., Monasterio, O., & Lagos, R. (2002). The expression of genes involved in microcin maturation regulates the production of active microcin E492. Biochimie, 84(5-6), 539-544.

David, S., Reuter, S., Harris, S. R., Glasner, C., Feltwell, T., Argimon, S., ... & Grundmann, H. (2019). Epidemic of carbapenem-resistant *Klebsiella pneumoniae* in Europe is driven by nosocomial spread. Nature microbiology, 4(11), 1919-1929.

De Lorenzo, V. (1984). Isolation and characterization of microcin E 492 from *Klebsiella pneumoniae*. Archives of microbiology, 139(1), 72-75.

De Lorenzo, V., & Pugsley, A. P. (1985). Microcin E492, a low-molecular-weight peptide antibiotic which causes depolarization of the *Escherichia coli* cytoplasmic membrane. Antimicrobial agents and chemotherapy, 27(4), 666-669.

De Lorenzo, V., Martínez, J. L., & Asensio, C. (1984). Microcin-mediated Interactions Between *Klebsiella pneumoniae* and *Escherichia coli* Strains. Microbiology, 130(2), 391-400.

De Lorenzo, Víctor (1985). Factors affecting microcin E492 production. The Journal of antibiotics, 38(3), 340-345.

Diancourt, L., Passet, V., Verhoef, J., Grimont, P. A., & Brisse, S. (2005). Multilocus sequence typing of *Klebsiella pneumoniae* nosocomial isolates. Journal of clinical microbiology, 43(8), 4178-4182.

Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E., & Crook, D. W. (2012). Transforming clinical microbiology with bacterial genome sequencing. Nature Reviews Genetics, 13(9), 601-612.

Dobrindt, U., Blum-Oehler, G., Nagy, G., Schneider, G., Johann, A., Gottschalk, G., & Hacker, J. (2002). Genetic structure and distribution of four pathogenicity islands (PAI I536 to PAI IV536) of uropathogenic *Escherichia coli* strain 536. Infection and immunity, 70(11), 6365-6372.

Dobrindt, U., Hochhut, B., Hentschel, U., & Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. Nature Reviews Microbiology, 2(5), 414-424.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC bioinformatics, 5(1), 1-19.

Escolar, L., Pérez-Martín, J., & de Lorenzo, V. (1998). Binding of the fur (ferric uptake regulator) repressor of *Escherichia coli* to arrays of the GATAAT sequence. Journal of molecular biology, 283(3), 537-547.

Follador, R., Heinz, E., Wyres, K. L., Ellington, M. J., Kowarik, M., Holt, K. E., & Thomson, N. R. (2016). The diversity of *Klebsiella pneumoniae* surface polysaccharides. Microbial genomics, 2(8).

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics, 28(23), 3150-3152.

Gilchrist, C. A., Turner, S. D., Riley, M. F., Petri Jr, W. A., & Hewlett, E. L. (2015). Whole-genome sequencing in outbreak analysis. Clinical microbiology reviews, 28(3), 541-563.

Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., & Bossuyt, P. M. (2003). The diagnostic odds ratio: a single indicator of test performance. Journal of clinical epidemiology, 56(11), 1129-1135.

Goetz, D. H., Holmes, M. A., Borregaard, N., Bluhm, M. E., Raymond, K. N., & Strong, R. K. (2002). The neutrophil lipocalin NGAL is a bacteriostatic agent that interferes with siderophore-mediated iron acquisition. Molecular cell, 10(5), 1033-1043.

Gu, D., Dong, N., Zheng, Z., Lin, D., Huang, M., Wang, L., ... & Chen, S. (2018). A fatal outbreak of ST11 carbapenem-resistant hypervirulent *Klebsiella pneumoniae* in a Chinese hospital: a molecular epidemiological study. The Lancet infectious diseases, 18(1), 37-46.

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics, 32(18), 2847-2849.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29(8), 1072-1075.

Hacker, J., Bender, L., Ott, M., Wingender, J., Lund, B., Marre, R., & Goebel, W. (1990). Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and

in vivo in various extra intestinal *Escherichia coli* isolates. Microbial pathogenesis, 8(3), 213-225.

Hamon, Y., & Peron, Y. (1960). Study of the method of fixation of colicins and pyocins on sensitive bacteria. Comptes rendus hebdomadaires des seances de l'Academie des sciences, 251, 1840-1842.

Hantke, K., Nicholson, G., Rabsch, W., & Winkelmann, G. (2003). Salmochelins, siderophores of *Salmonella enterica* and uropathogenic *Escherichia coli* strains, are recognized by the outer membrane receptor IroN. Proceedings of the National Academy of Sciences, 100(7), 3677-3682.

Hetz, C., Bono, M. R., Barros, L. F., & Lagos, R. (2002). Microcin E492, a channel-forming bacteriocin from *Klebsiella pneumoniae*, induces apoptosis in some human cell lines. Proceedings of the National Academy of Sciences, 99(5), 2696-2701.

Holt, K. E., Wertheim, H., Zadoks, R. N., Baker, S., Whitehouse, C. A., Dance, D., ... & Thomson, N. R. (2015). Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. Proceedings of the National Academy of Sciences, 112(27), E3574-E3581.

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics, 11(1), 1-11.

Inouye, M., Dashnow, H., Raven, L. A., Schultz, M. B., Pope, B. J., Tomita, T., ... & Holt, K. E. (2014). SRST2: rapid genomic surveillance for public health and hospital microbiology labs. Genome medicine, 6(11), 1-16.

Ito, R., Mustapha, M. M., Tomich, A. D., Callaghan, J. D., McElheny, C. L., Mettus, R.

T., ... & Doi, Y. (2017). Widespread fosfomycin resistance in Gram-negative bacteria attributable to the chromosomal fosA gene. MBio, 8(4), e00749-17.

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., & Daszak, P. (2008). Global trends in emerging infectious diseases. Nature, 451(7181), 990-993.

Katz, L. S., Griswold, T., Morrison, S. S., Caravas, J. A., Zhang, S., den Bakker, H. C., ... & Carleton, H. A. (2019). Mashtree: a rapid comparison of whole genome sequence files. Journal of Open Source Software, 4(44), 1762.

Koebnik, R. (2005). TonB-dependent trans-envelope signalling: the exception or the rule?. Trends in microbiology, 13(8), 343-347.

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. Nature biotechnology, 37(5), 540-546.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., & Phillippy, A. M. (2016). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. bioRxiv.

Lagos, R., Baeza, M., Corsini, G., Hetz, C., Strahsburger, E., Castillo, J. A., ... & Monasterio, O. (2001). Structure, organization and characterization of the gene cluster involved in the production of microcin E492, a channel-forming bacteriocin. Molecular microbiology, 42(1), 229-243.

Lagos, R., Tello, M., Mercado, G., García, V., & Monasterio, O. (2009). Antibacterial and antitumorigenic properties of microcin E492, a pore-forming bacteriocin. Current pharmaceutical biotechnology, 10(1), 74-85.

Lagos, R., Villanueva, J. E., & Monasterio, O. (1999). Identification and properties of the

genes encoding microcin E492 and its immunity protein. Journal of bacteriology, 181(1), 212-217.

Lagos, R., Wilkens, M., Vergara, C., Cecchi, X., & Monasterio, O. (1993). Microcin E492 forms ion channels in phospholipid bilayer membranes. FEBS letters, 321(2-3), 145-148.

Lai, Y. C., Lin, A. C., Chiang, M. K., Dai, Y. H., Hsu, C. C., Lu, M. C., ... & Chen, Y. T. (2014). Genotoxic *Klebsiella pneumoniae* in Taiwan. PloS one, 9(5), e96292.

Lam, M. M., Wick, R. R., Wyres, K. L., Gorrie, C. L., Judd, L. M., Jenney, A. W., ... & Holt, K. E. (2018). Genetic diversity, mobilisation and spread of the yersiniabactin-encoding mobile element ICE*Kp* in *Klebsiella pneumoniae* populations. Microbial genomics, 4(9).

Lam, M. M., Wyres, K. L., Duchêne, S., Wick, R. R., Judd, L. M., Gan, Y. H., ... & Holt, K. E. (2018). Population genomics of hypervirulent *Klebsiella pneumoniae* clonal-group 23 reveals early emergence and rapid global dissemination. Nature communications, 9(1), 1-10.

Lam, M. M., Wyres, K. L., Judd, L. M., Wick, R. R., Jenney, A., Brisse, S., & Holt, K. E. (2018). Tracking key virulence loci encoding aerobactin and salmochelin siderophore synthesis in *Klebsiella pneumoniae*. Genome medicine, 10(1), 1-15.

Lam, M., Wick, R. R., Watts, S. C., Cerdeira, L. T., Wyres, K. L., & Holt, K. E. (2021). A genomic surveillance framework and genotyping tool for *Klebsiella pneumoniae* and its related species complex. Nature communications, 12(1), 1-16.

Langmead, B. (2010). Aligning short sequencing reads with Bowtie. Current protocols in bioinformatics, 32(1), 11-7.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), 357-359.

Larsen, M. V., Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R. L., ... & Lund, O. (2012). Multilocus sequence typing of total-genome-sequenced bacteria. Journal of clinical microbiology, 50(4), 1355-1361.

Lee, C. A. (1996). Pathogenicity islands and the evolution of bacterial pathogens. Infectious agents and disease, 5(1), 1-7.

Lee, I. R., Molton, J. S., Wyres, K. L., Gorrie, C., Wong, J., Hoh, C. H., ... & Gan, Y. H. (2016). Differential host susceptibility and bacterial virulence factors driving *Klebsiella* liver abscess in an ethnically diverse population. Scientific reports, 6(1), 1-12.

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic acids research, 49(W1), W293-W296.

Levin, B. R. (2002). Models for the spread of resistant pathogens. The Netherlands journal of medicine, 60(7 Suppl), 58-64.

Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics, 32(14), 2103-2110.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34(18), 3094-3100.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078-2079.

Li, J., Zhang, H., Ning, J., Sajid, A., Cheng, G., Yuan, Z., & Hao, H. (2019). The nature and epidemiology of OqxAB, a multidrug efflux pump. Antimicrobial Resistance & Infection Control, 8(1), 1-13.

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large

sets of protein or nucleotide sequences. Bioinformatics, 22(13), 1658-1659.

Lin, T. L., Lee, C. Z., Hsieh, P. F., Tsai, S. F., & Wang, J. T. (2008). Characterization of integrative and conjugative element ICE Kp1-associated genomic heterogeneity in a *Klebsiella pneumoniae* strain isolated from a primary liver abscess. Journal of bacteriology, 190(2), 515-526.

Long, S. W., Olsen, R. J., Eagar, T. N., Beres, S. B., Zhao, P., Davis, J. J., ... & Musser, J. M. (2017). Population genomic analysis of 1,777 extended-spectrum beta-lactamase-producing *Klebsiella pneumoniae* isolates, Houston, Texas: unexpected abundance of clonal group 307. MBio, 8(3), e00489-17.

MacCannell, D. (2016). Next generation sequencing in clinical and public health microbiology. Clinical Microbiology Newsletter, 38(21), 169-176.

Mao, C., Bhardwaj, K., Sharkady, S. M., Fish, R. I., Driscoll, T., Wower, J., ... & Williams, K. P. (2009). Variations on the tmRNA gene. RNA biology, 6(4), 355-361.

Marcoleta, A. E., Berríos-Pastén, C., Nuñez, G., Monasterio, O., & Lagos, R. (2016). *Klebsiella pneumoniae* asparagine tDNAs are integration hotspots for different genomic islands encoding microcin E492 production determinants and other putative virulence factors present in hypervirulent strains. Frontiers in microbiology, 7, 849.

Marcoleta, A. E., Gutiérrez-Cortez, S., Hurtado, F., Argandoña, Y., Corsini, G., Monasterio, O., & Lagos, R. (2018). The Ferric uptake regulator (Fur) and iron availability control the production and maturation of the antibacterial peptide microcin E492. PloS one, 13(8), e0200835.

Marcoleta, A., Gutiérrez-Cortez, S., Maturana, D., Monasterio, O., & Lagos, R. (2013). Whole-genome sequence of the microcin E492-producing strain *Klebsiella pneumoniae*

RYC492. Genome announcements, 1(3), e00178-13.

Marcoleta, A., Marín, M., Mercado, G., Valpuesta, J. M., Monasterio, O., & Lagos, R. (2013). Microcin E492 amyloid formation is retarded by posttranslational modification. Journal of bacteriology, 195(17), 3995-4004.

Marr, C. M., & Russo, T. A. (2019). Hypervirulent *Klebsiella pneumoniae*: a new public health threat. Expert review of anti-infective therapy, 17(2), 71-73.

Massip, C., Chagneau, C. V., Boury, M., & Oswald, E. (2020). The synergistic triad between microcin, colibactin, and salmochelin gene clusters in uropathogenic *Escherichia coli*. Microbes and infection, 22(3), 144-147.

Maurelli, A. T., Fernández, R. E., Bloch, C. A., Rode, C. K., & Fasano, A. (1998). “Black holes” and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli.* Proceedings of the National Academy of Sciences, 95(7), 3943-3948.

McInerney, J. O., McNally, A., & O'connell, M. J. (2017). Why prokaryotes have pangenomes. Nature microbiology, 2(4), 1-5.

McKenzie, G. J., & Rosenberg, S. M. (2001). Adaptive mutations, mutator DNA polymerases and genetic change strategies of pathogens. Current opinion in microbiology, 4(5), 586-594.

Méndez-Vilas, A. (Ed.). (2013). Microbial pathogens and strategies for combating them: science, technology and education. Formatex Research Center.

Mercado, G., Tello, M., Marín, M., Monasterio, O., & Lagos, R. (2008). The production in vivo of microcin E492 with antibacterial activity depends on salmochelin and EntF. Journal of bacteriology, 190(15), 5464-5471.

Moran, N. A. (2002). Microbial minimalism: genome reduction in bacterial pathogens. Cell, 108(5), 583-586.

Motro, Y., & Moran-Gilad, J. (2017). Next-generation sequencing applications in clinical bacteriology. Biomolecular detection and quantification, 14, 1-6.

Müller, C. W. (2001). Transcription factors: global and detailed views. Current opinion in structural biology, 11(1), 26-32.

Nadimpalli, M. L., de Lauzanne, A., Phe, T., Borand, L., Jacobs, J., Fabre, L., ... & Stegger, M. (2019). *Escherichia coli* ST410 among humans and the environment in Southeast Asia. International journal of antimicrobial agents, 54(2), 228-232.

Nassif, X. A. V. I. E. R., Fournier, J. M., Arondel, J., & Sansonetti, P. J. (1989). Mucoid phenotype of *Klebsiella pneumoniae* is a plasmid-encoded virulence factor. Infection and immunity, 57(2), 546-552.

Navon-Venezia, S., Kondratyeva, K., & Carattoli, A. (2017). *Klebsiella pneumoniae*: a major worldwide source and shuttle for antibiotic resistance. FEMS microbiology reviews, 41(3), 252-275.

Nazos, P. M., Antonucci, T. K., Landick, R., & Oxender, D. L. (1986). Cloning and characterization of *livH*, the structural gene encoding a component of the leucine transport system in *Escherichia coli*. Journal of bacteriology, 166(2), 565-573.

Nelson, J. T., Lee, J., Sims, J. W., & Schmidt, E. W. (2007). Characterization of SafC, a catechol 4-O-methyltransferase involved in saframycin biosynthesis. Applied and environmental microbiology, 73(11), 3575-3580.

Nolan, E. M., Fischbach, M. A., Koglin, A., & Walsh, C. T. (2007). Biosynthetic tailoring of microcin E492m: post-translational modification affords an antibacterial siderophore−

peptide conjugate. Journal of the American Chemical Society, 129(46), 14336-14347.

Nougayrède, J. P., Homburg, S., Taieb, F., Boury, M., Brzuszkiewicz, E., Gottschalk, G., ... & Oswald, E. (2006). *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells. Science, 313(5788), 848-851.

Ochman, H., Lawrence, J. G., & Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. nature, 405(6784), 299-304.

Oliver, J. D. (2000). The public health significance of viable but nonculturable bacteria. Nonculturable microorganisms in the environment, 277-300.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome biology, 17(1), 1-14.

Orellana, C., & Lagos, R. (1996). The activity of microcin E492 from *Klebsiella pneumoniae* is regulated by a microcin antagonist. FEMS microbiology letters, 136(3), 297-303.

Paczosa, M. K., & Mecsas, J. (2016). *Klebsiella pneumoniae*: going on the offense with a strong defense. Microbiology and Molecular Biology Reviews, 80(3), 629-661.

Pansegrau, W., Balzer, D., Kruft, V., LuRz, R. U. D. I., & Lanka, E. (1990). In vitro assembly of relaxosomes at the transfer origin of plasmid RP4. Proceedings of the National Academy of Sciences, 87(17), 6555-6559.

Patzer, S., Baquero, M. R., Bravo, D., Moreno, F., & Hantke, K. (2003). The colicin G, H and X determinants encode microcins M and H47, which might utilize the catecholate siderophore receptors FepA, Cir, Fiu and IroN. Microbiology, 149(9), 2557-2570.

Petrosillo, N., Taglietti, F., & Granata, G. (2019). Treatment options for colistin resistant

*Klebsiella pneumoniae*: present and future. Journal of clinical medicine, 8(7), 934.

Postle, K., & Kadner, R. J. (2003). Touch and go: tying TonB to transport. Molecular microbiology, 49(4), 869-882.

Potter, R. F., Lainhart, W., Twentyman, J., Wallace, M. A., Wang, B., Burnham, C. A. D., ... & Dantas, G. (2018). Population structure, antibiotic resistance, and uropathogenicity of *Klebsiella variicola*. MBio, 9(6), e02481-18.

Pugsley, A. P. (1985). *Escherichia coli* K12 strains for use in the identification and characterization of colicins. Microbiology, 131(2), 369-376.

Pugsley, A. P., Moreno, F., & De Lorenzo, V. (1986). Microcin-E492-insensitive mutants of *Escherichia coli* K12. Microbiology, 132(12), 3253-3259.

Qi, Y., Wei, Z., Ji, S., Du, X., Shen, P., & Yu, Y. (2011). ST11, the dominant clone of KPC-producing *Klebsiella pneumoniae* in China. Journal of Antimicrobial Chemotherapy, 66(2), 307-312.

Rambaut, A. (2009). FigTree. Tree figure drawing tool. http://tree.bio.ed.ac.uk/software/figtree/.

Randall-Hazelbauer, L., & Schwartz, M. (1973). Isolation of the bacteriophage lambda receptor from *Escherichia coli*. Journal of bacteriology, 116(3), 1436-1446.

Rasko, D. A., Myers, G. S., & Ravel, J. (2005). Visualization of comparative genomic analyses by BLAST score ratio. BMC bioinformatics, 6(1), 1-7.

Ribeiro-Gonçalves, B., Francisco, A. P., Vaz, C., Ramirez, M., & Carriço, J. A. (2016). PHYLOViZ Online: web-based tool for visualization, phylogenetic inference, analysis and sharing of minimum spanning trees. Nucleic acids research, 44(W1), W246-W251.

Rice, L. B. (2008). Federal funding for the study of antimicrobial resistance in nosocomial

pathogens: no ESKAPE. The Journal of infectious diseases, 197(8), 1079-1081.

Ristuccia, P. A., & Cunha, B. A. (1984). *Klebsiella*. Infection Control & Hospital Epidemiology, 5(7), 343-347.

Rodrigues, C., Passet, V., Rakotondrasoa, A., Diallo, T. A., Criscuolo, A., & Brisse, S. (2019). Description of *Klebsiella africanensis sp. nov*., *Klebsiella variicola subsp. tropicalensis subsp. nov*. and *Klebsiella variicola subsp. variicola subsp. nov*. Research in microbiology, 170(3), 165-170.

Schirmer, T., Keller, T. A., Wang, Y. F., & Rosenbusch, J. P. (1995). Structural basis for sugar translocation through maltoporin channels at 3.1 Å resolution. Science, 267(5197), 512-514.

Schoeniger, J. S., Hudson, C. M., Bent, Z. W., Sinha, A., & Williams, K. P. (2016). Experimental single-strain mobilomics reveals events that shape pathogen emergence. Nucleic acids research, 44(14), 6830-6839.

Schubert, H. L., Blumenthal, R. M., & Cheng, X. (2003). Many paths to methyltransfer: a chronicle of convergence. Trends in biochemical sciences, 28(6), 329-335.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics, 30(14), 2068-2069.

Shon, A. S., Bajwa, R. P., & Russo, T. A. (2013). Hypervirulent (hypermucoviscous) *Klebsiella pneumoniae*: a new and dangerous breed. Virulence, 4(2), 107-118.

Silva, M., Machado, M. P., Silva, D. N., Rossi, M., Moran-Gilad, J., Santos, S., ... & Carriço, J. A. (2018). chewBBACA: a complete suite for gene-by-gene schema creation and strain identification. Microbial genomics, 4(3).

Sintchenko, V., & Holmes, E. C. (2015). The role of pathogen genomics in assessing

disease transmission. Bmj, 350.

Siu, L. K., Fung, C. P., Chang, F. Y., Lee, N., Yeh, K. M., Koh, T. H., & Ip, M. (2011). Molecular typing and virulence analysis of serotype K1 *Klebsiella pneumoniae* strains isolated from liver abscess patients and stool samples from noninfectious subjects in Hong Kong, Singapore, and Taiwan. Journal of clinical microbiology, 49(11), 3761-3765.

Siu, L. K., Yeh, K. M., Lin, J. C., Fung, C. P., & Chang, F. Y. (2012). *Klebsiella pneumoniae* liver abscess: a new invasive syndrome. The Lancet infectious diseases, 12(11), 881-887.

Skovgaard, N. (2007). New trends in emerging pathogens. International journal of food microbiology, 120(3), 217-224.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., ... & Birney, E. (2002). The Bioperl toolkit: Perl modules for the life sciences. Genome research, 12(10), 1611-1618.

Strahsburger, E., Baeza, M., Monasterio, O., & Lagos, R. (2005). Cooperative uptake of microcin E492 by receptors FepA, Fiu, and Cir and inhibition by the siderophore enterochelin and its dimeric and trimeric hydrolysis products. Antimicrobial agents and chemotherapy, 49(7), 3083-3086.

Struve, C., Roe, C. C., Stegger, M., Stahlhut, S. G., Hansen, D. S., Engelthaler, D. M., ... & Krogfelt, K. A. (2015). Mapping the evolution of hypervirulent *Klebsiella pneumoniae.* MBio, 6(4), e00630-15.

Sullivan, M. J., Petty, N. K., & Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. Bioinformatics, 27(7), 1009-1010.

Tanwar, J., Das, S., Fatima, Z., & Hameed, S. (2014). Multidrug resistance: an emerging

crisis. Interdisciplinary perspectives on infectious diseases, 2014.

Team, R. C. (2000). R language definition. Vienna, Austria: R foundation for statistical computing.

Thomas, X., Destoumieux-Garzón, D., Peduzzi, J., Afonso, C., Blond, A., Birlirakis, N., ... & Rebuffat, S. (2004). Siderophore peptide, a new type of post-translationally modified antibacterial peptide with potent activity. Journal of Biological Chemistry, 279(27), 28233-28242.

Tooke, C. L., Hinchliffe, P., Bragginton, E. C., Colenso, C. K., Hirvonen, V. H., Takebayashi, Y., & Spencer, J. (2019). β-Lactamases and β-Lactamase Inhibitors in the 21st Century. Journal of molecular biology, 431(18), 3472-3500.

Trakhanov, S., Vyas, N. K., Luecke, H., Kristensen, D. M., Ma, J., & Quiocho, F. A. (2005). Ligand-free and-bound structures of the binding protein (LivJ) of the *Escherichia coli* ABC leucine/isoleucine/valine transport system: trajectory and dynamics of the interdomain rotation and ligand specificity. Biochemistry, 44(17), 6597-6608.

Vaser, R., & Šikić, M. (2021). Time-and memory-efficient genome assembly with Raven. Nature Computational Science, 1(5), 332-336.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., ... & Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS one, 9(11), e112963.

Walker, K. A., Miner, T. A., Palacios, M., Trzilova, D., Frederick, D. R., Broberg, C. A., ... & Miller, V. L. (2019). A *Klebsiella pneumoniae* regulatory mutant has reduced capsule expression but retains hypermucoviscosity. MBio, 10(2), e00089-19.

Walker, K. A., Treat, L. P., Sepúlveda, V. E., & Miller, V. L. (2020). The small protein

RmpD drives hypermucoviscosity in *Klebsiella pneumoniae*. MBio, 11(5), e01750-20.

Wall, L. (1994). The Perl programming language.

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., & Barton, G. J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics, 25(9), 1189-1191.

Wick, R. R., & Holt, K. E. (2020). rrwick/Minipolish: Minipolish v0. 1.3.

Wick, R. R., Judd, L. M., Cerdeira, L. T., Hawkey, J., Méric, G., Vezina, B., ... & Holt, K. E. (2021). Trycycler: consensus long-read assemblies for bacterial genomes. Genome biology, 22(1), 1-17.

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS computational biology, 13(6), e1005595.

Wiener, M. C. (2005). TonB-dependent outer membrane transport: going for Baroque?. Current opinion in structural biology, 15(4), 394-400.

Wilson, G. S., & Miles, A. A. (1975). Topley and Wilson's principles of bacteriology, virology and immunity (Vol. 2, No. 6th edition).

Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L. H., ... & Achtman, M. (2006). Sex and virulence in *Escherichia coli*: an evolutionary perspective. Molecular microbiology, 60(5), 1136-1151.

World Health Organization. (2017). Prioritization of pathogens to guide discovery, research and development of new antibiotics for drug-resistant bacterial infections, including tuberculosis (No. WHO/EMP/IAU/2017.12). World Health Organization.

Wyres, K. L., & Holt, K. E. (2016). *Klebsiella pneumoniae* population genomics and

antimicrobial-resistant clones. Trends in microbiology, 24(12), 944-956.

Wyres, K. L., & Holt, K. E. (2018). *Klebsiella pneumoniae* as a key trafficker of drug resistance genes from environmental to clinically important bacteria. Current opinion in microbiology, 45, 131-139.

Wyres, K. L., Gorrie, C., Edwards, D. J., Wertheim, H. F., Hsu, L. Y., Van Kinh, N., ... & Holt, K. E. (2015). Extensive capsule locus variation and large-scale genomic recombination within the *Klebsiella pneumoniae* clonal group 258. Genome biology and evolution, 7(5), 1267-1279.

Wyres, K. L., Lam, M., & Holt, K. E. (2020). Population genomics of *Klebsiella pneumoniae*. Nature Reviews Microbiology, 18(6), 344-359.

Wyres, K. L., Wick, R. R., Judd, L. M., Froumine, R., Tokolyi, A., Gorrie, C. L., ... & Holt, K. E. (2019). Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of *Klebsiella pneumoniae*. PLoS genetics, 15(4), e1008114.

Xie, Z., & Tang, H. (2017). ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. Bioinformatics, 33(21), 3340-3347.

Xu, M., Fu, Y., Fang, Y., Xu, H., Kong, H., Liu, Y., ... & Li, L. (2019). High prevalence of KPC-2-producing hypervirulent *Klebsiella pneumoniae* causing meningitis in Eastern China. Infection and drug resistance, 12, 641.

Yang, X., Chan, E. W. C., Zhang, R., & Chen, S. (2019). A conjugative plasmid that augments virulence in *Klebsiella pneumoniae*. Nature microbiology, 4(12), 2039-2043.

Zhang, R., Lin, D., Chan, E. W. C., Gu, D., Chen, G. X., & Chen, S. (2016). Emergence of carbapenem-resistant serotype K1 hypervirulent *Klebsiella pneumoniae* strains in China. Antimicrobial agents and chemotherapy, 60(1), 709-711.