Genomic and functional analysis of colibactin harboring *Escherichia coli*

Thesis submitted to the University of Hyderabad
for the degree of

DOCTOR OF PHILOSOPHY

By

Arya Suresh
(Reg. No: 14LTPM02)

Department of Biotechnology and Bioinformatics
School of Life Sciences
University of Hyderabad
Hyderabad, 500046
India
June, 2021

University of Hyderabad
(A Central University by an Act of Parliament)
Department of Biotechnology and Bioinformatics
School of Life Sciences
P.O. Central University, Gachibowli, Hyderabad-500046

DECLARATION

The research work presented in the thesis entitled **"Genomic and functional analysis of colibactin harboring *Escherichia coli*"** has been carried out by me at the Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad, Hyderabad, under the guidance of Prof. Niyaz Ahmed. I hereby declare that this work is original and has not been submitted in part or full for any other degree or diploma of any other university or institution.

Date: 09|06|2021

Signature:

Name: Arya Suresh
Reg. No.: 14LTPM02

University of Hyderabad
(A Central University by an Act of Parliament)
Department of Biotechnology and Bioinformatics
School of Life Sciences
P.O. Central University, Gachibowli, Hyderabad-500046

CERTIFICATE

This is to certify that the thesis entitled **"Genomic and functional analysis of colibactin harboring *Escherichia coli*"** submitted by **Ms. Arya Suresh** bearing registration number **14LTPM02** in partial fulfilment of the requirements for the award of Doctor of Philosophy in the Department of Biotechnology and Bioinformatics, School of Life Sciences is a bonafide work carried out by her under my supervision and guidance.

This thesis is free from plagiarism and has not been submitted previously in part or in full to this or any other University or Institution for the award of any degree or diploma.

Parts of this thesis have been:

A. Published in the following journal:

1. **Molecular genetic and functional analysis of *pks*-harboring, extra-intestinal pathogenic *Escherichia coli* from India**

   Arya Suresh, Amit Ranjan, Savita Jadhav, Arif Hussain, Sabiha Shaik, Munirul Alam, Ramani Baddam, Lothar H. Wieler, Niyaz Ahmed; 2018; *Frontiers in Microbiology;* 9(2631); doi:10.3389/fmicb.2018.02631.

2. **Evolutionary dynamics based on comparative genomics of pathogenic *Escherichia coli* lineages harboring polyketide synthase (*pks*) island.**

   Arya Suresh, Sabiha Shaik, Ramani Baddam, Amit Ranjan, Shamsul Qumar, Savita Jadhav, Torsten Semmler, Irfan A. Ghazi, Lothar H. Wieler, Niyaz Ahmed; *mBio; 12*(1), pp.e03634-20; doi:10.1128/mBio.03634-20.

B. Presented in the following international and national conferences:

1- Participated and presented a poster in "14th Asian Conference on Diarrhoeal Disease &Nutrition (ASCODD)" hosted by icddr,b and RGCB from 30th October to 1st November 2017 in Kochi, India

2- Participated and presented a poster "59th Annual Conference of Association of Microbiologist of India (AMI -2018)" organized by University of Hyderabad, Hyderabad, India

Further, the student has passed the following courses towards the fulfilment of course work requirements for the award of the Ph.D. degree

Sl. No.	Course code	Subject	Credits	Remarks
1	BT 801	Research Methodology/Analytical Techniques	4	Pass
2	BT 802	Research ethics, biosafety, data analysis and biostatistics	4	Pass
3	BT 803	Research Proposal and scientific writing	4	Pass

Prof. Niyaz Ahmed

Research supervisor

Dept. of Biotechnology and Bioinformatics

Professor Dr. Niyaz Ahmed, PhD
Department of Biotechnology & Bioinformatics
University of Hyderabad, Hyderabad, India

Head

Dept. of Biotechnology and Bioinformatics

अध्यक्ष / Head
जैव प्रौद्योगिकी एवं जैव सूचना विज्ञान विभाग
Department of Biotechnology & Bioinformatics
हैदराबाद विश्वविद्यालय
University of Hyderabad
हैदराबाद / Hyderabad - 500 046.

Dean

School of Life Sciences

DEAN
School of Life Sciences
University of Hyderabad
Hyderabad - 500 046.

Dedicated to

My beloved family

Acknowledgement

I would like to start by thanking my Ph.D. supervisor Prof. Niyaz Ahmed for his constant support and motivation throughout my tenure of research work. I would like to heartfully thank him for introducing me to genomics and giving me the freedom to think and work independently and always encouraging me to pursue new ideas. I fondly cherish his wonderful classroom lectures that could make any topic interesting and all the scientific discussions which I had with him throughout the research period that were intriguing and motivating at the same time.

I would like to thank my doctoral committee members, Dr. Nooruddin Khan and Dr. Mohd. Akif for their valuable suggestions, support and encouragement. I would like to thank the present and former Heads of the Department of Biotechnology and Bioinformatics and the present and former Deans of School of Life Sciences for their continued help and support and enabling me to use the Departmental/School facilities. I would like to acknowledge the faculty members of the School of Life Sciences for their lectures during my M.Sc. programme and valuable suggestions during my study.

I would also like to acknowledge all the departmental and school non-teaching staffs for their help in doing all the official works in the department/school throughout this period of work.

I would like to thank our collaborators Prof. Lothar Wieler and Dr. Torsten Semmler, Robert Koch Institute, Berlin, Germany for their valuable suggestions during the study and also for the help with the whole genome sequencing of the isolates.

I would like to thank the members of my dearest PBL family, who created a wonderful and productive lab environment. I am grateful for the bond we share despite the distance and time zones, on both professional and personal fronts. I would like to thank Dr. Amit with whom I started working during my M.Sc. dissertation, and further continuing into my Ph.D. research. I would like to thank him for training me in all experimental works, for instilling interest in the topic and for the long discussions we had which always left me with many questions to address. I would like to thank Dr. Sabiha and Dr. Ramani for being such a great team to work with, particularly during our genomics part of the work. I am grateful to both of them for always being there for me during my professional and personal challenges, and I hope we continue our friendship always. I would like to thank Dr. Arif and Dr. Nishanth for all their valuable scientific advice, discussions and suggestions. I have to especially mention Aditya for his valuable help in making me comfortable

with Linux and all the software which have been instrumental in my work, the long skype discussions spent in troubleshooting, and for always being a good friend. I would also like to thank my dear friends Sumeet, and Dr. Kshitij for constantly encouraging and supporting me throughout the research work. I would cherish our learning together, trouble-shootings, and the brainstorming sessions. I am grateful to my PBL family members Dr. Kishore, Dr. Savita, Dr. Vidyullatha, Dr. Narender, Dr. Shivendra, Dr. Shankar, Dr. Shamsul, Priya, Dr. Majjid Qaria, Dr. Mohammad Majid, Naveen and Anuradha for their immense support, motivation and providing a memorable time in the lab which I will always cherish.

I would also like to thank my faculties from my bachelor's degree programme Dr. Deepthi, Dr. Lini and Dr. Bindu for suggesting me to apply for this Int. M.Sc. Ph.D. programme and for their continued encouragement. I miss the presence of my dear teacher Late Dr. Thomson Kuruvilla, whose empathy, love and support for his students will always be my motivation.

I would like to thank all my friends from UoH for all the wonderful times and memories.

I am immensely grateful to my parents for being there for me every time and for believing in me and my education. Words will not suffice for the sacrifices my parents have made for me to support my journey, and their constant encouragement and unconditional love has made it so much easier. I also thank my brother for always encouraging me and my family for all their support.

Last, but not least, I would like to thank my husband Sarath, without whose unconditional support and patience this would not have been possible for me. I thank him for always being understanding and for being my constant motivator through the ups and downs of the journey, and for having faith in me.

- Arya

Table of Contents

CHAPTER 3: RESULTS

List of Abbreviations

bp	**base pair**
BAPS	**Bayesian Analysis of Population Structure**
BLAST	**Basic Local Alignment Search Tool**
BRIG	**Blast Ring Image Generator**
CARD	**Comprehensive Antibiotic Resistance Database**
CDS	**Coding sequences**
CDT	**Cyclolethal Distending Toxin**
CNF	**Cytotoxic Necrotizing Factor**
cif	**Cycle Inhibiting Factor**
µg	**micro gram**
µL	**micro liter**
CFU	**Colony Forming Units**
CLA	**Contig Layout Authenticator**
CLSI	**Clinical and Laboratory Standards Institute**
COG	**Clusters of Orthologous groups**
DAEC	**Diffusely adherent *Escherichia coli***
ddH_2O	**Double distilled water**
DEC	**Diarrheagenic *Escherichia coli***
dNTP	**Deoxyribonucleotide triphosphate**
E. coli	***Escherichia coli***
EAEC	**Enteroaggregative *Escherichia coli***
ECOR	***E. coli* reference strain collection**
EDTA	**Ethylene diamine tetra acetate**

EHEC	**Enterohaemorrhagic *Escherichia coli***
EIEC	**Enteroinvasive *Escherichia coli***
EPEC	**Enteropathogenic *Escherichia coli***
ERIC	**Enterobacterial Repetitive Intragenic Consensus sequence**
ESBL	**Extended spectrum of beta-lactamase**
ETEC	**Enterotoxigenic *Escherichia coli***
ExPEC	**Extraintestinal pathogenic *Escherichia coli***
iTOL	**Interactive Tree of Life**
IGR	**Intergenic Regions**
MATE	**Multidrug and toxic compound extrusion**
MDR	**Multidrug resistance**
MGE	**Mobile genetic elements**
MLEE	**Multi-locus enzyme electrophoresis**
MLST	**Multi-locus strain typing**
MPEC	**Mammary Pathogenic *Escherichia coli***
NCBI	**National Centre for Biotechnology Information**
NGS	**Next generation sequencing**
NMEC	**Neonatal meningitis *Escherichia coli***
NRPS	**Non-ribosomal peptide synthase**
OD	**Optical density**
PCR	**Polymerase chain reaction**
PFGE	**Pulse field gel electrophoresis**
pks	**polyketide synthase**
RAPD	**Random Amplified Polymorphic DNA**

RM	**Restriction Modification**
SBF	**Specific Biofilm Formation**
SEPEC	**Sepsis associated pathogenic *Escherichia coli***
SSTI	**Skin and soft tissue infection**
ST	**Sequence type**
STEC	**Shigatoxigenic *Escherichia coli***
TAE	**Tris acetic acid Ethylene diamine tetra acetate**
UPEC	**Uropathogenic *Escherichia coli***
UPGMA	**Unweighted Pair Group Method with Arithmetic Mean**
UTI	**Urinary tract infection**
VF	**Virulence factors**
VFDB	**Virulence factor database**
WGS	**Whole genome sequencing**

Chapter 1

INTRODUCTION

Escherichia coli

Escherichia coli (Class: *Gamma proteobacteria;* Family: *Enterobacteriaceae)* is a ubiquitous and highly versatile Gram-negative, rod-shaped microorganism, which colonizes the human gut within a few hours of birth, further exhibiting interactions ranging from harmless commensalism (sometimes beneficial by offering colonization resistance against other invading pathogens) to severe forms of pathogenicity. The mucus layer of the mammalian caecum and colon forms the successful niche for commensal *E. coli*, which can rarely turn pathogenic in case of the immunocompromised state of the host and when the integrity of the gastrointestinal barrier is lost (Kaper et al., 2004). Analysis of *E. coli* genomes revealed that the genome sizes of pathogenic *E. coli* were observed to be 1.0 Mb in excess than the commensal *E. coli* strains, which could mainly be attributed to the presence of the genes which encode different virulence factors like toxins, adhesins, siderophores and invasins, that are unlikely to be present or absent in the commensal strains (Croxen and Finlay, 2010). These virulence factors are observed to be found mostly associated with pathogenicity islands and phages, and are capable of undergoing horizontal gene transfer (HGT) which can disseminate these traits, hence offering the recipient organisms various fitness advantages (Ahmed et al., 2008). This virulence gene repertoire can further integrate with the main chromosome, the successful combinations of which are capable of persistence and evolution into specific pathotypes capable of causing intestinal and extra intestinal infections in hosts (Kaper et al., 2004). There are six types of intestinal pathogenic *E. coli* (Nataro and Kaper, 1998; Kaper et al., 2004; Chaudhuri and Henderson, 2012) that have been well-characterized:

1. **Enteropathogenic *E. coli* (EPEC):** distinguished by characteristic attaching and effacing (A/E) lesions in histopathology.
2. **Enterohemorrhagic *E. coli* (EHEC**): associated with hemorrhagic colitis and haemolytic uremic syndrome, and produces shigatoxin (stx), also known as verocytotoxin.

3. **Enterotoxigenic *E. coli* (ETEC)**: encodes heat-stable/labile enterotoxins and is associated with infantile and traveller's diarrhoea.
4. **Enteroinvasive *E. coli* (EIEC)**: displays similarity with Shigella at biochemical, genetic and pathogenic levels, but manifests lesser clinical severity of infections.
5. **Enteroaggregative *E. coli* (EAEC)**: exhibits "stacked brick" adherence configuration, also known as auto aggregation.
6. **Diffusely adherent *E. coli* (DAEC)**: induces cytopathic effect on the enterocytes to form structures that resemble long finger-like projections, that can wrap around the diffusely adherent bacteria.

The **extra intestinal pathogenic *E. coli*** (ExPEC) are the *E. coli* (non-commensal) which can cause extra intestinal infections mainly includes sepsis-associated *E. coli* (SEPEC), neonatal meningitis associated *E. coli* (NMEC), and uropathogenic *E. coli* (UPEC) (Johnson and Russo, 2002; Croxen and Finlay, 2010). Avian pathogenic *E. coli* (APEC) forms another category of ExPEC that is associated with septicemia, respiratory infections and pericarditis in poultry (Kaper et al., 2004). Mammary pathogenic *Escherichia coli* (MPEC), also a subset of ExPEC was observed to cause mastitis in cattle, a common disease in cattle farms (Shpigel et al., 2008).

In addition, *E. coli* populations have also shown to be naturalized to persist for extended periods in sites/niches outside the intestinal tract such as sand, soil, sediments and also in association with algae and periphyton exhibiting genotypes distinct from animal origin ones, constituting the environmental *E. coli* (Jang et al., 2017)

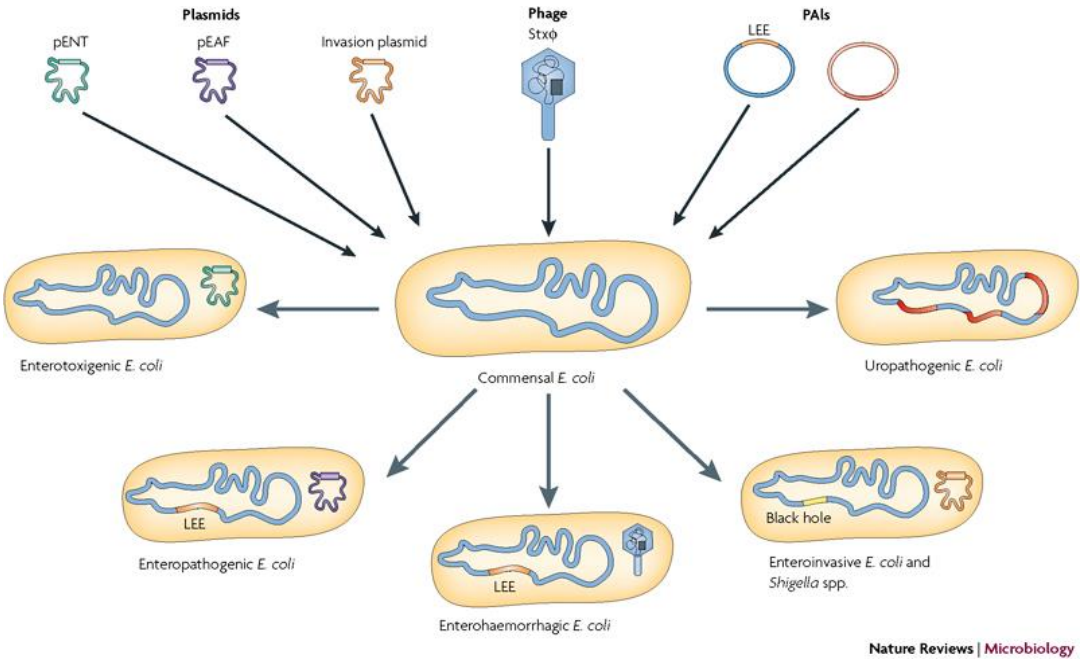


Figure 1.1: Acquisition of different mobile genetic elements through horizontal gene transfer which can contribute to the evolution of different types of pathogenic *Escherichia coli*. (Ahmed et al., 2008).

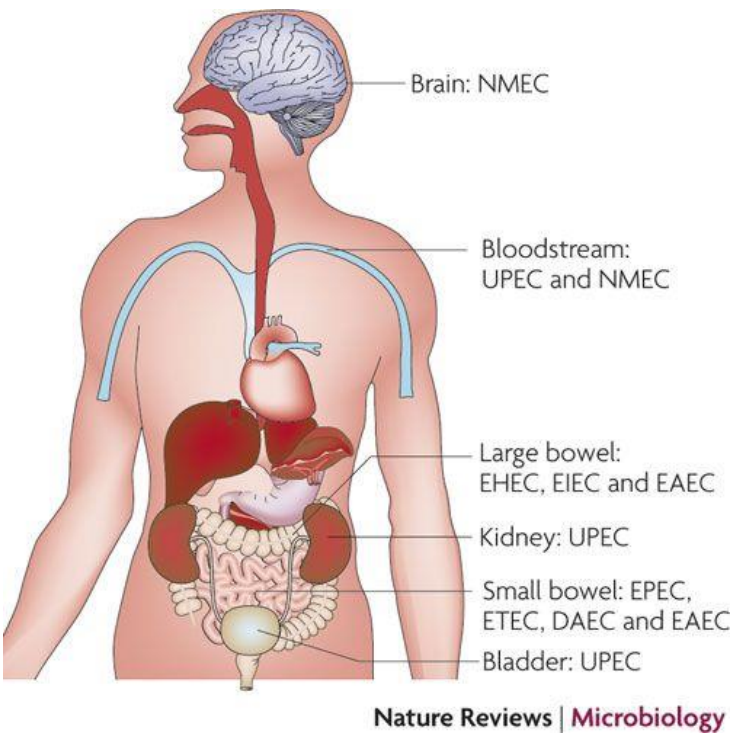


Figure 1.2: Different sites of colonization of *E. coli* pathovars in the human body which results in infections ranging from diarrhoea to septicemia and meningitis (Croxen and Finlay, 2010).

Methods of *E. coli* subtyping

Several methods of subtyping *E. coli* of varied approaches and specificities have been pivotal in understanding the molecular diversity, clonal lineages and phylogeny of the bacterium (Tenaillon et al., 2010; Chaudhuri and Henderson, 2012; Dale and Woodford, 2015). Non-random associations between 173 O (somatic), 80 K (capsular) and 56 H (flagellar) antigens led to the development of serotyping techniques which delineates *E. coli* into stable lineages (clones) (Tenaillon et al., 2010). Apart from the traditional serotyping employing antibodies to detect surface antigens O, H and K (Orskov et al., 1977), polymerase chain reaction (PCR) and whole genome sequencing (WGS) based methods enable rapid and more accurate serotyping (Fratamico et al., 2016). Multilocus Enzyme Electrophoresis (MLEE) which supports the neutral theory of molecular evolution characterizes isolates by relative electrophoretic mobility of several housekeeping enzymes which designates the strains to five important phylogenetic groups (also designated as phylogroups) i.e. A, B1, B2, D and E (Selander et al., 1986; Dale and Woodford, 2015). Quadruplex PCR based phylogrouping method developed by Clermont et al., rapidly assigns the *E. coli* strains to one of the seven phylogroups A, B1, B2, C, D, E and F and *Escherichia* cryptic clade I using the primers for the genes *chuA, yjaA, arpA* and TspE4.C2 to detect their presence/absence which forms the basis for the phylogroup assignment (Clermont et al., 2013). Similar to MLEE, Multilocus Sequence Typing (MLST) where nucleotide sequences of selected housekeeping alleles (n = 6 to 8) are analyzed for its nucleotide sequences and the allelic profile thus obtained is used as the basis of assignment of a sequence type to the strain (Maiden et al., 1998; Sullivan et al., 2005; Larsen et al., 2012). Currently, there are three schemes of MLST available for *E. coli*, each using different combinations of housekeeping genes, of which the Achtman scheme (http://mlst.warwick.ac.uk/mlst/dbs/Ecoli) is the most widely used (Clermont et al., 2015). In addition, a new subtyping method called CH typing which uses genes 489-nucleotide internal fragment of type 1 fimbrial adhesin gene *fimH* and 469-nucleotide internal *fumC* fragment

employed as a standard MLST locus, which enables sequence types to be further delineated to clonal subgroups, was also developed (Weissman et al., 2012). Pulse-field gel electrophoresis and whole genome sequencing (WGS) offers high-resolution phylogeny and epidemiological concordance (Dale and Woodford, 2015). The superior resolution, scalability, and declining costs associated with whole genome sequencing are increasingly making the approach most desirable for epidemiological surveillance and comprehensive study of the pathogen including genome-wide phylogeny, typing and profiling for virulence, resistance and phage coordinates (Clermont et al., 2015). The availability of numerous standalone and web-based software and well-curated data sharing repositories enable high -throughput genome analysis for transmission, population structure and evolution studies.

Virulence factors in *E. coli*

Pathogenic *Escherichia coli* possess an arsenal of virulence factors in varied repertoire which are capable of subverting the host cell mechanisms to enable persistence in otherwise protected environments in the host and exhibit mild to severe forms of pathogenesis (Croxen and Finlay, 2010). These virulence factors are categorized based on the type of mechanisms that it attributes to the pathogen during its multistep scheme of colonizing the host. Adhesins facilitate the colonization of the bacterium by forming characteristic structures like fimbriae and are also involved in signal transduction pathways or cytoskeletal rearrangements that facilitate pathogenesis (Kaper et al., 2004). Protectins offer protection against phagocytic engulfment and complement-mediated humoral immune response thereby evading host responses (Emody et al., 2003). Siderophores, the high-affinity iron chelators, enable the bacteria to competitively acquire iron from the host (Saha et al., 2013). Secreted toxins like heat-labile/stable enterotoxins, haemolysins, shigatoxins, genotoxins like cytolethal distending toxin *(CDT)*, cycle inhibiting factor *(cif)*, cytotoxic necrotizing factor *(CNF)*, colibactin etc. confer the bacterium ability to interfere with fundamental cellular pathways, including cell cycle of the host,

thereby damaging the host cell (Kaper et al., 2004; Dubois et al., 2010). These genotoxins are transported into the host cells using different types of secretion systems and autotransporters (Kaper et al., 2004).

Genomic islands comprise of large genomic regions (>10kb), often flanked by repeat structures, carrying cryptic or functional mobility factors (integrases, transposases etc.), display association with tRNA genes and possess distinct G+C content (Hacker and Kaper, 2000). A subset of genomic islands called pathogenicity islands (PAIs) confer "quantum leaps" in the evolution of bacterial virulence by carrying numerous virulence-associated factors and enables adaptive evolution of bacteria through its horizontal gene transfer (Groisman and Ochman, 1996; Dobrindt et al., 2004). Loss and gain of virulence genes have established different *E. coli* pathovars such as EPEC, EHEC, ETEC, ExPEC, DAEC etc. (Croxen and Finlay, 2010). EPEC harbors a characteristic 35 kb pathogenicity island which encodes for type III secretion system (T3SS) and other effector proteins, known as the locus of enterocyte effacement (LEE) (Mcdaniel et al., 1995). Phage encoded shigatoxin (verocytotoxin) forms the defining character of EHEC, which also harbors a 92 kb virulence plasmid that encodes for adhesin ToxB, and LEE which encodes for T3SS and effector proteins similar to EPEC (Nataro and Kaper, 1998). ETEC encodes for colonization factor (CF) which enables its engagement with host epithelial cells of the small intestine, along with outer membrane proteins *tia* and *tibA* (Turner et al., 2006). Heat stable and/or labile enterotoxins produced by the pathogen causes diarrhoea in ETEC infections (Turner et al., 2006). EIEC are obligate intracellular pathogens devoid of flagella and adherence factors where virulence is mostly mediated by a 220 kb plasmid which encodes for T3SS that facilitate invasion and cell survival in macrophages (Ogawa et al., 2008). Other key effectors which contribute to EIEC invasion and virulence include *ipaC, ipgD ipaA* etc (Ogawa et al., 2008). The 100 kb pAA plasmid of EAEC encodes for aggregative adherence fimbriae (AAFs) which contribute to the characteristic adherence in epithelial cells (Nataro and Kaper, 1998). Fimbrial and afimbrial

adhesins collectively designated as *afa-dir* adhesins along with the secreted autotransporter *(sat)* interact with the intestinal and urinary epithelial cells in DAEC infections (Servin, 2014) The virulence factors in different extraintestinal pathogenic *E. coli* pathovars have been tabulated in Table 1.1.

Table 1.1: Virulence factors associated with different ExPEC pathotypes (Sarowska et al., 2019)

Virulence genes		Function	ExPEC pathotype
Adhesins			
fim	Type 1 fimbriae	Gene involved in extraintestinal infections for colonization and biofilm formation	SEPEC, UPEC, APEC, NMEC
afa	Afimbrial adhesin	This non-fibrous adhesin binds to the DAF receptor present in the epithelial cell surface and also displays hemagglutination capability	UPEC
dra	Dr fimbriae	Binds to the DAF receptor present in the epithelial cell surface and mediates the internalization of the bacteria into host cells	UPEC
pap	P fimbriae	Involved in extraintestinal infections as a colonization factor.	UPEC, APEC, SEPEC
sfa	S fimbriae	Adherence factor which promotes bacterial adhesion onto the intestinal epithelial cells, lower urinary tract cells, as well as kidney cells	NMEC, UPEC
foc	F1C fimbriae	Adherence of renal epithelial cells and endothelial cells of the kidney and bladder	UPEC
iha	Iha	Adhesion factor, iron regulated	UPEC
mat	Mat	Meningitis associate fimbirae, temperature regulated	NMEC
crl, csg	Curli fiber gene	Involved in curli fiber formation during biofilm production, thereby enhancing bacterial pathogenicity	UPEC, APEC, SEPEC
agn43(flu)	Antigen43	Autotransporter protein involved in biofim formation and adhesion	UPEC
Invasins			
ibeA,B,C	Ibe ABC	Invasion into the cells of the host tissues	SEPEC, APEC, NMEC

Virulence genes		Function	ExPEC pathotype
Iron uptake			
iuc, aer	Aerobactin	Siderophores which acquire iron from the host system	UPEC, APEC
irp	Iron repressible protein	Synthesis of Yersiniabactin	NMEC
iroN	Salmochelin	Siderophore receptor	NMEC, UPEC, APEC, SEPEC
chu, hma	ChuA, Hma	Capable of uptake and utilization of Iron from the heamoglobin of the host	SEPEC, UPEC
sitA,B,C	SitABC	Iron and Manganese transportation	UPEC, APEC
Protectins/serum resistance			
traT	Transfer protein	Inhibit classical pathway of complement system	SEPEC, NMEC, APEC
KpsMI-neuA, KpsMII	Capsular Antigen	Protection of bacteria from phagocytosis by the host immune system	SEPEC, NMEC
omp	Outer membrane protein	Enables the bacteria to evade host immune system and facilitate intracellular survival	NMEC, UPEC
iss	Increased serum survival	Phagocytosis evasion	APEC, NMEC, SEPEC
colV, cvaC	ColV, CvaC	Colonization factor	NMEC, APEC, SEPEC
Toxins			
pic	Serine protease autotransporter	Damages the host cell membrane, facilitates colonization of epithelium, and performs mucin degradation	UPEC
sat	Secreted autotransporter toxin	A proteolytic toxin, effect cytotoxic—influences on cell vacuolization	UPEC
vat	Vacuolating autotransporter toxin	A proteolytic toxin, induces host cell vacuolization	UPEC, APEC
hlyA	Hemolysin A	Pore formation in host cell membranes which leads to cell lysis	UPEC

Virulence genes		Function	ExPEC pathotype
cnf	Cytotoxic necrotizing factor	Necrosis of host cells	SEPEC, UPEC
cdt	Cytolethal distending toxin	Cytolethal distending factor	SEPEC

Cyclomodulins: A class of bacterial toxins

Cyclomodulins are a class of bacterial genotoxins and effectors that can inflict genomic insults to the host DNA and interfere with the cell cycle which could lead to tumor initiation and progress (Nougayrède et al., 2005a; Gagnaire et al., 2017; Faïs et al., 2018). Four types of such toxins that have been well studied in *E. coli* have been described below:

Cytolethal distending toxin (CDT)

Cytolethal distending toxin (CDT) has been observed to be present in different members of *E. coli* and few other members of *Enterobacteriaceae.* CDT is a tripartite holotoxin in which CDT B forms the active enzyme and CDT A and CDT C are involved in its delivery to host cells (except in *Salmonella typhi* where it is directly internalized) (Lara-Tejero and Galán, 2001; Haghjoo and Galán, 2004; Nougayrède et al., 2005a). CDT B undergoes receptor-mediated endocytosis followed by translocation to the nucleus mediated by nuclear localization sequence further inflicting host DNA damage (Nougayrède et al., 2005a). This results in a cascade of DNA damage responses eventually resulting in the cell cycle arrest at the G2-M phase. CDT was observed to induce apoptosis in the lymphoid cells and senescence in epithelial, endothelia and mesenchymal cells (Grasso and Frisan, 2015; Martin et al., 2016). CDT intoxication is observed to contribute to the formation of actin stress fibres and characteristic distending phenotype of the host cells; prolonged survival of such cells could potentially

result in carcinogenesis (Nougayrède et al., 2005b). CDT has been shown to enhance host inflammatory response through damage-associated molecular patterns, production of IL-6, IL-8 and TNF-alpha (Shenker et al., 2001). Cytotoxicity of CDT towards B and T lymphocytes, inhibition of IFN-gamma secretion and apoptosis and inhibition of the activity of dendritic cells could render immunosuppressive functions to CDT (Martin et al., 2016). The contribution of CDT in colonization and long-term persistence of *Campylobacter jejuni, Helicobacter hepaticus* and *Salmonella typhimurium* in hosts have also been documented (Martin et al., 2016). CDT producing bacteria are overrepresented in colorectal cancer and IBD patient samples (Arthur et al., 2012; Buc et al., 2013a).

Cytotoxic necrotizing factor (CNF)

Cytotoxic necrotizing factor (CNF) is a transglutaminase that can deaminate Gln63 of Rho and Gln 61 of Rac and CDC42 thereby resulting in focal adhesions, the assembly of actin stress fibres, increased DNA synthesis and cytokinesis inhibition resulting in micropinocytosis and multinucleated giant cells (Horiguchi, 2001; Nougayrède et al., 2005a). CNF-1 and CNF-2 are homologs of the necrotizing toxin encoded by chromosome and plasmid respectively (Falbo et al., 1993). CNF-1 was also observed to block cell cycle at G2/M transition phase (Falzano et al., 2006) and properties similar to the well-studied *H. pylori* toxin CagA (Travaglione et al., 2008).

Cycle inhibiting factor (cif)

The Cycle Inhibiting Factor *(Cif)* is a cyclomodulin that is introduced into eukaryotic cells using the Type III secretion system (T3SS) of enteropathogenic and enterohemorrhagic *Escherichia coli* (EPEC and EHEC) (Jubelin et al., 2009). The effects of this toxin were first observed by De Rycke et al (De Rycke et al., 1997) in HeLa cells that were transiently infected with EPEC strains. *Cif* was observed to block the cell cycle at G2M (Nougayrède et al., 2001) and G1S (Samba-Louaka et al., 2008) transition phase depending upon the stage of the cell cycle during infection. The cytopathic effect demonstrated

by *cif* includes characteristic cell enlargement, focal adhesions, stress fibre formation with endoreduplication and impaired cytokinesis (Nougayrède et al., 2005b) and is dependent on the LEE in EPEC and EHEC (Kaper et al., 2004) although the gene encoding the toxin is located on a lambdoid prophage outside the LEE (Marchès et al., 2003; Taieb et al., 2011). This toxin could enable the pathogen to evade host immune responses by interfering with the dendritic cell functions and enable the pathogen colonization and persistence by delaying the epithelial cell renewal of the gut (Taieb et al., 2011).

Colibactin

Colibactin, another class of genotoxic, non-ribosomal peptide-polyketide secondary metabolite was discovered by Eric Oswald and coworkers in 2006 in uropathogenic, commensal and neonatal meningitis strains of *E. coli*. These strains were observed to induce double-stranded breaks when infected on cultured HeLa cells and transposon mutagenesis indicated a 54kb genomic island called *pks* (polyketide synthase) island responsible for the production of the genotoxin colibactin. *E. coli* strains harboring *pks* island demonstrated the ability to induce cytopathic effect characterized by megalocytosis and blocking the cell cycle at G_2M transition of mitosis (Brzuszkiewicz et al., 2006). It was observed that the transfer of the entire island to a *pks* negative strain rendered the strain capable of exhibiting genotoxicity (Brzuszkiewicz et al., 2006). The direct contact of the bacterium with the host cell is a prerequisite for its genotoxic activity and was not observed when the cells were treated with culture supernatants or cell lysates (Brzuszkiewicz et al., 2006). The genomic island encodes for three non-ribosomal peptide megasynthases (NRPS), three polyketide megasynthases (*PKS*), three hybrid NRPS/*PKS* and nine accessory tailoring and editing enzymes which constitute the machinery for colibactin synthesis(Brzuszkiewicz et al., 2006).

In vivo infection experiments induced DNA damage responses in mouse models and further cell divisions were accompanied by signs of incomplete DNA repair, increased mutations, anchorage independence and other forms of chromosome instability.(Cuevas-Ramos et al., 2010) Experimental evidence also supports that *pks* island enables long term persistence of *E. coli* in the human gut possibly due to retarded rate of renewal of enterocytes, with possibilities of carcinogenesis(Nowrouzian and Oswald, 2012). Colibactin has also found to elicit metabolic reprogramming in infected cells, producing ROS and proinflammatory molecules, inducing premature cellular senescence. Through soluble and extracellular matrix-associated factors, *pks*+ *E. coli* induced a "bystander" effect in the nearby cells, displaying a pro-tumorigenic activity. Transient infection studies revealed that the effect prevailed in the absence of bacteria. (Secher et al., 2013) Enhancement of tumor was confirmed in xenograft and mouse models, where the senescent intestinal epithelial cells produced factors like hepatocyte growth factor (HGF) which promoted tumor growth (Cougnoux et al., 2014). Colibactin promotes senescence by destabilizing the SUMOylation process under the control of miRNA by downregulating SENP1, thereby modulating tumor microenvironment favoring cancer progression(Dalmasso et al., 2015a). Colibactin was also found to enhance lymphopenia in sepsis mouse models and also impaired the chances to survive through antibiotics and rehydration treatment, providing the first experimental evidence of colibactin as a virulence factor in *E. coli* (McCarthy et al., 2015). Analysis of colorectal cancer biopsies has also shown a high prevalence (55-66.7%) of *pks*+ *E. coli,* which is consistent with the above experimental observations of the role of colibactin in cancer progression (Putze et al., 2009). *pks* island genes were also observed to be enriched in colorectal cancer metagenomes (Wirbel et al., 2019). Infection with *pks* positive *E. coli* was observed to transform the normal murine colon epithelial cells and organoids into a pre-malignant state through chromosomal aberrations (Iftekhar et al., 2021). Colibactin was demonstrated to induce covalent modification in host DNA, formation of adenine-colibactin adducts and alkylate DNA *in vitro* and *in vivo* (Wilson et

al., 2019). The toxin also leaves a characteristic mutational signature which can be detected in CRC patients (Pleguezuelos-Manzano et al., 2020).

This genomic island is observed among pathogenic, commensal and even probiotic bacterial strains (Massip et al., 2019). The epidemiological information pertaining to *pks*+ *E. coli* has been summarized in Table 1.2. The *pks* island was also observed to be prevalent in other bacteria belonging to the family Enterobacteriaceae, other than *E. coli* like *Klebsiella pneumoniae, Klebsiella aerogenes, Citrobacter koseri,* (Putze et al., 2009). The island was found to be integrated into the *asn* tRNA locus, a hotspot for DNA insertion and recombination and together with P4-like bacteriophage integrase genes, and is flanked by direct repeats of 16bp and possesses an increased GC content compared to the *E. coli* core genome. These integrative elements function to transfer these determinants to other members of *Enterobacteriaceae.* Until the discovery of colibactin, iron chelators yersiniabactin and enterobactin were the only known polyketide/non-ribosomal peptide hybrid and non-ribosomal peptide respectively in the family *Enterobacteriaceae. pks* island shows distribution similar to that of yersiniabactin and indicates a possible linkage between both the determinants. It suggests the presence of a chromosomal integration event resulting in the acquisition of these pathogenicity island followed by further dissemination via homologous recombination. GC content analysis indicates the possible origin of the *pks* island from *C. koseri,* but not have been yet confirmed. (Putze et al., 2009)

Colibactin biosynthesis is carried out by an assembly line machinery located in the 54 kb *pks* genomic island which consists of 19 genes comprising of non-ribosomal peptide megasynthases (NRPS: *clbH, clbJ* and *clbN),* polyketide megasynthases (*PKS*: *clbC, clbI* and *clbO)*, two-hybrid NRPS-*PKS* megasynthases (*clbB* and *clbK)* and nine accessory and tailoring enzymes. A recent study has described the regulatory role of *clbR,* a LuxR-type DNA-binding helix-turn-helix (HTH) domain as a key transcriptional activator involved in the expression of the colibactin biosynthetic gene cluster

(Wallenstein et al., 2020). Systematic mutagenesis of the genes deciphered that all the genes except one accessory enzyme gene were required to induce the cytopathic effect on host cells. Inactive precolibactin is transported to periplasm space using *clbM* (MATE transporter) where *clbP* peptidase removes the N-myristoyl-D-asn to release mature functional colibactin(Mousa et al., 2016, 2017). This prodrug activation mechanism, in addition to colibactin resistance protein *clbS* whose cyclopropane hydrolase activity could render colibactin into an innocuous compound, are involved in the protection of pathogen's self-DNA from the genotoxic activity of colibactin (Bossuet-Greif et al., 2015, 2016; Tripathi et al., 2017).

The unstable nature and trace amounts in which colibactin is produced by the cells have made this compound elusive to isolation and characterization. Different multidisciplinary studies have been undertaken to determine the biosynthesis, activity and cellular trafficking of colibactin metabolite and the structure of colibactin has been successfully deduced recently using genetics, tandem mass spectrometry, chemical synthesis and isotope labelling techniques using colibactin crosslinked to two DNA nucleobases isolated from bacterial extracts (Xue et al., 2019b). It was determined that colibactin is formed through the coupling of two complex intermediates which generates a nearly symmetrical structure possessing two electrophilic cyclopropane warheads which form the site of DNA adduction and crosslinking (Xue et al., 2019a).

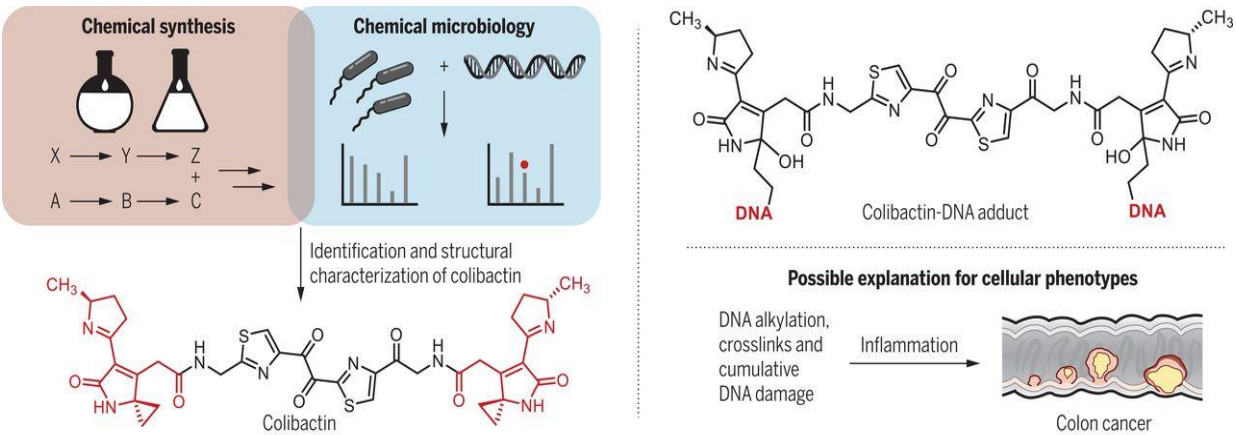


Figure 1.3: Colibactin structural elucidation through a combined approach of chemical synthesis and tandem mass spectrometry of isolated colibactin-DNA crosslinks. The elements in the colibactin structure highlighted in red are the cyclopropane motifs, which form DNA cross-links that bind to DNA forming adducts. (Xue et al., 2019b).

Table 1.2: Epidemiological data on the prevalence of *pks* positive bacteria

Reference	Organism	Total no. of isolates	*pks* positive prevalence	Details of isolates	Geographical location
(Putze et al., 2009)	*E. coli*	1092	9.5% (104/1092)	Commensal and clinical ExPEC	Germany
	K. pneumoniae	141	3.5% (5/141)	Positives were extraintestinal pathogenic isolates	
	E. aerogenes	11	27.3% (3/11)		
	C. koseri	1	100% (1/1)		
(Nougayrède et al., 2006)	*E. coli*	185	0% (intestinal *E. coli)* 53% (ExPEC) 34% (Fecal *E. coli)*	55 intestinal *E. coli,* 92 ExPEC and 32 from healthy individuals	Germany
(Johnson et al., 2008)	ExPEC	131	44% (58/131)	62 bloodstream isolates and 69 fecal isolates	USA
(Nowrouzian and Oswald, 2012)	*E. coli*	130	33%	Rectal swabs and fecal samples from infants	Sweden
(Buc et al., 2013b)	*E. coli*	69	55% CRC 19.3% diverticulosis	38 CRC biopsies 31 diverticulosis	France
(Sarshar et al., 2017)	*E. coli*	1500	88/1500	1500 *E. coli* isolates from 20 biopsies of precancerous lesions and healthy controls. No *pks* positives detected in the control	Rome
(Shimpoh et al., 2017)	*E. coli*	98	43%	Colonic lavage of CRC	Japan
(Chen et al., 2017)	*K. pneumoniae*	400	16.7% (67/400)	Blood, respiratory, urine and other (100 samples each)	Taiwan
(Payros et al., 2014)	*E. coli*	184	26.9%	Stool samples from healthy neonates	France
(Micenková et al., 2017)	*E. coli*	314	31.4%	Blood samples	Czechia and Slovakia
(Krieger et al., 2011)	*E. coli*	18	72.2%	Urine	USA
(Lai et al., 2014)	*K. pneumoniae*	207	25.6%	Liver and non-hepatic abscesses and non-abscesses related cases	Taiwan
(Lan et al., 2019)	*K. pneumoniae*	190	26.8% (51/190)	Blood stream isolates	China
(Iyadorai et al., 2020)	*E. coli*	48	16.7% (8/48)	Resection tissues from CRC patients	Malaysia
		23	4.35% (1/23)	Biopsy from colonoscopy in healthy control	

Reference	Organism	Total no. of isolates	*pks* positive prevalence	Details of isolates	Geographical location
(Lee and Lee, 2018)	*E. coli*	146	17.8% (26/146)	Blood samples	Korea
(Shimpoh et al., 2017)	*E. coli*	35	43% (15/35)	Colonic lavage from CRC patient	Japan
		37	51% (19/37)	Colonic lavage from adenoma patient	
		26	46% (12/26)	Colonic lavage from healthy control	

So far, only a few studies have attempted to understand the pattern of transfer and evolution of the *pks* pathogenic island and its co-evolution with the genome that harbors it. Enterobacterial repetitive intergenic consensus (ERIC) and Random amplified polymorphic DNA (RAPD) based genetic fingerprinting of *pks*-positive *E. coli* obtained from human intestinal polyps showed diverse clustering patterns implying the potential ability to colonize different environments (Sarshar et al., 2017). Another study performed bioinformatic analyses which displayed the high prevalence of *pks* island among *Escherichia* species with close similarity of *pks* island of *E. coli* with *K. aerogenes, K. pneumoniae* and *C. koseri* (Morgan et al., 2019). The combination of *in-silico* and *in vitro* study on ECOR collection by Messerer, Fischer and Schubert, 2017 demonstrated that the immobile PAI group i.e., those devoid of any transfer or mobility regions, comprising of the high-pathogenicity island (HPI), *pks*, and *serU* undergoes horizontal gene transfer *"en bloc"* along with neighboring chromosomal backbone; which was observed to be F' mediated transfer (Messerer et al., 2017). The high homology within *pks* island sequences also conveyed its recent acquisition (Messerer et al., 2017)

Rationale and Objectives

Mobile genetic element enabled horizontal gene transfer (HGT), inactivation of antivirulence genes (Bliven and Maurelli, 2012) and point mutation derived functional alterations significantly contribute to the evolution of virulence in *Escherichia coli* (Denamur et al., 2020). Virulence factors like toxins,

adhesins, capsules and iron acquisition systems could potentially be carried on or shuttled through mobile genetic elements like genomic islands, phages and plasmids. These genes are capable of undergoing horizontal gene transfer among compatible organisms (Clermont et al., 2000; Ahmed et al., 2008) and observed abundantly distributed in extraintestinal pathogenic (ExPEC) strains. Previous reports from our group (Ranjan et al., 2015b) (Hussain et al., 2014a) (Hussain et al., 2012b) (Jadhav et al., 2011) have shown that the ExPEC infections are endemic and antibiotic resistance within these strains, a major problem worldwide. Our studies have revealed the presence of the most successful pandemic clone ST131 in India and its possible evolutionary success worldwide(Ranjan et al., 2015b) (Hussain et al., 2014a).

Considering altogether the serious concern of ExPEC infections in India this study was undertaken to decipher the prevalence of highly virulent ExPEC strains from our collection that could produce the genotoxin colibactin to gain insights into the risks towards severe infection and its possible epidemiological links. This study further aimed to look into the antibiotic resistance and the virulence determinant of such isolates and characterize these in terms of their genotypic and phenotypic properties to gain the first insights in India where the infection burden is quite high.

Whole genome sequencing and genome-wide analysis are increasingly being undertaken as a popular method in epidemiological investigations owing to the extraordinary resolution up to single base differences and the ability to capture the genome-wide signature of the organism in a cost-effective manner (Parkhill and Wren, 2011). In order to gain a high-resolution understanding of the acquisition, maintenance and evolution of the *pks* island, we also performed a large-scale high throughput pan-genome and phylogenetic analysis. A combinatorial approach involving functional molecular epidemiology and high-throughput phylogenomic analysis was undertaken for this work to comprehensively study and contribute insights to the distribution and evolutionary dynamics of this

pathogenic island of clinical significance and the three broad objectives of the study are as mentioned below:

1. Molecular epidemiology and characterization of *pks* island harboring *E. coli* from the Indian population
2. Whole genome comparative analysis of *pks* island harboring *Escherichia coli* from the Indian population
3. Deciphering the evolution and transmission of *pks* island harboring *Escherichia coli*

Chapter 2

METHODOLOGY

Bacterial isolates

The molecular epidemiology and functional characterization of *pks* positive *E. coli* were carried out using 462 ExPEC isolates obtained from Dr D. Y. Patil University Hospital, Pune, India as a part of their routine diagnostic screening during the years 2009–2015, as described in our previous study (Ranjan et al., 2016). Out of the 462 isolates, 370 isolates cultured from urine, 63 were cultured from pus and the remaining 29 were obtained from other extra intestinal clinical samples. Standard microbiological laboratory methods were undertaken for the identification and preservation of these isolates (Jadhav et al., 2011). All the isolates were collected, preserved and handled as per standard biosafety guidelines and according to the approvals of the Institutional Biosafety Committee (IBSC) of the University of Hyderabad (Ref. UH/IBSC/NA/12/7 dated 09/4/2012 and NA-N-32 dated 27/8/2015). The clinical details of the *pks* positive isolates is tabulated in Table 3.1.

Preparation of Lysates (Heat Lysis Method)

Heat killed bacterial cell lysates were prepared which contained genomic DNA and this lysate was used as the template for PCR amplification (Ranjan et al., 2017). Micropipette tips containing scrapped bacterial culture was mixed thoroughly in 200 µL Milli-Q in a 0.2 mL PCR tube. The tubes were kept in a thermal cycler for 95°C for 25 minutes for the cell lysis to take place. The PCR tubes were centrifuged at 6000 RPM for 5 minutes and the supernatant was transferred to fresh Eppendorf tubes.

Detection of *pks* genomic island

The presence of *pks* island in the isolates was ascertained by PCR mediated screening using primers corresponding to the four representative genes of the island, ie., two flanking (*clbB* and *clbQ*) and two internal (*clbA* and *clbN*), also denoting the integrity of the island (Johnson et al., 2008). PCR was carried out using heat killed bacterial lysates as templates in 30 cycles at the following reaction conditions:

initial denaturation at 95^{0}C for 4 minutes, denaturation at 94^{0}C for 30 seconds, annealing and extension conditions varied depending upon the gene, and the final extension was carried out at 68^{0}C for 10 minutes. Gene-specific reaction conditions were set according to Table 2.1.

Table 2.1: PCR conditions for the screening of *pks* island

Gene	Annealing temperature(^{0}C)	Annealing time (seconds)	Extension temperature (^{0}C)	Extension time (seconds)	PCR product size (bp)
clbA	45	30	68	60	981
clbN	60	30	68	50	711
clbB	48	30	68	40	556
clbQ	65	30	68	50	797

Phylogroup identification by tetraplex PCR

Multiplex-PCR amplification of the four genes ie., *arpA, TspE4.C2*, *chuA,* and *yjaA* was employed to perform the *in silico* phylogrouping of the *pks* positive isolates (Clermont et al., 2013). Initial denaturation was carried out at 95^{0}C for 5 minutes, denaturation at 94^{0}C for 45 seconds, annealing at 55^{0}C for 45 minutes, extension at 68^{0}C for one minute, final extension at 68^{0}C for seven minutes. Based on the presence of genes the strains were grouped into one of the eight phylogroups (Clermont et al., 2013).

Antibiotic susceptibility and ESBL production

pks positive isolates were subjected to antibiotic susceptibility test (AST) using the Kirby-Bauer disc diffusion method, on bacterial lawns cultured in Mueller Hinton agar plates. Antimicrobial discs (Himedia, India) for ten different antibiotics i.e., clarithromycin (15 µg), chloramphenicol (30 µg), fosfomycin (200 µg), tetracycline (30 µg), co-trimoxazole (20 µg), gentamicin (10 µg), doxycycline (30 µg), nalidixic acid (30 µg), colistin (10 µg) and ciprofloxacin (5 µg) belonging to eight different classes

were used for AST of the isolates. Sterile cotton swabs on wooden applicators were used to pick up a single colony of the bacterial strains and dissolved in 3mL LB media. A fresh swab was used to streak the entire agar surface of the plate three times, turning the plate at 60^0 angles between each streaking. Discs were aseptically deposited with sterile forceps. The plates were inoculated at 37^0C overnight and the zone of inhibition around each disc was measured using Zone Scales (HiMedia). Isolates that displayed resistance against three or more classes of antimicrobials were categorized to b multi-drug resistant (MDR).

Disc synergy between clavulanic acid and indicator cephalosporins, CAZ (ceftazidime) and CTX (cefotaxime) was determined to study the ESBL production capabilities of the *pks* positive isolates. The bacterial lawn was spread evenly on Mueller Hinton Agar plates and the antimicrobial disc with and without the inhibitor clavulanic acid was deposited on the agar surface using sterile forceps. The plates were incubated at 37^0C overnight and the zone of inhibition around each disc was calculated. The bacterial strains were designated positive for ESBL production if the diameter of the zone of inhibition around disc containing antibiotic and clavulanic acid was greater than the indicator antibiotic alone. AST and ESBL production assays were carried out in accordance with the guidelines of Clinical Laboratory Standards Institute (CLSI) (CLSI, 2013).

Virulence and antimicrobial resistance genotyping

The *pks* positive isolates were screened for various virulence genes encoding toxins (*usp, cvaC, sat),* adhesins (*fimH, afa* and *sfaD/E),* protectant (*ibeA)* and siderophore (*iucD)* using PCR with reaction conditions and primers as described in a previous study (Jadhav et al., 2011; Ranjan et al., 2017). ESBL gene $bla_{\text{CTX-M-15}}$, (Monstein et al., 2007), tetracycline resistance gene (*tetA),* sulfonamides (*sul1*) and aminoglycoside acetyl transferases (*aac(6')-Ib)* were also screened for using primers and PCR conditions as described previously (Jadhav et al., 2011; Ranjan et al., 2016b, 2017). The *pks* positive isolates were

also screened for the gene *TEM* using its generic primers, and the gene variants were determined by the sequencing of the amplified products from PCR.

The bacterial colony lysate was prepared and used as the template and PCR was carried out in 30 cycles. Initial denaturation at 95^0C for 4 minutes, denaturation at 94^0C for 30 seconds, annealing and extension conditions varied depending upon the gene, the final extension was carried out at 68^0C for 10 minutes. Gene-specific reaction conditions were set according to Tables 2.2 and 2.3.

Table 2.2: PCR conditions for the genotypic resistance profiling

Gene	**Annealing temperature (^{0}C)**	**Annealing time (seconds)**	**Extension temperature (^{0}C)**	**Extension time (seconds)**	**PCR product size (bp)**
tetA	65	30	68	60	500
aac6'	56	30	68	45	600
sul1	65	30	68	60	750
$bla_{CTX\text{-}M\text{-}15}$	50	30	68	60	500
TEM	53	30	68	45	1000

Table 2.3: PCR conditions for the genotypic virulence profiling

Gene	Annealing temperature (^{0}C)	Annealing time (seconds)	Extension temperature (^{0}C)	Extension time (seconds)	PCR product size (bp)
fimH	56	30	68	60	1000
sfaD/E	61	30	68	30	500
usp	63	30	68	60	1000
afa	57	30	68	60	600
ibeA	50	30	68	30	750
cvaC	56	30	68	40	750
iucD	56	30	68	45	750
sat	55	30	68	45	600

Determination of siderophore production

The 35 *pks*-positive isolates subjected to siderophore production assay using Chrome Azurol S Blue agar plates (Schwyn and Neilands, 1987) to determine their iron acquisition capabilities. 100mL of Kings Broth, 30.2g of PIPES and 18.2g of Agar were dissolved in 750mL of double-distilled water, pH was maintained to 6.8 and volume was made up to 900mL. The dye solution was prepared by dissolving 60.5mg Chrome Azurol S dye, 10 mL of 1mM $FeCl_3$ in 10mM HCl, 72.9 mg of CTAB in double distilled water to a final volume of 100mL. The media and dye solution was autoclaved separately. After the media cooled to 50 ^{o}C, the dye was added and poured onto plates. The pure colony of the isolates were streaked separately on the plates, which were incubated at 37 ^{o}C for overnight. The isolates whose colonies showed the characteristic orange halos after incubation were designated to be siderophore producers (Schwyn and Neilands, 1987).

Biofilm formation assay

All the *pks*-positive isolates were subjected to the assay to determine their biofilm-forming capabilities as described previously (Nandanwar et al., 2014). Flat-bottom 96-well microtiter plates were used for the biofilm formation assay, which was performed twice in technical triplicates. The bacterial cultures were grown overnight and the OD at 600 nm was determined for the same. The cultures of the isolates were diluted to an OD of 0.05 in a fresh M63 medium (minimal medium). From the diluted cultures, 200 µL was pipetted in triplicates into the sterile 96-well microtiter plates. Initial OD at 600 nm ($OD_{600\ (0h)}$) was obtained and these plates were then covered by breathable sealing films. The plates were further incubated at a stationary condition for 48 hours at 28°C. After 48 hours of incubation, OD was obtained at 600 nm ($OD_{600\ (48h)}$). Media was aspirated from the wells and the wells were washed three times with 300 µL of deionized water. The plates were airdried and the bacteria were fixed using 250 µL of 99% methanol for 15 minutes. 0.1 % crystal violet solution was used for 30 minutes for staining the bacteria. The wells were further washed three times with deionized water to remove excess stain and the plates were air-dried. Inorder to solubilize the attached and stained bacteria, a solution containing 300 µL of Ethanol : Acetone (80:20) was added to the wells, and the plates were incubated for 30 minutes at 100 rpm. The OD at 570 nm was read in a microtiter plate reader.

To calculate the SBF (Specific Biofilm Formation) value; the following formula was used:

$$\text{SBF (specific biofilm formation)} = (AB-CW)/G$$

Where AB= OD at 570 nm of attached and stained bacteria, CW= OD of control wells (wells containing bacterium free media) at 570 nm and G= [$OD_{600\ (48h)}$ - $OD_{600\ (0h)}$], corresponding to the bacterial growth.

Serum Resistance Assay

All the 35 the *pks* positive *E. coli* in the collection were subjected to the Serum Resistance Assay, wherein the isolated were analyzed for their capabilities of exhibiting resistance against human serum (50%) (Hussain et al., 2014b). Overnight culture (5 µL) was added to the LB broth (495 µL) and placed on a shaking incubator (200 rpm) for one hour incubation at 37 °C. The cultures were further pelleted and resuspended in 1X PBS (1mL). In a 96 well mircotiter plate, 30 µL of this inoculum was added to 270 µL of 50% human serum in triplicate wells. The initial sample was collected before the incubation, which was serially diluted and on the LB agar plates to enumerate the colony-forming units (CFU) at 0-hour. The 96-well plates were further subjected to incubation at 100 rpm for 3 hours at 37 °C. After the 3 hours of incubation, the samples were serially diluted, and further plated on the LB agar plates to determine the 3-hour CFU counts. Strains which showed equal or higher number of colony-forming units (CFU) at 3 hours as compared to 0 hours were considered to be resistant to human serum. The experiment was performed in technical triplicates and repeated twice.

Statistical Analysis

GraphPad Prism (version 5.01) was employed for the statistical analysis. The Nonparametric Mann-Whitney U test was performed for serum resistance assay and p-values ≤ 0.05 were considered to be significant.

Enterobacterial Repetitive Intergenic Consensus-PCR

Genetic relatedness of the *pks*-positive isolates was performed by Enterobacterial Repetitive Intergenic Consensus Sequence (ERIC) based PCR using ERIC1R (5' ATGTAAGCTCCTGGGGATTCAC 3') and ERIC2 (5' AAGTAAGTGACTGGGGTGAGCG 3') primers (Versalovic et al., 1991). The PCR was performed for 30 cycles, as per the reaction conditions described below. Agarose gel (1.5%) was used to run the amplicons and the gel image was analyzed using BioNumerics software (version 7.1,

Applied Maths Belgium). A dendrogram was obtained by dice similarity index based on the Unweighted Pair Group Method with Arithmetic mean (UPGMA) algorithm.

Table 2.4: Reaction conditions for ERIC PCR

Reaction Step	Temperature (^{0}C)	Time(seconds)
Initial Denaturation	95	120
Denaturation	94	30
Annealing	47	60
Extension	68	480
Final Extension	68	600

Whole genome sequencing, assembly and annotation

Genomic DNA from 25 *pks* positive *E. coli* isolates were isolated and purified of any RNA contamination using Qiagen DNeasy blood and tissue kit (Qiagen, Germany) and sequenced by using the Illumina MiSeq platform. The paired-end reads were obtained in FASTQ format which was further subjected to NGS QC Toolkit (Patel and Jain, 2012) for their quality control. The high-quality reads greater than 150 bp were selected and Fastx trimmer was used to trim the filtered reads at positions based on a Phred Score cutoff value of 20. The trimmed filtered reads after quality control were used for further assembly. The high-quality reads of the 25 strains were further subjected to *de novo* assembly using SPAdes Genome Assembler (v3.6.1) with the coverage cutoff value of 20 (Bankevich et al., 2012) and QUAST (Gurevich et al., 2013) was used to determine the assembly statistics. Two out of the 25 strains were discarded from further analysis due to poor quality. The contigs were further ordered and scaffolded using C-L-Authenticator using *E. coli* ATCC25922 complete genome as reference and the scaffolds were annotated using PROKKA (Seemann, 2014). Artemis (Rutherford et

al., 2000) was used to determine the genome statistics and sequence types of the isolates were identified using the *in silico* MLST pipeline using in-house scripts and the publicly available MLST pipeline (https://github.com/tseemann/mlst), which uses the PubMLST database (https://pubmlst.org/) (Jolley and Maiden, 2010). ECTyper (https://github.com/phac-nml/ecoli_serotyping) was used to determine the serotypes of the genomes *in silico*.

Analysis of the genomes for the resistance and virulence determinants

Amino acid sequence files of the annotated genomes were used to determine the virulence and resistance gene profiles by performing BLASTp against the Virulence Factor Database (Chen et al., 2005) and Comprehensive Antibiotic Resistance Database (McArthur et al., 2013), respectively. Percentage identity of 70% and query coverage of 75% were used as thresholds for the analysis of the genomes for the presence of the respective genes. The heat plots depicting the presence-absence status of the genes was generated using gplot (https://github.com/talgalili/gplots) package of R. The virulence and resistance profiles of 23 in-house *pks*-positive genomes were also compared with those of 23 in-house *pks*-negative genomes using the same methodology (accession IDs are listed in Table A2 in the appendix).

Whole genome comparative analysis and visualization

E. coli IHE3034 (Accession: CP001969.1) complete genome was used as the reference genome in the analysis using Blast Ring Image Generator (BRIG) wherein the assembled genomes of *pks* island carrying in-house strains (n=23) were compared to the reference (Alikhan et al., 2011) to determine their genetic relatedness. The coordinates of *pks* island were also annotated in the BRIG image.

The trimmed and filtered reads of the genomes were mapped and aligned to the reference *pks* island sequence along with flanking regions obtained from NCBI (Accession No.: AM229678.1) using

SAMtools (Li et al., 2009) and Bowtie 2 (Langmead and Salzberg, 2013). The mapped reads were assembled *de novo* using SPAdes and the sequences of assembled contigs were obtained, which corresponded to *pks* island sequences from each genome. These *pks* island sequences were further subjected to BRIG using a complete *pks* island sequence (Accession No.: AM229678.1) as the reference to visualize the integrity of the genomic island.

Identification of *pks*-island containing genomes in the public domain

E. coli genome sequences were downloaded from the public database using in-house scripts and the genomes which were greater than 4.8 Mb size and which had less than 200 contigs each, were used for the study. A total of 3784 draft and 306 complete genomes were selected after the data curation step as the final database and were employed for further downstream analysis. Complete *pks* island along with flanking regions of *E. coli* IHE3034 and its coding sequences were obtained from NCBI (Accession No.: AM229678.1) as the reference sequence of the genomic island. The genomes were screened for the presence of *pks* island genes obtained from the above reference using BLASTn (Camacho et al., 2009) with query coverage and identity thresholds of 85%, respectively.

Genome annotation, *in-silico* MLST and phylogrouping

The genomes of both NCBI and newly sequenced in-house isolates (n=23) were subjected to annotation using the software Prokka (Seemann, 2014). All the genomes were subjected to *in-silico* phylogrouping and Multi Locus Sequence Typing (MLST) to determine their phylogroups and sequence types using *in-silico* MLST pipeline using the in-house scripts (Hussain et al., 2017; Shaik et al., 2017) and MLST pipeline (https://github.com/tseemann/mlst) which uses PubMLST database (https://pubmlst.org/) (Jolley and Maiden, 2010). ECTyper (https://github.com/phac-nml/ecoli_serotyping) was used to determine the serotypes of the genomes using *in-silico* serotyping.

ST95 pan-genome analysis

ST95 genomes were used as a model dataset to study the distribution and evolutionary pattern of *pks* island because of the availability of both *pks*-positive (n = 110) and negative genomes (n=49). The ST95 genomes (n=159) were annotated using Prokka (Seemann, 2014), and the pangenome analysis was performed using Roary (Page et al., 2015), with percentage identity and e-value cutoffs of 85% and 0.00001, respectively, for the determination of orthologous gene clusters. Genes shared by all the 159 isolates constitute the core genome, and the core genes identified from pangenome analysis were subjected to COG classification using eggNOG (Huerta-Cepas et al., 2016), and the COG groups were tabulated. The genomes were further analyzed using CHTyper (Roer et al., 2018) which uses *fumC* and *fimH* allele sequences for the *in-silico* determination of CH Types.

ST95 core genome phylogeny

The core genes determined using Roary (Page et al., 2015) were subjected to PRANK (Löytynoja, 2014) for nucleotide alignment, and the resultant core genome alignment was subjected to trimAL (Capella-Gutiérrez et al., 2009) (using --strict flag) for trimming and refining the alignment. The alignment was also used as input for hierBAPS (Cheng et al., 2013) to perform Bayesian method based hierarchical clustering based on sequence variations. The refined alignment was further subjected to IQ-TREE (Nguyen et al., 2015) with Model Finder (Kalyaanamoorthy et al., 2017) to optimize the best nucleotide substitution model and to construct a maximum likelihood phylogenetic tree with 500 bootstrap replicates. The maximum likelihood phylogenetic tree of the core genome thus obtained was further subjected to ClonalFrameML (Didelot and Wilson, 2015) to handle recombination regions and the resultant final tree was visualized using Interactive Tree of Life (iTOL) (Letunic and Bork, 2019). The branches of the core genome phylogenetic tree were colour-coded according to the presence/absence of *pks* island, and the BAPS cluster, serogroup and CH type information were also

annotated in the tree using data strips. A core genome phylogenetic tree including an outgroup ED1a (Accession no.: GCA_000026305.1) was also additionally constructed using the methodology which has been mentioned above.

ST95 IGR phylogeny

The GFF files derived from the annotation of ST95 (n=159) genomes using Prokka (Seemann, 2014) and the gene presence-absence output file from the pan-genome analysis by Roary (Page et al., 2015) were used to perform the intergenic region analysis using Piggy (Thorpe et al., 2018). The core IGRs, which constitute the intergenic regions (IGRs) which were shared by all the genomes, were extracted and aligned using Prank (Löytynoja, 2014), followed by trimming using trimAL (Capella-Gutiérrez et al., 2009) to refine the alignment by removing spurious and poorly aligned regions. IQ-TREE (Nguyen et al., 2015) was employed along with Model Finder (Kalyaanamoorthy et al., 2017) (-MFP flag) for construction of IGR phylogeny with 500 bootstrap replicates, followed by recombination region analysis using ClonalFrameML (Didelot and Wilson, 2015) to produce maximum likelihood phylogeny of the core intergenic regions of ST95 genomes. The resultant phylogenetic tree was visualized using iTOL (Letunic and Bork, 2019) and annotated with CH Type and serotype information of the isolates as obtained previously.

Pan-genome wide analysis using Scoary

Pan-genome wide association study comparing the genomes belonging to *pks*-positive and mixed clades of the core genome and core intergenic phylogenetic trees with genomes belonging to exclusively *pks*-negative clade was performed using Scoary (Brynildsrud et al., 2016) with the help of the output gene_presence_absence.csv file of Roary (Page et al., 2015). The *pks*-positive genomes (n=110) and *pks*-negative genomes belonging to the mixed clade (n=20) were grouped together and designated with trait value "1", and the *pks*-negative genomes forming the exclusive *pks*-negative clade

(n=29) were designated with trait value "0" in the Scoary (Brynildsrud et al., 2016) input. The analysis was extrapolated to screen the *pks*-positive (n=530) and *pks*-negative (n=3583) genomes prevalence of the various differentially enriched genes among the two groups using BLASTn (Camacho et al., 2009) with identity and query coverage thresholds of 85% respectively.

RM system analysis

The Restriction Modification (RM) gene profiling of the genomes was performed by BLAST (Camacho et al., 2009) analysis using REBASE Gold Standard Database (Roberts et al., 2015). The REBASE database was clustered by using U-Clust (Edgar, 2010) with an identity threshold defined at 90%. The database which was thus curated was used for the detection of various RM systems in *E. coli* genomes. BLASTn was performed using the curated database against the genomes with identity and query coverage thresholds of 85% to determine the pattern of distribution of RM systems among *pks*-positive and negative genomes. The BLAST analysis of the curated RM systems database was performed against the three datasets; namely, ST95 genomes (n=159), *pks*-positive genomes (n=530), and *pks*-negative genomes (n=3583). In case of conditions where the genomes were observed to carry the modification and recognition subunits; the sequences of their cognate restriction enzymes obtained from REBASE were also separately analyzed if they were not already included in the gold standard database.

***pks* island phylogeny**

The *pks*-positive genomes were subjected to BLAST analysis (Camacho et al., 2009) against the *pks* island sequence and the genomes which demonstrated identity and query coverage values greater than 95% and 85%, respectively for *pks* island were used to obtain genomes harboring *pks* island within a single contig. The core genome phylogenetic tree of the genomes selected by this way (n=247) was constructed using methods as described in the previous sections. The *pks* island sequences were

extracted from the genomes using the locus information of BLAST outputs using the EMBOSS extract-align program (http://emboss.sourceforge.net/apps/cvs/emboss/apps/extractalign.html) and the in-house scripts were employed to handle reverse complements. The island sequences were aligned using PRANK (Löytynoja, 2014) and refined using trimAL (with --strict flag) (Capella-Gutiérrez et al., 2009). Bayesian clustering of the sequence alignment was performed using Bayesian Analysis of Population Structure (BAPS) (Corander et al., 2005). IQ-TREE (Nguyen et al., 2015) with 1000 bootstrap replicates was used for the construction of maximum likelihood phylogeny after determining the optimal nucleotide substitution model using Model Finder (Kalyaanamoorthy et al., 2017). The resultant core genome phylogenetic tree was visualized using iTOL (Letunic and Bork, 2019) along with annotations for sequence type (ST) and serotype information. The core genome and *pks* island sequence phylogenetic trees were also compared using the "connect taxa" functionality of Dendroscope (v3.7.3)(Huson and Scornavacca, 2012).

Chapter 3

RESULTS

Results of Objective 1

Molecular epidemiology and characterization of *pks* island harboring *E. coli* from the Indian population

Screening for *pks* island and phylogenetic grouping

The ExPEC (n=462) isolates were screened for the presence of *pks* island, out of which 35 isolates were observed to harbor all the four targeted genes (*clbA, clbN, clbB,* and *clbQ)* which belonging to the flanking and internal regions. 30 of these *pks* positive isolates were originally cultured from urine, 4 were cultured from pus and one was obtained from blood (Table 3.1). The prevalence of *pks* positive isolates was determined to be 7.6% of the total isolates studied (Table 3.1). Identification of *E. coli* phylogenetic groups was performed using multiplex PCR and it was observed that 34 out of the 35 isolates belonged to phylogroup B2, while one isolate was assigned to phylogroup D.

Table 3.1: Clinical details of *pks* positive isolates

Strain	Gender	Age	Disease status	Source
NA035	Male	65	Septicaemia	Urine
NA100	Female	45	Septicaemia	Urine
NA147	Female	34	UTI	Urine
NA150	Female	19	Pyelonephritis	Urine
NA159	Female	22	Cystitis	Urine
NA172	Male	60	Prostitis	Urine
NA247	Male	20	Cystitis	Urine
NA258	Male	55	UTI	Urine
NA266	Female	20	UTI	Urine
NA280	Male	35	Septicaemia	Urine
NA281	Female	20	Cystitis	Urine
NA310	Female	26	UTI	Urine
NA313	Female	17	UTI	Urine
NA334	Female	26	UTI	Urine
NA336	Female	30	UTI	Urine
NA608	Male	2	Chronic UTI	Urine
NA611	Male	2	Chronic UTI and severe abdominal Pain	Blood
NA623	Male	75	UTI	Urine
NA626	Female	8	Renal calculi	Urine
NA651	Female	55	Surgical site infection	Pus
NA664	Male	36	UTI	Urine
NA666	Female	28	Unknown	Urine

Strain	Gender	Age	Disease status	Source
NA675	Male	60	UTI	Urine
NA690	Female	32	Unknown	Urine
NA695	Female	30	Unknown	Urine
NA697	Female	23	Unknown	Urine
NA698	Female	39	Unknown	Urine
NA706	Male	50	Unknown	Pus
NA714	Male	11	UTI	Urine
NA731	Female	35	UTI	Urine
NA733	Female	22	Renal calculi	Urine
NA744	Female	35	Haematuria	Urine
NA749	Female	20	Unknown	Urine
NA786	Female	40	Omphitis	Pus
NA792	Male	60	Diabetic Foot	Pus

Antimicrobial Susceptibility Testing and Determination of ESBL Production

AST performed against ten different antimicrobial drugs belonging to eight different antibiotic classes demonstrated that the *pks* positive isolates (n=35) were only moderately resistant to the antibiotics. The results indicated that the maximum resistance was observed against clarithromycin (100%), followed by nalidixic acid (71.4%). The *pks* positive isolates were observed to demonstrate less resistance to tetracycline (22.86%), co-trimoxazole (14.29%), doxycycline (11.43%), gentamicin (5.71%), and ciprofloxacin (5.71%). All the *pks* positive isolates were sensitive to fosfomycin, chloramphenicol, and colistin. Out of the 35 *pks* positive isolates 4 (11.42%) were observed to be multidrug-resistant and 13 (37.14%) were found to be ESBL producers. The results of AST and the resistance profile of the isolates against the antibiotics tested are described in Table 3.2.

Virulence and Resistance Genotyping

The *pks*-positive isolates were subjected to PCR based genotyping for various virulence and resistance markers. It was observed that the *pks* positive isolates were possessed a higher number of virulence genes compared to the resistance genes as detected by PCR. Among the bacterial adhesin genes which

were screened, *sfaD/E* and *fimH* were present in all the isolates (100%), while *afa* was absent in all the isolates. PCR based the virulence genes that belonged to the category of toxins revealed the prevalence of the *usp* gene among all the *pks* positive isolates (100%), whereas *sat* and *cvaC* showed a percentage prevalence of 34.29% and 42.86%, respectively. The protectant gene *ibeA,* and the siderophore system gene *iucD*, were observed in 31.43% and 51.43% of the isolates, respectively (Table 3.2).

PCR based genotyping for antibiotic resistance determinants demonstrated that the ESBL gene $bla_{\text{CTX-M-15}}$ was present in 25.71% (n = 9) of the *pks* positive isolates. All these nine isolates were also ESBL producers as observed phenotypically by the double-disk synergy test as described in the previous section. Further, eight isolates (22.86%) showed the prevalence of the gene bla_{TEM}, and BLAST analysis of the PCR product sequences identified all of the amplicons to be $bla_{\text{TEM-1}}$ sequences. 20% of the *pks*-positive isolates carried the gene *aac(6')-Ib* which confer aminoglycoside resistance. Sulfonamide resistance gene *sul1* and tetracycline resistance gene *tetA* was observed to be prevalent in 11.43% and 5.71% of the isolates respectively (Table 3.2). Overall, a low prevalence of resistance determinants among the *pks* positive isolates was observed, which was in concordance with the phenotypic observations based on double disk diffusion assays.

Phenotypic determination of siderophore production

The Chrome Azurol S Agar was used for the determination of siderophores in the strains. A ternary complex formed between chrome azurol dye, CTAB and ferric ions will render a blue color to the media. The bacteria capable of accumulating the iron in ferric form from the media will grow with orange halos around its colonies. All *pks*-positive isolates were observed to be siderophore producers as indicated by the orange halos formed on the Chrome Azurol S plates (100%).

Table 3.2: Phylogroups, virulence and resistance genotypes, and antimicrobial resistance of *pks*-positive *E. coli* isolates

Genotypic Characterization		**No. (%) of positive isolates**
Phylogroup		
B2		34 (97.14)
D		1 (2.86)
Virulence factors: Genotypic determinant		
Adhesins	*fimH*	35 (100)
	sfaD/E	35 (100)
	afa	0 (0)
Toxins	*usp*	35 (100)
	sat	12 (34.29)
	cvaC	15 (42.86)
Protectins	*ibeA*	11 (31.43)
Iron acquisition	*iucD*	18 (51.43)
Resistance factors: Antibiotic class		
Tetracyclines	*tetA*	2 (5.71)
Fluoroquinolones	*aac6'*	7 (20)
Sulfonamides	*sul1*	4 (11.43)
ESBL	bla_{TEM}	8 (22.86)
	$bla_{\text{CTX-M-15}}$	9 (25.71)
Antimicrobial class or phenotype	**Specific Drug**	**No. (%) of resistant isolates**
Aminoglycoside	Gentamicin	2 (5.71)
Tetracyclines	Tetracycline	8 (22.86)
	Doxycycline	4 (11.43)
Sulfonamide/trimethoprim	Co-trimoxazole	5 (14.29)
Phenicol	Chloramphenicol	0 (0)
Phosphonic acid derivative	Fosfomycin	0 (0)
Fluoroquinolone	Ciprofloxacin	2 (5.71)
	Nalidixic Acid	25 (71.43)
Macrolide	Clarithromycin	35 (100)
Antibacterial peptide	Colistin	0 (0)
Multidrug Resistance		11 (31.43)
ESBL		13 (37.14)

Biofilm formation assay

All the 35 *pks* positive isolates were subjected to biofilm formation assay in order to document their capabilities of biofilm formation by determining their specific biofilm formation (SBF) values. Isolates that displayed the SBF values greater than 1 were categorized as strong biofilm formers. Moderate biofilm formers comprised of the isolates with SBF values between 0.5 and1.0 and weak biofilm formers comprised of isolates with SBF values less than 0.5 (Figure 3.1). The majority (21out of 35) of the *pks* positive were identified as strong biofilm formers, while ten isolates were moderate and four isolates were weak biofilm formers.

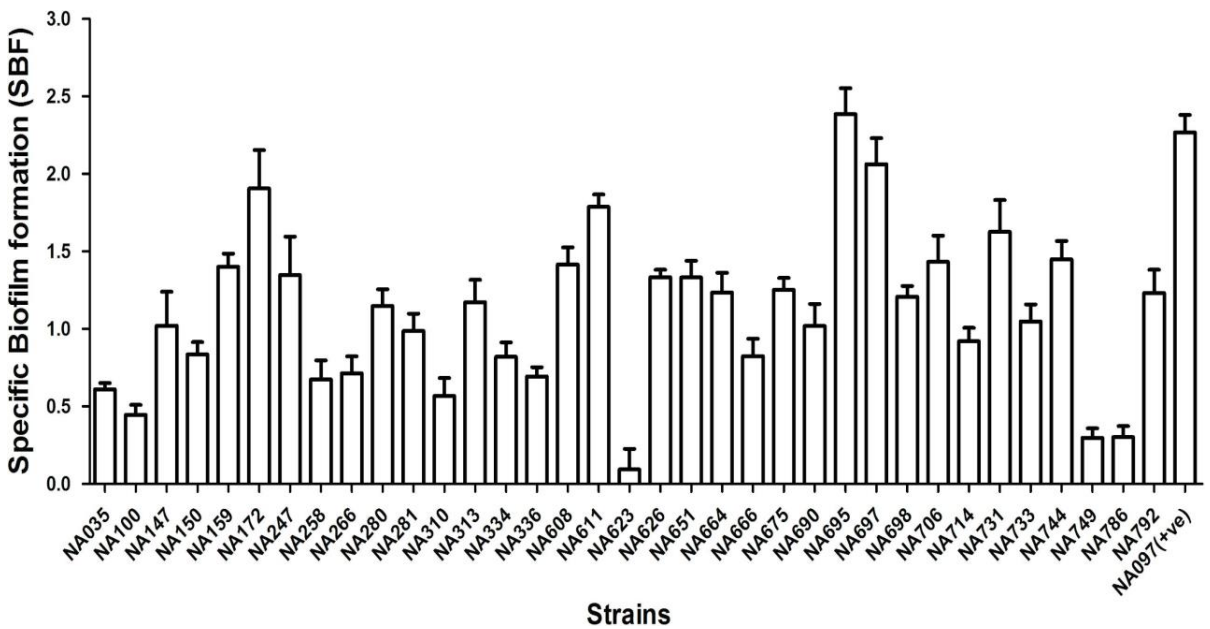


Figure 3.1: Results of biofilm formation assay performed among 35 *pks* positive *E. coli* isolates in M63 medium. The mean of the values of specific biofilm formation (SBF) are indicated in the graph. The majority of strains i.e., 21 out of the 35 strains were strong biofilm formers, and the remaining 10 and 4 strains showed moderate biofilm formation and weak biofilm formation, respectively. NA097 was employed as a positive control in the biofilm formation assay.

Serum Resistance Assay

Serum Resistance Assay was used to determine the resistance of the *pks* positive isolates to the bactericidal activity of human serum. All the isolates exhibited significantly higher value of CFU after 3-hour incubation as compared to the initial readings indicating that all *pks* positive isolates were the were resistant to the bactericidal activity of human serum. The assay was performed in technical triplicates twice (Figure 3.2). Mann–Whitney *U* test was used for the statistical analysis and the *p*-values were calculated.

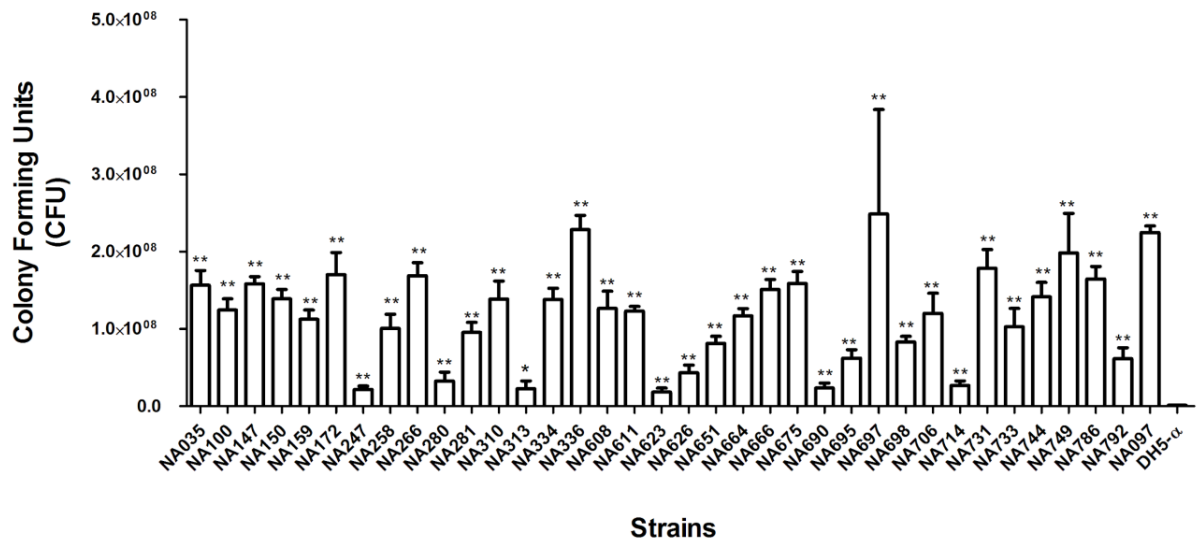


Figure 3.2: Results of the serum resistance assay of *pks*-positive strains in human serum. Mann-Whitney U test was employed for calculating the significant differences. Significant differences are indicated by asterisks and $p \leq 0.05$ was considered to be significant. * = p-value ≤ 0.05, ** = p-value ≤ 0.01. NA097 was taken as the positive control, while DH5-α served as the negative control.

Clonality of *pks*-positive *E. coli* isolates

ERIC PCR of 34 *pks*-positive strains reflected high clonality within these strains. At 85% similarity, eight small clusters (A-H) were observed that represented 19 of 34 isolates (Figure 3.3). As a majority of isolates were from urine, we observed them in seven out of the eight clades (B-F). However, one pus isolate clustered with urine isolate(s) in the remaining clade (clade A). The lone isolate from blood did not cluster with any of the observed clonal clades. Overall, we observed a high degree of clonality among *E. coli* strains of phylogroup B2 harbouring *pks* pathogenicity island irrespective of temporal and source origin.

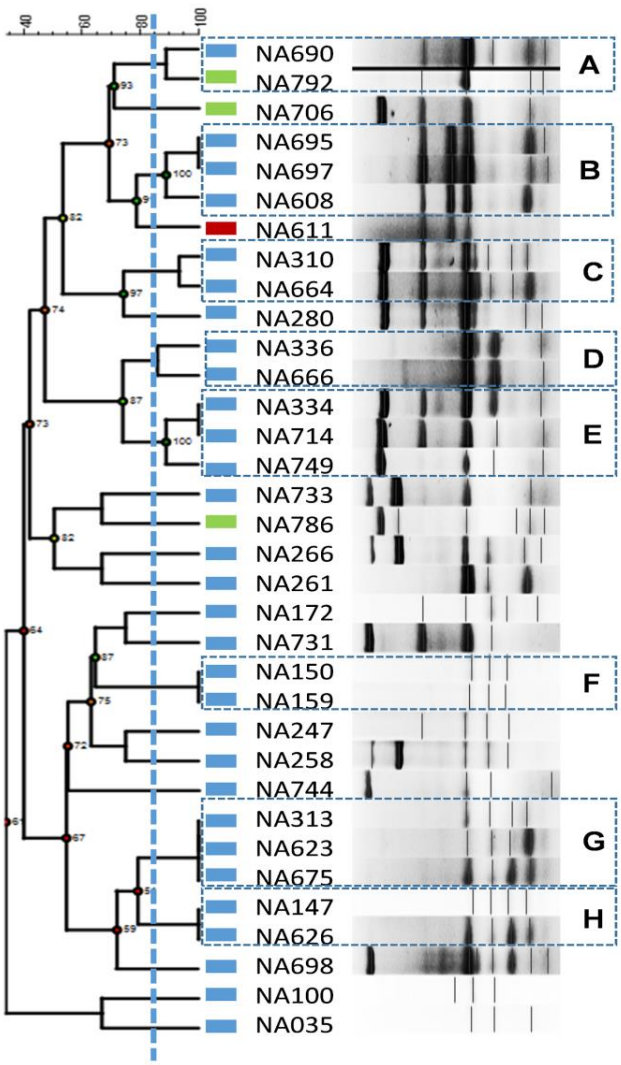


Figure 3.3: ERIC-PCR based dendrogram of *pks*-positive *E. coli* isolates generated using UPGMA algorithm based on dice similarity coefficient by Bionumerics® software. Close genetic relationships were observed among the *pks*-positive *E. coli* in the form of eight clonal clades (A-H) at the level of 85% similarity index. Blue boxes represent isolates from urine, green from pus and red from blood.

Results of Objective 2

Whole genome comparative analysis of *pks* island harboring *Escherichia coli* from the Indian population

Genome characteristics

Whole genome sequencing of 23 *pks*-positive *E. coli* isolates was performed for performing the whole genome comparative analysis. The genomes displayed an approximate size of 5.1 Mb with an average G+C content of 50.4%. The average number of coding sequence (CDS) was ~5000, and a coding percentage of 87% was displayed (Table 3.4). *In silico* sequence typing revealed that the 23 *pks*-positive genomes were showing distribution among the sequence types ST12 (n=6), ST73 (n=4), ST827 (n=3), ST14 (n=3), ST998 (n=3), ST1057 (n=2), ST83 (n=1) and ST127 (n=1) (Table 3.4). The assembly statistics and genome sequence characteristics are summarized in tables 3.3 and 3.4. The GenBank accession numbers of the 23 newly sequenced genomes have also been listed in table 3.4.

Whole genome comparison using BRIG

A whole genome comparison of the 23 in-house *pks*-positive genomes was performed using BRIG (Alikhan et al., 2011) with complete genome IHE3034 as the reference (Figure 3.4 (a)). Results from the analysis indicated that the genomes share a high degree of similarity and variable regions were mostly identified as phages (denoted as black arcs) (Figure 3.4(a)). *pks* island (labelled as a red arc) was also found to be conserved throughout the genomes (Figure 3.4(a)). The island sequences which were reconstructed from the respective genomes were used as the query and *pks* island from IHE3034 was used as the reference in BRIG (Alikhan et al., 2011) (Figure 3.4(b)). The *pks* island sequence was also annotated and positions of the individual genes of the island were depicted in the outermost ring (Figure 3.4(b)). It was observed that the island sequences showed a high degree of conservation among the genomes with variations only in the flanking regions in few cases.

Table 3.3: Assembly and scaffolding statistics of the *pks* positive genomes

Isolate	Avg. Genome coverage	No. of filtered contigs	Number of scaffolds	Total bp
NA147	89.077	104	93	5001103
NA150	59.886	100	97	5007721
NA258	44.567	179	174	5163691
NA266	76.36	131	108	5212596
NA280	63.49	150	137	5274649
NA310	58.63	171	161	5269727
NA334	51.42	162	150	5159446
NA336	61.98	159	137	5259602
NA608	183.89	78	51	5178873
NA611	47.284	82	71	5107459
NA623	71.59	178	168	5200798
NA651	76.34	63	53	5177187
NA664	63.17	183	169	5306725
NA666	65.95	132	118	5105662
NA675	49.8	98	88	5159592
NA695	69.16	86	74	5135536
NA698	62.03	176	162	5266727
NA706	62.46	149	136	5315252
NA733	70.07	117	106	5226616
NA744	79.51	112	95	5195215
NA749	80.54	144	128	5221527
NA786	78.78	157	140	5395297
NA792	82.78	78	61	4999821

Table 3.4: Genome characteristics of *pks* positive isolates

Isolate	No. of CDS	Avg. CDS length	Coding %	GC%	No. of rRNAs	No. of tRNAs	ST	Serotype	Accession Number
NA147	4825	908	87.6	50.49	9	73	ST12	O4:H1	JADBJB000000000
NA150	4810	911	87.5	50.5	9	73	ST12	O4:H1	JADBJA000000000
NA258	5042	891	87	50.59	8	69	ST827	O4:H1	JADNRJ000000000
NA266	5098	892	87.2	50.46	9	85	ST827	O4:H1	JADBIZ000000000
NA280	5164	886	86.8	50.49	14	89	ST14	O18:H5	JADBIY000000000
NA310	5194	883	87.1	50.53	9	78	ST1057	O75:H5	JADBIX000000000
NA334	4995	893	86.5	50.31	11	68	ST73	O6:H1	JADBIW000000000
NA336	5173	882	86.7	50.39	10	69	ST73	O6:H1	JADBIV000000000
NA608	5013	905	87.6	50.52	10	74	ST998	O2:H6	JADBIU000000000
NA611	4917	911	87.7	50.54	10	68	ST998	O2:H6	JADBIT000000000
NA623	5081	889	86.9	50.45	9	78	ST12	O4:H5	JADBIS000000000
NA651	5038	900	87.6	50.41	10	73	ST83	O6:H5	JADBIR000000000
NA664	5219	883	86.8	50.52	13	77	ST14	O18:H5	JADBIQ000000000
NA666	4956	894	86.8	50.44	12	74	ST73	O6:H1	JADBIP000000000
NA675	5035	896	87.4	50.52	13	78	ST12	O4:H1	JADBIO000000000
NA695	4934	910	87.5	50.52	12	74	ST998	O2:H6	JADBIN000000000
NA698	5168	886	87	50.5	9	76	ST12	O4:H1	JADBIM000000000
NA706	5223	886	87	50.5	10	86	ST14	O18:H5	JADBIL000000000
NA733	5106	893	87.3	50.5	9	75	ST827	O4:H1	JADBIK000000000
NA744	5069	893	87.1	50.4	10	73	ST12	O4:H1	JADBIJ000000000
NA749	5081	892	86.8	50.34	8	77	ST73	O6:H1	JADBII000000000
NA786	5347	881	87.3	50.5	11	81	ST1057	O75:H5	JADBIH000000000
NA792	4801	913	87.7	50.47	11	78	ST127	O6:H31	JADBIG00000000

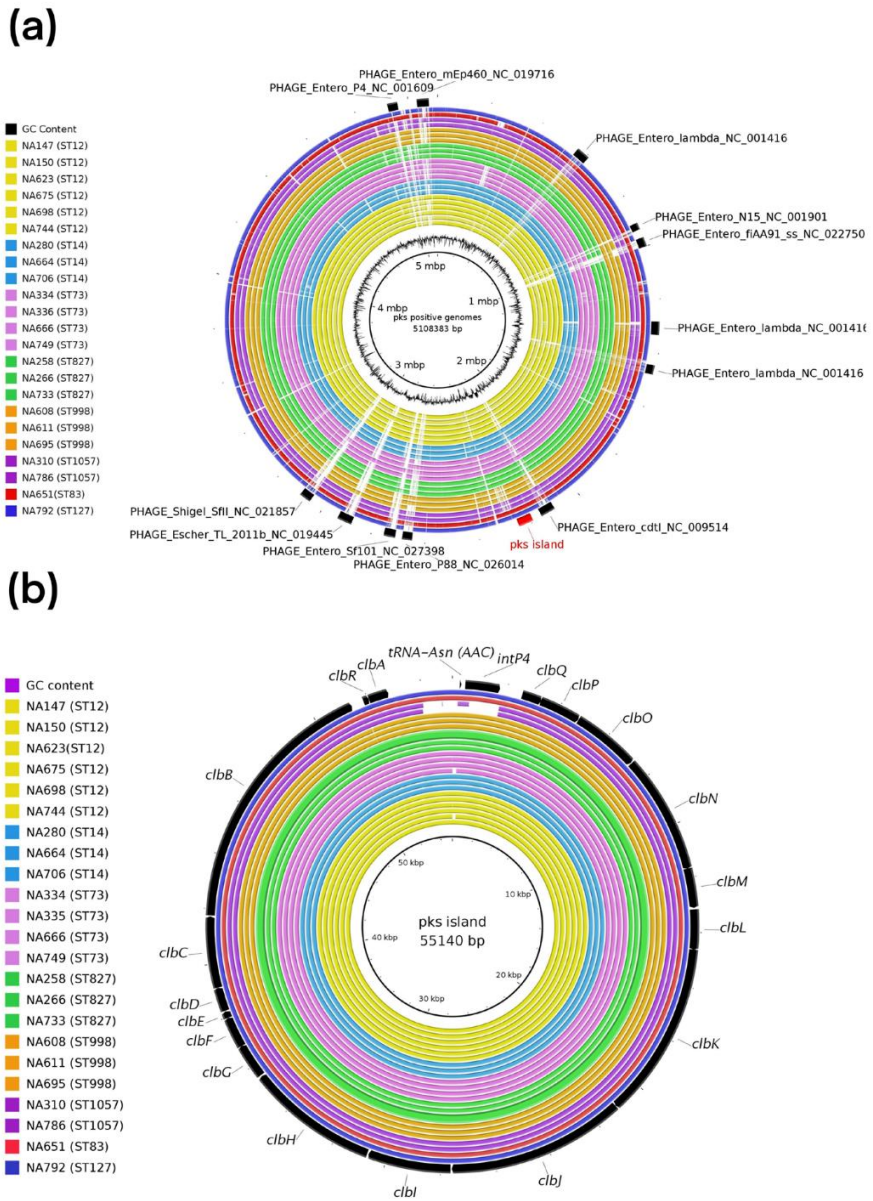


Figure 3.4(a): Whole genome comparative analysis of 23 *pks* positive genomes using BRIG, with IHE3034 as the reference. Each ring represents a genome and the rings have been color coded based on the sequence type and the genome names along with sequence types have been labelled. In addition, phages and *pks* island have also been annotated in the outermost ring. **Figure 3.4(b):** *pks* island from

23 genomes reconstructed from the aligned reads were used as query against *pks* island along with flanking regions from IHE3034 as reference. Each ring represents an island sequence from each *pks* positive genome and the rings have been colour-coded based on the sequence type of the genome. The genes of the island have also been annotated and represented in the outermost ring.

Virulome profiling of in-house *pks*-positive genomes

The *pks*-positive in-house genomes were screened for the prevalence of various virulence and antibiotic resistance gene coordinates to determine the pathogenic potential of these ExPEC isolates. *In-silico* virulence profiling using the Virulence factor database (VFDB) (Chen et al., 2005) demonstrated these *pks*-positive isolates to possess an abundance of adherence factors, siderophores and type VI secretion systems as depicted in the heat map (Figure 3.5). Among the category of adherence genes, the *Csg(A-G)* gene complex involved in the production of curli fibres and its assembly and transport, *E. coli* common pilus *ECP(A-E)* and type I fimbrial protein *fim (A-I)* were observed to be distributed in most of the genomes. Also, peritrichous flagellar protein *(flg, fli and flh),* chemotaxis protein *Che A/B/R/W/Y/Z* and flagellar motor protein *mot A* and *B* were found to be distributed in the majority of genomes. Invasin protein genes *ibe B/C* and *tia* were observed to be distributed in nearly 23 isolates. Among the secretion systems, genes encoding Type VI secretion systems (98 out of the total 111 genes belonging to the category of secretion systems) were observed to be most in abundance, followed by the genes encoding for general secretory pathway proteins *gsp (C-M),* nearly in all the genomes. Among the Type VI secretion systems, *aec (7, 16-19, 23-32), c3386, c3401, c3402* and *ECABU_c310170* were present in 18 or more genomes out of the 23 *pks* positive genomes. Yersiniabactin siderophore system genes *ybt (A, E, X, P-U)* and *irp (1/2)* were observed to be distributed in all the isolates. Most of the genomes harbored other siderophore systems like *chu (A, S-Y),* ferrienterobactin transporter *fep (A-E, G)*, enterobactin synthase *ent (A-F, S)*, enterobactin esterase *fes*, and salmochelin genes *iro (B-E, N).* It was also observed that 17 genomes out of the 23

pks positives harbored aerobactin siderophore synthesis system; *iuc (A-D)* and *iut A*. Among toxin genes, *hly A-D*, which encodes hemolysin, uropathogenic specific protein *usp*, and haemoglobin protease *vat* were present in nearly all the 23 genomes. In addition, cyclomodulin cytotoxic necrotizing factor *cnf-1* was present in 10/23 isolates. BLAST analysis using the VFDB database also confirmed the presence of *pks* island genes in all the 23 genomes indicating the integrity of the island within the genomes (Figure 3.5). Further, in a comparative analysis between *pks*-positive and *pks*-negative genomes; *pks* positive genomes were observed to possess a larger virulence repertoire than *pks* negative genomes (data is available in the following link: https://mbio.asm.org/content/mbio/12/1/e03634-20/DC4/embed/inline-supplementary-material-4.xlsx?download=true).

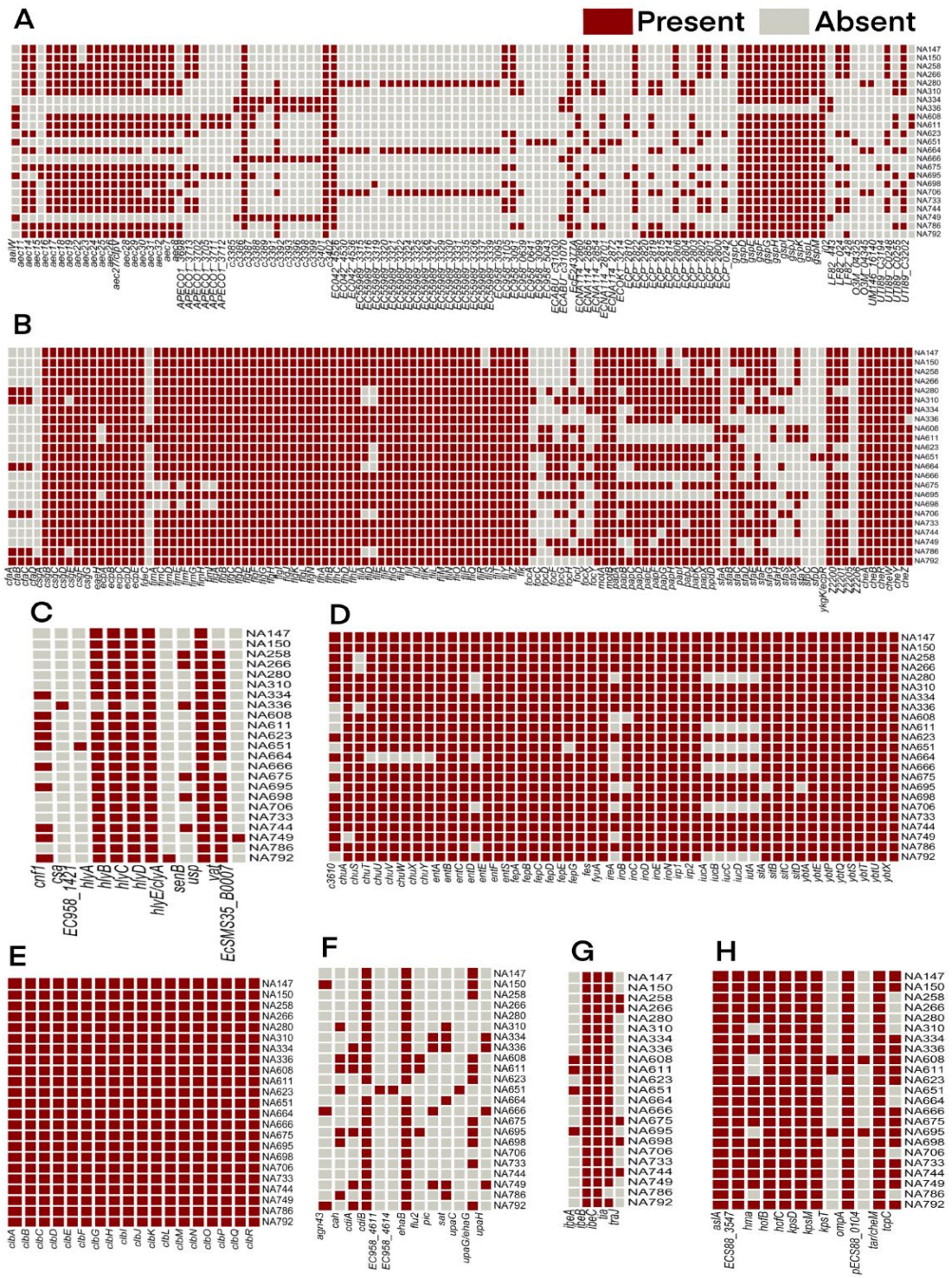


Figure 3.5: Heat map depicting the virulence profile of 23 in-house *pks*-positive isolates, depicting the presence and absence of 333 virulence genes belonging to different categories; A) Secretory

system, B) Adherence factors, C) Toxins, D) Siderophores/Iron acquisition systems, E) *pks* island genes, F) Autotransporters, G) Invasins and H) Others (genes corresponding to columns, from left to right, may be read in a sequential reading frame, such that each gene name aligns correctly with a single, corresponding column).

Resistome profiling of in-house *pks*-positive genomes

The whole genome i*n silico* antimicrobial gene profiling demonstrated that the majority of the resistance genes harbored by *pks*-positive in-house isolates belonged to the non-specific antibiotic efflux pumps category (Figure 3.6). The majority of the efflux pumps including aminoglycoside efflux pump (*acr),* global regulator (*CRP),*two-component regulatory system (*baeSR),* multiple antibiotic resistance family *mar,* and electrochemical gradient powered transporter *emr* was found to be prevalent in most of the genomes. Multidrug efflux system *mdt,* coupled with *gadX/W,* which offer resistance to penams, fluoroquinolones and macrolides were also observed in most *pks*-positive isolates. In the category of antibiotic inactivation, *ampC*, a class C beta-lactamase that encodes for resistance against penicillins and cephalosporins were observed in all the isolates. Other beta-lactamases like OXA-1 (n=3), CTX-M-15 (n=4), and TEM-1 (n=5) were detected in few isolates. Genes involved in antibiotic target replacement like the coordinates from gene family phosphoethanolamine transferase (*ugd/PmrE, eptA/PmrC and PmrF)* and bacitracin resistance gene *BacA* offering resistance against cationic antimicrobial peptides observed to be distributed in all the genomes (Figure 3.6). Further, in a comparative BLAST analysis of antibiotic resistance genes between *pks*-positive and *pks*-negative genomes; the *pks* positive genomes possessed smaller antibiotic resistance repertoire with a lesser number of specific antibiotic resistance determinants as compared to *pks* negative genomes (data is available in the following link: https://mbio.asm.org/content/mbio/12/1/e03634-20/DC5/embed/inline-supplementary-material-5.xlsx?download=true)

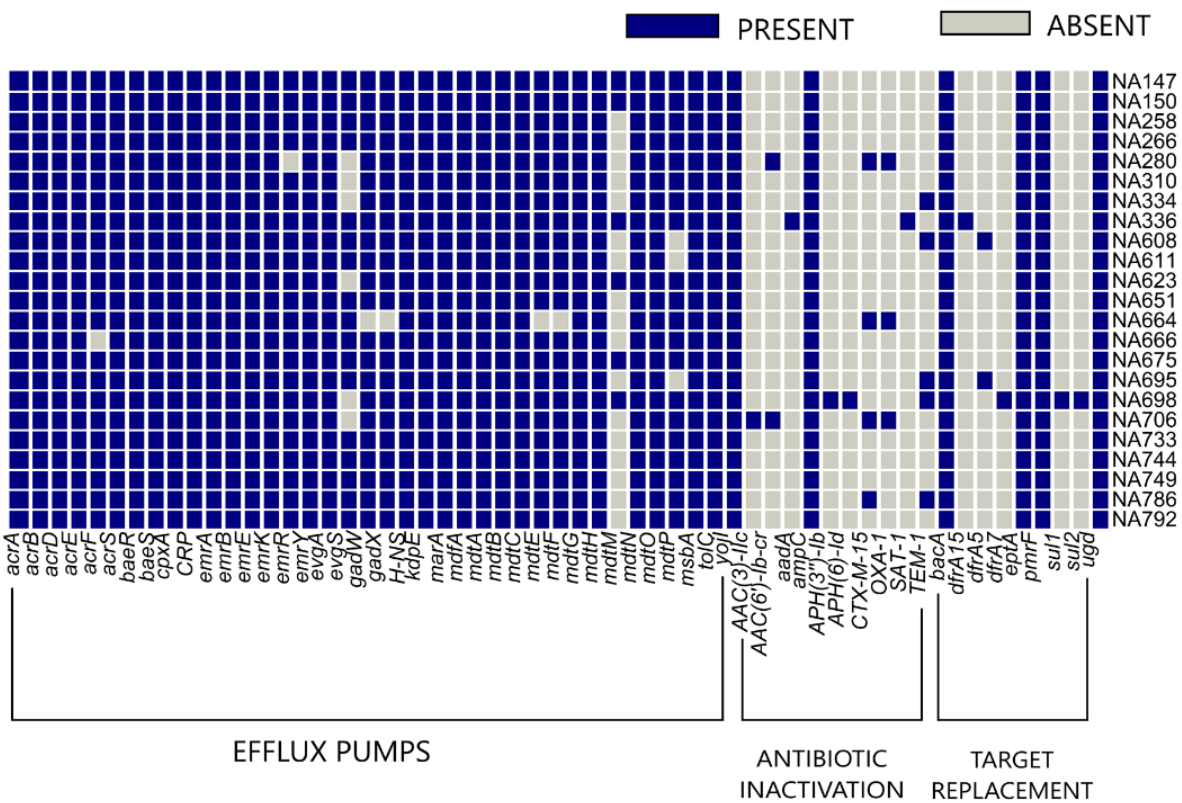


Figure 3.6: Heat map showing the presence and absence of 57 antibiotic resistance genes in 23 in-house *pks*-positive genomes. Blue and grey boxes indicate the presence and absence of the resistance gene, respectively. Gene names have been represented on X-axis and isolate names on the Y-axis. The genes have also been categorized according to the mechanism of action against antimicrobials (gene names correspond to columns from left to right; they may be read in a sequential reading frame beginning *acrA* aligning to the left-most column to *ugd* aligning with the right-most column, in the heat map).

Results of Objective 3

Deciphering the evolution and transmission of *pks* island harboring *Escherichia coli*

Prevalence and distribution of *pks*-positive *E. coli*

A total of 4113 genomes of *E. coli* were analyzed, out of which 306 were complete genomes and 3753 were draft genomes downloaded from NCBI, 31 were in-house genomes sequenced for our previous studies and 23 were the newly sequenced *pks* positive genomes for the present work. A total of 530 *pks* positive genomes were observed, including the 23 newly sequenced genomes. Among the 530 genomes, 247 genomes carried all the 19 genes of the island, while 184 carried 18 genes, and 80 genomes carried 17 genes of the *pks* island. The remaining 19 genomes carried less than 17 genes. Out of these 530 *pks*-positive genomes, in 247 genomes, the *pks* island (54kb in size) was observed to be present in a single contig. All the 530 *pks*-positive genomes carried 14 to 19 *pks* island genes and the rest of the genomes did not harbor any of the *pks* island genes and were hence designated as *pks*-negative.

In-silico MLST analysis revealed that there is a higher prevalence of the *pks* genomic island in sequence types ST73 (n=179) and ST95 (n=110), followed by ST127 (n=52) and ST12 (n=48) (Table 3.5). Interestingly, all the ST73 isolates were observed to carry *pks* island (Table 3.5) and none of the highly successful clonal group ST131 genomes harbored the island sequence. The percentage prevalence of the *pks*-positive genomes among sequence types ST95, ST127 and ST12 were 69.18% (110/159), 78.78% (52/66) and 97.9% (48/49), respectively. *In-silico* phylogrouping demonstrated that the majority of the isolates (82%) belonged to phylogroup B2; indicating a strong association of *pks* island harboring *E. coli* to the B2 phylogroup (Table 3.5). This observation was similar to the results from PCR based phylogrouping of the *pks* positive in-house isolates that showed the majority (97%) of the isolates to belong to the B2 phylogroup (Table 3.2)

In-silico identification of serotypes using EC Typer demonstrated a higher prevalence of *pks*-positives in certain serotypes. O6:H1 serotype was observed to have the highest number of *pks*-positive genomes (n=110), followed by O6:H31 (n=48), O4:H5 (n=46) and O18:H7 (n=40). The prevalence

pattern of *pks*-positive genomes in different serotypes have been described in Table 3.5. Sequence types and serotypes with fewer than 10 genomes were grouped under "miscellaneous".

Table 3.5: Sequence type, phylogroup and serotype distribution of *pks*-positive genomes (n=530) obtained from NCBI

Subtype	Percentage (Number)
Phylogroup	
B2	81.69% (433)
A	0.56% (3)
Unknown	17.73% (94)
Sequence Type	
ST73	33.8% (179)
ST95	20.7% (110)
ST127	9.8% (52)
ST12	9.1% (48)
ST141	3.6% (19)
ST998	3.02% (16)
ST404	2.07% (11)
ST80	1.7% (9)
Miscellaneous	12.6% (67)
Unknown	3.6% (19)
Serogroup	
O6:H1	20.7% (110)
O6:H31	9% (48)
O4:H5	8.6% (46)
O18:H7	7.5% (40)
O2:H6	6.8% (36)
O1:H7	6.4% (34)
O2:H1	6% (32)
O2:H7	6% (32)
O75:H5	4.3% (23)
O22:H1	3.7% (20)
O4:H1	3.7% (20)
O25:H1	3.2% (17)
O2:H4	3% (16)
O18:H1	2.2% (12)
Miscellaneous	7.5% (40)

Unidentified	0.75 (4)

Pangenome analysis of ST95 genomes

The sequence type ST95 was observed to possess both *pks*-positive (n=110) as well as *pks*-negative (n=49) genomes and was hence considered to be a suitable model dataset for this study. Comparison between the *pks*-positive and *pks*-negative genomes from ST95 was considered to help provide insights into the potential acquisition and maintenance of *pks*-island. A total of 3057 genes constituted the core of 159 ST95 genomes which included 110 *pks*-positives and 49 *pks*-negatives.

Table 3.6: Table showing the COG classification of core genes from 159 ST95 genomes

COG CLASSIFICATION	Number of genes
Cellular processes and signalling	
Cell membrane/wall/envelope biogenesis (M)	177
Cell division, chromosome partitioning, and cell cycle control (D)	32
Protein turnover, post-translational modification, and chaperones (O)	119
Motility of the cells (N)	22
Vesicular transport, secretion, and intracellular trafficking (U)	52
Mechanisms of signal transduction (T)	89
Cellular mechanisms of defense (V)	40
Information processing	
Transcription (K)	232
Processing and modification of RNA (A)	2
Replication, repair, and recombination (L)	102
Translation, ribosomal biogenesis and structure (J)	153
Metabolism	
Production and conversion of energy (C)	228
Transportation and metabolism of amino acids (E)	219
Transportation and metabolism of carbohydrates (G)	280
Transportation and metabolism of nucleotides (F)	103

Transportation and metabolism of inorganic ions (P)	216
Transportation and metabolism of lipids (I)	71
Transportation and metabolism of coenzymes (H)	124
Biosynthesis, transportation, and catabolism of secondary metabolites (Q)	24
Poorly characterized	
Unknown Function (S)	649
Unidentified	71
Multiple classes	52

Core genome phylogeny of ST95

Maximum likelihood core genome phylogeny obtained from IQ-TREE (Nguyen et al., 2015) demonstrated 5 different clades, where the green branches denote *pks*-positives and red branches denote *pks*-negatives (Figure 3.7). Clades I and II were observed to comprise both *pks*-positive and negative genomes with a mixed cladding pattern (Figure 3.7). Clades III and V were observed to predominantly carry *pks*-positive genomes, except for one *pks*-negative genome each, whereas clade IV consisted of only *pks*-negative genomes (Figure 3.7). The distinct clustering of *pks*-positive and negative isolates in a core genome-based phylogeny hints towards the role of the core genome in the acquisition and maintenance of *pks* island which forms a part of its accessory genome. All the major clades of the ST95 core genome phylogenetic tree (Figure 3.7) had bootstrap support values ranging from 89% to 100%. The core genome phylogeny of 159 ST95 genomes which was constructed along with an outgroup ED1a is depicted in Figure 3.8.

The ST95 isolates were grouped into six different clusters using hierBAPS (Cheng et al., 2013) based on the first level of clustering. The BAPS clusters were in concordance with the clades of the maximum likelihood phylogeny obtained from IQ-TREE(Nguyen et al., 2015) and the BAPS clusters were represented in the first data strip of Figure 3.7. The BAPS clusters 5 and 6 belonged to the *pks*-positive clades V and III, respectively; whereas BAPS clusters 2 and 4 belonged to the *pks*-negative

clade IV (Figure 3.7). Genomes constituting BAPS clusters 1 and 3 mostly belonged to clades I and II which showed intermixed cladding of both *pks*-positives and negative isolates (Figure 3.7).

In-silico serotyping using EC Typer classified the ST95 isolates into eight different serotypes. The branching pattern in the ML phylogeny was revealed to be mostly based on the serotype of *E. coli* and also showed association to the island's prevalence (Figure 3.7). The serotypes O18:H7 and O2:H7 comprised of mostly *pks*-positive isolates and the mixed clade belonged to O2:H4. Interestingly O1:H7 isolates were observed to form two separate clades and belonged to two sets of BAPS clusters, one of which was *pks*-positive and the other was *pks*-negative (Figure 3.7).

In-silico CH Typing (Weissman et al., 2012) demonstrated that all the ST95 genomes belonged to the same C-Type 38, and the variations were observed within the Type I fimbrial gene *fimH*. This sequence variation classified the 159 genomes into nine different CH Types (Figure 3.7). Clades I and II which comprises the *pks*-positive and negative mixed cluster were observed to belong to CH Type 38-27. *pks*-positive clade III (except 95N035) carried genomes belonging to CH-types 38-18, 38-15, 38-16 and 38-107. The *pks*-negative clade IV comprised of genomes belonging to CH-types 38-54 and 38-30. Clade V which was predominantly *pks*-positive belonged to CH Type 38-41, except the genomes 95N044, *pks*p129 and *pks*p176, which belonged to CH Types 38-54, 3827, and 38-43, respectively (Figure 3.7).

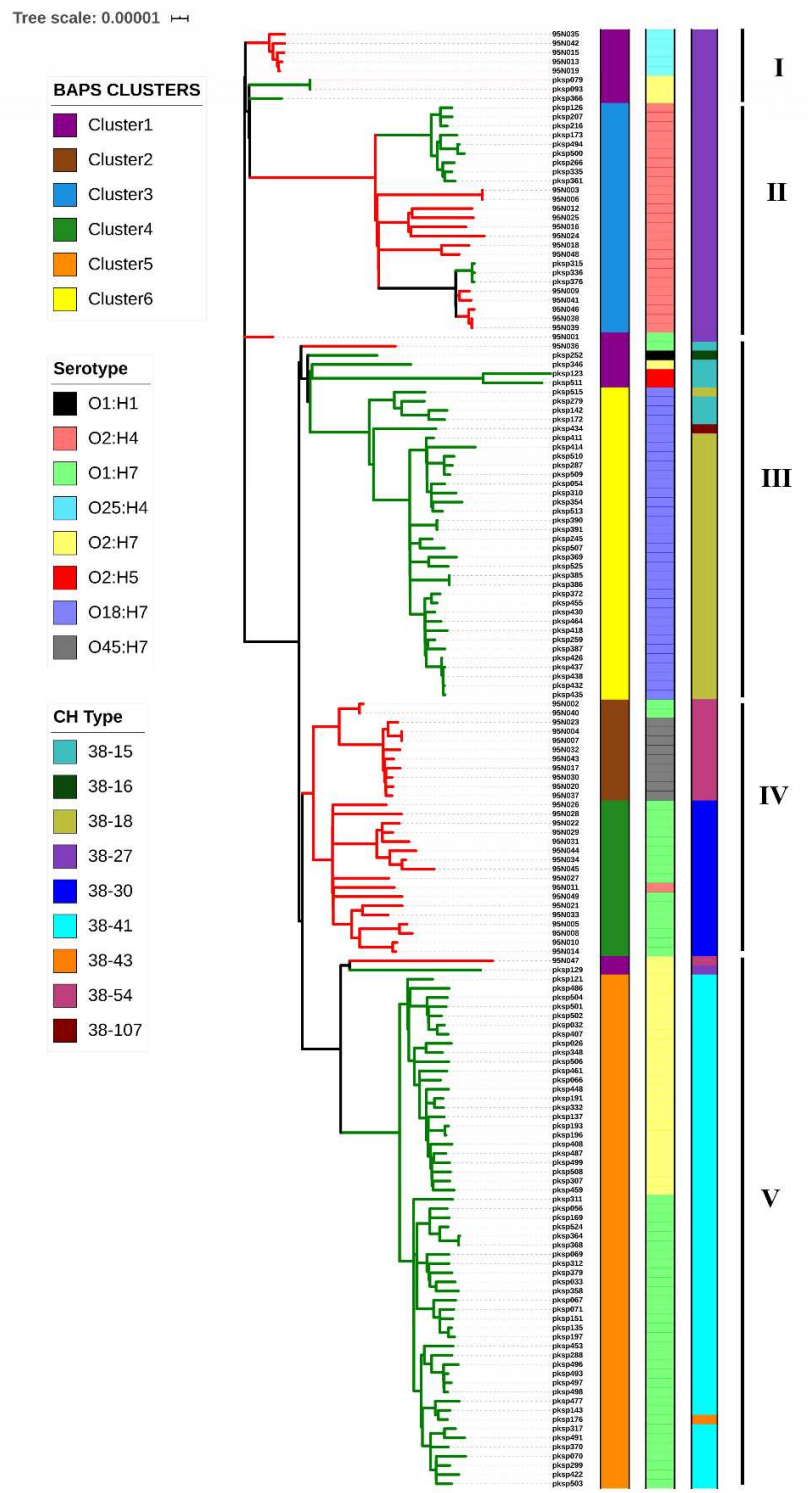


Figure 3.7: Maximum likelihood core genome phylogeny of 159 ST95 isolates constructed using IQ-TREE, and ClonalFrameML and visualized using iTOL. A total of five clades were observed (I-V).

Green and red branches represent genomes positive and negative for *pks* island, respectively. The first data strip represents the BAPS clusters, the second data strip represents the serotypes of the genomes as identified by ECTyper and the third represents CH types of the genomes, respectively.

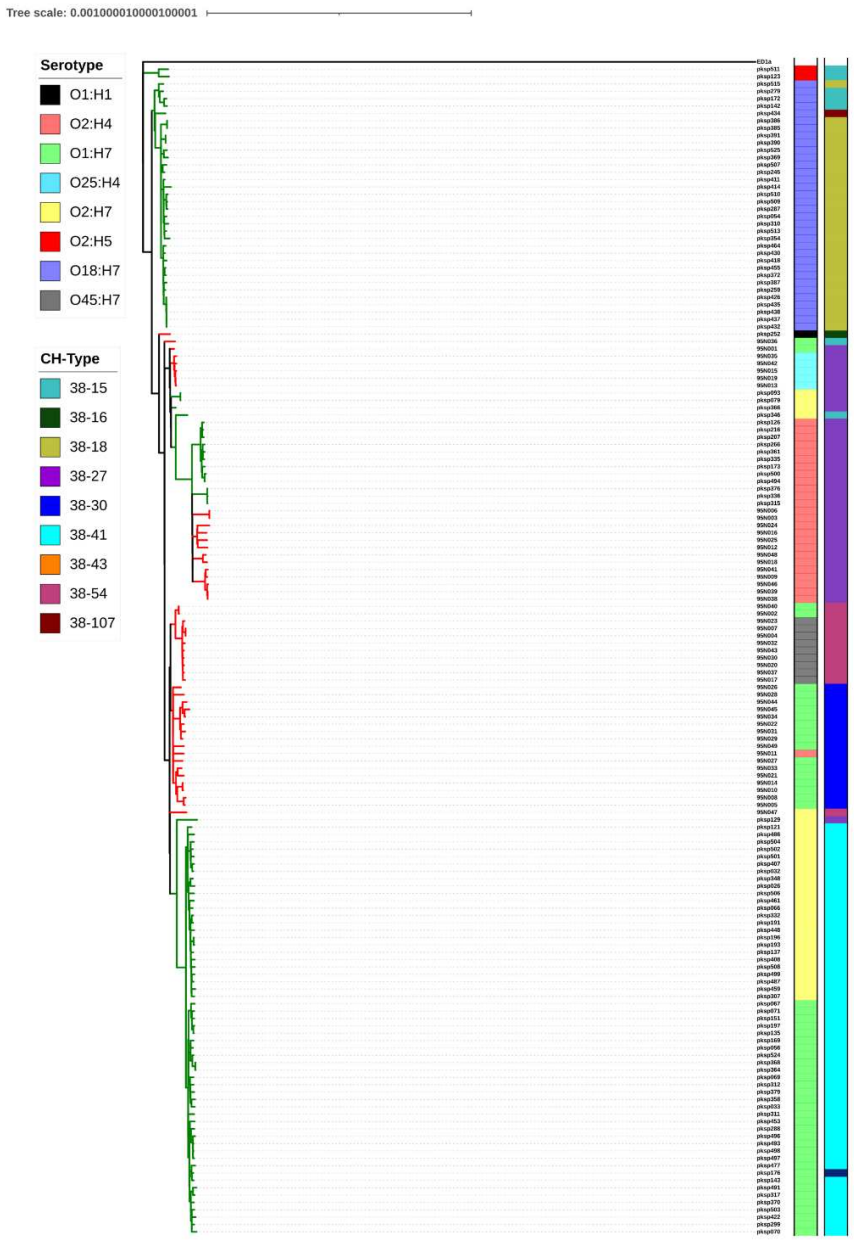


Figure 3.8: Maximum likelihood core genome phylogeny of 159 ST95 isolates, along with one outgroup genome ED1a constructed using IQ-TREE and ClonalFrameML, and visualized using

iTOL. Green and red branches represent genomes positive and negative for *pks* island, respectively. The first data strip represents the serotypes of the genomes as identified by ECTyper and the second represents CH types of the genomes, respectively.

Intergenic region (IGR) analysis of ST95 genomes

Intergenic regions (IGR), albeit comprising of non-coding DNA sequences, form an important part of the bacterial genome with abundantly distributed regulatory regions which play a crucial role in the phenotypic variations in the bacteria (Oren et al., 2014). The analysis of core IGR regions, in addition to the coding counterpart of the genome provides an improved resolution to the evolutionary analysis of bacteria. The analysis of core IGR phylogeny was performed to ascertain the correlation between the sequence variation of the intergenic region to *pks* island distribution pattern. The core IGR phylogeny was constructed using core IGR sequences extracted by PIGGY (Thorpe et al., 2018)(Figure 3.9) and showed cladding pattern reflective of the carriage of *pks* island more distinctly compared to the core genome phylogeny with the *pks*-negative cluster found cladding separately from *pks*-positive and mixed clades. The IGR clades from O1:H7 *pks*-positive and *pks*-negative genomes were also observed to be more distinct compared to core genome phylogeny. All the major clades of the ST95 IGR phylogenetic tree (Figure 3.9) had bootstrap values ranging from 92.7% to 100%.

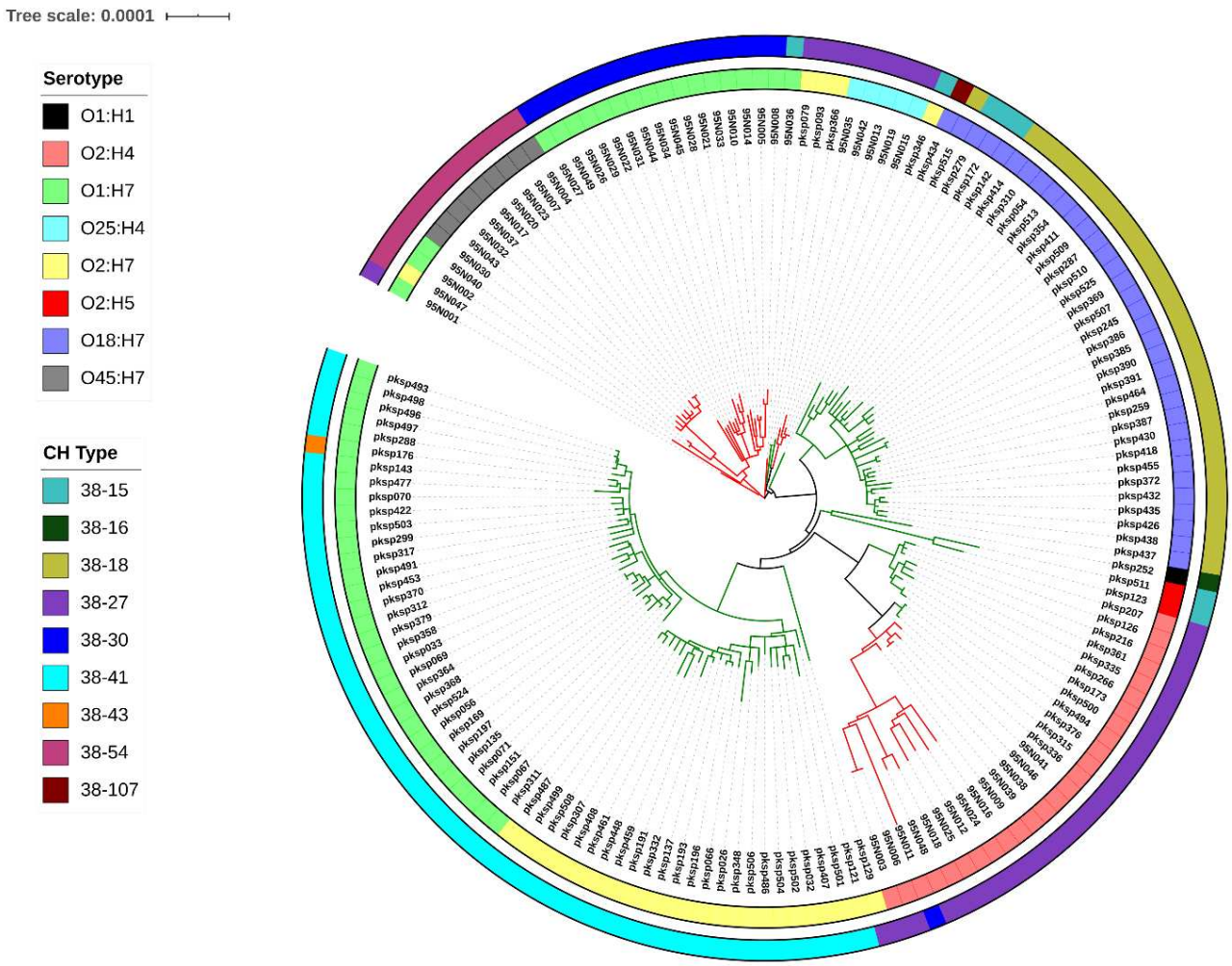


Figure 3.9: Maximum likelihood intergenic region phylogeny of 159 ST95 isolates constructed using IQ-TREE and visualized using iTOL. Green and red branches represent *pks* island positive and negative, respectively. The inner ring represents the serotype of the genomes as identified by ECTyper and the outer ring represents the CH type of the genomes obtained from CH Typer.

Pan-genome wide analysis using Scoary for ST95 genomes

Pan-genome wide analysis of accessory genes was performed using Scoary (Brynildsrud et al., 2016) to identify genes that could have a potential correlation to the prevalence of *pks* island within the genome (Table 3.7). Genomes that belonged to Clade IV (comprising exclusively of *pks* negative

genomes) of the core genome phylogenetic tree, also belonging to the *pks* negative cluster of the core IGR phylogenetic tree, were compared to the rest of the genomes which belonged to *pks*-positive and mixed clades. The genes which displayed differential prevalence and enrichment in the two sets of genomes; ie., which were completely absent in Clade IV *pks*-negatives and present in almost all the other genomes and vice versa are documented in Table 3.7 along with their prevalence details and functional annotations. Putative acetyltransferase *yjgM,* toxin-antitoxin biofilm protein *tabA_2,* and ornithine carbamoyltransferase chain I *argI_1* were observed to be present as two different orthologs due to their sequence variation in each of these two groups of genomes analyzed. The prevalence of these genes across the *pks*-positive (n=530) and *pks*-negative (n=3583) genomes were also evaluated and the results are displayed in Table 3.7.

Table 3.7: Results of pan-genome wide analysis of ST95 genomes using Scoary between *pks*-positive and mixed clades (Clades I, II, III and V) in comparison with exclusively *pks* negative clade (Clade IV). The prevalence of the differentially enriched genes in the entire dataset of *pks*-positive (n=530) and *pks*-negative genomes (n=3583) has also been shown.

Sl. No.	Gene	Non-unique gene name	Annotation	ST95 Scoary results		Prevalence analysis among the entire dataset	
				Prevalence among *pks*-positives (n=110) and *pks*-negative genomes (n=20) from mixed clade	Prevalence among genomes from *pks*-negative clade(n=29)	Prevalence among *pks* positives (n=530)	Prevalence among *pks* negatives (n=3583)
1.	*ybcF_2*		putative carbamate kinase	130	0	476 (89.8%)	261 (7.28%)
2.	*argI_1*		ornithine carbamoyltransferase chain I	130	0	477 (90%)	269 (7.50%)
3.	*tabA_2*		toxin-antitoxin biofilm protein	130	0	477 (90%)	264 (7.36%)

4.	*yjgM*		putative acetyltransferase	130	0	522 (98.4%)	261 (7.28%)
5.	*arcA*		Arginine deiminase	130	0	477 (90%)	259 (7.2%)
6.	*arg*R_*2*		ArgR-arg	129	0	476 (89.9%)	264 (7.3%)
7.	*idnO*		5-keto-D-gluconate 5-reductase	128	0	440 (83%)	1128 (31.4%)
8.	*idnD*		L-idonate 5-dehydrogenase	128	0	440 (83%)	1129 (31.5%)
9.	*idn*R		IdnR transcriptional regulator	128	0	440 (83%)	1126 (31.4%)
10.	*idnK*		D-gluconate kinase, thermosensitive	128	0	440 (83%)	1131 (31.56%)
11.	*idnT*		L-idonate / 5-ketogluconate / gluconate transporter IdnT	128	0	440 (83%)	1133 (31.62%)
12.	group_8070		hypothetical protein	128	0	482 (90.9%)	358 (9.9%)
13.	*yfcC_2*		putative inner membrane protein; putative S-transferase	127	0	477 (90%)	263 (7.34%)
14.	group_212	*tabA_2*	toxin-antitoxin biofilm protein	0	29	39 (7.3%)	3298 (92.04%)
15.	*bdcA*		c-di-GMP binding protein involved in biofilm dispersal	0	29	39 (7.3%)	3244 (90.53%)
16.	group_1863		hypothetical protein	0	28	41 (7.73%)	3153 (87.99%)
17.	*bdc*R		putative transcriptional regulator	0	28	39 (7.3%)	3271 (91.29%)
18.	group_3605	*yjgM*	putative acetyltransferase	0	28	50 (9.4%)	3563 (99.44%)
19.	group_7174	*hsd*R	Type-1 restriction enzyme R protein	0	28	122 (23%)	484 (13.5%)
20.	group_803		hypothetical protein	0	28	146 (27.5%)	989 (27.6%)
21.	group_2253	*mdtM*	Multidrug efflux transporter	0	28	135 (25%)	3201 (89.3%)
22.	group_2075	*hsdM*	host modification; DNA methylase M	0	27	122 (23%)	484 (13.5%)

RM system analysis

The REBASE (Roberts et al., 2015) gold standard database consisted of 3211 genes, which were subjected to clustering by using U-Clust (Edgar, 2010), and the dataset which was hence curated consisted of 2171 genes which were used as the database for RM system analysis of the genomes. The prevalence pattern of RM systems showed correlation to the maximum likelihood phylogenetic clades and the serotype distribution in the case of the ST95 core genome phylogenetic tree (Figure 3.10). The RM systems of particular interest which were observed included M.Eco9001I, S.Eco9281I, S.Eco9001I.

The genomes which belonged to the *pks*-negative exclusive clade harbored M.Eco9001I (except 95N045 which was observed to carry the truncated gene). They also carried either one of S.Eco9281I or S.Eco9001I and Eco9001IP when analyzed separately, as the gene encoding the main restriction enzyme subunit was not included in the REBASE (Roberts et al., 2015) gold standard database. O2:H4 and O25:H4 which also comprised of *pks*-negative genomes did not carry the above-mentioned genes, and O1:H7 *pks*-positives and three O1:H7 *pks*-negatives (95N039, 95N001 and 95N036) which clustered differently from the main O1:H7 *pks*-negative clade, were also observed not to carry these genes (Figure 3.10).

RM system distribution patterns of the 530 *pks*-positives and 3583 *pks*-negatives were also analyzed to decipher their prevalence in these genomes and the ones which showed specific prevalence patterns ie., Type-I RM systems Eco9001I/9281I and EcoCFTI, and Type-III RM system Eco.CFTII are described in Table 3.8. The modification and recognition genes of these RM systems were part of the REBASE (Roberts et al., 2015) gold standard database, while their cognate restriction subunit gene sequences (Eco9001IP/9281IP, Eco.CFTIP and Eco.CFTIIP) were separately analyzed. While analyzing the sequence types and serotypes of the genomes carrying these RM systems, it was observed

that the genomes with the complete EcoCFTI system interestingly belonged to the ST73 complex, but showed no serogroup specificity.

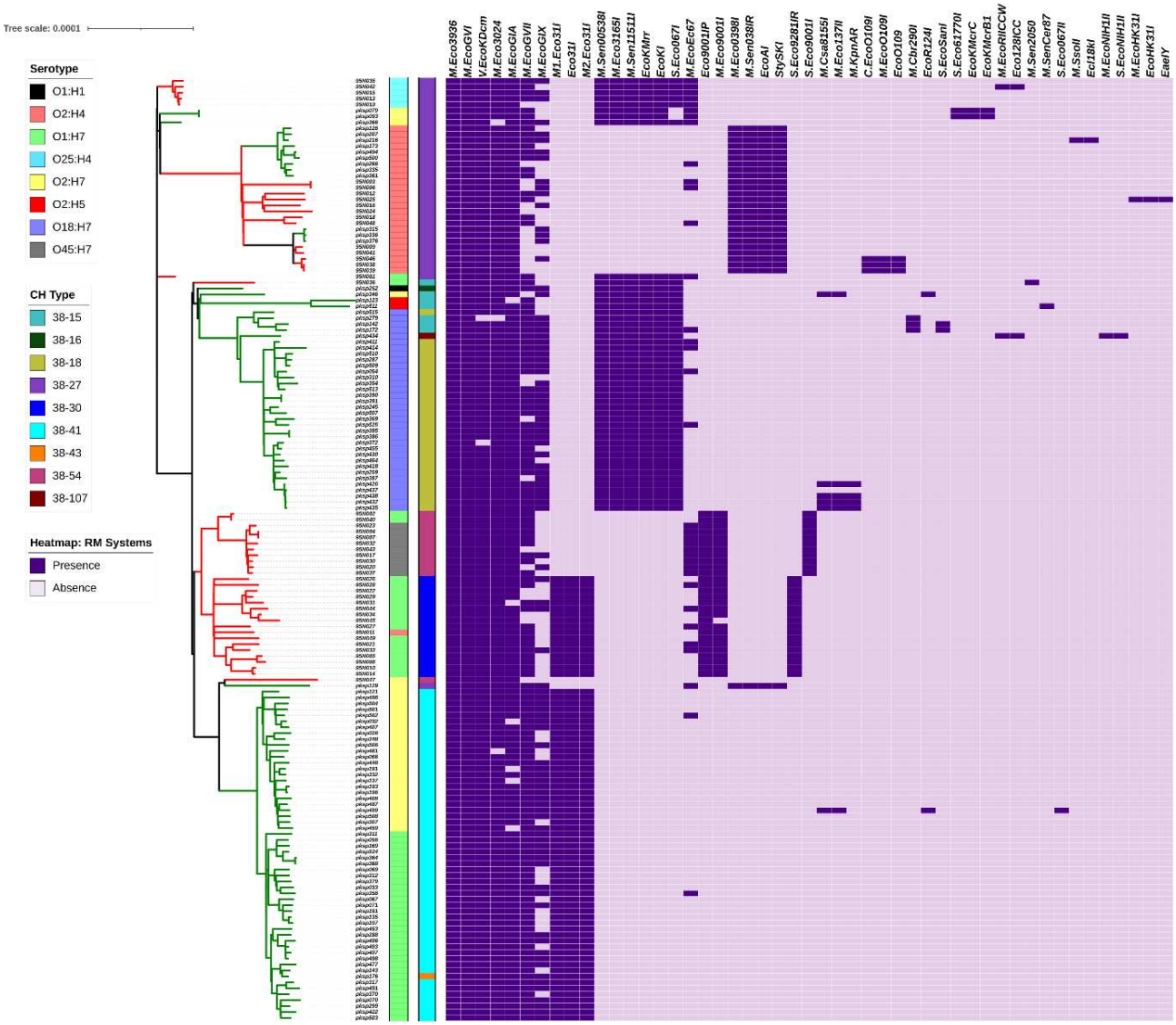


Figure 3.10: RM system prevalence pattern of ST95 genomes plotted as heat map along with core genome phylogeny with serotype and CH-type labelled. The prevalence pattern showed accordance with the cladding pattern of the phylogeny as well as the serotype distribution of the genomes.

Table 3.8: Comparison of prevalence of selected RM systems among *pks*-positive and *pks*-negative genomes

RM System	Genes	Prevalence (percentage) among *pks*-positives (n=530)	Prevalence (percentage) among *pks*-negatives (n=3583)
Eco900I/928I (TYPE-I-RM)	*Eco9001P/9281IP*	124 (23.4%)	484 (13.5%)
	M.Eco9001I/9281I	124 (23.4%)	252 (7.03%)
	S.Eco9001I/9281I	43 (8.11%)	43 (1.2%)
Eco.CFTI (TYPE-I RM)	*Eco.CFTIP*	257 (48.5%)	761 (21.23%)
	M.EcoCFTI	257 (48.5%)	761 (21.23%)
	S.EcoCFTI	156 (29.43%)	7 (0.19%)
Eco.CFTII (TYPE-III RM)	*Eco.CFTIIP*	159 (30%)	4 (0.11%)
	M.EcoCFTII	157 (29.62%)	0 (0%)

pks **island phylogeny**

The core genome phylogeny (number of core genes = 2579) and *pks* island sequence phylogeny of 247 genomes which contained *pks* island in a single contig were compared to study the effect of the pattern of evolution of the island sequences with respect to the core genome and the subtypes (sequence type and serotype). The core genome phylogeny of the core genes from 247 genomes which was constructed using IQTree (Nguyen et al., 2015) showed a cladding pattern reflective of the sequence type and serotype with few exceptions (Figure 3.11A). This core genome phylogeny (Figure 3.11A) was compared with that of the phylogeny of *pks* island sequences derived from the 247 genomes (Figure 3.11B) using Dendroscope (Huson and Scornavacca, 2012) as shown in (Figure 3.12). Hierarchical BAPS clustering (Cheng et al., 2013) of the island alignment provided 3 clusters at the first level, which are depicted using different clade colours (black, purple and orange) in the phylogenetic tree (Figure 3.11B). The cladding pattern was in agreement with the obtained BAPS clusters. Cluster-1 (black clade) consisted of only 6 genomes, which formed a distinct clade compared to Cluster-2 and Cluster-3 which comprised the rest of the genomes analyzed (Figure 3.11B). The clustering pattern of the island sequences were not fully reflective of the sequence type of their respective genomes in contrast to the core genome phylogeny, ie., same sequence types were clustering in multiple clades, except for *pks* island of genomes from ST998 which was observed to cluster together. This lack of concordance with the core genome clustering pattern could indicate HGT being the possible mode of transfer of this genomic island. Islands from ST12, ST73 and ST95 genomes were found to display intermixed pattern, indicating the possibility of horizontal gene transfer of the island across these sequence types (Figure 3.11B). The bootstrap support values of the major clades of the core genome phylogeny of 247 *pks*-positive genomes (Figure 3.11A) ranged from 98.6% to 100% and that of *pks*-island phylogeny (Figure 3.11B) ranged from 84.6%-100%.

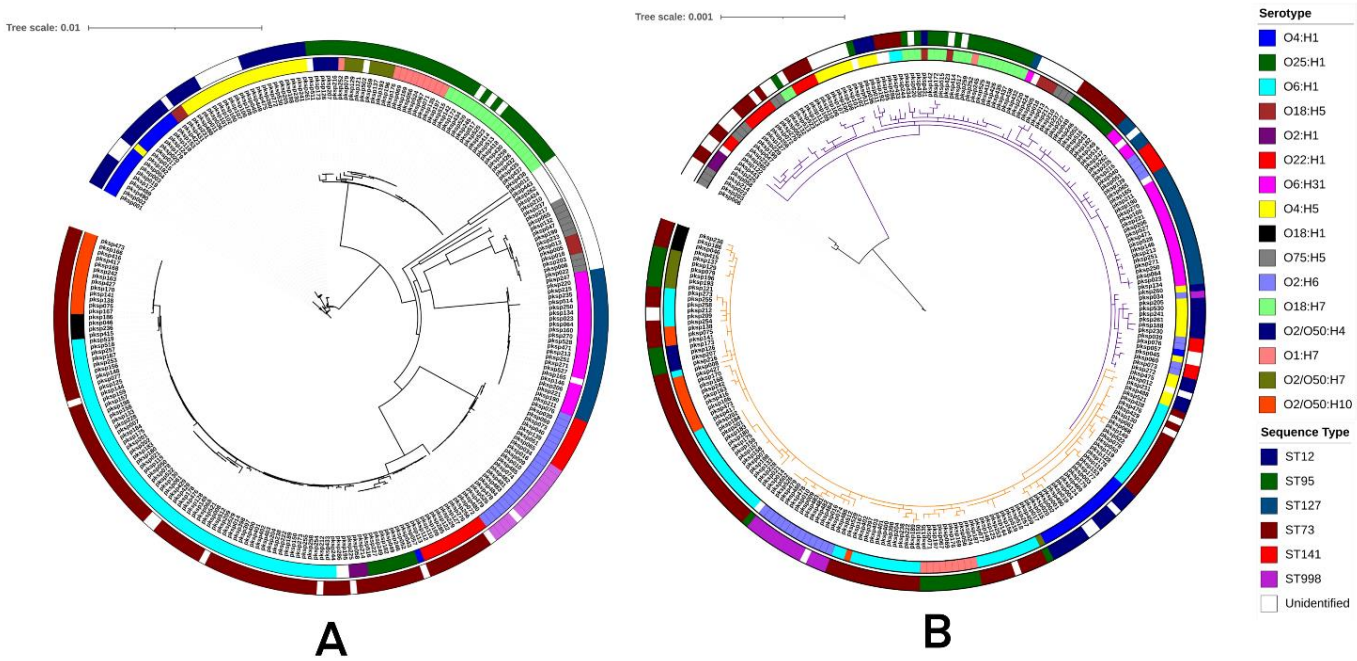


Figure 3.11: A) Core genome phylogeny of 247 *pks*-positive genomes constructed using IQTree and visualized using iTOL. B) Phylogeny of *pks* islands from 247 genomes constructed using IQ-TREE and visualized by iTOL. Clade colors depict the three BAPS clusters ie., Cluster-1 (Black), Cluster-2 (Purple) and Cluster-3 (Orange). In both A and B, the inner ring denotes serotypes and the outer ring denotes sequence types of the genomes from which the island was derived and legends for the same have been provided in the figure.

Figure 3.12: Comparison of the core genome phylogeny of 247 *pks*-positive genomes (first panel) along with their *pks* island sequence phylogeny (second panel) using the connect taxa functionality of Dendroscope.

Discussion

Colibactin, a bacterial genotoxin which is capable of inducing host DNA damage is synthesized as a secondary metabolite by the *pks* genomic island. The genotoxin could contribute to increased virulence and severe clinical outcomes in the host. The *pks* island found in certain members of Enterobacteriaceae is emerging as an important virulence marker in colorectal cancer progression, meningitis and septicemia (Faïs et al., 2018). Several studies have described the role of colibactin in CRC (Arthur et al., 2012; Dejea et al., 2018; Chagneau et al., 2019; Lopès et al., 2020), including the synergy between microbiota and host cells in CRC progression (Chagneau et al., 2019), making the genotoxin an important virulence factor which requires urgent attention owing to its clinical implications. The *pks* island shows wide distribution among neonatal *E. coli* K1 isolates and was observed to have a major role in the fully virulent phenotype of the bacteria in a neonatal systemic infection model (McCarthy et al., 2015). Colibactin was identified and described by Nougayrède *et al.* in 2006 for the first time, many studies have been undertaken to comprehensively elucidate this bacterial genotoxin (Bossuet-Greif et al., 2018; Faïs et al., 2018). However, information on the epidemiology and characterization of the *pks*-positive *E. coli* isolates from the Southern World is scarce. 35 out of the 462 clinical ExPEC isolates screened in the current study showed the prevalence of the four representative genes, thereby demonstrating the presence of complete *pks* island in their genomes which might have the potential to synthesize functional colibactin. Hence, the prevalence of *pks* island was recorded as 7.6% and this constitutes the first epidemiological information from India describing *pks* island harboring *E. coli*. The reports from other countries like from James R. Johnson *et al.*, 2008 and Shimpoh *et al.*, 2017 showed contrasting observations, wherein a high *pks* prevalence was demonstrated among the clinical *E. coli*.

Our previous studies indicated high genetic diversity among the majority of the clinical *E. coli* from India which did not harbour *pks* island (*pks* negative) and such isolates demonstrated high antibiotic resistance (AMR); specifically a clonally evolving pandemic sequence type 131 *E. coli* isolates were

identified and characterized from India (Jadhav et al., 2011; Ranjan et al., 2016b; Hussain et al., 2017; Shaik et al., 2017). Herein, the emergence of lineage-specific virulence of colibactin gene cluster harboring *E. coli*, which also demonstrated comparatively antibiotic resistance has been shown for the first time from India. Analysis of these strains would be of high clinical significance, as these genotoxic lineages could become a public health concern, also given the high burden of infectious diseases in India. Thus, the findings of this study have can contribute to a better understanding of these pathogenic *E. coli* in human diseases and their clinical significance.

The genotoxic *E. coli* which carry *pks* island was also observed to show strong association to bacteremia and colorectal tumors in humans (Johnson et al., 2008; Buc et al., 2013b). Significant lower rates of survival was observed in mice infected with colibactin positive *E. coli,* as compared with those infected with isogenic colibactin-negative mutants of *E. coli* (Marcq et al., 2014). *pks*-positive *E. coli* infection induces senescence in the host cells, and concurrently produces tumor growth promoting growth factors (Secher et al., 2013; Cougnoux et al., 2014; Dalmasso et al., 2015b). In our current study, *pks*-positive *E. coli* isolates were detected in various specimens like urine, pus and blood. Hence the *pks*-positive *E. coli* can potentially exhibit invasive and non-invasive infections at various anatomical sites.

Previous studies have reported high prevalence of *pks* positive isolates to B2 phylogroup, which is comprised largely of extraintestinal pathogenic *E. coli* (Taieb et al., 2016; Sarshar et al., 2017). The isolates characterized in the current study also belonged predominantly to phylogroup B2 (97%), showing concordance with the previous reports, except for one isolate which was observed to belong to D phylogroup (3%) (Table 3.2). The phylogroups B2 and D are reported to harbour pathogenic strains of *E. coli* with a large virulence gene repertoire, as compared to strains belonging to B1 and A phylogroups (Picard et al., 1999; Johnson et al., 2008).

A strong correlation between large repertoire of virulence genes and the pathogenic spectrum of *E. coli* strains has been demonstrated by numerous studies (Bien et al., 2012). ExPEC-associated virulence genes can be classified into multiple categories like secretory toxins, invasins, adhesins and iron aquisition systems (Johnson, 1991). In the virulence gene screening using PCR, 100% of the *pks* positive isolates cariied *fimH* (D mannose-specific adhesin of minor fimbrial component), *usp* (uropathogenic specific protein) and *sfaD/E* (s-fimbrial adhesin), while ≥ 30% of them harbored *sat* (secreted autotransporter vacuolating cytotoxin), *iucD* (aerobactin synthesis enzyme) *ibeA* (invasin), and *cvaC* (precursor of colicin-V). Gene *afa*, an afimbrial adhesin gene, was found to be completely absent among the *pks* positive isolates (Table 3.2). Previous studies have also indicated the association between *pks* island and iron acquistion systems among the B2 *E. coli* strains (Martin et al., 2017), and we also observed concordance with these observations as all the *pks* positive strains demonstrated siderophore production. Reports have indicated the localization of the *pks* island within High-Pathogenicity Island (HPI) and association with iron acquisition system gene clusters in other members of *Enterobacteriaceae* (Putze et al., 2009). We observed majority of the *pks*-positive *E. coli* isolates to be strong biofilm formers in the M63 medium (Figure 3.1) and all the *pks*-positive isolates to demonstrate resistance to bactericidal activity of human serum (Figure 3.2). We hypothesise that these phenotypic and genotypic virulence traits possessed by the *pks*-positive *E. coli* could render fitness traits to facilitate successful colonization in the host niches.

AST of the *pks*-positive isolates demonstrated that the *pks* positive isolates were uniquely exhibiting low antibiotic resistance. This finding is in concordance with the previous studies which have suggested that similar trends of low AMR in *pks* island harboring *E. coli* (Chen et al., 2017; Sarshar et al., 2017). Detailed characterization and analysis of of such isolates from different samples is needed to delineate the reason behind such an observation. ESBL production and multidrug resistance (MDR) were observed in only 37% and 11.42% of the *pks* positive isolates, respectively (Table 3.2). High

MDR rates of 95% for clinical ST131 strains, and 91% for clinical and stool non-ST131 strains were demonstrated in our previous studies (Hussain et al., 2014a). Metallo-(β-lactamase (MBL) producing *E. coli* isolates displayed 100% (Ranjan et al., 2016b) and *E. coli* isolates from skin and soft tissue infection (SSTI) displayed 67% (Ranjan et al., 2017) as the rates of multidrug resistance in our previous reports. We observed that *pks* positive isolates showed contrastingly lesser rates of AMR as compared to the *pks* negative isolates which were characterized in our previous studies from similar settings. A low prevalence of *tetA* (tetracycline resistance), *sul1* (sulphonamide resistance), *aac(6')-Ib* (aminoglycoside resistance), $bla_{\text{TEM-1}}$ (broad-spectrum-β-lactamases) genes and $bla_{\text{CTX-M-15}}$ (extended-spectrum-β-lactamases) genes were observed in the resistance genotyping, in concordance with our phenotypic antimicrobial resistance profiling (Table 3.2). We observed that all the $bla_{\text{CTXM-15}}$ positive *E. coli* isolates were observed to show ESBL production in phenotypic double-disk synergy tests(Table 3.2), although further screening of other CTX-M genes and groups, as well as ESBL classes are warranted.

We also performed high throughput phylogenomic comparison of *pks* island harboring *E. coli* genomes from the in-house culture collection and publicly available ones from NCBI were used to draw insights into the island's acquisition and evolution. Whole genome-based virulome and resistome analysis revealed that the in-house *pks*-positive genomes possessed a high number of genes contributing to virulence (Figure 3.5). Genes conferring antimicrobial resistance prevalent in the *pks*-positive genomes mostly constituted of efflux pumps and only a few specific antibiotic resistance determinants were observed (Figures 3.6). These findings were in line with the phenotypic observation of reduced antibiotic resistance and increased functional virulent characteristics displayed by the *pks*-positive isolates (Table 3.2, Figure 3.1, 3.2), compared to the frequently observed multidrug-resistant *pks*-negative ExPEC clones obtained from the Indian population (Hussain et al., 2012a, 2014c; Ranjan et al., 2015a, 2016a, 2017). Our previous genomic studies on the *pks*-negative ExPEC collection displayed

a higher prevalence of specific antibiotic resistance genes and a relatively lower prevalence of virulence genes (Ranjan et al., 2016a, 2017; Shaik et al., 2017) compared to our current analysis on *pks*-positive genomes. Notably, all the *pks* positive genomes harbored *BacA* gene, which is involved in resistance against the cyclic polypeptide antibiotic, bacitracin (El Ghachi et al., 2004), and genes *Pmr (E,C,F)* involved in the binding of cationic antimicrobial peptides like polymyxin (Olaitan et al., 2014). The large virulence gene repertoire in *pks*-positive isolates is consistent with the previous report based on PCR based observations on bacteremia isolates (Johnson et al., 2008), implying its clinical significance. Adhesins and type VI secretion systems showed abundance, and there was an increased prevalence of genes belonging to different siderophore production systems (Figure 3.5) in concordance with the phenotypic observations of siderophore production assay and other reports which indicate potential associations between *pks* island and iron acquisition systems (Martin et al., 2017). A previous study reported that the *pks* island encoded peptidase ClbP is involved in the genotoxin activation as well as renders antimicrobial activity either through microcins (Mcc) biosynthesis or secretion independently, or in cooperation with glucosyltransferase, thus reflecting the crucial co-selection of these islands in the evolution of pathogenic phylogroup B2(Massip et al., 2019). In a recent study, the microcin, salmochelin, and colibactin have also been indicated as a triad that could potentially provide a selective advantage for the bacterial colonization in the rectal reservoir with minimal genetic cost (Massip et al., 2020). The abundance of the siderophore systems like yersiniabactin, enterobactin, salmochelin and *chu A,S-Y* genes along with *pks* island could potentially play a role in the successful colonization and persistence of these isolates (Figure 3.5). The virulence factor profiling also showed an increased prevalence of haemolysin system (*hly)* in *pks*-positive isolates (Figure 3.5)*,* the association which was indicated as a risk factor for colorectal cancer (Yoshikawa et al., 2020).

The study was further expanded to screen 4090 genomes of *E. coli* obtained from NCBI, out of which *pks* island was detected in 507 genomes, in addition to 23 in-house genomes. The 530 positive

genomes were further subjected to various *in-silico* typing methods to identify distribution patterns among various *E. coli* subtypes (Table 3.5). All the ST73 *E. coli* isolates were observed to harbor *pks* island, in contrast to the most prevalent (Table 3.5) and highly successful ExPEC pandemic clone ST131 dataset which was notably completely *pks*-negative. It is also interesting to note that the prominent STs with *pks* island positive genomes, i.e., ST73 and ST95 have been previously reported to show low antibiotic resistance (Bengtsson et al., 2012; Roer et al., 2017; Denamur et al., 2020), which along with genotypic and phenotypic observations and CARD based genome analysis in the could indicate the association of *pks* island to isolates having a reduced antimicrobial resistance profile. *In silico* phylogrouping showed a strong association of *pks* island to B2 phylogroup, similar to the multiplex PCR based observations (Table 3.2) and in line with the previous studies (Johnson et al., 2008; Putze et al., 2009; Dubois et al., 2010; Kohoutova et al., 2014; Sarshar et al., 2017) as well.

ST95 is a successful ExPEC clonal complex that displays functional virulence properties of host adhesion, invasion, biofilm and serum resistance (Nandanwar et al., 2014) and clinical implications like UTI, newborn meningitis and a predominant avian and companion animal pathogen (Denamur et al., 2020). The ST95 dataset was used as a model for studying *pks* island as it was the only sequence type that carried a comparable number of *pks*-positives and negatives. A total of five clades were obtained (Figure 3.7) which were comparable to a previous study on the analysis of STc95 genomes which identified 5 subgroups within the STc95 complex (Gordon et al., 2017). Clades-I, II, III, IV and V of the core genome phylogeny in our study (Figure 3.7) showed correspondence to subgroup C, E, B, D and A, respectively based on the similar serotype and *fimH* type (Gordon et al., 2017) (Clade III additionally carried O2:H5, O1:H7, O1:H1 and O2:H7 genomes in small numbers). The prevalence pattern of the *pks* island sequence was in line with the previously observed prevalence of the *clbB* gene in the same study (Gordon et al., 2017). Differential cladding patterns observed in ST95 core genome phylogeny with separate positive, negative and mixed clusters indicate the potential role of the core

genome in horizontal gene transfer and integration of the island (Figure 3.7). Core intergenic regions phylogeny showed a cladding pattern more reflective of *pks* island carriage (Figure 3.9). A previous study has demonstrated that the patterns of polymorphism of the intergenic region *o454-nlpD* displayed concordance with the phylogenetic background as well as some important virulence-associated genes in *E. coli* (Ewers et al., 2014). Studies win ST131 *E. coli* have described core intergenic region substitutions to show association with the acquisition of accessory genome (McNally et al., 2016) and the analysis of ST95 genomes with respect to *pks* island shows a similar pattern. Although most of the clustering patterns of ST95 core genome phylogeny reflected the serotypes, O1:H7 showed a peculiar distribution into different clades containing *pks*-positive and negative genomes, and they also had a distinct *FimH* type (Figure 3.9). Scoary (Brynildsrud et al., 2016) was used for the pan-genome wide analysis of accessory genes, where the positive and mixed clusters were used as a combined dataset (Clades I, II, III and V) to compare with genomes belonging to the completely *pks*-negative clade (Clade IV) and the genes that showed differential enrichment among the groups were listed in Table 3.7. It was interesting to note that genes *idn O,D,R,K,T* belonging to the subsidiary system for L-Idonic acid catabolism which may provide a metabolic advantage for colonization (Bausch et al., 1998), were present in all genomes belonging to the *pks*-positive and mixed cluster while being completely absent in the members of the exclusively negative clade (Clade IV) (Table 3.7). The Type-1 restriction enzyme R protein *hsd*R and DNA methylase *hsdM* were observed to be present only among the genomes belonging to the *pks*-negative exclusive clade (Table 3.7).

As RM systems are shown to be involved in the regulation of HGT and recombination (Oliveira et al., 2014), their prevalence was studied among *pks*-positive and negative datasets as a preliminary analysis to determine their putative role in transfer or incompatibility of the acquisition of the *pks* island. A previous study has indicated the potential role of restriction-modification systems in the acquisition of resistance plasmids in ST95 O1:H7 isolates (Stephens et al., 2017). Since the *pks* island

showed clade-specific distribution patterns within the ST95 core genome phylogenetic tree, the tree topology was compared with its RM system prevalence data as a model to study the RM system diversity and finer distribution pattern (Figure 3.10). The analysis is limited to the RM systems in the curated Gold Standard database of REBASE and their selected cognate restriction enzyme subunit counterparts of the systems. When overlaid with the core genome phylogeny, the topology of the RM prevalence pattern showed relation to the sub-clades, reflective of their serotypes (Figure 3.10). This observation is similar to the results from a previous study describing the methyltransferase diversity among 95 ST131 *E. coli* isolates in which the RM system profiles were observed to show relation to their phylogenetic clusters (Forde et al., 2015). Another study in *Burkholderia pseudomallei* showed the clade-specific complement of the RM system, which potentially caused the clade-specific patterns in the DNA methylome (Nandi et al., 2015). The population structure of *Neisseria meningitidis* was also observed to coincide with its RM system distribution, suggesting the role of RM systems as a barrier in DNA exchange driving the formation of distinct phylogenetic lineages (Budroni et al., 2011). Similar sub-lineage correlations based on serovars and phylogenetic cladding of genomes with identical RM profile was observed in a previous study involving *Salmonella enterica* (Roer et al., 2016). Based on this evidence and our observations we hypothesize that the RM system profiles of the isolates might have a potential role in shaping the phylogenetic lineages, guiding the DNA exchange and thus playing a role in the horizontal acquisition of the genomic island. Notably in the analysis of RM system profile in the entire *pks*-positive and *pks*-negative dataset, the type III RM system EcoCFTII showed higher prevalence in *pks*-positive genomes compared to the *pks*-negative genomes (Table 3.8). However, the limitation of a small number of curated candidates available for RM system analysis is to be noted and careful interpretation is mandated. Based on these preliminary observations from prevalence analysis of RM systems among *pks*-positive and *pks*-negative genomes; the question on their probable role in

acquisition and maintenance of the mobile genetic elements will be interesting to pursue in future studies.

The phylogeny of *pks* island, in contrast to the core genome phylogeny, shows intermixed distribution among the various sequence types and serotypes (except in certain groups) indicative of probable frequent horizontal gene transfer across the sequence types (Figures 3.11, 3.12). This observation is in line with the evidence from a previous study where the comparison between phylogenetic trees of the core genome and *pks* island sequences within ECOR collection displayed different clustering pattern indicating the transmission of the island to be horizontal (Messerer et al., 2017). This along with other observations of prevalence patterns demonstrate that certain sequence types of *E. coli* like ST73, ST95 and ST12 show increased capability of acquiring the island, and frequent horizontal exchanges of the island could occur across these subtypes.

Summary and Conclusions

Extraintestinal pathologies caused by highly virulent strains of *E. coli* amount to clinical implications with high morbidity and mortality rates. The genotoxin colibactin encoded by the *pks* island of *E. coli*, mainly belonging to the phylogroup B2, has been reported as an important determinant of bacterial pathogenicity. Pathogenic strains of *E. coli* are continuously evolving with the HGT based acquisition of mobile genetic elements, such as pathogenicity islands for instance the *pks* island, which produces the genotoxin colibactin, resulting in severe clinical outcomes, including colorectal cancer progression. The current study encompasses molecular epidemiology and, high-throughput comparative genomics to address questions pertaining to the acquisition and evolutionary pattern of the *pks* genomic island within *E. coli* subtypes. It is crucial to gain insights into the distribution, transfer, and maintenance of pathogenic islands, as they harbor multiple virulence genes involved in pathogenesis and clinical implications of the infection.

The present study was carried out to detect the *pks* pathogenicity island in extraintestinal pathogenic *E. coli* (ExPEC) isolated from a tertiary hospital in Pune, India. Of the 462 isolates screened, *pks* genomic island was detected in 35 (7.6%) isolates, which predominantly belonged to pathogenic phylogroup B2 (97%), and harbored virulence genes such as *fimH*, *sfaD/E*, and *usp*. Biofilm formation assay revealed 21 of the 35 *pks*-carrying isolates to be strong, 10 isolates to be moderate and 4 as weak (SBF < 0.5) biofilm formers. All of the *pks*-carrying isolates were resistant against the bactericidal activity of human serum. Assays carried out to detect antimicrobial susceptibility revealed 11% of these isolates to be multidrug-resistant, 37% ESBL producers and 25% were positive for $bla_{\text{CTX-M-15}}$.

We further employed a high-throughput whole-genome comparison and phylogenetic analysis of such pathogenic *E. coli* isolates to gain insights into the patterns of distribution, horizontal transmission, and evolution of this island. For this analysis, 23 *pks*-positive ExPEC genomes were newly sequenced, and their virulome and resistome profiles indicated a preponderance of virulence encoding genes and

a reduced number of genes for antimicrobial resistance. In addition, 4,090 *E. coli* genomes from the public domain were also analysed for large-scale screening for *pks*-positive genomes, out of which a total of 530 *pks*-positive genomes were characterised to understand the subtype-based distribution pattern(s). The *pks* island showed a significant association with the B2 phylogroup (82.2%) and a high prevalence in sequence type 73 (ST73; n = 179) and ST95 (n=110) and O6:H1(n = 110) serotype. ST 95 was selected as a model data set because it had both *pks*-positive and *pks*-negative genomes, to understand the evolution and transmission dynamics of *pks* island. Maximum-likelihood (ML) phylogeny of the core genome and intergenic regions (IGRs) of the ST95 genomes displayed clustering in relation to their carriage of the *pks* island. Further prevalence patterns of genes encoding RM systems in the *pks*-positive and *pks*-negative genomes were also analysed to determine their potential role in acquisition and capability to maintain the genomes. The maximum-likelihood phylogeny based on the core genome and *pks* island sequences from 247 genomes with an intact *pks* island demonstrated horizontal gene transfer of the island across sequence types and serotypes, with few exceptions. This study contributes vitally to the understanding of different subtypes and that have a higher propensity to carry the *pks* island-encoded genotoxin and henceforth can impart possible clinical implications.

As genotoxin colibactin harbored by extraintestinal pathogenic *E. coli* (ExPEC) and other members of the Enterobacteriaceae has been increasingly reported worldwide, the present study was undertaken to understand the distribution and characteristics of such *E. coli* in Indian settings, wherein a moderate prevalence of *pks* harboring *E. coli* was observed among the among clinical ExPEC isolates in our collection. The *pks* positive isolates demonstrated relatively low antimicrobial resistance and was observed to carry multiple virulence factors. The findings from our study perhaps provide the first essential baseline data regarding the functional molecular infection epidemiology of these genotoxic *E. coli* and their clinical significance in India. Our study is perhaps the first one to perform large scale,

whole genome-based investigations of the genotoxic *pks* island to understand its distribution among different *E. coli* population and their evolutionary relationships. The preferential distribution pattern of such genotoxin carrying *E. coli* was demonstrated using different methods of *in-silico* subtyping and these observations could supplement diagnostic settings with information regarding the potential genotoxic nature of the isolates. The *pks*-island phylogeny indicates horizontal transmission and the ability to exchange between compatible *E. coli* subtypes. Investigation of the ST95 model depicted the higher prevalence of *pks* island in specific serotypes and CH types, showing the nature of horizontal gene transfer and finer evolution within a particular ST. The core genome and core intergenic region phylogeny provided better insights to gain a comprehensive understanding of the clade-specific pattern of distribution of the island, which is otherwise a part of the accessory genome. The potential role of RM systems in shaping the lineages and driving the acquisition of the island among compatible isolates needs to be performed at higher resolution in further studies and these could provide interesting insights into the HGT and evolution of *E. coli.*

In conclusion, our study is perhaps the first one to report the epidemiological information on the prevalence of colibactin harboring *Escherichia coli* from India population and also to perform a large-scale, whole-genome-based investigations with respect to the distribution of the *pks* island(s) among different *E. coli* populations and the consequent evolutionary relationships. The preferential distribution pattern of the *pks* (encoded genotoxin)-harboring *E. coli* was studied using different computational methods of subtyping. These observations may be able to provide support to the diagnostic systems or health care modalities aimed at understanding the clinical implications of the potential genotoxic nature of *pks*-positive isolates. The observed colibactin genomic island harboring pathogenic *E. coli* advocate the urgent need for broader surveillance to decipher and prevent transmission of these ExPEC in community and hospital settings. It will also be interesting to understand the distribution and evolutionary relationships of *pks* island in other members of the family

Enterobacteriaceae, and to understand the potential horizontal gene transfer across the different species. Further, high-throughput studies involving other cyclomodulins like CDT, CNF and *cif* could also provide crucial insights into the nature of evolution of such genotoxic bacteria.

Bibliography

Ahmed, N., Dobrindt, U., Hacker, J., and Hasnain, S. E. (2008). Genomic fluidity and pathogenic bacteria: applications in diagnostics, epidemiology and intervention. *Nat. Rev. Microbiol.* 6, 387–394. doi:10.1038/nrmicro1889.

Alikhan, N. F., Petty, N. K., Ben Zakour, N. L., and Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): Simple prokaryote genome comparisons. *BMC Genomics*. doi:10.1186/1471-2164-12-402.

Arthur, J. C., Perez-Chanona, E., Mühlbauer, M., Tomkovich, S., Uronis, J. M., Fan, T.-J., et al. (2012). Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* 338, 120–3. doi:10.1126/science.1224820.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* doi:10.1089/cmb.2012.0021.

Bausch, C., Peekhaus, N., Utz, C., Blais, T., Murray, E., Lowary, T., et al. (1998). Sequence analysis of the GntII (Subsidiary) system for gluconate metabolism reveals a novel pathway for L-idonic acid catabolism in Escherichia coli. *J. Bacteriol.* doi:10.1128/jb.180.14.3704-3710.1998.

Bengtsson, S., Naseer, U., Sundsfjord, A., Kahlmeter, G., and Sundqvist, M. (2012). Sequence types and plasmid carriage of uropathogenic Escherichia coli devoid of phenotypically detectable resistance. *J. Antimicrob. Chemother.* doi:10.1093/jac/dkr421.

Bien, J., Sokolova, O., and Bozko, P. (2012). Role of uropathogenic escherichia coli virulence factors in development of urinary tract infection and kidney damage. *Int. J. Nephrol.* doi:10.1155/2012/681473.

Bliven, K. A., and Maurelli, A. T. (2012). Antivirulence genes: Insights into pathogen evolution through gene loss. *Infect. Immun.* doi:10.1128/IAI.00740-12.

Bossuet-Greif, N., Dubois, D., Petit, C., Tronnet, S., Martin, P., Bonnet, R., et al. (2015). Escherichia coli ClbS is a colibactin resistance protein. *Mol. Microbiol.* doi:10.1111/mmi.13272.

Bossuet-Greif, N., Dubois, D., Petit, C., Tronnet, S., Martin, P., Bonnet, R., et al. (2016). Escherichia coliClbS is a colibactin resistance protein. *Mol. Microbiol.* 99, 897–908. doi:10.1111/mmi.13272.

Bossuet-Greif, N., Vignard, J., Taieb, F., Mirey, G., Dubois, D., Petit, C., et al. (2018). The colibactin genotoxin generates DNA interstrand cross-links in infected cells. *MBio.* doi:10.1128/mBio.02393-17.

Brynildsrud, O., Bohlin, J., Scheffer, L., and Eldholm, V. (2016). Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* doi:10.1186/s13059-016-1108-8.

Brzuszkiewicz, E., Gottschalk, G., Buchrieser, C., Dobrindt, U., and Oswald, E. (2006). Escherichia coli induces DNA Double-Strand Breaks in Eukaryotic Cells. *Science (80-.).* 313, 848–851.

Buc, E., Dubois, D., Sauvanet, P., Raisch, J., Delmas, J., Darfeuille-Michaud, A., et al. (2013a). High Prevalence of Mucosa-Associated E. coli Producing Cyclomodulin and Genotoxin in Colon Cancer. *PLoS One*. doi:10.1371/journal.pone.0056964.

Buc, E., Dubois, D., Sauvanet, P., Raisch, J., Delmas, J., Darfeuille-Michaud, A., et al. (2013b). High Prevalence of Mucosa-Associated E. coli Producing Cyclomodulin and Genotoxin in Colon Cancer. *PLoS One* 8, e56964. doi:10.1371/journal.pone.0056964.

Budroni, S., Siena, E., Dunning Hotopp, J. C., Seib, K. L., Serruto, D., Nofroni, C., et al. (2011). Neisseria meningitidis is structured in clades associated with restriction modification systems that

modulate homologous recombination. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.1019751108.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*. doi:10.1186/1471-2105-10-421.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. doi:10.1093/bioinformatics/btp348.

Chagneau, C. V., Garcie, C., Bossuet-Greif, N., Tronnet, S., Brachmann, A. O., Piel, J., et al. (2019). The Polyamine Spermidine Modulates the Production of the Bacterial Genotoxin Colibactin. *mSphere* 4, 1–11. doi:10.1128/msphere.00414-19.

Chaudhuri, R. R., and Henderson, I. R. (2012). The evolution of the Escherichia coli phylogeny. *Infect. Genet. Evol.* doi:10.1016/j.meegid.2012.01.005.

Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., et al. (2005). VFDB: A reference database for bacterial virulence factors. *Nucleic Acids Res.* doi:10.1093/nar/gki008.

Chen, Y.-T., Lai, Y.-C., Tan, M.-C., Hsieh, L.-Y., Wang, J.-T., Shiau, Y.-R., et al. (2017). Prevalence and characteristics of pks genotoxin gene cluster-positive clinical Klebsiella pneumoniae isolates in Taiwan. *Sci. Rep.* 7, 43120. doi:10.1038/srep43120.

Cheng, L., Connor, T. R., Sirén, J., Aanensen, D. M., and Corander, J. (2013). Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol. Biol. Evol.* doi:10.1093/molbev/mst028.

Clermont, O., Bonacorsi, S., and Bingen, E. (2000). Rapid and simple determination of the Escherichia

coli phylogenetic group. *Appl. Environ. Microbiol.* doi:10.1128/AEM.66.10.4555-4558.2000.

Clermont, O., Christenson, J. K., Denamur, E., and Gordon, D. M. (2013). The Clermont Escherichia coli phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environ. Microbiol. Rep.* 5, 58–65. doi:10.1111/1758-2229.12019.

Clermont, O., Gordon, D., and Denamur, E. (2015). Guide to the various phylogenetic classification schemes for escherichia coli and the correspondence among schemes. *Microbiol. (United Kingdom).* doi:10.1099/mic.0.000063.

CLSI (2013). Performance Standards for Antimicrobial Disk and Dilution Susceptibility Tests for Bacteria Isolated From Animals; Approved Standard. VET01-A4. *Clin. Lab. Stand. Inst. - Fourth Ed.*

Corander, J., Marttinen, P., Sirén, J., and Tang, J. (2005). BAPS: Bayesian analysis of population structure. *Man. ver.*

Cougnoux, A., Dalmasso, G., Martinez, R., Buc, E., Delmas, J., Gibold, L., et al. (2014). Bacterial genotoxin colibactin promotes colon tumour growth by inducing a senescence-associated secretory phenotype. *Gut*, 1–11. doi:10.1136/gutjnl-2013-305257.

Croxen, M. A., and Finlay, B. B. (2010). Molecular mechanisms of Escherichia coli pathogenicity. *Nat. Rev. Microbiol.* 8, 26–38. doi:10.1038/nrmicro2265.

Cuevas-Ramos, G., Petit, C. R., Marcq, I., Boury, M., Oswald, E., and Nougayrède, J.-P. (2010). Escherichia coli induces DNA damage in vivo and triggers genomic instability in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* 107, 11537–11542. doi:10.1073/pnas.1001261107.

Dale, A. P., and Woodford, N. (2015). Extra-intestinal pathogenic Escherichia coli (ExPEC): Disease,

carriage and clones. *J. Infect.* doi:10.1016/j.jinf.2015.09.009.

Dalmasso, G., Cougnoux, A., Delmas, J., Darfeuille-Michaud, A., and Bonnet, R. (2015a). The bacterial genotoxin colibactin promotes colon tumor growth by modifying the tumor microenvironment. *Gut Microbes* 5, 675–680. doi:10.4161/19490976.2014.969989.

Dalmasso, G., Cougnoux, A., Delmas, J., Darfeuille-Michaud, A., and Bonnet, R. (2015b). The bacterial genotoxin colibactin promotes colon tumor growth by modifying the tumor microenvironment. *Gut Microbes*. doi:10.4161/19490976.2014.969989.

De Rycke, J., Comtet, E., Chalareng, C., Boury, M., Tasca, C., and Milon, A. (1997). Enteropathogenic Escherichia coli O103 from rabbit elicits actin stress fibers and focal adhesions in hela epithelial cells, cytopathic effects that are linked to an analog of the locus of enterocyte effacement. *Infect. Immun.* doi:10.1128/iai.65.7.2555-2563.1997.

Dejea, C. M., Fathi, P., Craig, J. M., Boleij, A., Taddese, R., Geis, A. L., et al. (2018). Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria. *Science (80-.).* 359, 592–597. doi:10.1126/science.aah3648.

Denamur, E., Clermont, O., Bonacorsi, S., and Gordon, D. (2020). The population genetics of pathogenic Escherichia coli. *Nat. Rev. Microbiol.*, 1–18. doi:10.1038/s41579-020-0416-x.

Didelot, X., and Wilson, D. J. (2015). ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLoS Comput. Biol.* doi:10.1371/journal.pcbi.1004041.

Dobrindt, U., Hochhut, B., Hentschel, U., and Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nat. Rev. Microbiol.* doi:10.1038/nrmicro884.

Dubois, D., Delmas, J., Cady, A., Robin, F., Sivignon, A., Oswald, E., et al. (2010). Cyclomodulins in

urosepsis strains of Escherichia coli. *J. Clin. Microbiol.* 48, 2122–2129. doi:10.1128/JCM.02365-09.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. doi:10.1093/bioinformatics/btq461.

El Ghachi, M., Bouhss, A., Blanot, D., and Mengin-Lecreulx, D. (2004). The bacA gene of Escherichia coli encodes an undecaprenyl pyrophosphate phosphatase activity. *J. Biol. Chem.* doi:10.1074/jbc.M401701200.

Emody, L., Kerényi, M., and Nagy, G. (2003). Virulence factors of uropathogenic Escherichia coli. *Int. J. Antimicrob. Agents* 22. doi:10.1016/S0924-8579(03)00236-X.

Ewers, C., Dematheis, F., Singamaneni, H. D., Nandanwar, N., Fruth, A., Diehl, I., et al. (2014). Correlation between the genomic o454-nlpD region polymorphisms, virulence gene equipment and phylogenetic group of extraintestinal Escherichia coli (ExPEC) enables pathotyping irrespective of host, disease and source of isolation. *Gut Pathog.* doi:10.1186/s13099-014-0037-x.

Faïs, T., Delmas, J., Barnich, N., Bonnet, R., and Dalmasso, G. (2018). Colibactin: More than a new bacterial toxin. *Toxins (Basel).* doi:10.3390/toxins10040151.

Falbo, V., Pace, T., Picci, L., Pizzi, E., and Caprioli, A. (1993). Isolation and nucleotide sequence of the gene encoding cytotoxic necrotizing factor 1 of Escherichia coli. *Infect. Immun.* doi:10.1128/iai.61.11.4909-4914.1993.

Falzano, L., Filippini, P., Travaglione, S., Miraglia, A. G., Fabbri, A., and Fiorentini, C. (2006). Escherichia coli cytotoxic necrotizing factor 1 blocks cell cycle G 2/M transition in uroepithelial cells. *Infect. Immun.* doi:10.1128/IAI.01413-05.

Forde, B. M., Phan, M. D., Gawthorne, J. A., Ashcroft, M. M., Stanton-Cook, M., Sarkar, S., et al. (2015). Lineage-specific methyltransferases define the methylome of the globally disseminated escherichia coli ST131 clone. *MBio*. doi:10.1128/mBio.01602-15.

Fratamico, P. M., DebRoy, C., Liu, Y., Needleman, D. S., Baranzoni, G. M., and Feng, P. (2016). Advances in molecular serotyping and subtyping of Escherichia coli. *Front. Microbiol.* doi:10.3389/fmicb.2016.00644.

Gagnaire, A., Nadel, B., Raoult, D., Neefjes, J., and Gorvel, J. P. (2017). Collateral damage: Insights into bacterial mechanisms that predispose host cells to cancer. *Nat. Rev. Microbiol.* doi:10.1038/nrmicro.2016.171.

Gordon, D. M., Geyik, S., Clermont, O., O'Brien, C. L., Huang, S., Abayasekara, C., et al. (2017). Fine-Scale Structure Analysis Shows Epidemic Patterns of Clonal Complex 95, a Cosmopolitan Escherichia coli Lineage Responsible for Extraintestinal Infection. *mSphere*. doi:10.1128/msphere.00168-17.

Grasso, F., and Frisan, T. (2015). Bacterial genotoxins: Merging the DNA damage response into infection biology. *Biomolecules*. doi:10.3390/biom5031762.

Groisman, E. A., and Ochman, H. (1996). Pathogenicity islands: Bacterial evolution in quantum leaps. *Cell*. doi:10.1016/S0092-8674(00)81985-6.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*. doi:10.1093/bioinformatics/btt086.

Hacker, J., and Kaper, J. B. (2000). Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol.* doi:10.1146/annurev.micro.54.1.641.

Haghjoo, E., and Galán, J. E. (2004). Salmonella typhi encodes a functional cytolethal distending toxin that is delivered into host cells by a bacterial-internalization pathway. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.0400932101.

Horiguchi, Y. (2001). Escherichia coli cytotoxic necrotizing factors and Bordetella dermonecrotic toxin: The dermonecrosis-inducing toxins activating Rho small GTPases. *Toxicon*. doi:10.1016/S0041-0101(01)00149-0.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., et al. (2016). EGGNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* doi:10.1093/nar/gkv1248.

Huson, D. H., and Scornavacca, C. (2012). Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* doi:10.1093/sysbio/sys062.

Hussain, A., Ewers, C., Nandanwar, N., Guenther, S., Jadhav, S., Wieler, L. H., et al. (2012a). Multi-resistant uropathogenic Escherichia coli from an endemic zone of urinary tract infections in India: genotypic and phenotypic characteristics of ST131 isolates of the CTX-M-15 Extended-Spectrum-Beta-Lactamase producing lineage. *Antimicrob. Agents Chemother.* doi:10.1128/AAC.01099-12.

Hussain, A., Ewers, C., Nandanwar, N., Guenther, S., Jadhav, S., Wieler, L. H., et al. (2012b). Multiresistant uropathogenic Escherichia coli from a region in India where urinary tract infections are endemic: genotypic and phenotypic characteristics of sequence type 131 isolates of the CTX-M-15 extended-spectrum-β-lactamase-producing lineage. *Antimicrob. Agents Chemother.* 56, 6358–65. doi:10.1128/AAC.01099-12.

Hussain, A., Ranjan, A., Nandanwar, N., Babbar, A., Jadhav, S., and Ahmed, N. (2014a). Genotypic

and phenotypic profiles of Escherichia coli isolates belonging to clinical sequence type 131 (ST131), clinical non-ST131, and fecal non-ST131 lineages from India. *Antimicrob. Agents Chemother.* 58, 7240–9. doi:10.1128/AAC.03320-14.

Hussain, A., Ranjan, A., Nanwar, N., Babbar, A., Jadhav, S., and Ahmed, N. (2014b). Genotypic and phenotypic profiles of escherichia coli isolates belonging to clinical sequence type 131 (ST131), clinical non-ST131, and fecal non-ST131 lineages from India. *Antimicrob. Agents Chemother.* 58, 7240–7249. doi:10.1128/AAC.03320-14.

Hussain, A., Ranjan, A., Nanwar, N., Babbar, A., Jadhav, S., and Ahmed, N. (2014c). Genotypic and phenotypic profiles of escherichia coli isolates belonging to clinical sequence type 131 (ST131), clinical non-ST131, and fecal non-ST131 lineages from India. *Antimicrob. Agents Chemother.* doi:10.1128/AAC.03320-14.

Hussain, A., Shaik, S., Ranjan, A., Nandanwar, N., Tiwari, S. K., Majid, M., et al. (2017). Risk of transmission of antimicrobial resistant Escherichia coli from commercial broiler and free-range retail chicken in India. *Front. Microbiol.* doi:10.3389/fmicb.2017.02120.

Iftekhar, A., Berger, H., Bouznad, N., Heuberger, J., Boccellato, F., Dobrindt, U., et al. (2021). Genomic aberrations after short-term exposure to colibactin-producing E. coli transform primary colon epithelial cells. *Nat. Commun.* 12. doi:10.1038/s41467-021-21162-y.

Iyadorai, T., Mariappan, V., Vellasamy, K. M., Wanyiri, J. W., Roslani, A. C., Lee, G. K., et al. (2020). Prevalence and association of pks+ Escherichia coli with colorectal cancer in patients at the University Malaya Medical Centre, Malaysia. *PLoS One.* doi:10.1371/journal.pone.0228217.

Jadhav, S., Hussain, A., Devi, S., Kumar, A., Parveen, S., Gandham, N., et al. (2011). Virulence characteristics and genetic affinities of multiple drug resistant uropathogenic Escherichia coli

from a semi urban locality in India. *PLoS One* 6, e18063. doi:10.1371/journal.pone.0018063.

Jang, J., Hur, H. G., Sadowsky, M. J., Byappanahalli, M. N., Yan, T., and Ishii, S. (2017). Environmental Escherichia coli: ecology and public health implications—a review. *J. Appl. Microbiol.* doi:10.1111/jam.13468.

Johnson, J. R. (1991). Virulence factors in Escherichia coli urinary tract infection. *Clin. Microbiol. Rev.* doi:10.1128/CMR.4.1.80.Updated.

Johnson, J. R., Johnston, B., Kuskowski, M. A., Nougayrede, J. P., and Oswald, E. (2008). Molecular epidemiology and phylogenetic distribution of the Escherichia coli pks genomic island. *J. Clin. Microbiol.* 46, 3906–3911. doi:10.1128/JCM.00949-08.

Johnson, J. R., and Russo, T. A. (2002). Extraintestinal pathogenic Escherichia coli: "The other bad E coli." *J. Lab. Clin. Med.* 139, 155–162. doi:10.1067/mlc.2002.121550.

Jolley, K. A., and Maiden, M. C. J. (2010). BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*. doi:10.1186/1471-2105-11-595.

Jubelin, G., Varela Chavez, C., Taieb, F., Banfield, M. J., Samba-Louaka, A., Nobe, R., et al. (2009). Cycle Inhibiting Factors (CIFs) are a growing family of functional cyclomodulins present in invertebrate and mammal bacterial pathogens. *PLoS One*. doi:10.1371/journal.pone.0004855.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods*. doi:10.1038/nmeth.4285.

Kaper, J. B., Nataro, J. P., and Mobley, H. L. (2004). Pathogenic Escherichia coli. *Nat. Rev. Microbiol.* 2, 123–140. doi:10.1038/nrmicro818.

Kohoutova, D., Smajs, D., Moravkova, P., Cyrany, J., Moravkova, M., Forstlova, M., et al. (2014). Escherichia coli strains of phylogenetic group B2 and D and bacteriocin production are associated with advanced colorectal neoplasia. *BMC Infect. Dis.* doi:10.1186/s12879-014-0733-7.

Krieger, J. N., Dobrindt, U., Riley, D. E., and Oswald, E. (2011). Acute Escherichia coli prostatitis in previously health young men: Bacterial virulence factors, antimicrobial resistance, and clinical outcomes. *Urology*. doi:10.1016/j.urology.2010.12.059.

Lai, Y. C., Lin, A. C., Chiang, M. K., Dai, Y. H., Hsu, C. C., Lu, M. C., et al. (2014). Genotoxic Klebsiella pneumoniae in Taiwan. *PLoS One*. doi:10.1371/journal.pone.0096292.

Lan, Y., Zhou, M., Jian, Z., Yan, Q., Wang, S., and Liu, W. (2019). Prevalence of pks gene cluster and characteristics of Klebsiella pneumoniae-induced bloodstream infections. *J. Clin. Lab. Anal.* doi:10.1002/jcla.22838.

Langmead, B., and Salzberg, S. (2013). Bowtie2. *Nat. Methods*. doi:10.1038/nmeth.1923.Fast.

Lara-Tejero, M., and Galán, J. E. (2001). CdtA, CdtB, and CdtC form a tripartite complex that is required for cytolethal distending toxin activity. *Infect. Immun.* doi:10.1128/IAI.69.7.4358-4365.2001.

Larsen, M. V., Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R. L., et al. (2012). Multilocus sequence typing of total-genome-sequenced bacteria. *J. Clin. Microbiol.* doi:10.1128/JCM.06094-11.

Lee, E., and Lee, Y. (2018). Prevalence of Escherichia coli carrying pks islands in bacteremia patients. *Ann. Lab. Med.* doi:10.3343/alm.2018.38.3.271.

Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new

developments. *Nucleic Acids Res.* doi:10.1093/nar/gkz239.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* doi:10.1093/bioinformatics/btp352.

Lopès, A., Billard, E., Casse, A. H., Villéger, R., Veziant, J., Roche, G., et al. (2020). Colibactin-positive Escherichia coli induce a procarcinogenic immune environment leading to immunotherapy resistance in colorectal cancer. *Int. J. Cancer.* doi:10.1002/ijc.32920.

Löytynoja, A. (2014). Phylogeny-aware alignment with PRANK. *Methods Mol. Biol.* doi:10.1007/978-1-62703-646-7_10.

Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., et al. (1998). Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.95.6.3140.

Marchès, O., Ledger, T. N., Boury, M., Ohara, M., Tu, X., Goffaux, F., et al. (2003). Enteropathogenic and enterohaemorrhagic Escherichia coli deliver a novel effector called Cif, which blocks cell cycle G2/M transition. *Mol. Microbiol.* 50, 1553–1567. doi:10.1046/j.1365-2958.2003.03821.x.

Marcq, I., Martin, P., Payros, D., Cuevas-Ramos, G., Boury, M., Watrin, C., et al. (2014). The genotoxin colibactin exacerbates lymphopenia and decreases survival rate in mice infected with septicemic Escherichia coli. *J. Infect. Dis.* 210, 285–294. doi:10.1093/infdis/jiu071.

Martin, O. C. B., Frisan, T., and Mihaljevic, B. (2016). "Bacterial Genotoxins as the Interphase Between DNA Damage and Immune Response," in doi:10.1007/978-94-007-6725-6_14-1.

Martin, P., Tronnet, S., Garcie, C., and Oswald, E. (2017). Interplay between siderophores and colibactin genotoxin in Escherichia coli. *IUBMB Life* 69, 435–441. doi:10.1002/iub.1612.

Massip, C., Branchu, P., Bossuet-Greif, N., Chagneau, C. V., Gaillard, D., Martin, P., et al. (2019). Deciphering the interplay between the genotoxic and probiotic activities of Escherichia coli Nissle 1917. *PLoS Pathog.* doi:10.1371/journal.ppat.1008029.

Massip, C., Chagneau, C. V., Boury, M., and Oswald, E. (2020). The synergistic triad between microcin, colibactin, and salmochelin gene clusters in uropathogenic Escherichia coli. *Microbes Infect.*, 1–4. doi:10.1016/j.micinf.2020.01.001.

McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., et al. (2013). The comprehensive antibiotic resistance database. *Antimicrob. Agents Chemother.* doi:10.1128/AAC.00419-13.

McCarthy, A. J., Martin, P., Cloup, E., Stabler, R. A., Oswald, E., and Taylor, P. W. (2015). The Genotoxin Colibactin Is a Determinant of Virulence in Escherichia coli K1 Experimental Neonatal Systemic Infection. *Infect. Immun.* 83, 3704–11. doi:10.1128/IAI.00716-15.

Mcdaniel, T. K., Jarvis, K. G., Donnenberg, M. S., and Kaper, J. B. (1995). A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.92.5.1664.

McNally, A., Oren, Y., Kelly, D., Pascoe, B., Dunn, S., Sreecharan, T., et al. (2016). Combined Analysis of Variation in Core, Accessory and Regulatory Genome Regions Provides a Super-Resolution View into the Evolution of Bacterial Populations. *PLoS Genet.* doi:10.1371/journal.pgen.1006280.

Messerer, M., Fischer, W., and Schubert, S. (2017). Investigation of horizontal gene transfer of pathogenicity islands in Escherichia coli using next-generation sequencing. *PLoS One.* doi:10.1371/journal.pone.0179880.

Micenková, L., Beňová, A., Frankovičová, L., Bosák, J., Vrba, M., Ševčíková, A., et al. (2017). Human Escherichia coli isolates from hemocultures: Septicemia linked to urogenital tract infections is caused by isolates harboring more virulence genes than bacteraemia linked to other conditions. *Int. J. Med. Microbiol.* doi:10.1016/j.ijmm.2017.02.003.

Monstein, H. J., Östholm-Balkhed, Å., Nilsson, M. V., Nilsson, M., Dornbusch, K., and Nilsson, L. E. (2007). Multiplex PCR amplification assay for the detection of blaSHV, blaTEM and blaCTX-M genes in Enterobacteriaceae. *APMIS* 115, 1400–1408. doi:10.1111/j.1600-0463.2007.00722.x.

Morgan, R. N., Saleh, S. E., Farrag, H. A., and Aboulwafa, M. M. (2019). Prevalence and pathologic effects of colibactin and cytotoxic necrotizing factor-1 (Cnf 1) in Escherichia coli: Experimental and bioinformatics analyses. *Gut Pathog.* 11, 1–18. doi:10.1186/s13099-019-0304-y.

Mousa, J. J., Newsome, R. C., Yang, Y., Jobin, C., and Bruner, S. D. (2017). ClbM is a versatile, cation-promiscuous MATE transporter found in the colibactin biosynthetic gene cluster. *Biochem. Biophys. Res. Commun.* 482, 1233–1239. doi:10.1016/j.bbrc.2016.12.018.

Mousa, J. J., Yang, Y., Tomkovich, S., Shima, A., Newsome, R. C., Tripathi, P., et al. (2016). MATE transport of the E. Coli-derived genotoxin colibactin. *Nat. Microbiol.* doi:10.1038/nmicrobiol.2015.9.

Nandanwar, N., Janssen, T., Kühl, M., Ahmed, N., Ewers, C., and Wieler, L. H. (2014). Extraintestinal pathogenic Escherichia coli (ExPEC) of human and avian origin belonging to sequence type complex 95 (STC95) portray indistinguishable virulence features. *Int. J. Med. Microbiol.* 304, 835–842. doi:10.1016/j.ijmm.2014.06.009.

Nandi, T., Holden, M. T. G., Didelot, X., Mehershahi, K., Boddey, J. A., Beacham, I., et al. (2015). Burkholderia pseudomallei sequencing identifies genomic clades with distinct recombination,

accessory, and epigenetic profiles. *Genome Res.* doi:10.1101/gr.177543.114.

Nataro, J. P., and Kaper, J. B. (1998). Diarrheagenic Escherichia coli. *Clin. Microbiol. Rev.* 11, 142–201. doi:file://Z:\References\Text Files\00000004469.txt.

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* doi:10.1093/molbev/msu300.

Nougayrède, J.-P., Homburg, S., Taieb, F., Boury, M., Brzuszkiewicz, E., Gottschalk, G., et al. (2006). Escherichia coli induces DNA double-strand breaks in eukaryotic cells. *Science* 313, 848–51. doi:10.1126/science.1127059.

Nougayrède, J.-P., Taieb, F., De Rycke, J., and Oswald, E. (2005a). Cyclomodulins: bacterial effectors that modulate the eukaryotic cell cycle. *Trends Microbiol.* 13, 103–10. doi:10.1016/j.tim.2005.01.002.

Nougayrède, J.-P., Taieb, F., Rycke, J. De, and Oswald, E. (2005b). Cyclomodulins: bacterial effectors that modulate the eukaryotic cell cycle. *Trends Microbiol.* 13, 103–110. doi:10.1016/j.tim.2005.01.002.

Nougayrède, J. P., Boury, M., Tasca, C., Marchès, O., Milon, A., Oswald, E., et al. (2001). Type III secretion-dependent cell cycle block caused in HeLa cells by enteropathogenic Escherichia coli O103. *Infect. Immun.* doi:10.1128/IAI.69.11.6785-6795.2001.

Nowrouzian, F. L., and Oswald, E. (2012). Escherichia coli strains with the capacity for long-term persistence in the bowel microbiota carry the potentially genotoxic pks island. *Microb. Pathog.* 53, 180–182. doi:10.1016/j.micpath.2012.05.011.

Ogawa, M., Handa, Y., Ashida, H., Suzuki, M., and Sasakawa, C. (2008). The versatility of Shigella effectors. *Nat. Rev. Microbiol.* doi:10.1038/nrmicro1814.

Olaitan, A. O., Morand, S., and Rolain, J. M. (2014). Mechanisms of polymyxin resistance: Acquired and intrinsic resistance in bacteria. *Front. Microbiol.* doi:10.3389/fmicb.2014.00643.

Oliveira, P. H., Touchon, M., and Rocha, E. P. C. (2014). The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.* doi:10.1093/nar/gku734.

Oren, Y., Smith, M. B., Johns, N. I., Zeevi, M. K., Biran, D., Ron, E. Z., et al. (2014). Transfer of noncoding DNA drives regulatory rewiring in Bacteria. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.1413272111.

Orskov, I., Orskov, F., Jann, B., and Jann, K. (1977). Serology, chemistry, and genetics of O and K antigens of Escherichia coli. *Bacteriol. Rev.* doi:10.1128/mmbr.41.3.667-710.1977.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., et al. (2015). Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. doi:10.1093/bioinformatics/btv421.

Parkhill, J., and Wren, B. W. (2011). Bacterial epidemiology and biology - lessons from genome sequencing. *Genome Biol.* doi:10.1186/gb-2011-12-10-230.

Patel, R. K., and Jain, M. (2012). NGS QC toolkit: A toolkit for quality control of next generation sequencing data. *PLoS One*. doi:10.1371/journal.pone.0030619.

Payros, D., Secher, T., Boury, M., Brehin, C., M??nard, S., Salvadorcartier, C., et al. (2014). Maternally acquired genotoxic Escherichia coli alters offspring's intestinal homeostasis. *Gut Microbes* 5.

doi:10.4161/gmic.28932.

Picard, B., Garcia, J. S., Gouriou, S., Duriez, P., Brahimi, N., Bingen, E., et al. (1999). The link between phylogeny and virulence in Escherichia coli extraintestinal infection? *Infect. Immun.*

Pleguezuelos-Manzano, C., Puschhof, J., Rosendahl Huber, A., van Hoeck, A., Wood, H. M., Nomburg, J., et al. (2020). Mutational signature in colorectal cancer caused by genotoxic pks + E. coli. *Nature* 580. doi:10.1038/s41586-020-2080-8.

Putze, J., Hennequin, C., Nougayrède, J. P., Zhang, W., Homburg, S., Karch, H., et al. (2009). Genetic structure and distribution of the colibactin genomic island among members of the family Enterobacteriaceae. *Infect. Immun.* 77, 4696–4703. doi:10.1128/IAI.00522-09.

Ranjan, A., Shaik, S., Hussain, A., Nandanwar, N., Semmler, T., Jadhav, S., et al. (2015a). Genomic and functional portrait of a highly virulent, CTX-M-15-producing H30-Rx subclone of Escherichia coli sequence type (ST) 131. *Antimicrob. Agents Chemother.* Available at: http://aac.asm.org/content/early/2015/07/14/AAC.01447-15.abstract.

Ranjan, A., Shaik, S., Hussain, A., Nandanwar, N., Semmler, T., Jadhav, S., et al. (2015b). Genomic and Functional Portrait of a Highly Virulent, CTX-M-15-Producing H30-Rx Subclone of Escherichia coli Sequence Type 131. *Antimicrob. Agents Chemother.* 59, 6087–95. doi:10.1128/AAC.01447-15.

Ranjan, A., Shaik, S., Mondal, A., Nandanwar, N., Hussain, A., Semmler, T., et al. (2016a). Molecular epidemiology and genome dynamics of New Delhi metallo-beta-lactamase (NDM) producing extraintestinal pathogenic E. coli (ExPEC) strains from India. *Antimicrob. Agents Chemother.* doi:10.1128/AAC.01345-16.

Ranjan, A., Shaik, S., Mondal, A., Nandanwar, N., Hussain, A., Semmler, T., et al. (2016b). Molecular epidemiology and genome dynamics of New Delhi Metallo-β-Lactamase-producing extraintestinal pathogenic Escherichia coli strains from India. *Antimicrob. Agents Chemother.* doi:10.1128/AAC.01345-16.

Ranjan, A., Shaik, S., Nandanwar, N., Hussain, A., Tiwari, S. K., Semmler, T., et al. (2017). Comparative genomics of escherichia coli isolated from skin and soft tissue and other extraintestinal infections. *MBio* 8. doi:10.1128/mBio.01070-17.

Roberts, R. J., Vincze, T., Posfai, J., and Macelis, D. (2015). REBASE-a database for DNA restriction and modification: Enzymes, genes and genomes. *Nucleic Acids Res.* doi:10.1093/nar/gku1046.

Roer, L., Hansen, F., Frølund Thomsen, M. C., Knudsen, J. D., Hansen, D. S., Wang, M., et al. (2017). WGS-based surveillance of third-generation cephalosporin-resistant Escherichia coli from bloodstream infections in Denmark. *J. Antimicrob. Chemother.* doi:10.1093/jac/dkx092.

Roer, L., Hendriksen, R. S., Leekitcharoenphon, P., Lukjancenko, O., Kaas, R. S., Hasman, H., et al. (2016). Is the Evolution of Salmonella enterica subsp. enterica Linked to Restriction-Modification Systems? *mSystems.* doi:10.1128/msystems.00009-16.

Roer, L., Johannesen, T. B., Hansen, F., Stegger, M., Tchesnokova, V., Sokurenko, E., et al. (2018). CHTyper, a web tool for subtyping of extraintestinal pathogenic Escherichia coli based on the fumC and fimH alleles. *J. Clin. Microbiol.* doi:10.1128/JCM.00063-18.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., et al. (2000). Artemis: Sequence visualization and annotation. *Bioinformatics.* doi:10.1093/bioinformatics/16.10.944.

Saha, R., Saha, N., Donofrio, R. S., and Bestervelt, L. L. (2013). Microbial siderophores: A mini review.

J. Basic Microbiol. doi:10.1002/jobm.201100552.

Samba-Louaka, A., Nougayrède, J.-P., Watrin, C., Jubelin, G., Oswald, E., and Taieb, F. (2008). Bacterial cyclomodulin Cif blocks the host cell cycle by stabilizing the cyclin-dependent kinase inhibitors p21 and p27. *Cell. Microbiol.* 10, 2496–508. doi:10.1111/j.1462-5822.2008.01224.x.

Sarowska, J., Futoma-Koloch, B., Jama-Kmiecik, A., Frej-Madrzak, M., Ksiazczyk, M., Bugla-Ploskonska, G., et al. (2019). Virulence factors, prevalence and potential transmission of extraintestinal pathogenic Escherichia coli isolated from different sources: Recent reports. *Gut Pathog.* doi:10.1186/s13099-019-0290-0.

Sarshar, M., Scribano, D., Marazzato, M., Ambrosi, C., Aprea, M. R., Aleandri, M., et al. (2017). Genetic diversity, phylogroup distribution and virulence gene profile of pks positive Escherichia coli colonizing human intestinal polyps. *Microb. Pathog.* 112, 274–278. doi:10.1016/j.micpath.2017.10.009.

Schwyn, B., and Neilands, J. B. (1987). Universal chemical assay for the detection and determination of siderophores. *Anal. Biochem.* 160, 47–56. doi:10.1016/0003-2697(87)90612-9.

Secher, T., Samba-Louaka, A., Oswald, E., and Nougayrède, J. P. (2013). Escherichia coli Producing Colibactin Triggers Premature and Transmissible Senescence in Mammalian Cells. *PLoS One* 8. doi:10.1371/journal.pone.0077157.

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*. doi:10.1093/bioinformatics/btu153.

Selander, R. K., Caugant, D. A., Ochman, H., Musser, J. M., Gilmour, M. N., and Whittam, T. S. (1986). Methods of multilocus enzyme electrophoresis for bacterial population genetics and

systematics. *Appl. Environ. Microbiol.* doi:10.1128/aem.51.5.873-884.1986.

Servin, A. L. (2014). Pathogenesis of human diffusely adhering Escherichia coli expressing Afa/Dr adhesins (Afa/Dr DAEC): Current insights and future challenges. *Clin. Microbiol. Rev.* doi:10.1128/CMR.00036-14.

Shaik, S., Ranjan, A., Tiwari, S. K., Hussain, A., Nandanwar, N., Kumar, N., et al. (2017). Comparative genomic analysis of globally dominant ST131 clone with other epidemiologically successful extraintestinal pathogenic escherichia coli (ExPEC) lineages. *MBio.* doi:10.1128/mBio.01596-17.

Shenker, B. J., Hoffmaster, R. H., Zekavat, A., Yamaguchi, N., Lally, E. T., and Demuth, D. R. (2001). Induction of Apoptosis in Human T Cells by Actinobacillus actinomycetemcomitans Cytolethal Distending Toxin Is a Consequence of G 2 Arrest of the Cell Cycle . *J. Immunol.* doi:10.4049/jimmunol.167.1.435.

Shimpoh, T., Hirata, Y., Ihara, S., Suzuki, N., Kinoshita, H., Hayakawa, Y., et al. (2017). Prevalence of pks-positive Escherichia coli in Japanese patients with or without colorectal cancer. *Gut Pathog.* doi:10.1186/s13099-017-0185-x.

Shpigel, N. Y., Elazar, S., and Rosenshine, I. (2008). Mammary pathogenic Escherichia coli. *Curr. Opin. Microbiol.* 11. doi:10.1016/j.mib.2008.01.004.

Stephens, C. M., Adams-Sapper, S., Sekhon, M., Johnson, J. R., and Riley, L. W. (2017). Genomic Analysis of Factors Associated with Low Prevalence of Antibiotic Resistance in Extraintestinal Pathogenic Escherichia coli Sequence Type 95 Strains. *mSphere.* doi:10.1128/msphere.00390-16.

Sullivan, C. B., Diggle, M. A., and Clarke, S. C. (2005). Multilocus sequence typing: Data analysis in clinical microbiology and public health. *Mol. Biotechnol.* doi:10.1385/MB:29:3:245.

Taieb, F., Nougayrède, J. P., and Oswald, E. (2011). Cycle inhibiting factors (Cifs): Cyclomodulins that usurp the ubiquitin-dependent degradation pathway of host cells. *Toxins (Basel).* doi:10.3390/toxins3040356.

Taieb, F., Petit, C., Nougayrède, J.-P., and Oswald, E. (2016). The Enterobacterial Genotoxins: Cytolethal Distending Toxin and Colibactin. *EcoSal Plus* 7. doi:10.1128/ecosalplus.ESP-0008-2016.

Tenaillon, O., Skurnik, D., Picard, B., and Denamur, E. (2010). The population genetics of commensal Escherichia coli. *Nat.Rev.Microbiol.* 8, 207–217. doi:10.1038/nrmicro2298.

Thorpe, H. A., Bayliss, S. C., Sheppard, S. K., and Feil, E. J. (2018). Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. *Gigascience.* doi:10.1093/gigascience/giy015.

Travaglione, S., Fabbri, A., and Fiorentini, C. (2008). The Rho-activating CNF1 toxin from pathogenic E. coli: A risk factor for human cancer development? *Infect. Agent. Cancer.* doi:10.1186/1750-9378-3-4.

Tripathi, P., Shine, E. E., Healy, A. R., Kim, C. S., Herzon, S. B., Bruner, S. D., et al. (2017). ClbS Is a Cyclopropane Hydrolase That Confers Colibactin Resistance. *J. Am. Chem. Soc.* 139, 17719–17722. doi:10.1021/jacs.7b09971.

Turner, S. M., Scott-Tucker, A., Cooper, L. M., and Henderson, I. R. (2006). Weapons of mass destruction: Virulence factors of the global killer enterotoxigenic Escherichia coli. *FEMS Microbiol. Lett.* doi:10.1111/j.1574-6968.2006.00401.x.

Versalovic, J., Koeuth, T., and Lupski, J. R. (1991). Distribution of repetitive DNA sequences in

eubacteria and application to fingerprinting of bacterial genomes. *Nucleic Acids Res.* 19, 6823–31. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=329316&tool=pmcentrez&rendertype=abstract [Accessed April 21, 2016].

Wallenstein, A., Rehm, N., Brinkmann, M., Selle, M., Bossuet-Greif, N., Sauer, D., et al. (2020). ClbR Is the Key Transcriptional Activator of Colibactin Gene Expression in Escherichia coli Downloaded from. doi:10.1128/mSphere.

Weissman, S. J., Johnson, J. R., Tchesnokova, V., Billig, M., Dykhuizen, D., Riddell, K., et al. (2012). High-resolution two-locus clonal typing of extraintestinal pathogenic Escherichia coli. *Appl. Environ. Microbiol.* doi:10.1128/AEM.06663-11.

Wilson, M. R., Jiang, Y., Villalta, P. W., Stornetta, A., Boudreau, P. D., Carrá, A., et al. (2019). The human gut bacterial genotoxin colibactin alkylates DNA. *Science (80-.).* 363. doi:10.1126/science.aar7785.

Wirbel, J., Pyl, P. T., Kartal, E., Zych, K., Kashani, A., Milanese, A., et al. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* 25. doi:10.1038/s41591-019-0406-6.

Xue, M., Kim, C. S., Healy, A. R., Wernke, K. M., Wang, Z., Frischling, M. C., et al. (2019a). Structure elucidation of colibactin. *bioRxiv* 1000, 1–31. doi:10.1101/574053.

Xue, M., Kim, C. S., Healy, A. R., Wernke, K. M., Wang, Z., Frischling, M. C., et al. (2019b). Structure elucidation of colibactin and its DNA cross-links. *Science (80-.).* 365. doi:10.1126/science.aax2685.

Yoshikawa, Y., Tsunematsu, Y., Matsuzaki, N., Hirayama, Y., Higashiguchi, F., Sato, M., et al. (2020). Characterization of colibactin-producing *Escherichia coli* isolated from Japanese patients with colorectal cancer. *Jpn. J. Infect. Dis.* doi:10.7883/yoken.jjid.2020.066.

Appendix

Table A1: Accession numbers and strain names of 530 *pks* positive NCBI genomes used in the study. (Genomes denoted in italics (*pks*p001 to *pks*p023) represents the newly sequenced genomes for the study. Genomes belonging to ST95 (n=110) are denoted with * in their IDs.)

ID	Strain	Assembly
pksp001	*NA147*	JADBJB000000000
pksp002	*NA150*	JADBJA000000000
pksp003	*NA258*	JADNRJ000000000
pksp004	*NA266*	JADBIZ000000000
pksp005	*NA280*	JADBIY000000000
pksp006	*NA310*	JADBIX000000000
pksp007	*NA334*	JADBIW000000000
pksp008	*NA336*	JADBIV000000000
pksp009	*NA608*	JADBIU000000000
pksp010	*NA611*	JADBIT000000000
pksp011	*NA623*	JADBIS000000000
pksp012	*NA651*	JADBIR000000000
pksp013	*NA664*	JADBIQ000000000
pksp014	*NA666*	JADBIP000000000
pksp015	*NA675*	JADBIO000000000
pksp016	*NA695*	JADBIN000000000
pksp017	*NA698*	JADBIM000000000
pksp018	*NA706*	JADBIL000000000
pksp019	*NA733*	JADBIK000000000
pksp020	*NA744*	JADBIJ000000000
pksp021	*NA749*	JADBII000000000
pksp022	*NA786*	JADBIH000000000
pksp023	*NA792*	JADBIG00000000
*pks*p024	83972	GCA_000159295.1
*pks*p025	MS 45-1	GCA_000164295.1
*pks*p026*	MS 110-3	GCA_000164415.1
*pks*p027	MS 153-1	GCA_000164435.1
*pks*p028	MS 200-1	GCA_000164535.1
*pks*p029	MS 185-1	GCA_000164575.1
*pks*p030	MS 60-1	GCA_000164595.1
*pks*p031	NC101	GCA_000179795.1
*pks*p032*	H263	GCA_000190915.1
*pks*p033*	H397	GCA_000241975.1
*pks*p034	9.1649	GCA_000194475.2

ID	Strain	Assembly
*pks*p035	LCT-EC106	GCA_000259695.1
*pks*p036	J96	GCA_000295775.2
*pks*p037	LCT-EC52	GCA_000331615.1
*pks*p038	LCT-EC59	GCA_000317395.1
*pks*p039	KTE15	GCA_000350745.1
*pks*p040	KTE16	GCA_000350765.1
*pks*p041	KTE39	GCA_000350865.1
*pks*p042	KTE187	GCA_000350945.1
*pks*p043	KTE188	GCA_000350965.1
*pks*p044	KTE189	GCA_000350985.1
*pks*p045	KTE191	GCA_000351005.1
*pks*p046	KTE201	GCA_000351045.1
*pks*p047	KTE205	GCA_000351085.1
*pks*p048	KTE206	GCA_000351105.1
*pks*p049	KTE214	GCA_000351205.1
*pks*p050	KTE220	GCA_000351245.1
*pks*p051	KTE224	GCA_000351265.1
*pks*p052	KTE230	GCA_000351305.1
*pks*p053	KTE53	GCA_000351485.1
*pks*p054*	KTE55	GCA_000351505.1
*pks*p055	KTE57	GCA_000351545.1
*pks*p056*	KTE58	GCA_000351565.1
*pks*p057	KTE60	GCA_000351585.1
*pks*p058	KTE67	GCA_000351645.1
*pks*p059	KTE72	GCA_000351665.1
*pks*p060	KTE86	GCA_000351805.1
*pks*p061	KTE87	GCA_000351825.1
*pks*p062	KTE93	GCA_000351845.1
*pks*p063	KTE169	GCA_000352025.1
*pks*p064	KTE8	GCA_000352085.1
*pks*p065	KTE43	GCA_000352225.1
*pks*p066*	KTE22	GCA_000352265.1
*pks*p067*	KTE59	GCA_000352365.1
*pks*p068	KTE63	GCA_000352385.1

ID	Strain	Assembly
*pk*sp069*	KTE65	GCA_000352405.1
*pk*sp070*	KTE118	GCA_000352545.1
*pk*sp071*	KTE123	GCA_000352565.1
*pk*sp072	KTE141	GCA_000352645.1
*pk*sp073	KTE183	GCA_000352905.1
*pk*sp074	KTE207	GCA_000353005.1
*pk*sp075	KTE209	GCA_000353025.1
*pk*sp076	KTE215	GCA_000353065.1
*pk*sp077	KTE218	GCA_000353105.1
*pk*sp078	KTE223	GCA_000353125.1
*pk*sp079*	KTE229	GCA_000353165.1
*pk*sp080	KTE104	GCA_000353185.1
*pk*sp081	KTE106	GCA_000326165.1
*pk*sp082	KTE124	GCA_000326225.1
*pk*sp083	KTE129	GCA_000326265.1
*pk*sp084	KTE131	GCA_000326285.1
*pk*sp085	KTE133	GCA_000326305.1
*pk*sp086	KTE137	GCA_000326325.1
*pk*sp087	KTE145	GCA_000326705.1
*pk*sp088	KTE153	GCA_000326385.1
*pk*sp089	KTE160	GCA_000326405.1
*pk*sp090	KTE167	GCA_000326905.1
*pk*sp091	KTE168	GCA_000326445.1
*pk*sp092	KTE174	GCA_000326605.1
*pk*sp093*	KTE179	GCA_000326485.1
*pk*sp094	KTE180	GCA_000326805.1
*pk*sp095	KTE85	GCA_000326885.1
*pk*sp096	KTE88	GCA_000326625.1
*pk*sp097	KTE97	GCA_000326545.1
*pk*sp098	KTE99	GCA_000326665.1
*pk*sp099	TOP379	GCA_000397225.1
*pk*sp100	TOP382-1	GCA_000397245.1
*pk*sp101	TOP382-2	GCA_000397265.1
*pk*sp102	TOP382-3	GCA_000397285.1
*pk*sp103	TOP291	GCA_000397305.1
*pk*sp104	TOP293-2	GCA_000397345.1
*pk*sp105	TOP498	GCA_000397405.1
*pk*sp106	TOP550-2	GCA_000397445.1
*pk*sp107	TOP550-3	GCA_000397465.1
*pk*sp108	TOP550-4	GCA_000397485.1
*pk*sp109	TOP2652	GCA_000397625.1
*pk*sp110	TOP2662-1	GCA_000397645.1

ID	Strain	Assembly
*pk*sp111	TOP2662-2	GCA_000397665.1
*pk*sp112	TOP2662-3	GCA_000397685.1
*pk*sp113	TOP2662-4	GCA_000397705.1
*pk*sp114	HM27	GCA_000387825.2
*pk*sp115	HM65	GCA_000387785.2
*pk*sp116	ATCC 25922	GCA_000401755.1
*pk*sp117	KTE182	GCA_000408065.1
*pk*sp118	KTE195	GCA_000408125.1
*pk*sp119	KTE226	GCA_000408285.1
*pk*sp120	KTE89	GCA_000408505.1
*pk*sp121*	HVH 1 (4-6876161)	GCA_000456005.1
*pk*sp122	HVH 2 (4-6943160)	GCA_000456025.1
*pk*sp123*	HVH 3 (4-7276001)	GCA_000456045.1
*pk*sp124	HVH 4 (4-7276109)	GCA_000456065.1
*pk*sp125	HVH 7 (4-7315031)	GCA_000456125.1
*pk*sp126*	HVH 12 (4-7653042)	GCA_000494955.1
*pk*sp127	HVH 13 (4-7634056)	GCA_000456185.1
*pk*sp128	HVH 16 (4-7649002)	GCA_000456205.1
*pk*sp129*	HVH 19 (4-7154984)	GCA_000456265.1
*pk*sp130	HVH 20 (4-5865042)	GCA_000456285.1
*pk*sp131	HVH 21 (4-4517873)	GCA_000456305.1
*pk*sp132	HVH 26 (4-5703913)	GCA_000456385.1
*pk*sp133	HVH 27 (4-7449267)	GCA_000456405.1
*pk*sp134	HVH 28 (4-0907367)	GCA_000456425.1
*pk*sp135*	HVH 30 (4-2661829)	GCA_000456465.1
*pk*sp136	HVH 31 (4-2602156)	GCA_000456485.1
*pk*sp137*	HVH 35 (4-2962667)	GCA_000456545.1

ID	Strain	Assembly
*pks*p138	HVH 37 (4-2773848)	GCA_000456565.1
*pks*p139	HVH 38 (4-2774682)	GCA_000456585.1
*pks*p140	HVH 39 (4-2679949)	GCA_000456605.1
*pks*p141	HVH 40 (4-1219782)	GCA_000456625.1
*pks*p142*	HVH 42 (4-2100061)	GCA_000456665.1
*pks*p143*	HVH 48 (4-2658593)	GCA_000456765.1
*pks*p144	HVH 51 (4-2172526)	GCA_000456785.1
*pks*p145	HVH 55 (4-2646161)	GCA_000456825.1
*pks*p146	HVH 56 (4-2153033)	GCA_000456845.1
*pks*p147	HVH 58 (4-2839709)	GCA_000456865.1
*pks*p148	HVH 61 (4-2736020)	GCA_000456905.1
*pks*p149	HVH 68 (4-0888028)	GCA_000456965.1
*pks*p150	HVH 74 (4-1034782)	GCA_000457045.1
*pks*p151*	HVH 76 (4-2538717)	GCA_000457065.1
*pks*p152	HVH 77 (4-2605759)	GCA_000457085.1
*pks*p153	HVH 78 (4-2735946)	GCA_000457105.1
*pks*p154	HVH 80 (4-2428830)	GCA_000457145.1
*pks*p155	HVH 86 (4-7026218)	GCA_000494975.1
*pks*p156	HVH 89 (4-5885604)	GCA_000457265.1
*pks*p157	HVH 92 (4-5930790)	GCA_000457325.1
*pks*p158	HVH 95 (4-6074464)	GCA_000457345.1
*pks*p159	HVH 96 (4-5934869)	GCA_000457385.1
*pks*p160	HVH 100 (4-2850729)	GCA_000457405.1

ID	Strain	Assembly
*pks*p161	HVH 103 (4-5904188)	GCA_000457435.1
*pks*p162	HVH 107 (4-5860571)	GCA_000457495.1
*pks*p163	HVH 109 (4-6977162)	GCA_000457515.1
*pks*p164	HVH 111 (4-7039018)	GCA_000457555.1
*pks*p165	HVH 112 (4-5987253)	GCA_000457575.1
*pks*p166	HVH 114 (4-7037740)	GCA_000457615.1
*pks*p167	HVH 116 (4-6879942)	GCA_000457675.1
*pks*p168	HVH 117 (4-6857191)	GCA_000457695.1
*pks*p169*	HVH 118 (4-7345399)	GCA_000457715.1
*pks*p170	HVH 120 (4-6978681)	GCA_000457755.1
*pks*p171	HVH 125 (4-2634716)	GCA_000457815.1
*pks*p172*	HVH 126 (4-6034225)	GCA_000457835.1
*pks*p173*	HVH 127 (4-7303629)	GCA_000457855.1
*pks*p174	HVH 128 (4-7030436)	GCA_000457875.1
*pks*p175	HVH 132 (4-6876862)	GCA_000457915.1
*pks*p176*	HVH 137 (4-2124971)	GCA_000457995.1
*pks*p177	HVH 138 (4-6066704)	GCA_000458015.1
*pks*p178	HVH 142 (4-5627451)	GCA_000458095.1
*pks*p179	HVH 143 (4-5674999)	GCA_000458115.1
*pks*p180	HVH 144 (4-4451937)	GCA_000458135.1
*pks*p181	HVH 149 (4-4451880)	GCA_000458215.1
*pks*p182	HVH 156 (4-3206505)	GCA_000458335.1
*pks*p183	HVH 157 (4-3406229)	GCA_000458355.1

ID	Strain	Assembly
*pks*p184	HVH 159 (4-5818141)	GCA_000458395.1
*pks*p185	HVH 160 (4-5695937)	GCA_000458415.1
*pks*p186	HVH 161 (4-3119890)	GCA_000458435.1
*pks*p187	HVH 169 (4-1075578)	GCA_000458535.1
*pks*p188	HVH 171 (4-3191958)	GCA_000458575.1
*pks*p189	HVH 172 (4-3248542)	GCA_000458605.1
*pks*p190	HVH 185 (4-2876639)	GCA_000458765.1
*pks*p191*	HVH 192 (4-3054470)	GCA_000458895.1
*pks*p192	HVH 197 (4-4466217)	GCA_000458995.1
*pks*p193*	HVH 199 (4-5670322)	GCA_000459035.1
*pks*p194	HVH 204 (4-3112802)	GCA_000459135.1
*pks*p195	HVH 207 (4-3113221)	GCA_000459195.1
*pks*p196*	HVH 210 (4-3042480)	GCA_000459255.1
*pks*p197*	HVH 211 (4-3041891)	GCA_000459275.1
*pks*p198	HVH 212 (3-9305343)	GCA_000459295.1
*pks*p199	HVH 213 (4-3042928)	GCA_000459315.1
*pks*p200	HVH 216 (4-3042952)	GCA_000459355.1
*pks*p201	HVH 218 (4-4500903)	GCA_000459395.1
*pks*p202	HVH 220 (4-5876842)	GCA_000459415.1
*pks*p203	HVH 225 (4-1273116)	GCA_000459495.1
*pks*p204	HVH 227 (4-2277670)	GCA_000459515.1
*pks*p205	HVH 228 (4-7787030)	GCA_000459535.1
*pks*p206	KOEGE 30 (63a)	GCA_000459615.1
*pks*p207*	KOEGE 32 (66a)	GCA_000459635.1
*pks*p208	KOEGE 43 (105a)	GCA_000459695.1
*pks*p209	KOEGE 44 (106a)	GCA_000459715.1
*pks*p210	KOEGE 56 (169a)	GCA_000459735.1
*pks*p211	KOEGE 58 (171a)	GCA_000459755.1
*pks*p212	KOEGE 61 (174a)	GCA_000459775.1
*pks*p213	KOEGE 70 (185a)	GCA_000459835.1
*pks*p214	UMEA 3014-1	GCA_000459955.1
*pks*p215	UMEA 3022-1	GCA_000459975.1
*pks*p216*	UMEA 3041-1	GCA_000460015.1
*pks*p217	UMEA 3053-1	GCA_000460055.1
*pks*p218	UMEA 3087-1	GCA_000460095.1
*pks*p219	UMEA 3088-1	GCA_000460115.1
*pks*p220	UMEA 3097-1	GCA_000460135.1
*pks*p221	UMEA 3113-1	GCA_000460175.1
*pks*p222	UMEA 3121-1	GCA_000460215.1
*pks*p223	UMEA 3122-1	GCA_000460235.1
*pks*p224	UMEA 3159-1	GCA_000460415.1
*pks*p225	UMEA 3161-1	GCA_000460455.1
*pks*p226	UMEA 3172-1	GCA_000460515.1
*pks*p227	UMEA 3173-1	GCA_000460535.1
*pks*p228	UMEA 3175-1	GCA_000460575.1
*pks*p229	UMEA 3178-1	GCA_000460615.1
*pks*p230	UMEA 3185-1	GCA_000460655.1
*pks*p231	UMEA 3193-1	GCA_000460695.1
*pks*p232	UMEA 3208-1	GCA_000460815.1
*pks*p233	UMEA 3215-1	GCA_000460855.1
*pks*p234	UMEA 3216-1	GCA_000460875.1
*pks*p235	UMEA 3217-1	GCA_000460895.1
*pks*p236	UMEA 3220-1	GCA_000460915.1
*pks*p237	UMEA 3221-1	GCA_000460935.1
*pks*p238	UMEA 3222-1	GCA_000460955.1
*pks*p239	UMEA 3230-1	GCA_000460975.1
*pks*p240	UMEA 3233-1	GCA_000460995.1
*pks*p241	UMEA 3244-1	GCA_000461035.1
*pks*p242	UMEA 3257-1	GCA_000461055.1

ID	Strain	Assembly
*pks*p243	UMEA 3264-1	GCA_000461075.1
*pks*p244	UMEA 3268-1	GCA_000461095.1
*pks*p245*	UMEA 3298-1	GCA_000461155.1
*pks*p246	UMEA 3337-1	GCA_000461275.1
*pks*p247	UMEA 3341-1	GCA_000461295.1
*pks*p248	UMEA 3391-1	GCA_000461335.1
*pks*p249	UMEA 3490-1	GCA_000461355.1
*pks*p250	UMEA 3585-1	GCA_000461375.1
*pks*p251	UMEA 3617-1	GCA_000461435.1
*pks*p252*	UMEA 3632-1	GCA_000461455.1
*pks*p253	UMEA 3652-1	GCA_000463605.1
*pks*p254	UMEA 3687-1	GCA_000461555.1
*pks*p255	UMEA 3694-1	GCA_000461575.1
*pks*p256	UMEA 3705-1	GCA_000461635.1
*pks*p257	UMEA 3707-1	GCA_000461655.1
*pks*p258	UMEA 3821-1	GCA_000461715.1
*pks*p259*	UMEA 3834-1	GCA_000461735.1
*pks*p260	UMEA 3955-1	GCA_000461815.1
*pks*p261	UMEA 4075-1	GCA_000461835.1
*pks*p262	UMEA 4076-1	GCA_000461855.1
*pks*p263	UMEA 4207-1	GCA_000461875.1
*pks*p264	907391	GCA_000488315.1
*pks*p265	907892	GCA_000488475.1
*pks*p266*	908675	GCA_000488755.1
*pks*p267	910096-2	GCA_000488795.1
*pks*p268	A25922R	GCA_000488815.1
*pks*p269	A35218R	GCA_000488835.1
*pks*p270	UMEA 3426-1	GCA_000488075.1
*pks*p271	UMEA 3290-1	GCA_000488095.1
*pks*p272	UMEA 3693-1	GCA_000488115.1
*pks*p273	UMEA 3342-1	GCA_000488155.1
*pks*p274	LAU-EC6	GCA_000506445.2
*pks*p275	HVH 23 (4-6066488)	GCA_000507605.1
*pks*p276	HVH 83 (4-2051087)	GCA_000507625.1
*pks*p277	JCM 5491	GCA_000614625.1
*pks*p278	Nissle 1917	GCA_000333215.1
*pks*p279*	A192PP	GCA_001245225.1
*pks*p280	7996-1	GCA_000699365.1
*pks*p281	UCD_JA17	GCA_000599745.2
*pks*p282	UCD_JA23	GCA_000599765.2

ID	Strain	Assembly
*pks*p283	BIDMC 83	GCA_000633655.1
*pks*p284	2009-46	GCA_000696545.1
*pks*p285	UCD_JA17_pb	GCA_000714915.1
*pks*p286	UCD_JA23_pb	GCA_000715035.1
*pks*p287*	SCB12	GCA_000817355.1
*pks*p288*	BIDMC 65	GCA_000692475.1
*pks*p289*	3-105-05_S4_C2	GCA_000700145.1
*pks*p290	4-203-08_S1_C1	GCA_000700705.1
*pks*p291	8-415-05_S4_C1	GCA_000711455.1
*pks*p292	8-415-05_S4_C2	GCA_000711365.1
*pks*p293	8-415-05_S4_C3	GCA_000711435.1
*pks*p294	4-203-08_S1_C2	GCA_000713945.1
*pks*p295	4-203-08_S1_C3	GCA_000713975.1
*pks*p296	8-415-05_S3_C3	GCA_000713455.1
*pks*p297	8-415-05_S3_C1	GCA_000713495.1
*pks*p298	8-415-05_S3_C2	GCA_000713585.1
*pks*p299	upec-98	GCA_000776315.1
*pks*p300	upec-93	GCA_000776695.1
*pks*p301	upec-91	GCA_000776855.1
*pks*p302	upec-9	GCA_000776795.1
*pks*p303	upec-87	GCA_000776745.1
*pks*p304	upec-85	GCA_000776455.1
*pks*p305	upec-84	GCA_000776215.1
*pks*p306	upec-80	GCA_000776035.1
*pks*p307*	upec-8	GCA_000776195.1
*pks*p308	upec-79	GCA_000776415.1
*pks*p309	upec-77	GCA_000776155.1
*pks*p310*	upec-76	GCA_000776235.1
*pks*p311*	upec-75	GCA_000776655.1
*pks*p312*	upec-73	GCA_000776375.1
*pks*p313	upec-7	GCA_000776175.1
*pks*p314	upec-65	GCA_000776615.1
*pks*p315*	upec-61	GCA_000776565.1

ID	Strain	Assembly
*pks*p316	upec-60	GCA_000776505.1
*pks*p317*	upec-51	GCA_000776965.1
*pks*p318	upec-48	GCA_000777025.1
*pks*p319	upec-40	GCA_000777135.1
*pks*p320	upec-39	GCA_000777165.1
*pks*p321	upec-38	GCA_000777195.1
*pks*p322	upec-36	GCA_000777215.1
*pks*p323	upec-289	GCA_000777415.1
*pks*p324	upec-288	GCA_000777435.1
*pks*p325	upec-287	GCA_000777455.1
*pks*p326	upec-285	GCA_000777495.1
*pks*p327	upec-277	GCA_000777605.1
*pks*p328	upec-276	GCA_000777625.1
*pks*p329	upec-261	GCA_000777845.1
*pks*p330	upec-260	GCA_000777895.1
*pks*p331	upec-258	GCA_000777975.1
*pks*p332*	upec-255	GCA_000778035.1
*pks*p333	upec-253	GCA_000778075.1
*pks*p334	upec-251	GCA_000778095.1
*pks*p335*	upec-250	GCA_000778105.1
*pks*p336*	upec-249	GCA_000778135.1
*pks*p337	upec-244	GCA_000778215.1
*pks*p338	upec-237	GCA_000778335.1
*pks*p339	upec-236	GCA_000778355.1
*pks*p340	upec-232	GCA_000778415.1
*pks*p341	upec-230	GCA_000778435.1
*pks*p342	upec-229	GCA_000776715.1
*pks*p343	upec-228	GCA_000778815.1
*pks*p344	upec-226	GCA_000778685.1
*pks*p345	upec-225	GCA_000778955.1
*pks*p346*	upec-209	GCA_000778465.1
*pks*p347	upec-201	GCA_000779255.1
*pks*p348*	upec-197	GCA_000779425.1
*pks*p349	upec-193	GCA_000779545.1
*pks*p350	upec-186	GCA_000779585.1
*pks*p351	upec-184	GCA_000779715.1
*pks*p352	upec-181	GCA_000779795.1
*pks*p353	upec-172	GCA_000779995.1
*pks*p354*	upec-169	GCA_000780095.1
*pks*p355	upec-166	GCA_000780115.1
*pks*p356	upec-161	GCA_000780155.1
*pks*p357	upec-158	GCA_000780195.1

ID	Strain	Assembly
*pks*p358*	upec-157	GCA_000780215.1
*pks*p359	upec-156	GCA_000780235.1
*pks*p360	upec-153	GCA_000780335.1
*pks*p361*	upec-144	GCA_000780595.1
*pks*p362	upec-140	GCA_000780735.1
*pks*p363	upec-14	GCA_000780675.1
*pks*p364*	upec-139	GCA_000780755.1
*pks*p365	upec-138	GCA_000780775.1
*pks*p366*	upec-136	GCA_000780695.1
*pks*p367	upec-135	GCA_000780795.1
*pks*p368*	upec-131	GCA_000780875.1
*pks*p369*	upec-129	GCA_000780925.1
*pks*p370*	upec-124	GCA_000781035.1
*pks*p371	upec-123	GCA_000781045.1
*pks*p372*	upec-120	GCA_000781095.1
*pks*p373	upec-117	GCA_000781175.1
*pks*p374	upec-115	GCA_000781215.1
*pks*p375	upec-109	GCA_000781355.1
*pks*p376*	upec-106	GCA_000781385.1
*pks*p377	upec-10	GCA_000785355.1
*pks*p378	blood-11-0041	GCA_000779495.1
*pks*p379*	blood-11-0031	GCA_000780275.1
*pks*p380	blood-10-1386	GCA_000779615.1
*pks*p381	blood-10-1310	GCA_000778765.1
*pks*p382	blood-10-1308	GCA_000778915.1
*pks*p383	blood-10-1126	GCA_000779125.1
*pks*p384	blood-10-1105	GCA_000779025.1
*pks*p385*	blood-10-0687	GCA_000781555.1
*pks*p386*	blood-10-0686	GCA_000781575.1
*pks*p387*	blood-09-0751	GCA_000782055.1
*pks*p388	blood-08-1203	GCA_000782635.1
*pks*p389	blood-08-0997	GCA_000782655.1
*pks*p390*	blood-08-0654	GCA_000782695.1
*pks*p391*	blood-08-0493	GCA_000782735.1
*pks*p392	blood-08-0379	GCA_000782755.1
*pks*p393	blood-08-0215	GCA_000782775.1
*pks*p394	UPEC_011	GCA_001651725.1
*pks*p395	UPEC_008	GCA_001651625.1
*pks*p396	UPEC_001	GCA_001651715.1
*pks*p397	GSK25213	GCA_000807565.1
*pks*p398	GSK2528	GCA_000807635.1
*pks*p399	SCB11	GCA_000817375.1

ID	Strain	Assembly
*pk*sp400	GSK2522	GCA_000807575.1
*pk*sp401	GSK2524	GCA_000807555.1
*pk*sp402	GSK252FU	GCA_000807655.1
*pk*sp403	GSK252BU	GCA_000800675.1
*pk*sp404	932_ECOL	GCA_001059575.1
*pk*sp405	696_ECOL	GCA_001057995.1
*pk*sp406	502_ECOL	GCA_001057065.1
*pk*sp407*	417_ECOL	GCA_001056665.1
*pk*sp408*	121_ECOL	GCA_001054095.1
*pk*sp409	1187_ECOL	GCA_001076105.1
*pk*sp410	11_ECOL	GCA_001052125.1
*pk*sp411*	RS218	GCA_000817345.1
*pk*sp412	VACI-14	GCA_001448025.1
*pk*sp413	M17 - 1	GCA_001010195.1
*pk*sp414*	LSPQ A134697	GCA_001262455.1
*pk*sp415	BWH59	GCA_001030285.1
*pk*sp416	MGH122	GCA_001030435.1
*pk*sp417	BIDMC97	GCA_001030445.1
*pk*sp418*	BIDMC114	GCA_001030665.1
*pk*sp419	UCD-JA09	GCA_001306575.1
*pk*sp420	UCD-JA19	GCA_001306585.1
*pk*sp421	UCD-JA30	GCA_001306685.1
*pk*sp422*	UCD-JA38	GCA_001306635.1
*pk*sp423	50639799	GCA_001463205.1
*pk*sp424	50870281	GCA_001463455.1
*pk*sp425	STEC 1528	GCA_001608125.1
*pk*sp426*	GN02005	GCA_001519135.1
*pk*sp427	GN02007	GCA_001519115.1
*pk*sp428	GN02009	GCA_001519125.1
*pk*sp429	GN02045	GCA_001519215.1
*pk*sp430*	GN02099	GCA_001519715.1
*pk*sp431	GN02137	GCA_001519675.1
*pk*sp432*	GN02148	GCA_001519755.1
*pk*sp433	GN02163	GCA_001519475.1
*pk*sp434*	GN02165	GCA_001519285.1
*pk*sp435*	GN02172	GCA_001519235.1
*pk*sp436	GN02183	GCA_001519315.1
*pk*sp437*	GN02254	GCA_001519735.1
*pk*sp438*	GN02260	GCA_001519555.1
*pk*sp439	GN02289	GCA_001521215.1
*pk*sp440	GN02314	GCA_001521195.1
*pk*sp441	GN02323	GCA_001519595.1

ID	Strain	Assembly
*pk*sp442	GN02350	GCA_001521225.1
*pk*sp443	GN02370	GCA_001520015.1
*pk*sp444	GN02392	GCA_001520055.1
*pk*sp445	GN02411	GCA_001520895.1
*pk*sp446	GN02529	GCA_001520215.1
*pk*sp447	GN02547	GCA_001521355.1
*pk*sp448*	GN02627	GCA_001520195.1
*pk*sp449	GN02639	GCA_001519645.1
*pk*sp450	GN02766	GCA_001520815.1
*pk*sp451	GN02787	GCA_001521155.1
*pk*sp452	GN02867	GCA_001521015.1
*pk*sp453*	GN03324	GCA_001521575.1
*pk*sp454	GN03398	GCA_001519485.1
*pk*sp455*	GN03409	GCA_001520775.1
*pk*sp456	GN03545	GCA_001520715.1
*pk*sp457	GN03661	GCA_001521115.1
*pk*sp458	GN03786	GCA_001521315.1
*pk*sp459*	GN04262	GCA_001521455.1
*pk*sp460	GN02748	GCA_001518355.1
*pk*sp461*	GN02487	GCA_001524905.1
*pk*sp462	UM149	GCA_001571585.1
*pk*sp463	UM131	GCA_001571575.1
*pk*sp464*	UM141	GCA_001571565.1
*pk*sp465	UC37	GCA_001571745.1
*pk*sp466	JPH264	GCA_001562835.1
*pk*sp467	sheep1	GCA_001615225.1
*pk*sp468	sheep6	GCA_001614495.1
*pk*sp469	sheep17	GCA_001616475.1
*pk*sp470	GN04499	GCA_001620985.1
*pk*sp471	GN04772	GCA_001621225.1
*pk*sp472	GN05109	GCA_001621345.1
*pk*sp473	GN05681	GCA_001621675.1
*pk*sp474	GN05963	GCA_001621885.1
*pk*sp475	GN05992	GCA_001621915.1
*pk*sp476	GN06113	GCA_001621995.1
*pk*sp477*	GN06168	GCA_001622105.1
*pk*sp478	NGF2	GCA_001683595.1
*pk*sp479	NGF3	GCA_001683585.1
*pk*sp480	NGF4	GCA_001683575.1
*pk*sp481	1409150006	GCA_001692775.1
*pk*sp482	1408270010	GCA_001692865.1
*pk*sp483	1512290008	GCA_001692805.1

ID	Strain	Assembly
*pks*p484	1512290026	GCA_001692785.1
*pks*p485	Fec 67	GCA_001865185.1
*pks*p486*	SF-384	GCA_001877815.1
*pks*p487*	SF-452	GCA_001877805.1
*pks*p488	No.12	GCA_001865915.1
*pks*p489	80//6	GCA_001865925.1
*pks*p490	B-11870	GCA_001865985.1
*pks*p491*	SF-491	GCA_001881225.1
*pks*p492	SF-495	GCA_001881235.1
*pks*p493*	SF-518	GCA_001881245.1
*pks*p494*	SF-522	GCA_001881055.1
*pks*p495	SF-523	GCA_001881275.1
*pks*p496*	SF-560	GCA_001881305.1
*pks*p497*	SF-567	GCA_001881315.1
*pks*p498*	SF-572	GCA_001881355.1
*pks*p499*	SF-596	GCA_001881345.1
*pks*p500*	SF-626	GCA_001881075.1
*pks*p501*	SF-095	GCA_001881385.1
*pks*p502*	SF-126	GCA_001881105.1
*pks*p503*	MVAST0098	GCA_001881125.1
*pks*p504*	MVAST0176	GCA_001881395.1
*pks*p505	MVAST0234	GCA_001881425.1
*pks*p506*	USVAST184	GCA_001881155.1
*pks*p507*	USVAST245	GCA_001881435.1
*pks*p508*	USVAST267	GCA_001881165.1
*pks*p509*	USVAST356	GCA_001881465.1
*pks*p510*	USVAST406	GCA_001881205.1
*pks*p511*	F-18	GCA_001854565.1
*pks*p512	CFT073	GCA_000007445.1
*pks*p513*	UTI89	GCA_000013265.1
*pks*p514	536	GCA_000013305.1
*pks*p515*	IHE3034	GCA_000025745.1
*pks*p516	ABU 83972	GCA_000148365.1
*pks*p517	UM146	GCA_000148605.1
*pks*p518	clone D i2	GCA_000233875.1
*pks*p519	clone D i14	GCA_000233895.1
*pks*p520	PMV-1	GCA_000493595.1
*pks*p521	Nissle 1917	GCA_000714595.1
*pks*p522	ATCC 25922	GCA_000743255.1
*pks*p523	RS218	GCA_000800845.2
*pks*p524*	SF-166	GCA_001280385.1
*pks*p525*	SF-173	GCA_001280405.1

ID	Strain	Assembly
*pks*p526	NGF1	GCA_001660585.1
*pks*p527	ECONIH2	GCA_001675145.1
*pks*p528	K-15KW01	GCA_001683435.1
*pks*p529	UPEC 26-1	GCA_001693315.1
*pks*p530	D8	GCA_001900395.1

Table A2: Isolate names and accession IDs of in-house *pks* negative genomes used *in silico* virulence/resistance gene profiling (Sl. No.s 1 to 23) and ST95 *pks* negative genomes from NCBI (Sl. No.s 24 to 72)

Sl. No.	ID	Isolate	Assembly
1.	NA1004	NA1004	GCA_002918085.1
2.	NA023	NA023	GCA_002224755.1
3.	NA057	NA057	GCA_002224705.1
4.	NA081	NA081	GCA_002224715.1
5.	NA101	NA101	GCA_002224795.1
6.	NA112	NA112	GCA_002224745.1
7.	NA447	NA447	GCA_002224785.1
8.	NA084	NA084	GCA_001713585.1
9.	NA086	NA086	GCA_001713575.1
10.	NA099	NA099	GCA_001713555.1
11.	NA703	NA703	GCA_001713545.1
12.	NA724	NA724	GCA_001713625.1
13.	NA114	NA114	GCA_000214765.3
14.	NA1001	NA1001	GCA_002918015.1
15.	NA1002	NA1002	GCA_002918005.1
16.	NA1003	NA1003	GCA_002918075.1
17.	NAEC6	NAEC6	GCA_002918095.1
18.	NAEC1	NAEC1	GCA_002918065.1
19.	NAEC2	NAEC2	GCA_002918175.1
20.	NAEC3	NAEC3	GCA_002918145.1
21.	NAEC4	NAEC4	GCA_002918155.1
22.	NAEC5	NAEC5	GCA_002918165.1
23.	NA090	NA090	GCA_002407385.1
24.	95N001	H252	GCA_000190895.1
25.	95N002	DSM 30083	GCA_000690815.1
26.	95N003	KTE4	GCA_000350645.1
27.	95N004	KTE5	GCA_000350665.1
28.	95N005	KTE62	GCA_000351605.1
29.	95N006	KTE3	GCA_000407685.1
30.	95N007	KTE7	GCA_000407705.1
31.	95N008	KTE27	GCA_000407885.1
32.	95N009	KTE240	GCA_000408305.1
33.	95N010	HVH 5 (4-7148410)	GCA_000456085.1
34.	95N011	HVH 32 (4-3773988)	GCA_000456505.1
35.	95N012	HVH 59 (4-1119338)	GCA_000456885.1
36.	95N013	HVH 73 (4-2393174)	GCA_000457025.1
37.	95N014	HVH 102 (4-6906788)	GCA_000465155.1
38.	95N015	HVH 104 (4-6977960)	GCA_000457455.1

Sl. No.	ID	Isolate	Assembly
39.	95N016	HVH 148 (4-3192490)	GCA_000495015.1
40.	95N017	HVH 170 (4-3026949)	GCA_000458555.1
41.	95N018	HVH 178 (4-3189163)	GCA_000495055.1
42.	95N019	HVH 180 (4-3051617)	GCA_000458685.1
43.	95N020	HVH 191 (3-9341900)	GCA_000458875.1
44.	95N021	HVH 201 (4-4459431)	GCA_000459075.1
45.	95N022	HVH 203 (4-3126218)	GCA_000459115.1
46.	95N023	HVH 217 (4-1022806)	GCA_000459375.1
47.	95N024	HVH 222 (4-2977443)	GCA_000459455.1
48.	95N025	UMEA 3140-1	GCA_000460295.1
49.	95N026	UMEA 3203-1	GCA_000460775.1
50.	95N027	UMEA 3206-1	GCA_000460795.1
51.	95N028	UMEA 3662-1	GCA_000461495.1
52.	95N029	UMEA 3702-1	GCA_000461595.1
53.	95N030	UMEA 3893-1	GCA_000461775.1
54.	95N031	597	GCA_000503475.1
55.	95N032	HVH 214 (4-3062198)	GCA_000507665.1
56.	95N033	AL505	GCA_001499595.1
57.	95N038	BIDMC 49b	GCA_000522365.1
58.	95N039	BIDMC 49a	GCA_000522385.1
59.	95N040	ATCC 11775	GCA_000734955.1
60.	95N041	upec-185	GCA_000779695.1
61.	95N042	50857972	GCA_001463405.1
62.	95N043	GN02476	GCA_001520875.1
63.	95N044	GN04665	GCA_001621085.1
64.	95N045	GN04676	GCA_001621125.1
65.	95N046	GN05696	GCA_001621665.1
66.	95N047	018PP2015	GCA_001700095.1
67.	95N048	SF-501	GCA_001881045.1
68.	95N049	MVAST0326	GCA_001881145.1
69.	95N034	SF-088	GCA_001280325.1
70.	95N035	SF-468	GCA_001280345.1
71.	95N036	APEC O1	GCA_000014845.1
72.	95N037	S88	GCA_000026285.1

Media Composition for microbiology experiments

M63 Media (1 Litre) (For biofilm formation assay)

Ammonium sulphate: 2.0 g

Potassium phosphate: 13.6 g/L

Ferrous sulphate: 0.5 mg

pH was adjusted to 7 and volume was made up to 1000 mL. The incomplete media so prepared was autoclaved and allowed to cool. 10 mL of 20% glycerol and 1mL 1M $MgSO_4$ was added.

King's Broth (1L) (For the preparation of chrome azurol S agar medium)

Peptone: 16 g

K_2HPO_4: 1.6 g

$MgSO_4$: 1.6 g

Glycerol: 10 mL

Luria Bertani Broth (For the growth and maintenance of *E. coli* cells)

25g for 1000mL (Himedia)

Ingredients (g/L)

Tryptone: 10g

Yeast Extract: 5g

Sodium Chloride: 10g

Final pH (at 25°C): 7.5±0.2

Luria Bertani Agar (For the growth and maintenance of *E. coli* cells)

40g for 1000mL (Himedia)

Ingredients (g/L)

Tryptone: 10g

Yeast Extract: 5g

Sodium Chloride: 10g

Agar: 15g

Final pH (at 25°C): 7.5±0.2

Mueller-Hinton Agar (Used for the determination of microbial susceptibility to antimicrobial agents)

38g for 1000mL (Himedia)

Ingredients (g/L)

Beef Extract: 300g

Casein acid hydrolysate: 17g

Starch: 1.5g

Agar: 17g

Final pH (at 25°C): 7.3±0.1

Materials for molecular biology experiments

Agarose gel (1.5%)

Agarose: 1.5 g

1x TAE Buffer: 100 ml

50x TAE Stock Solution (1 litre)

Tris Base: 242 g

Glacial Acetic Acid: 57.1 mL

0.5 M EDTA: 100 mL

Mix Tris with stir bar to dissolve in about 600 mL of ddH2O. Add the EDTA and Acetic Acid and bring the final volume to 1 L with ddH2O. Store at room temperature

1x TAE

50x TAE solution was diluted to 1X TAE using ddH2O.

6X DNA loading dye

50% Glycerol: 5 ml

0.1% Bromophenol Blue: 10 mg

0.5M EDTA solution: 2 ml

0.1 M Tris (pH 8.0): 0.1 mL

Distilled water 2.9 ml

TE Buffer (10 mMTris pH 8.0 with HCl, 1 mM EDTA)

Tris base: 1.21 g/L

EDTA $2H_2O$: 0.37 g/L (Dissolved in 800ml of ddH2O and pH adjusted to 8.0 with HCl)

Final volume adjusted to 1 liter with ddH2O

Reagents for Polymerase chain reaction (PCR)

10 x PCR Buffer

Taq DNA polymerase (New England Biolabs)

dNTPs PCR nucleotide Mix: dATP, dCTP, dGTP, dTTP(New England Biolabs)

Ethidium Bromide (20mg/mL)

Markers (Fermentas)

100 bp DNA ladder

1 kb DNA ladder

2-log DNA Ladder (0.1-10kb)

λ *Hind* III DNA Ladder

Publications

frontiers in Microbiology

ORIGINAL RESEARCH
published: 15 November 2018
doi: 10.3389/fmicb.2018.02631

Molecular Genetic and Functional Analysis of *pks*-Harboring, Extra-Intestinal Pathogenic *Escherichia coli* From India

*Arya Suresh[1], Amit Ranjan[1], Savita Jadhav[2], Arif Hussain[1], Sabiha Shaik[1], Munirul Alam[3], Ramani Baddam[3], Lothar H. Wieler[4] and Niyaz Ahmed[1,3]**

[1] Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, University of Hyderabad, Hyderabad, India, [2] Department of Microbiology, Dr. D. Y. Patil Medical College, Hospital and Research Centre (Dr. D. Y. Patil Vidyapeeth), Pune, India, [3] International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b), Dhaka, Bangladesh, [4] Robert Koch Institute, Berlin, Germany

OPEN ACCESS

Edited by:
Dongsheng Zhou, Beijing Institute of Microbiology and Epidemiology, China

Reviewed by:
Yasufumi Matsumura, Kyoto University, Japan
Maojun Zhang, National Institute for Communicable Disease Control and Prevention (China CDC), China
Shaofu Qiu, Institute of Disease Control and Prevention, Academy of Military Medical Sciences, China

***Correspondence:**
Niyaz Ahmed
niyaz.ahmed@uohyd.ac.in;
niyaz.ahmed@icddrb.org

Specialty section:
This article was submitted to Infectious Diseases, a section of the journal Frontiers in Microbiology

Received: *05 August 2018*
Accepted: *16 October 2018*
Published: *15 November 2018*

Citation:
Suresh A, Ranjan A, Jadhav S, Hussain A, Shaik S, Alam M, Baddam R, Wieler LH and Ahmed N (2018) Molecular Genetic and Functional Analysis of pks-Harboring, Extra-Intestinal Pathogenic Escherichia coli From India. Front. Microbiol. 9:2631. doi: 10.3389/fmicb.2018.02631

Colibactin, a genotoxin, encoded by the *pks* pathogenicity island of *Escherichia coli* belonging to the B2 phylogroup has been reported as a determinant of bacterial pathogenicity. The present study was carried out to detect the *pks* pathogenicity island in extraintestinal pathogenic *E. coli* (ExPEC) isolated from a tertiary hospital in Pune, India. Of 462 isolates analyzed, the *pks* genomic island was detected in 35 (7.6%) isolates, which predominantly belonged to pathogenic phylogroup B2 (97%), and harbored virulence genes such as *fimH*, *sfaD/E*, and *usp*. Biofilm formation assay revealed 21 of the 35 *pks*-carrying isolates to be strong (SBF $>$ 1.0), 10 isolates to be moderate (SBF = 0.5–1.0), and 4 as weak (SBF $<$ 0.5) biofilm formers. All of the *pks*-carrying isolates proved resistant against bactericidal activity of human serum. Assays carried out to detect antimicrobial susceptibility revealed 11% of these isolates to be multidrug resistant, 37% producing ESBL and 25% were positive for $bla_{CTX-M-15}$. The observed prevalence of multidrug resistance and colibactin producing characteristics among pathogenic *E. coli* belonging to phylogenetic group B2 advocate urgent need for broader surveillance in order to understand and prevent transmission of these ExPEC in community and hospital settings.

Keywords: genotoxins, *pks* island, colibactin, extraintestinal pathogenic *E. coli* (ExPEC), virulence

INTRODUCTION

Extraintestinal pathogenic *Escherichia coli* (ExPEC), apart from being a seasoned nosocomial pathogen is also an important cause of community-acquired (Johnson and Russo, 2002) and other infections such as urinary tract infection, sepsis, neonatal meningitis, and colibacillosis in humans and animals (Kaper et al., 2004; Smith et al., 2007). Genome analysis revealed that pathogenic *E. coli* have genome sizes in excess of about 1.0 Mb than the commensal strains mainly due to the presence of genes encoding multiple virulence factors, such as adhesins, toxins, invasins and siderophores that are absent or unlikely to be present in commensal strains (Croxen and Finlay, 2010). The virulence factors are mostly associated with phages and pathogenicity islands and undergo horizontal gene transfer which disseminates traits, thereby offering fitness advantages to recipient organisms (Ahmed et al., 2008).

Secreted toxins are the virulence factors that play an important role in long term colonization and pathogenesis of ExPEC. Some of these known secreted toxins comprise of important genotoxins, such as cytolethal distending toxins (CDT's), cycle inhibiting factors (*cif*'s) and cytotoxic necrotizing factors (CNF's) which can directly regulate the cell cycle of the host (Nougayrède et al., 2005). Colibactin is another important genotoxin produced by a 54-kb pathogenicity island known as *pks* island harbored by the members of *Enterobacteriaceae* (Nougayrède et al., 2006). This genomic island consists of a *clbA-S* gene cluster that encodes non-ribosomal peptides and polyketide synthases along with accessory and tailoring enzymes (Nougayrède et al., 2006). Colibactin acts as a cyclomodulin and blocks the eukaryotic cell cycle causing progressive enlargement of the nucleus as well as the cell body eventually leading to cell death (Cuevas-Ramos et al., 2010). The cytopathic effect of these genotoxins is mediated by live bacteria and requires a direct contact with the host cell (Nougayrède et al., 2006).

The *pks* island was first identified in sequenced genomes of ExPEC prototype strains and was detected predominantly in strains of phylogenetic group B2 (Johnson et al., 2008). The island was shown to display several signatures reminiscent of its horizontal acquisition (Putze et al., 2009; Messerer et al., 2017). The origin and prevalence of the colibactin island among enteric pathogens is largely unexplored. However, it was also found to be present in other members of *Enterobacteriaceae* such as *Citrobacter koseri, Klebsiella pneumoniae,* and *Enterobacter aerogenes* (Putze et al., 2009). Epidemiological studies demonstrated the prevalence of *pks* in ExPEC and their association with severe infections in different host populations of varied geographical locations (Johnson et al., 2008; Buc et al., 2013; Shimpoh et al., 2017).

Epidemiological studies on ExPEC in India have so far been focused on understanding antimicrobial resistance, evolution and presence of pandemic clones (Avasthi et al., 2011; Jadhav et al., 2011; Hussain et al., 2012, 2014; Shaik et al., 2017). However, a better understanding of ExPEC associated virulence factors may help in the development of therapeutic interventions, such as diagnostics and/or vaccines against ExPEC infections, and this would also facilitate risk assessment of ExPEC strains.

While colibactin is regarded as a virulence factor in ExPEC, not much is known in the context of its molecular epidemiology entailing Indian clinical isolates. Therefore, the present study was performed in order to investigate the prevalence and carriage of *pks* island, and to decipher the genotypic and functional characteristics of *pks* harboring clinical ExPEC isolates from India.

MATERIALS AND METHODS

Bacterial Isolates

A total of 462 isolates of ExPEC were harnessed for this study. These isolates were originally collected by SJ and her colleagues from Dr D. Y. Patil University Hospital, Pune, India as a part of their routine diagnostic screening during the years 2009–2015 and were also described in our previous study (Ranjan et al., 2016). The bacterial collection of our study, essentially a subset of the collection studied by Ranjan et al. (2016), comprised 370 isolates cultured from urine, 63 from pus and 29 from other extra intestinal clinical samples. Standard microbiological laboratory methods were employed for the identification and preservation of these isolates (Jadhav et al., 2011). All isolates were collected, preserved and handled as per standard biosafety guidelines and according to the approvals of the Institutional Biosafety Committee (IBSC) of the University of Hyderabad (Ref. UH/IBSC/NA/12/7 dated 09/4/2012 and NA-N-32 dated 27/8/2015). The clinical information of the isolates is described in the **Supplementary Table S1**.

Detection of *pks*-Genomic Island and Phylogroup Determination

The clinical isolates were screened for the presence of *pks* island by PCR using primers for the four representative genes of the genomic island so as to generate two flanking (*clbB* and *clbQ*) and two internal (*clbA* and *clbN*) amplicons in order to document presence of a complete island (Johnson et al., 2008). Heat killed bacterial lysates were used as DNA templates for PCR amplification, as described previously (Ranjan et al., 2017). PCR amplifications were carried out in 30 cycles at specific reaction conditions as described earlier (Johnson et al., 2008; Ranjan et al., 2017). The *pks*-positive isolates were assigned to one of the eight phylogroups based on multiplex-PCR amplification of four genes (*chuA, yjaA, arpA,* and *TspE4.C2*) as described elsewhere (Clermont et al., 2013).

Antibiotic Susceptibility and Extended Spectrum-β-Lactamase Production

Antibiotic susceptibility analysis was performed, as previously described, by Kirby-Bauer disk diffusion method, on Mueller Hinton agar plates (Qumar et al., 2017). Antimicrobial disks (Himedia, India) specific for fosfomycin (200 μg), chloramphenicol (30 μg), co-trimoxazole (20 μg), tetracycline (30 μg), gentamicin (10 μg), nalidixic acid (30 μg), doxycycline (30 μg), ciprofloxacin (5 μg), and colistin (10 μg) were used to determine the antibiotic susceptibility profile of the isolates. ESBL production was determined using disk synergy between clavulanic acid and indicator cephalosporins, CTX (cefotaxime) and CAZ (ceftazidime). Both the assays were performed in accordance with Clinical Laboratory Standards Institute (CLSI) guidelines (CLSI, 2013). Isolates exhibiting resistance to three or more antimicrobials were designated as multidrug resistant (MDR).

Virulence and Antimicrobial Resistance Genotyping

PCR based screening of virulence genes encoding bacterial adhesins (*fimH, sfaD/E, afa),* toxins (*usp, cvaC, sat),* iron acquisition system (*iucD)* and protectants (*ibeA)* were performed using primers and reaction conditions as previously described (Jadhav et al., 2011; Ranjan et al., 2017). ESBL gene $bla_{\text{CTX-M-15}}$, (Monstein et al., 2007), genes conferring resistance to tetracycline (*tetA),* sulfonamides (*sul1*) and aminoglycoside acetyl transferases

(*aac(6′)-Ib)* were also screened for using primers and PCR conditions as described in previous studies (Jadhav et al., 2011; Ranjan et al., 2016, 2017). The isolates were also screened by PCR using generic primers for TEM and amplified products were sequenced to identify the variants of the gene.

Determination of Siderophore Production, Biofilm Formation and Serum Resistance Assay

All the *pks*-positive isolates were screened for siderophore production using Chrome Azurol S Blue agar plates. A single colony of the bacterial isolate(s) was streaked on these plates and incubated overnight at 37°C. Colonies showing characteristic orange halos were identified as positive for siderophore production (Schwyn and Neilands, 1987).

All the 35 *pks*-positive isolates were analyzed for their biofilm forming capabilities as described previously (Nandanwar et al., 2014). Briefly, OD at 600 nm was taken for overnight grown bacterial cultures and all the isolates were diluted to an OD of 0.05 in fresh M63 minimal medium. An aliquot of 200 μL of the diluted culture was pipetted into flat-bottom 96 well sterile microtiter plates in triplicates. The plates were covered by a breathable sealing after obtaining OD at 600 nm [$OD_{600(0\ h)}$]. The plates were incubated for 48 h at stationary condition at 28°C. Following this, OD was obtained at 600 nm [$OD_{600(48\ h)}$]. Media was aspirated and wells washed thrice with 300 μL of deionized water. After air drying, bacteria were fixed using 250 μL of 99% methanol for 15 min and stained using 0.1% crystal violet solution for 30 min. Following staining, wells were washed thrice with deionized water and air dried. To solubilize the stained bacteria, 300 μL of Ethanol: Acetone (80:20) solution was added and incubated for 30 min at 100 rpm. OD at 570 nm was read in microtiter plate reader and specific biofilm formation was obtained using the following formula: SBF (specific biofilm formation) = (AB−CW)/G where AB = OD at 570 nm of attached and stained bacteria, CW = OD of control at 570 nm and G = $OD_{600(48\ h)} - OD_{600(0\ h)}$, representing bacterial growth. The experiment was repeated twice in technical triplicates.

Serum resistance was also determined for all the 35 *pks*-positive *E. coli* isolates *in vitro* using 50% human serum as described earlier (Hussain et al., 2014). Briefly, 5 μL of overnight culture was added to 495 μl of LB broth and incubated in a shaking incubator at 37°C for 1 h at 200 rpm. The bacterial cultures were pelleted and resuspended in 1mL of 1X PBS; 30 μL of this inoculum was added to 270 μL of 50% human serum in triplicates in a 96 well microtiter plate. In each case, an initial sample was collected and plated after dilution on LB agar plates for enumerating the colony forming units (CFU) at 0 h. The inoculated plate was incubated for 3 h at 37°C at 100 rpm. After 3 h, samples from each well were serially diluted and plated on LB agar plates for 3 h count. Isolates which had equal or higher CFU counts at 3 h compared to 0 h were considered resistant to human serum. Growth was obtained by subtracting CFU counts of 0 h from that of 3 h. The experiment was repeated two times in technical triplicate(s).

Statistical Analysis

All statistical calculations were performed using GraphPad Prism (version 5.01). Non-parametric Mann–Whitney *U* test was performed for serum resistance assay. *p*-values ≤ 0.05 were considered to be significant and were denoted in the graph.

RESULTS

Screening for *pks* Island and Phylogenetic Grouping

A total of 462 *E. coli* isolates from our collection were screened for the presence of *pks* island, of which 35 were found to be positive for all the four targeted genes (*clbA, clbB, clbN* and *clbQ)* amplified from flanking and internal regions. Of these, 30 were originally cultured from urine, four from pus and one from blood. The prevalence of *pks*-positive isolates was 7.6% of the total *E. coli* collection studied. Using multiplex PCR, identification of *E. coli* phylogenetic groups was performed and out of the 35 isolates, 34 belonged to phylogroup B2 while one was assigned to phylogroup D (**Table 1**).

Antimicrobial Susceptibility and ESBL Production

Antimicrobial susceptibility testing against nine different antimicrobial agents belonging to seven different antibiotic classes revealed that the isolates were only moderately resistant to antibiotics. Maximum resistance was observed against nalidixic acid (71.4%). The isolates were found to be less resistant to tetracycline (22.86%), co-trimoxazole (14.29%), doxycycline (11.43%), gentamicin (5.71%), and ciprofloxacin (5.71%) and were completely sensitive to fosfomycin, chloramphenicol, and colistin. Of the 35 *E. coli* isolates, 13 (37.14%) were found to be ESBL producers and 4 (11.42%) were multidrug resistant. The details of resistance profile(s) for each antibiotic tested are shown in **Table 1**.

Virulence and Resistance Genotyping

PCR based virulence and resistance genotyping of *pks*-positive isolates revealed their being relatively virulent as they possessed higher number of virulence genes compared to the resistance genes detected. Among bacterial adhesins tested, *fimH* and *sfaD/E* were 100% prevalent while *afa* was not detected in any of the isolates. Screening of toxin genes revealed presence of *usp* gene among all the isolates (100%), whereas *cvaC* and *sat* showed 42.86% and 34.29% prevalence, respectively. The genes *iucD* (iron acquisition system) and *ibeA* (protectant) were present in 51.43% and 31.43% of the isolates, respectively (**Table 1**).

Antimicrobial resistance genotyping revealed that the ESBL gene $bla_{CTX-M-15}$ was present in 25.71% ($n = 9$) of the isolates and all these isolates were phenotypically observed to be ESBL producers by double disk synergy test. Further, 22.86% ($n = 8$) of the isolates were positive for TEM, and sequencing of PCR

TABLE 1 | Phylogroups, virulence and resistance genotypes, and antimicrobial resistance of *pks*-positive *E. coli* isolates.

Genotypic characterization		No. (%) of positive isolates
Phylogenetic group		
B2		34 (97.14)
D		1 (2.86)
Virulence factors: Genotypic determinant		
Adhesins	*fimH*	35 (100)
	sfaD/E	35 (100)
	afa	0 (0)
Toxins	*usp*	35 (100)
	sat	12 (34.29)
	cvaC	15 (42.86)
Protectins	*ibeA*	11 (31.43)
Iron acquisition	*iucD*	18 (51.43)
Resistance factors: Antibiotic class		
Tetracyclines	*tetA*	2 (5.71)
Fluoroquinolones	*aac(6′)-Ib*	7 (20)
Sulfonamides	*sul1*	4 (11.43)
β-lactamases	bla_{TEM-1}	8 (22.86)
	$bla_{CTX-M-15}$	9 (25.71)
Antimicrobial class or phenotype	**Specific Drug**	**No. (%) of resistant isolates**
Aminoglycoside	Gentamicin	2 (5.71)
Tetracyclines	Tetracycline	8 (22.86)
	Doxycycline	4 (11.43)
Sulfonamide/trimethoprim	Co-trimoxazole	5 (14.29)
Phenicol	Chloramphenicol	0 (0)
Phosphonic acid derivative	Fosfomycin	0 (0)
Fluoroquinolone	Ciprofloxacin	2 (5.71)
	Nalidixic Acid	25 (71.43)
Antibacterial peptide	Colistin	0 (0)
Multidrug Resistance		4 (11.42)
ESBL		13 (37.14)

products followed by BLAST analysis identified all the amplicons to be entailing bla_{TEM-1}. The gene *aac(6′)-Ib*, which is involved in aminoglycoside resistance, was present in 20% of the *pks*-positive isolates. Occurrence of genes that confer sulfonamide (*sul1*) and tetracycline (*tetA*) resistance was observed to be at 11.43% and 5.71% of isolates, respectively (**Table 1**). Overall, we observed low prevalence of resistance genes in concordance with the phenotypic antibiotic resistances detected by disk diffusion assays.

Virulence Associated Phenotypes

The isolates were analyzed for siderophore production, biofilm formation, and serum resistance which are essential ExPEC virulence properties. All *pks*-positive isolates formed orange halos on Chrome Azurol S plates, confirming the production of siderophores. Biofilm formation assay was performed twice in triplicates for all the *pks* positive isolates in order to determine their biofilm forming capabilities and was documented by specific biofilm formation (SBF) values. Isolates showing SBF values greater than 1 were designated as strong biofilm formers, 0.5–1.0 as moderate biofilm formers and those showing less than 0.5 were considered as weak biofilm formers (**Figure 1**). Majority (21/35) of the isolates tested were strong biofilm formers while 10/35 isolates moderately formed the biofilm and 4/35 isolates were weak biofilm formers. Bacterial resistance to the bactericidal activity of human serum was also assessed and the number of CFU were found to be significantly higher for all the *pks* positive isolates after 3 h of incubation in the human serum as compared to the negative control, indicative of the fact that all the isolates were resistant to human serum. Serum resistance was performed twice in technical triplicate(s) (**Figure 2**) and *p*-values were obtained using Mann–Whitney *U* test.

DISCUSSION

The *pks* island encodes enzymes that are able to synthesize colibactin, a genotoxin that could induce host DNA damage and its presence may contribute to increased virulence and severe disease outcomes. Ever since the description of colibactin by Nougayrède et al. (2006), many studies were undertaken to develop a comprehensive understanding of this bacterial genotoxin (Bossuet-Greif et al., 2018; Faïs et al., 2018). However, epidemiological data on the prevalence of the same and factors

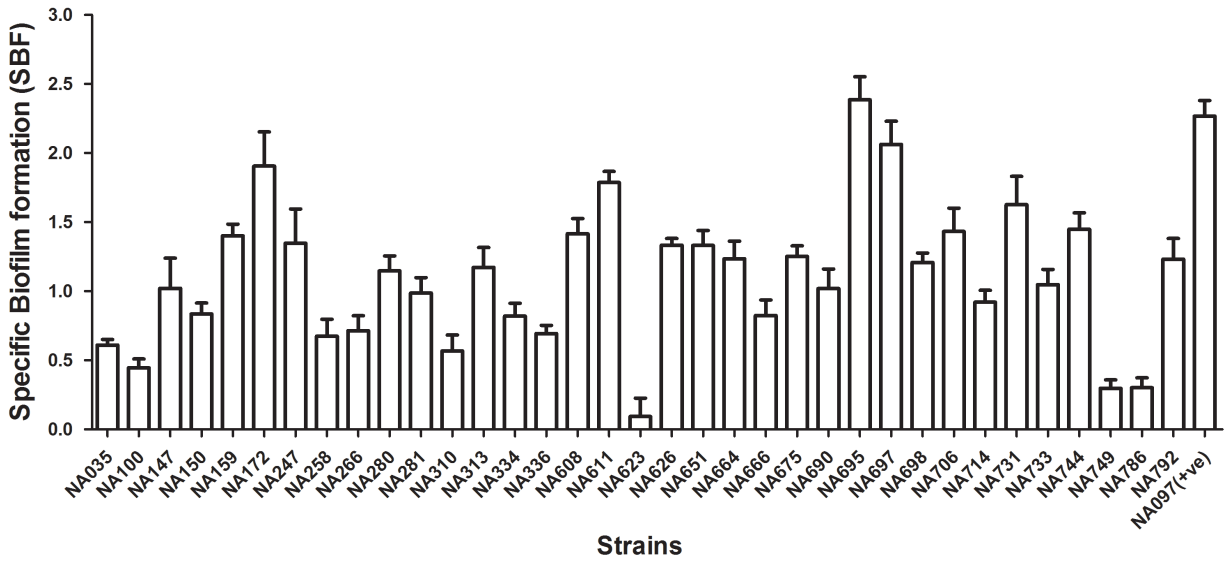


FIGURE 1 | Biofilm formation in 35 *pks* positive *E. coli* isolates in M63 medium. Values are shown as mean of specific biofilm formation. Isolates demonstrating SBF values > 1.0 were considered as strong, 0.5–1.0 as moderate and <0.5 as weak biofilm formers. Majority of isolates (21/35) demonstrated strong biofilm formation; the remaining 10 and 4 isolates showed moderate and weak biofilm formation, respectively. NA097 was employed as a positive control in the biofilm formation assay.

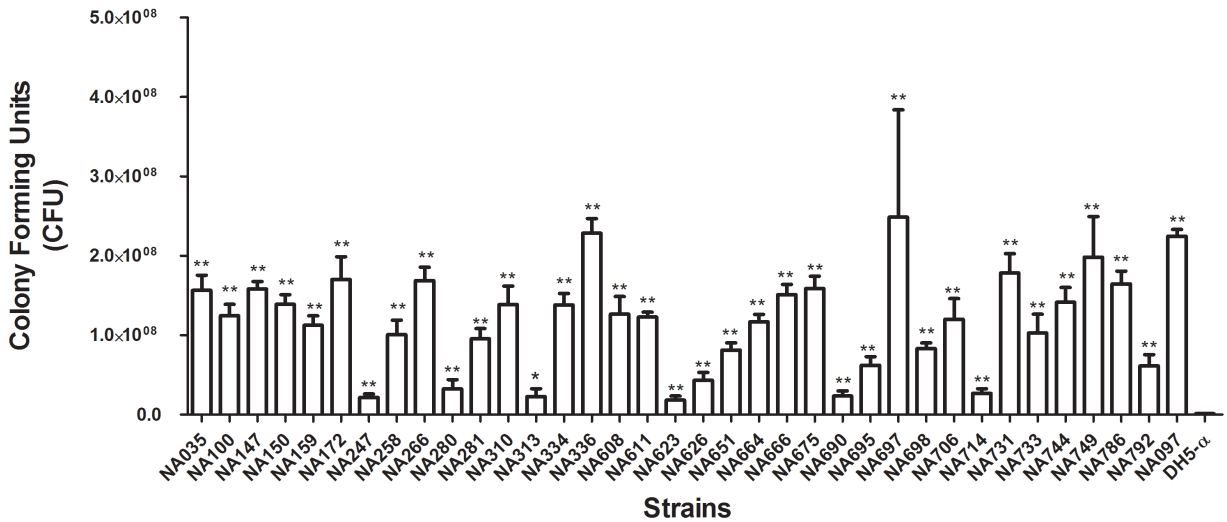


FIGURE 2 | Serum resistance assay of *pks*-positive isolates in human serum. Mann–Whitney *U* test was carried out for calculating the significant differences. Significant differences were indicated by asterisks and $p \leq 0.05$ was considered to be significant. *p-value ≤ 0.05, **p-value ≤ 0.01. NA097 was taken as the positive control, while DH5-α served as the negative control.

associated with *pks*-positive *E. coli* isolates, particularly from the Southern World have not been documented. The present study showed that 35 out of the 462 clinical ExPEC isolates were positive for all the four genes, indicating the presence of complete *pks* island(s) which might be able to synthesize functional colibactin. Thus, the overall prevalence was found to be 7.6% and this constitutes first epidemiological data on *pks* island harboring *E. coli* from India. In contrast, reports from other countries, such as those by Johnson et al. (2008) and Shimpoh et al. (2017) demonstrated high *pks* prevalence among clinical *E. coli*.

Our previous studies have suggested that majority of the *pks* negative clinical *E. coli* from India were genetically diverse and had a high prevalence of antibiotic resistance; these studies also identified and characterized the clonally evolving pandemic sequence type 131 *E. coli* isolates in India (Jadhav et al., 2011; Ranjan et al., 2016; Hussain et al., 2017; Shaik et al., 2017). However, the role of colibactin in the emergence of lineage specific virulence in *E. coli* that were comparatively less resistant to antibiotics has been shown herein for the first time from India. Such lineages could become a matter of public health concern and analysis of the underlying strains would be of great importance

given the high burden of infectious diseases in this region. Thus, the findings of this study have implications for better understanding of the epidemiological context of pathogenic *E. coli* in human diseases.

Escherichia coli harboring *pks* island was reported to be strongly associated with bacteremia and human colorectal tumors (Johnson et al., 2008; Buc et al., 2013). Recent reports suggest that mice infected with colibactin positive *E. coli* had significantly lower survival rates compared to those infected with isogenic colibactin-negative mutant(s) (Marcq et al., 2014). It has been further demonstrated that *pks*-positive *E. coli* infection induces cellular senescence and concurrently produces growth factors which promote tumor growth (Secher et al., 2013; Cougnoux et al., 2014; Dalmasso et al., 2015). The *pks*-positive *E. coli* isolates in the present study were detected in different specimen types including urine, blood and pus. Therefore, it can be surmised that the *pks*-positive *E. coli* might contribute in many invasive and non-invasive infections at different anatomic sites.

Previous studies reported that the *pks* island was majorly detected in *E. coli* phylogenetic group B2 strains, which are mainly documented as extraintestinal pathogens (Taieb et al., 2016; Sarshar et al., 2017). Our results were in line with these observations as the *pks*-positive isolates in our study also belonged predominantly to phylogroup B2 (97%), except for one isolate which belonged to phylogroup D (3%) (**Table 1**). The pathogenic strains of *E. coli* mainly belong to group B2 and, to a lesser extent, group D and frequently harbor higher number of virulence-factors than group A and group B1 strains (Picard et al., 1999; Johnson et al., 2008).

Several studies have demonstrated a strong correlation between the presence of virulence genes and the pathogenic spectrum of *E. coli* strains (Bien et al., 2012). These include multiple ExPEC-associated virulence genes such as adhesins, invasins, secretory toxins, and iron scavenging systems (Johnson, 1991). Accordingly, the virulent nature of *pks*-positive *E. coli* isolates was supported by our findings as 100% of them were positive for *fimH* (D mannose specific adhesin of minor fimbrial component), *sfaD/E* (s-fimbrial adhesin) and *usp* (uropathogenic specific protein) while $\geq$ 30% of them were positive for *iucD* (enzyme for siderophore aerobactin synthesis), *sat* (secreted autotransporter vacuolating cytotoxin), *cvaC* (colicin-V precursor) and *ibeA* (invasion protein). The afimbrial adhesin gene, *afa* was found to be completely absent (**Table 1**). These findings could be attributed to the pathogenic potential of the *pks*-positive strains along with the presence of many other virulence determinants. Furthermore, in the present study, siderophore production was detected in all *pks*-positive isolates; this observation was consistent with previous reports on the positive correlation between the presence of *pks* island and iron scavenging systems among the B2 *E. coli* strains (Martin et al., 2017). The localization of *pks* island within the High-Pathogenicity Island (HPI) and its physical association with siderophore biosynthesis gene cluster has also been described in other members of *Enterobacteriaceae* (Putze et al., 2009). A majority of *pks*-positive *E. coli* isolates demonstrated high biofilm forming capabilities in M63 medium (**Figure 1**) and all the *pks*-positive isolates tested were found to be resistant to serum bactericidal activity (**Figure 2**). We speculate that these genotypic and phenotypic virulence traits expressed by B2 *pks*-positive *E. coli* could act as fitness factors in order to colonize and initiate/establish infection in intestinal and extra-intestinal sites.

Antibiotic susceptibility of the *pks*-positive isolates was performed and the isolates were observed to be uniquely associated with low antimicrobial resistance, this finding is consistent with the previous reports which have suggested that *pks* island harboring *E. coli* exhibit reduced antibiotic resistance (Chen et al., 2017; Sarshar et al., 2017). This observation could likely be due to opportunist *E. coli* infections arising from intestinal microbiota. Additionally, detailed characterization of such isolates from fecal and clinical samples is needed to certainly understand the reason behind such an observation. Multidrug resistance and ESBL production was observed in only 11.42 and 37% isolates, respectively (**Table 1**). Previous studies from our group have shown MDR rates of 95% for clinical ST131 strains, and 91% for clinical and stool non-ST131 strains (Hussain et al., 2014). We have also previously reported MDR rates of 100% for metallo-(β-lactamase (MBL) producing *E. coli* isolates (Ranjan et al., 2016) and 67% for isolates from skin and soft tissue infections (Ranjan et al., 2017). The *pks* positive isolates in contrast demonstrated lesser rates of antimicrobial resistance compared to the *pks* negative isolates characterized from the similar settings in our previous studies. All nine $bla_{\mathrm{CTXM-15}}$ positive *E. coli* isolates were observed to be ESBL producers (**Table 1**), although further screening for other groups of CTX-M genes, and other ESBL classes are warranted. In concordance with the antimicrobial susceptibility results, molecular detection of antimicrobial resistance genes showed a low prevalence of *sul1* (sulphonamide resistance), *tetA* (tetracycline resistance), *aac(6′)-Ib* (aminoglycoside resistance), $bla_{\mathrm{CTX-M-15}}$ (extended spectrum-β-lactamases) and $bla_{\mathrm{TEM-1}}$ (broad spectrum-β-lactamases) genes (**Table 1**).

CONCLUSION

The prevalence of colibactin producing *E. coli* was found to be moderate among clinical *E. coli* isolates in our collection. These isolates harbored multiple virulence genes/traits and demonstrated relatively low antimicrobial resistance. These findings comprise essential baseline data required to understand aspects of functional molecular infection epidemiology of possibly genotoxic phenotypes of *E. coli* and their clinical significance. We hope to extend these studies at genomic and landscape scales to gain further insights into evolution, adaptation and dissemination of such isolates at clinical, community and ecosystem levels.

AUTHOR CONTRIBUTIONS

AS designed and performed all experiments with assistance from AR, SJ, AH, and SS. AS was responsible for all of the data and served as the guarantor for the manuscript. RB, MA, and LW participated in detailed discussions, advised on the interpretation

of some of the results and, contributed to the writing and editing of the manuscript. NA conceived the study and provided overarching supervision, laboratory facilities and resources, interpreted and discussed results, and wrote/edited the draft and final versions of the manuscript. All authors contributed to the development of the manuscript and its display items.

ACKNOWLEDGMENTS

AS acknowledges the junior research fellowship (JRF) from CSIR, India and would like to thank icddr,b for a short research stay support. We would like to thank all the members of Pathogen Biology Lab for their constructive comments on the manuscript and suggestions, especially to Nishant Nandanwar. icddr,b would like to thank its core donors: Sweden (SIDA), Bangladesh, Canada (CIDA and GAC), the United Kingdom (DFID).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.02631/full#supplementary-material

REFERENCES

Ahmed, N., Dobrindt, U., Hacker, J., and Hasnain, S. E. (2008). Genomic fluidity and pathogenic bacteria: applications in diagnostics, epidemiology and intervention. *Nat. Rev. Microbiol.* 6, 387–394. doi: 10.1038/nrmicro1889

Avasthi, T. S., Kumar, N., Baddam, R., Hussain, A., Nandanwar, N., Jadhav, S., et al. (2011). Genome of multidrug-resistant uropathogenic *Escherichia coli* strain NA114 from India. *J. Bacteriol.* 193, 4272–4273. doi: 10.1128/JB.05413-11

Bien, J., Sokolova, O., and Bozko, P. (2012). Role of uropathogenic *escherichia coli* virulence factors in development of urinary tract infection and kidney damage. *Int. J. Nephrol.* 2012:681473. doi: 10.1155/2012/681473

Bossuet-Greif, N., Vignard, J., Taieb, F., Mirey, G., Dubois, D., Petit, C., et al. (2018). The colibactin genotoxin generates DNA interstrand cross-links in infected cells. *mBio* 9:e02393-17. doi: 10.1128/mBio.02393-17

Buc, E., Dubois, D., Sauvanet, P., Raisch, J., Delmas, J., Darfeuille-Michaud, A., et al. (2013). High prevalence of mucosa-associated *E. coli* producing cyclomodulin and genotoxin in colon cancer. *PLoS One* 8:e56964. doi: 10.1371/journal.pone.0056964

Chen, Y.-T., Lai, Y.-C., Tan, M.-C., Hsieh, L.-Y., Wang, J.-T., Shiau, Y.-R., et al. (2017). Prevalence and characteristics of pks genotoxin gene cluster-positive clinical *Klebsiella pneumoniae* isolates in Taiwan. *Sci. Rep.* 7:43120. doi: 10.1038/srep43120

Clermont, O., Christenson, J. K., Denamur, E., and Gordon, D. M. (2013). The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environ. Microbiol. Rep.* 5, 58–65. doi: 10.1111/1758-2229.12019

CLSI (2013). *Performance Standards for Antimicrobial Disk and Dilution Susceptibility Tests for Bacteria Isolated From Animals; Approved Standard. VET01-A4*, 4 Edn. Wayne, PA: Clin. Lab. Stand. Inst.

Cougnoux, A., Dalmasso, G., Martinez, R., Buc, E., Delmas, J., Gibold, L., et al. (2014). Bacterial genotoxin colibactin promotes colon tumour growth by inducing a senescence-associated secretory phenotype. *Gut* 63, 1932–1942. doi: 10.1136/gutjnl-2013-305257

Croxen, M. A., and Finlay, B. B. (2010). Molecular mechanisms of *Escherichia coli* pathogenicity. *Nat. Rev. Microbiol.* 8, 26–38. doi: 10.1038/nrmicro2265

Cuevas-Ramos, G., Petit, C. R., Marcq, I., Boury, M., Oswald, E., and Nougayrède, J.-P. (2010). *Escherichia coli* induces DNA damage *in vivo* and triggers genomic instability in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.* 107, 11537–11542. doi: 10.1073/pnas.1001261107

Dalmasso, G., Cougnoux, A., Delmas, J., Darfeuille-Michaud, A., and Bonnet, R. (2015). The bacterial genotoxin colibactin promotes colon tumor growth by modifying the tumor microenvironment. *Gut Microbes* 5, 675–680. doi: 10.4161/19490976.2014.969989

Faïs, T., Delmas, J., Barnich, N., Bonnet, R., and Dalmasso, G. (2018). Colibactin: more than a new bacterial toxin. *Toxins* 10:151. doi: 10.3390/toxins10040151

Hussain, A., Ewers, C., Nandanwar, N., Guenther, S., Jadhav, S., Wieler, L. H., et al. (2012). Multi-resistant uropathogenic *Escherichia coli* from an endemic zone of urinary tract infections in India: genotypic and phenotypic characteristics of ST131 isolates of the CTX-M-15 extended-spectrum-beta-lactamase producing lineage. *Antimicrob. Agents Chemother.* 56, 6358–6365. doi: 10.1128/AAC.01099-12

Hussain, A., Ranjan, A., Nandanwar, N., Babbar, A., Jadhav, S., and Ahmed, N. (2014). Genotypic and phenotypic profiles of *Escherichia coli* isolates belonging to clinical sequence type 131 (ST131), clinical non-ST131, and fecal non-ST131 lineages from India. *Antimicrob. Agents Chemother.* 58, 7240–7249. doi: 10.1128/AAC.03320-14

Hussain, A., Shaik, S., Ranjan, A., Nandanwar, N., Tiwari, S. K., Majid, M., et al. (2017). Risk of transmission of antimicrobial resistant *Escherichia coli* from commercial broiler and free-range retail chicken in India. *Front. Microbiol.* 8:2120. doi: 10.3389/fmicb.2017.02120

Jadhav, S., Hussain, A., Devi, S., Kumar, A., Parveen, S., Gandham, N., et al. (2011). Virulence characteristics and genetic affinities of multiple drug resistant uropathogenic *Escherichia coli* from a semi urban locality in India. *PLoS One* 6:e18063. doi: 10.1371/journal.pone.0018063

Johnson, J. R. (1991). Virulence factors in *Escherichia coli* urinary tract infection. *Clin. Microbiol. Rev.* 4, 80–128. doi: 10.1128/CMR.4.1.80.Updated

Johnson, J. R., Johnston, B., Kuskowski, M. A., Nougayrede, J.-P., and Oswald, E. (2008). Molecular epidemiology and phylogenetic distribution of the *Escherichia coli* pks genomic island. *J. Clin. Microbiol.* 46, 3906–3911. doi: 10.1128/JCM.00949-08

Johnson, J. R., and Russo, T. A. (2002). Extraintestinal pathogenic *Escherichia coli*: "The other bad *E coli*". *J. Lab. Clin. Med.* 139, 155–162. doi: 10.1067/mlc.2002.121550

Kaper, J. B., Nataro, J. P., and Mobley, H. L. (2004). Pathogenic *Escherichia coli*. *Nat. Rev. Microbiol.* 2, 123–140. doi: 10.1038/nrmicro818

Marcq, I., Martin, P., Payros, D., Cuevas-Ramos, G., Boury, M., Watrin, C., et al. (2014). The genotoxin colibactin exacerbates lymphopenia and decreases survival rate in mice infected with septicemic *Escherichia coli*. *J. Infect. Dis.* 210, 285–294. doi: 10.1093/infdis/jiu071

Martin, P., Tronnet, S., Garcie, C., and Oswald, E. (2017). Interplay between siderophores and colibactin genotoxin in *Escherichia coli*. *IUBMB Life* 69, 435–441. doi: 10.1002/iub.1612

Messerer, M., Fischer, W., and Schubert, S. (2017). Investigation of horizontal gene transfer of pathogenicity islands in *Escherichia coli* using next-generation sequencing. *PLoS One* 12:e0179880. doi: 10.1371/journal.pone.0179880

Monstein, H. J., Östholm-Balkhed, Å, Nilsson, M. V., Nilsson, M., Dornbusch, K., and Nilsson, L. E. (2007). Multiplex PCR amplification assay for the detection of blaSHV, blaTEM and blaCTX-M genes in *Enterobacteriaceae*. *APMIS* 115, 1400–1408. doi: 10.1111/j.1600-0463.2007.00722.x

Nandanwar, N., Janssen, T., Kühl, M., Ahmed, N., Ewers, C., and Wieler, L. H. (2014). Extraintestinal pathogenic *Escherichia coli* (ExPEC) of human and avian origin belonging to sequence type complex 95 (STC95) portray indistinguishable virulence features. *Int. J. Med. Microbiol.* 304, 835–842. doi: 10.1016/j.ijmm.2014.06.009

Nougayrède, J.-P., Homburg, S., Taieb, F., Boury, M., Brzuszkiewicz, E., Gottschalk, G., et al. (2006). *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells. *Science* 313, 848–851. doi: 10.1126/science.1127059

Nougayrède, J.-P., Taieb, F., De Rycke, J., and Oswald, E. (2005). Cyclomodulins: bacterial effectors that modulate the eukaryotic cell cycle. *Trends Microbiol.* 13, 103–110. doi: 10.1016/j.tim.2005.01.002

Picard, B., Garcia, J. S., Gouriou, S., Duriez, P., Brahimi, N., Bingen, E., et al. (1999). The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection? *Infect. Immun.* 67, 546–553.

Putze, J., Hennequin, C., Nougayrède, J. P., Zhang, W., Homburg, S., Karch, H., et al. (2009). Genetic structure and distribution of the colibactin genomic island among members of the family Enterobacteriaceae. *Infect. Immun.* 77, 4696–4703. doi: 10.1128/IAI.00522-09

Qumar, S., Majid, M., Kumar, N., Tiwari, S. K., Semmler, T., Devi, S., et al. (2017). Genome dynamics and molecular infection epidemiology of multidrug-resistant *Helicobacter* pullorum isolates obtained from broiler and free-range chickens in India. *Appl. Environ. Microbiol.* 83:e02305-16. doi: 10.1128/AEM.02305-16

Ranjan, A., Shaik, S., Mondal, A., Nandanwar, N., Hussain, A., Semmler, T., et al. (2016). Molecular epidemiology and genome dynamics of New Delhi Metallo-β-Lactamase-producing extraintestinal pathogenic *Escherichia coli* strains from India. *Antimicrob. Agents Chemother.* 60, 6795–6805. doi: 10.1128/AAC.01345-16

Ranjan, A., Shaik, S., Nandanwar, N., Hussain, A., Tiwari, S. K., Semmler, T., et al. (2017). Comparative genomics of *Escherichia coli* isolated from skin and soft tissue and other extraintestinal infections. *mBio* 8:e01070-17. doi: 10.1128/mBio.01070-17

Sarshar, M., Scribano, D., Marazzato, M., Ambrosi, C., Aprea, M. R., Aleandri, M., et al. (2017). Genetic diversity, phylogroup distribution and virulence gene profile of pks positive *Escherichia coli* colonizing human intestinal polyps. *Microb. Pathog.* 112, 274–278. doi: 10.1016/j.micpath.2017.10.009

Schwyn, B., and Neilands, J. B. (1987). Universal chemical assay for the detection and determination of siderophores. *Anal. Biochem.* 160, 47–56. doi: 10.1016/0003-2697(87)90612-9

Secher, T., Samba-Louaka, A., Oswald, E., and Nougayrède, J. P. (2013). *Escherichia coli* producing colibactin triggers premature and transmissible senescence in mammalian cells. *PLoS One* 8:e77157. doi: 10.1371/journal.pone.0077157

Shaik, S., Ranjan, A., Tiwari, S. K., Hussain, A., Nandanwar, N., Kumar, N., et al. (2017). Comparative genomic analysis of globally dominant ST131 clone with other epidemiologically successful extraintestinal pathogenic *Escherichia coli* (ExPEC) lineages. *mBio* 8:e01596-17. doi: 10.1128/mBio.01596-17

Shimpoh, T., Hirata, Y., Ihara, S., Suzuki, N., Kinoshita, H., Hayakawa, Y., et al. (2017). Prevalence of *pks*-positive *Escherichia coli* in Japanese patients with or without colorectal cancer. *Gut Pathog.* 9:35. doi: 10.1186/s13099-017-0185-x

Smith, J. L., Fratamico, P. M., and Gunther, N. W. (2007). Extraintestinal pathogenic *Escherichia coli*. *Foodborne Pathog. Dis.* 4, 134–163. doi: 10.1089/fpd.2007.0087

Taieb, F., Petit, C., Nougayrède, J.-P., and Oswald, E. (2016). The Enterobacterial genotoxins: cytolethal distending toxin and colibactin. *EcoSal Plus.* doi: 10.1128/ecosalplus.ESP-0008-2016

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE
Clinical Science and Epidemiology

Evolutionary Dynamics Based on Comparative Genomics of Pathogenic *Escherichia coli* Lineages Harboring Polyketide Synthase (*pks*) Island

Arya Suresh,[a] Sabiha Shaik,[a] Ramani Baddam,[c] Amit Ranjan,[a] Shamsul Qumar,[a] Savita Jadhav,[b] Torsten Semmler,[c] Irfan A. Ghazi,[d] Lothar H. Wieler,[c] Niyaz Ahmed[a]

[a]Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, University of Hyderabad, Hyderabad, India
[b]Department of Microbiology, Dr. D. Y. Patil Medical College, Hospital and Research Centre (Dr. D. Y. Patil Vidyapeeth), Pune, India
[c]Robert Koch Institute, Berlin, Germany
[d]Department of Plant Sciences, University of Hyderabad, Hyderabad, India

ABSTRACT The genotoxin colibactin is a secondary metabolite produced by the polyketide synthase (*pks*) island harbored by extraintestinal pathogenic *E. coli* (ExPEC) and other members of the *Enterobacteriaceae* that has been increasingly reported to have critical implications in human health. The present study entails a high-throughput whole-genome comparison and phylogenetic analysis of such pathogenic *E. coli* isolates to gain insights into the patterns of distribution, horizontal transmission, and evolution of the island. For the current study, 23 *pks*-positive ExPEC genomes were newly sequenced, and their virulome and resistome profiles indicated a preponderance of virulence encoding genes and a reduced number of genes for antimicrobial resistance. In addition, 4,090 *E. coli* genomes from the public domain were also analyzed for large-scale screening for *pks*-positive genomes, out of which a total of 530 *pks*-positive genomes were studied to understand the subtype-based distribution pattern(s). The *pks* island showed a significant association with the B2 phylogroup (82.2%) and a high prevalence in sequence type 73 (ST73; $n = 179$) and ST95 ($n = 110$) and the O6:H1 ($n = 110$) serotype. Maximum-likelihood (ML) phylogeny of the core genome and intergenic regions (IGRs) of the ST95 model data set, which was selected because it had both *pks*-positive and *pks*-negative genomes, displayed clustering in relation to their carriage of the *pks* island. Prevalence patterns of genes encoding RM systems in the *pks*-positive and *pks*-negative genomes were also analyzed to determine their potential role in *pks* island acquisition and the maintenance capability of the genomes. Further, the maximum-likelihood phylogeny based on the core genome and *pks* island sequences from 247 genomes with an intact *pks* island demonstrated horizontal gene transfer of the island across sequence types and serotypes, with few exceptions. This study vitally contributes to understanding of the lineages and subtypes that have a higher propensity to harbor the *pks* island-encoded genotoxin with possible clinical implications.

IMPORTANCE Extraintestinal pathologies caused by highly virulent strains of *E. coli* amount to clinical implications with high morbidity and mortality rates. Pathogenic *E. coli* strains are evolving with the horizontal acquisition of mobile genetic elements, including pathogenicity islands such as the *pks* island, which produces the genotoxin colibactin, resulting in severe clinical outcomes, including colorectal cancer progression. The current study encompasses high-throughput comparative genomics and phylogenetic analyses to address the questions pertaining to the acquisition and evolution pattern of the genomic island in different *E. coli* subtypes. It is crucial to gain insights into the distribution, transfer, and maintenance of pathogenic islands, as they

Citation Suresh A, Shaik S, Baddam R, Ranjan A, Qumar S, Jadhav S, Semmler T, Ghazi IA, Wieler LH, Ahmed N. 2021. Evolutionary dynamics based on comparative genomics of pathogenic *Escherichia coli* lineages harboring polyketide synthase (*pks*) island. mBio 12: e03634-20. https://doi.org/10.1128/mBio.03634-20.

Editor Robert A. Bonomo, Louis Stokes Veterans Affairs Medical Center

Address correspondence to Niyaz Ahmed, niyaz.ahmed@uohyd.ac.in.

This article is a direct contribution from Niyaz Ahmed, a Fellow of the American Academy of Microbiology, who arranged for and secured reviews by Gerwald Koehler, OSU Center for Health Sciences; Sukhadeo Barbuddhe, ICAR National Research Centre on Meat, India; Philip Koshy, University of Malaya, Malaysia; and Dongsheng Zhou, Beijing Institute of Microbiology and Epidemiology, China.

Received 22 December 2020
Accepted 15 January 2021
Published 2 March 2021

harbor multiple virulence genes involved in pathogenesis and clinical implications of the infection.

KEYWORDS colibactin, *pks* island, polyketide synthase, genotoxins, *Escherichia coli*, *Escherichia* toxins, genomics, pathogenicity islands, phylogeny

Pathogenic *Escherichia coli* strains possess many different virulence factors in varied repertoires involved in subverting host cell mechanisms to enable persistence in otherwise protected environments of the host, with the capability to develop severe forms of pathogenesis that lead to high morbidity and mortality (1, 2). Mobile genetic element-enabled horizontal gene transfer (HGT), inactivation of antivirulence genes (3), and point mutation-derived functional alterations significantly contribute to the evolution of virulence in *E. coli* (2). Genes encoding different virulence factors, such as toxins, adhesins, iron acquisition systems, and capsules, could possibly be carried on or shuttled through mobile genetic elements, genomic islands, phages, and plasmids. These genes are capable of undergoing the horizontal gene transfer occurring among compatible organisms (4, 5) and could be abundantly distributed in extraintestinal pathogenic *E. coli* (ExPEC) strains. Genomic islands composed of large genomic regions (>10 kb), often flanked by repeat structures and carrying cryptic or functional mobility factors (integrases, transposases, etc.), display association with tRNA genes and possess distinct G+C contents (6). A subset of genomic islands called pathogenicity islands (PAIs) confer "quantum leaps" in the evolution of bacterial virulence by carrying numerous virulence-associated factors and enable adaptive evolution through horizontal gene transfer (7, 8). Colibactin is one such PAI-encoded genotoxic, nonribosomal peptide-polyketide secondary metabolite observed in uropathogenic, commensal, and neonatal meningitis-causing strains of *E. coli* (9). This metabolite was observed to induce double-stranded DNA breaks in eukaryotic cells, causing cell cycle arrest at the G_2-M phase and chromosomal aberrations (10, 11) and contributing to severe clinical manifestations like meningitis (12) and sepsis (9, 13).

Colibactin biosynthesis is carried out by an assembly line machinery located in the *pks* genomic island (54 kb) which consists of 19 genes comprising of nonribosomal peptide megasynthases (NRPS; *clbH*, *clbJ*, and *clbN*), polyketide megasynthases (PKS; *clbC*, *clbI*, and *clbO*), two hybrid NRPS-PKS (*clbB* and *clbK*), and nine accessory and tailoring enzymes (10). A recent study has described the regulatory role of *clbR*, a LuxR-type DNA-binding helix-turn-helix (HTH) domain as a key transcriptional activator involved in the expression of the colibactin biosynthetic gene cluster (14). The *pks* island, with an increased G+C content compared to the core genome, was reported to be integrated into the *asnW* tRNA locus and flanked by direct repeats of 16 bp together with P4-like bacteriophage integrase genes (10, 15). These integrative elements function to transfer genetic determinants to other members of *Enterobacteriaceae* (10, 15). The *pks* island is observed to be present in pathogenic, commensal, and even probiotic bacterial strains (16). It was also observed to be present in members of the *Enterobacteriaceae* other than *E. coli*, such as *Citrobacter koseri*, *Klebsiella pneumoniae*, and *Klebsiella aerogenes* (15). Colorectal cancer (CRC) biopsy samples were shown to display increased prevalence of the *pks* island-harboring *E. coli* (17, 18). *E. coli* isolates having *pks* islands were found in more than half of the patients with familial adenomatous polyps, and their colonic biofilms could enhance carcinogenesis through mucus degradation, followed by adherence and augmented colonization (19). In addition to their postulated role in CRC progression, numerous studies describe the *pks* islands as virulence factors with clinical implications entailing systemic infection, neonatal meningitis, and lymphopenia (12, 20–22).

So far, only a few studies have attempted to understand the pattern of transfer and evolution of the *pks* island and its coevolution with the genome that harbors it. Enterobacterial repetitive intergenic consensus (ERIC) and random amplified polymorphic DNA (RAPD)-based genetic fingerprinting of *pks*-positive *E. coli* isolates obtained

from human intestinal polyps showed diverse clustering patterns that implied their potential ability to colonize different environments (23). Another study performed bioinformatics analyses that unraveled the high prevalence of the *pks* island among *Escherichia* species, with close similarity of the *pks* island of *E. coli* with those of *K. aerogenes*, *K. pneumoniae*, and *C. koseri* (24). The combination of *in silico* and *in vitro* studies performed on the *Escherichia coli* Reference (ECOR) collection demonstrated that the immobile PAI group, i.e., those devoid of any transfer or mobility regions, comprising of high-pathogenicity island (HPI), *pks*, and *serU*, undergoes horizontal gene transfer "*en bloc*" along with the neighboring chromosomal backbone; this was observed to be F′-mediated transfer (25). The high homology within *pks* island sequences also conveyed the recent acquisition of the *pks* island (25). We attempted to employ a large-scale pangenome and phylogenetic analysis to comprehensively study and contribute insights to the distribution and evolutionary dynamics of this pathogenic island of clinical significance. The prevalence of *pks* island among ExPEC isolates from India and their genetic and functional characterization have been previously described by our group (26). The present study aims at describing the genome-wide comparisons and phylogenetic analysis of the *pks* island-carrying *E. coli* isolates from a previously described in-house collection, as well as the genome data obtained from the public domain. The study describes the distribution of *pks* island-harboring *E. coli* among phylogroups, sequence types, and serogroups, followed by pangenome and phylogenetic analyses with particular reference to genomes belonging to sequence type 95 (ST95) to understand the evolution and acquisition of this island. Phylogenetic analyses have also been performed to study the fine structure of island evolution with respect to the core genome and to understand the pattern of transfer and acquisition of the island. A preliminary study on the potential role of the distribution pattern of restriction modification systems towards the successful HGT and maintenance of *pks* islands has also been performed. We have employed large scale, whole-genome-based investigations for understanding the pathogenic *pks* islands with respect to their patterns of prevalence or preponderance and evolution among *E. coli* populations.

RESULTS

Genome characteristics. Whole-genome sequencing of 23 *pks*-positive *E. coli* isolates that were previously characterized (26) was performed in the current study. The genomes showed an approximate size of 5.1 Mb with an average G+C content of 50.4%. The average number of coding sequences (CDS) was ~5,000, displaying a coding percentage of 87%. The 23 *pks*-positive genomes analyzed here for the first time revealed distribution among the following different sequence types: ST12 ($n = 6$), ST73 ($n = 4$), ST827 ($n = 3$), ST14 ($n = 3$), ST998 ($n = 3$), ST1057 ($n = 2$), ST83 ($n = 1$), and ST127 ($n = 1$). The assembly statistics and genome sequence characteristics are summarized in Tables S1 and S2 in the supplemental material. The GenBank accession numbers of the 23 newly sequenced genomes have also been listed in Table S2. Whole-genome comparison of the 23 in-house *pks*-positive genomes was performed using BLAST Ring Image Generator (BRIG) (27) with the complete genome of strain IHE3034 as the reference (see Fig. S1a in the supplemental material). Results from the BRIG analysis indicated that the genomes shared a high degree of similarity, and variable regions were mostly identified as phages (denoted as black arcs). The *pks* island (denoted as a red arc) was also found to be conserved throughout the genomes. The island sequences reconstructed from the respective genomes were used as the query, along with the *pks* island sequence from IHE3034 as the reference in BRIG (27) (Fig. S1b). The island sequence was also annotated, and the individual genes of the island are depicted in the outermost ring (Fig. S1b). It was observed that the island sequences showed a high degree of conservation among the genomes, with variations only in the flanking regions in a few cases.

Virulome and resistome profiling of in-house *pks*-positive genomes. The in-house *pks*-positive genomes were screened to identify the prevalence of various virulence associated and antibiotic resistance conferring gene coordinates to determine the pathogenic potential of the corresponding ExPEC isolates. *In silico* virulence

profiling using the Virulence Factor Database (VFDB) (28) showed these *pks*-positive isolates to have an abundance of adherence factors, type VI secretion systems, and siderophores, as depicted in the heat map (Fig. 1). Among the category of adherence genes, the *csgABCDEFG* gene complex involved in curli fiber production, assembly and transport, *E. coli* common pilus genes (*ecpABCDE*), and the type I fimbrial protein genes *fimABCDEFGHI* were found distributed in most of the genomes. In addition, peritrichous flagellar proteins (encoded by *flg*, *fli*, and *flh*), flagellar motor proteins (*motA* and *motB*), and chemotaxis proteins (*cheABRWYZ*) were also found to be present in the majority of the genomes. The invasin protein genes *ibeB*, *ibeC* and *tia* were found in most of the 23 isolates studied. Among secretion systems, genes coding for type VI secretion systems (98 out of the total of 111 genes belonging to the category of secretion systems) were observed in the greatest abundance, followed by genes encoding general secretory pathway proteins (*gspCDEFGHIJKLM*). Among the type VI secretion systems, *aec7, aec16, aec 17, aec 18, aec 19, aec 23, aec 24, aec 25, aec 26, aec 27, aec 28, aec 29, aec 30, aec 31, aec 32*, *c3386*, *c3401*, *c3402*, and *ECABU_c310170* were present in 18 or more genomes out of the 23 *pks*-positive genomes. Yersiniabactin siderophore system genes *ybtAEXPQRSTU*, *irp1*, and *irp2* were found to be present in all the isolates. Most of the genomes harbored other siderophore systems like *chuASTUVWXY*, enterobactin synthase genes *entABCDEFS*, ferrienterobactin transporter genes *fepABCDEG*, enterobactin esterase gene *fes*, and salmochelin genes *iroBCDEN*. It was also observed that 17/23 genomes harbored the aerobactin siderophore synthesis system genes, *iucABCD* and *iutA*. Among toxin genes, hemolysin-encoding genes *hlyABCD*, the uropathogenic-specific protein gene *usp*, and the hemoglobin protease gene *vat* were present in ~23 genomes. In addition, the cyclomodulin cytotoxic necrotizing factor gene *cnf-1* was present in 10/23 isolates. Analysis using VFDB also confirmed the presence of *pks* island genes in all of the 23 genomes, indicating the integrity of the island in the genomes (Fig. 1). The comparison of the virulence profile of the in-house *pks*-positive genomes with that of the in-house *pks*-negative genomes has been described in Table S3 in the supplemental material.

In silico antimicrobial gene profiling revealed that the majority of the resistance genes carried by *pks*-positive, in-house isolates belonged to the nonspecific antibiotic efflux pumps category (Fig. 2). The majority of the efflux pumps, including the aminoglycoside efflux pump (*acr*), two-component regulatory system (*baeSR*), global regulator (*CRP*), electrochemical gradient-powered transporter *emr*, and multiple antibiotic resistance family *mar*, were found to be prevalent in most of the genomes. The multidrug efflux system *mdt*, coupled with *gadX* and *gadW*, which offer resistance to penams, fluoroquinolones, and macrolides, were also observed in most *pks*-positive isolates. In the category of antibiotic inactivation, *ampC*, a class C beta-lactamase that encodes resistance against penicillins and cephalosporins, was also found to be present in all the isolates. Other beta-lactamases like CTX-M-15 ($n = 4$), OXA-1 ($n = 3$), and TEM-1 ($n = 5$) were detected in a few isolates. Antibiotic target replacement genes like the bacitracin resistance gene *bacA* and the coordinates from the gene family encoding phosphoethanolamine transferase (*ugd* or *pmrE*, *eptA* or *pmrC*, and *pmrF*) offering resistance against cationic antimicrobial peptides were found distributed in all the genomes (Fig. 2). The comparison of the resistance profile of the in-house *pks*-positive genomes with that of the in-house *pks*-negative genomes has been described in the Table S4 in the supplemental material.

Prevalence and distribution of *pks*-positive *E. coli*. A total of 4,113 genomes of *E. coli* were analyzed, of which 306 were complete and 3,753 were draft genomes downloaded from NCBI; 31 genomes were in-house or sequenced as a part of previous studies, whereas 23 genomes, as described, were the newly sequenced genomes taken for the present work. A total of 530 genomes were found positive for the presence of *pks* island genes and were designated with in-house identifiers (IDs) (*pks*p001 to *pks*p530) (the genome list, in-house IDs, and the accession numbers of these genomes obtained from NCBI and used for further analyses have been described in Tables S5 and S6 in the

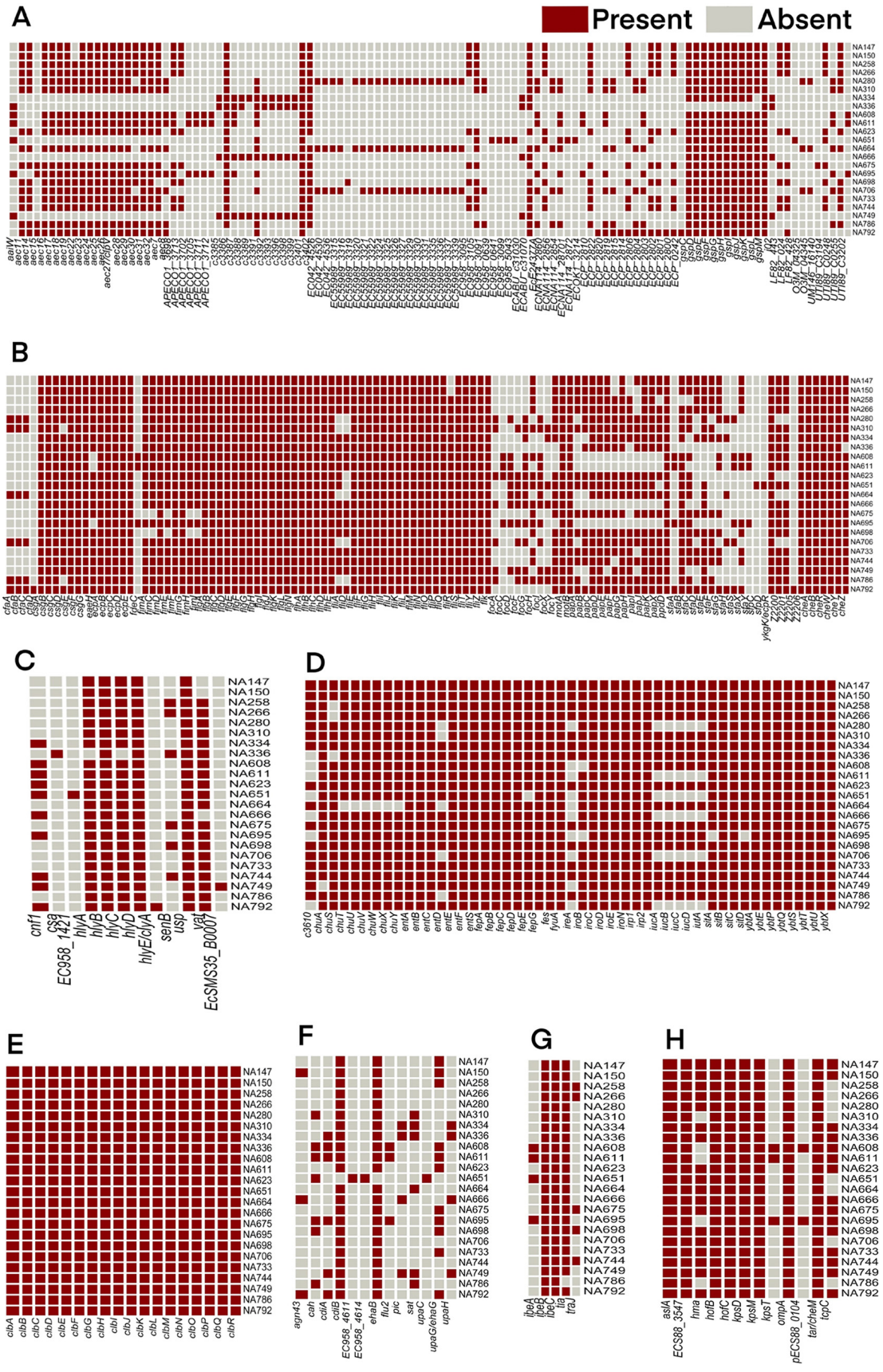


FIG 1 Heat map depicting the virulence profile of 23 in-house *pks*-positive isolates, depicting the presence and absence of 333 virulence genes belonging to different categories. (A) Secretory system, (B) adherence factors, (C) toxins, (D)

(Continued on next page)

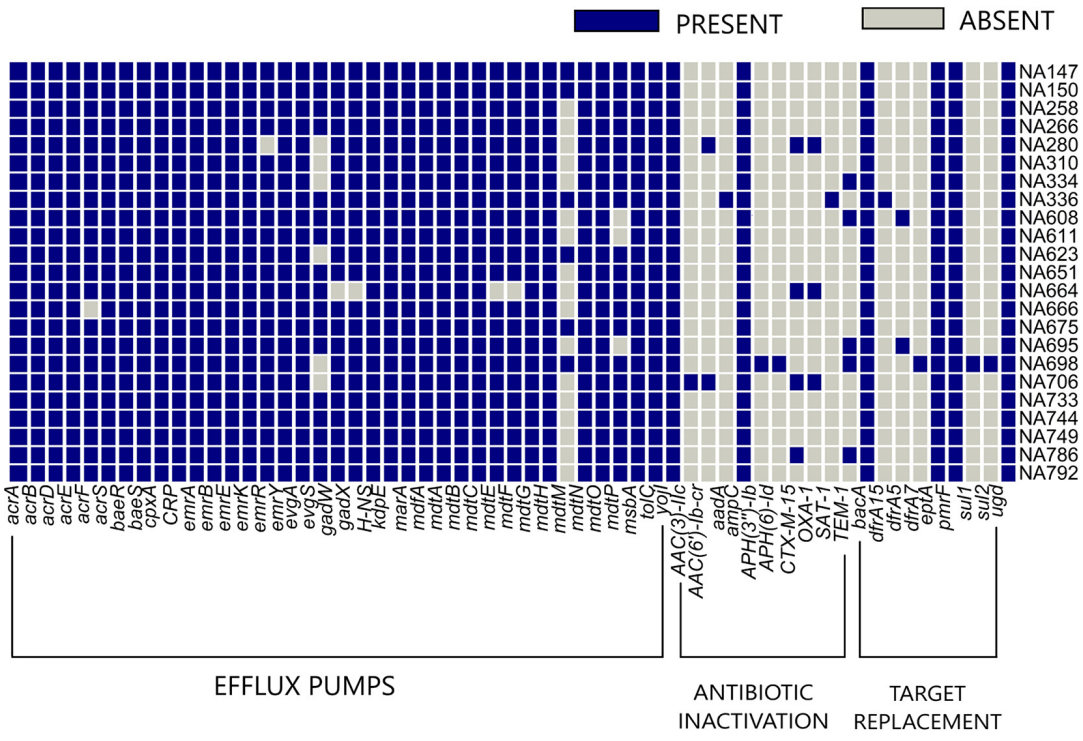


FIG 2 Heat map showing the presence and absence of 57 antibiotic resistance genes in 23 in-house *pks*-positive genomes. Blue and gray boxes indicate the presence and absence of the resistance gene, respectively. Gene names are represented on the *x* axis and isolate names on the *y* axis. The genes have also been categorized according to the mechanism of action against antimicrobials (gene names correspond to columns from left to right; they may be read in a sequential reading frame beginning *acrA* aligning to the left most column to *ugd* aligning with the right most column, in the heat map).

supplemental material). Among the 530 genomes, 247 genomes carried all 19 genes of the island, while 184 and 80 genomes carried 18 and 17 genes, respectively. The remaining 19 genomes carried fewer than 17 genes. Out of the 530 *pks*-positive genomes, the island (54 kb in size) was observed to be present in a single contig in 247 genomes. All of the *pks*-positive genomes carried 14 to 19 *pks* island genes, and the rest of the genomes did not harbor any of the *pks* island genes and hence were designated *pks*-negative.

In silico multilocus sequence typing (MLST) revealed that there is a higher prevalence of the *pks* genomic island in sequence types ST73 ($n = 179$) and ST95 ($n = 110$), followed by ST127 ($n = 52$) and ST12 ($n = 48$) (Table 1). Interestingly, all ST73 isolates were observed to harbor the *pks* island, and none of the genomes belonging to the highly successful clonal group ST131 carried the island sequence. The percent prevalence of *pks*-positive genomes among sequence types ST95, ST127, and ST12 was 69.18% (110/159), 78.78% (52/66), and 97.9% (48/49), respectively. *In silico* phylogrouping revealed that the majority of the isolates (82%) belonged to the B2 phylogroup, indicating a strong association of the *pks* island-harboring *E. coli* isolates with the B2 phylogroup (Table 1), similar to the observations from PCR-based phylogrouping of the in-house isolates in our previous study, which showed that the majority of the isolates belonged to the B2 phylogroup (97%) (26).

In silico determination of serotypes using ECTyper displayed a higher prevalence of *pks*-positives in certain serotypes. The O6:H1 serotype was shown to have the highest number of *pks*-positive genomes ($n = 110$), followed by O6:H31 ($n = 48$), O4:H5 ($n = 46$), and O18:H7 ($n = 40$). The prevalence pattern of *pks*-positive genomes in different sero-

FIG 1 Legend (Continued)
siderophores/iron acquisition systems, (E) *pks* island genes, (F) autotransporters, G) invasins, and (H) others (genes corresponding to columns, from left to right, may be read in a sequential reading frame, such that each gene name aligns correctly with a single, corresponding column).

TABLE 1 Sequence type, phylogroup, and serotype distribution of *pks*-positive genomes ($n = 530$) obtained from NCBI

Subtype	% (no.)
Phylogroup	
B2	81.69 (433)
A	0.56 (3)
Unknown	17.73 (94)
Sequence type	
ST73	33.8 (179)
ST95	20.7 (110)
ST127	9.8 (52)
ST12	9.1 (48)
ST141	3.6 (19)
ST998	3.02 (16)
ST404	2.07 (11)
ST80	1.7 (9)
Miscellaneous	12.6 (67)
Unknown	3.6 (19)
Serogroup	
O6:H1	20.7 (110)
O6:H31	9 (48)
O4:H5	8.6 (46)
O18:H7	7.5 (40)
O2:H6	6.8 (36)
O1:H7	6.4 (34)
O2:H1	6 (32)
O2:H7	6 (32)
O75:H5	4.3 (23)
O22:H1	3.7 (20)
O4:H1	3.7 (20)
O25:H1	3.2 (17)
O2:H4	3 (16)
O18:H1	2.2 (12)
Miscellaneous	7.5 (40)

types has been described in Table 1. Serotypes and sequence types with fewer than 10 genomes were grouped as "miscellaneous."

Pangenome analysis of ST95 genomes. The ST95 group was observed to have both *pks*-positive and *pks*-negative genomes and thus was considered a suitable model data set in this study. Comparison between the *pks*-positives and *pks*-negatives from ST95 could help in providing insights into the potential acquisition and maintenance of *pks* island. A total of 3,057 genes constituted the core of 159 ST95 genomes, which included 110 *pks* positives and 49 *pks* negatives. These genes were subjected to clusters of orthologous groups (COG) classification using EggNOG (29), where 2,337 out of 3,057 genes were assigned to different COG classes, and the results are depicted in Table S7 in the supplemental material.

Core genome phylogeny of ST95. Core genome maximum-likelihood (ML) phylogeny obtained from IQ-TREE (30) consisted of 5 different clades, where green branches denote *pks*-positives and red ones denote *pks*-negatives (Fig. 3). Clades I and II were observed to comprise both *pks*-positive and *pks*-negative genomes with mixed clading pattern(s). Clades III and V were found to consist predominantly of *pks*-positive genomes, except for one *pks*-negative genome each, whereas clade IV consisted of only *pks*-negative genomes. The distinct clustering of *pks*-positive and *pks*-negative isolates in a core genome-based phylogeny hinted towards the role of core genome in the acquisition and maintenance of the *pks* island (a part of accessory genome). All of the major clades of the ST95 core genome phylogenetic tree (Fig. 3) had bootstrap support values ranging from 89% to 100%. The core genome phylogeny of 159 ST95 genomes, along with an outgroup (ED1a), is depicted in Fig. S2 in the supplemental material.

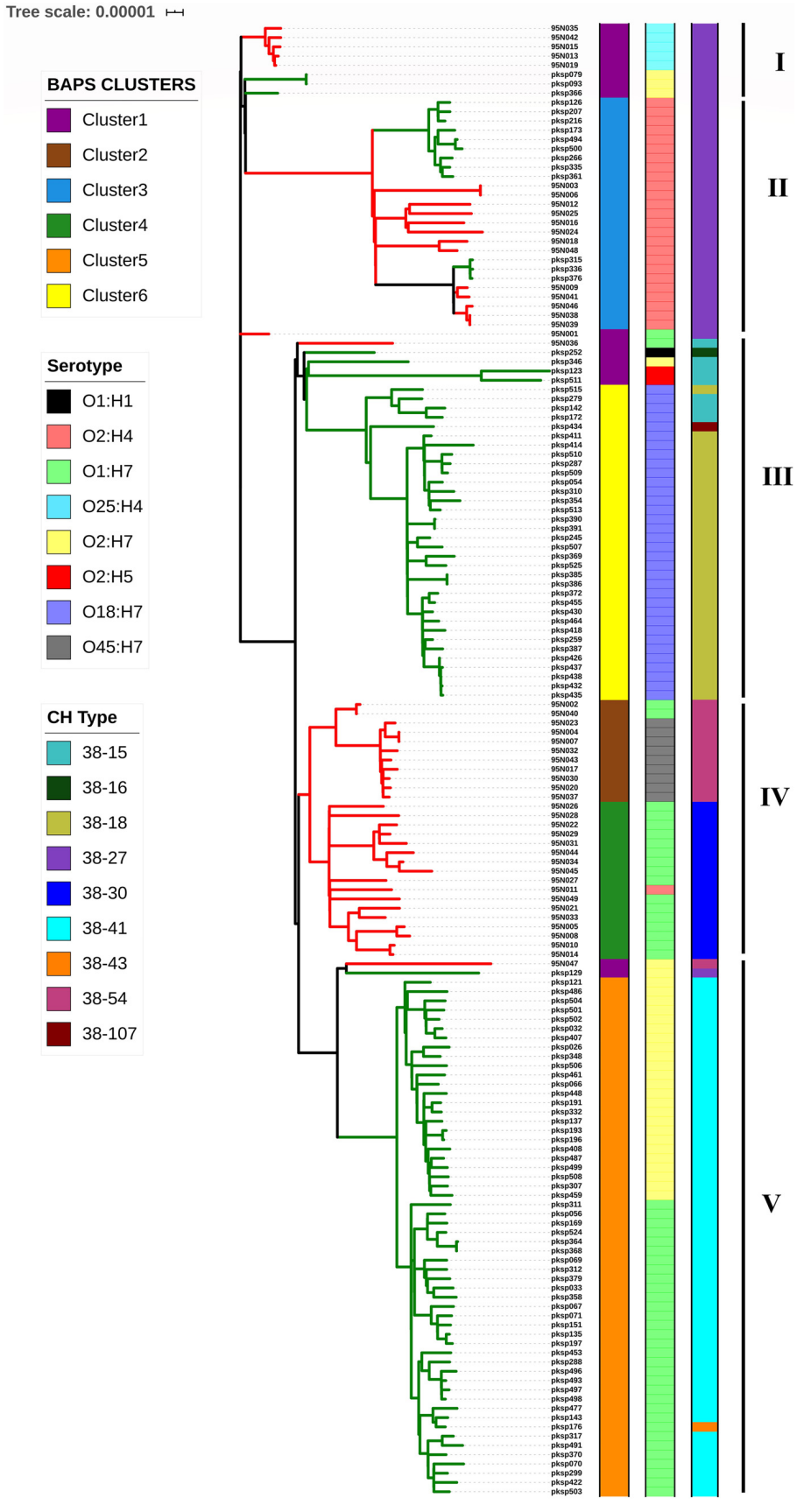


FIG 3 Maximum-likelihood core genome phylogeny of 159 ST95 isolates constructed using IQ-TREE and ClonalFrameML and visualized using iTOL. A total of five clades were observed (I to V). Green and

(Continued on next page)

TABLE 2 Results of pangenome-wide analysis of ST95 genomes using Scoary[a]

Sl. no.	Gene	Non-unique gene name	Annotation	ST95 Scoary results (no. of genomes)		Prevalence analysis [no. (%)] among the entire dataset	
				Prevalence among *pks*-positives ($n = 110$) and *pks*-negative genomes ($n = 20$) from mixed clade	Prevalence among genomes from *pks*-negative clade ($n = 29$)	Prevalence among *pks* positives ($n = 530$)	Prevalence among *pks* negatives ($n = 3,583$)
1	*ybcF_2*		Putative carbamate kinase	130	0	476 (89.8)	261 (7.28)
2	*argI_1*		Ornithine carbamoyltransferase chain I	130	0	477 (90)	269 (7.50)
3	*tabA_2*		Toxin-antitoxin biofilm protein	130	0	477 (90)	264 (7.36)
4	*yjgM*		Putative acetyltransferase	130	0	522 (98.4)	261 (7.28)
5	*arcA*		Arginine deiminase	130	0	477 (90)	259 (7.2)
6	*argR_2*		ArgR-*arg*	129	0	476 (89.9)	264 (7.3)
7	*idnO*		5-Keto-D-gluconate 5-reductase	128	0	440 (83)	1,128 (31.4)
8	*idnD*		L-Idonate 5-dehydrogenase	128	0	440 (83)	1,129 (31.5)
9	*idnR*		IdnR transcriptional regulator	128	0	440 (83)	1,126 (31.4)
10	*idnK*		D-Gluconate kinase, thermosensitive	128	0	440 (83)	1,131 (31.56)
11	*idnT*		L-Idonate/5-ketogluconate/ gluconate transporter IdnT	128	0	440 (83)	1,133 (31.62)
12	group_8070		Hypothetical protein	128	0	482 (90.9)	358 (9.9)
13	*yfcC_2*		Putative inner membrane protein; putative *S*-transferase	127	0	477 (90)	263 (7.34)
14	group_212	*tabA_2*	Toxin-antitoxin biofilm protein	0	29	39 (7.3)	3,298 (92.04)
15	*bdcA*		c-di-GMP binding protein involved in biofilm dispersal	0	29	39 (7.3)	3,244 (90.53)
16	group_1863		Hypothetical protein	0	28	41 (7.73)	3,153 (87.99)
17	*bdcR*		Putative transcriptional regulator	0	28	39 (7.3)	3,271 (91.29)
18	group_3605	*yjgM*	Putative acetyltransferase	0	28	50 (9.4)	3,563 (99.44)
19	group_7174	*hsdR*	Type 1 restriction enzyme R protein	0	28	122 (23)	484 (13.5)
20	group_803		Hypothetical protein	0	28	146 (27.5)	989 (27.6)
21	group_2253	*mdtM*	Multidrug efflux transporter	0	28	135 (25)	3,201 (89.3)
22	group_2075	*hsdM*	Host modification; DNA methylase M	0	27	122 (23)	484 (13.5)

[a]The analysis was performed between the *pks*-positive and mixed clades (clades I, II, III, and V) in comparison with the exclusively *pks*-negative clade (clade IV). The prevalence of the differentially enriched genes in the entire data set of *pks*-positive ($n = 530$) and *pks*-negative genomes ($n = 3,583$) is also shown.

The 159 isolates were grouped into six different clusters using hierBAPS (31) in the first level of clustering. The Bayesian analysis of population structure (BAPS) clusters were in concordance with maximum-likelihood phylogeny clades obtained from IQ-TREE (30) and were represented in the first data strip of Fig. 3. BAPS clusters 5 and 6 belonged to the *pks*-positive clades V and III, respectively, whereas BAPS clusters 2 and 4 belonged to the *pks*-negative clade IV. Genomes forming BAPS clusters 1 and 3 mostly belonged to clades I and II, which showed mixed clading of both *pks*-positive and *pks*-negative isolates (Fig. 3).

In silico serotyping using ECTyper classified the ST95 isolates into eight different serotypes. The branching pattern in the ML phylogeny was revealed to be mostly based on serovars of *E. coli* and also showed association with the island's prevalence. The serotypes O18:H7 and O2:H7 comprised of mostly *pks*-positive isolates and the mixed clade belonged to O2:H4. Interestingly, O1:H7 isolates formed two separate clades and BAPS clusters, one of which was *pks*-positive and the other *pks*-negative (Fig. 3).

In silico CH typing (32) revealed that all of the ST95 genomes belonged to the same C type, 38, and variations were shown in the type I fimbrial gene *fimH*, which classified the 159 genomes into nine different CH types (Fig. 3). Clades I and II, which comprised the *pks*-positive and *pks*-negative mixed cluster belonged to CH type 38-27. *pks*-posi-

FIG 3 Legend (Continued)
red branches represent genomes positive and negative for the *pks* island, respectively. The first data strip represents the BAPS clusters, the second data strip represents the serotypes of the genomes as identified by ECTyper, and the third represents the CH types of the respective genomes.

tive clade III (except 95N035) carried genomes belonging to CH types 38-18, 38-15, 38-16, and 38-107. The *pks*-negative clade IV comprised of genomes belonging to CH types 38-54 and 38-30. Clade V, which was predominantly *pks*-positive, belonged to CH type 38-41, except the genomes 95N044, *pks*p129 and *pks*p176, which belonged to CH types 38-54, 38-27, and 38-43, respectively (Fig. 3).

Pangenome-wide analysis using Scoary for ST95 genomes. A pangenome-wide analysis of accessory genes was performed using Scoary (33) to identify genes that could have a potential correlation to the *pks* island presence in the genome (Table 2). Genomes belonging to clade IV, which was an exclusively *pks*-negative clade, were compared to the rest of the genomes, which belonged to *pks*-positive and mixed clades. The genes that displayed differential prevalence and enrichment in the two sets of genomes, i.e., the ones which were completely absent in clade IV *pks*-negatives but were present in almost all the other genomes and vice versa are documented in Table 2, along with their prevalence details and functional annotations. Putative acetyltransferase gene *yjgM*, toxin-antitoxin biofilm protein gene *tabA_2*, and ornithine carbamoyltransferase chain I gene *argI_1* were each observed to be present as two different orthologs due to their sequence variation in each of these two groups of genomes analyzed. The prevalence of these genes across the *pks*-positive ($n = 530$) and *pks*-negative ($n = 3,583$) genomes were also evaluated, and the results are displayed in Table 2.

Intergenic region analysis of ST95 genomes. Intergenic regions (IGRs), although they comprise of noncoding DNA sequences, form an important part of the bacterial genome with abundantly distributed regulatory regions which play a crucial role in the phenotypic variations in the bacteria (34). The analysis of core IGR regions in addition to the coding counterpart of the genome provides an improved resolution to the evolutionary analysis of bacteria. The analysis of core IGR phylogeny was performed to ascertain the correlation of sequence variation of the intergenic region to *pks* island distribution pattern(s). The core IGR phylogeny constructed using core IGR sequences extracted by Piggy (35) (Fig. 4) showed a clading pattern reflective of the carriage of the *pks* island more distinctly compared to the core genome phylogeny, with the *pks*-negative cluster found to clade separately from *pks*-positive and mixed clades. The IGR clades from O1:H7 *pks*-positive and *pks*-negative genomes were also found to be more distinct compared to the core genome phylogeny. All of the major clades of the ST95 IGR phylogenetic tree (Fig. 4) had bootstrap values ranging from 92.7% to 100%.

RM system analysis. The REBASE (36) (Gold Standard Database) consisted of 3,211 genes, which were clustered using UCLUST (37), and the curated data set of 2,171 genes was used for restriction modification (RM) system analysis of the genomes. The prevalence pattern of RM systems showed a correlation to the phylogenetic clades and serogroup distribution in the ST95 core genome ML phylogeny (Fig. 5). RM systems of particular interest that were observed included M.Eco9001I, S.Eco9281I, and S.Eco9001I.

The genomes belonging to the exclusively *pks*-negative clade harbored M.Eco9001I (except 95N045, which carried the truncated gene). They also carried either one of S.Eco9281I or S.Eco9001I and Eco9001IP when separately analyzed, as the gene encoding the main restriction enzyme subunit was not included in the REBASE (36) Gold Standard Database. O2:H4 and O25:H4, which also comprised *pks*-negatives, did not carry the above-mentioned genes, and O1:H7 *pks*-positives and three O1:H7 *pks*-negatives (95N039, 95N001, and 95N036) that clustered differently from the main O1:H7 *pks*-negative clade, were also observed not to carry these genes (Fig. 5).

RM system patterns of 530 *pks*-positives and 3,583 *pks*-negatives were also analyzed to decipher their prevalence in these genomes and the ones which showed specific prevalence patterns, such as the type I RM systems Eco9001I/9281I and Eco.CFTI and the type III RM system Eco.CFTII described in Table 3. The modification and recognition genes of these RM systems were part of REBASE (36), while their cognate restriction subunit gene sequences (Eco9001IP/9281IP, Eco.CFTIP, and Eco.CFTIIP) were analyzed separately. While analyzing the sequence types and serotypes of the genomes carrying these RM systems, it was observed that the genomes with the complete Eco.CFTI system belonged, interestingly, to the ST73 complex, but showed no serogroup specificity.

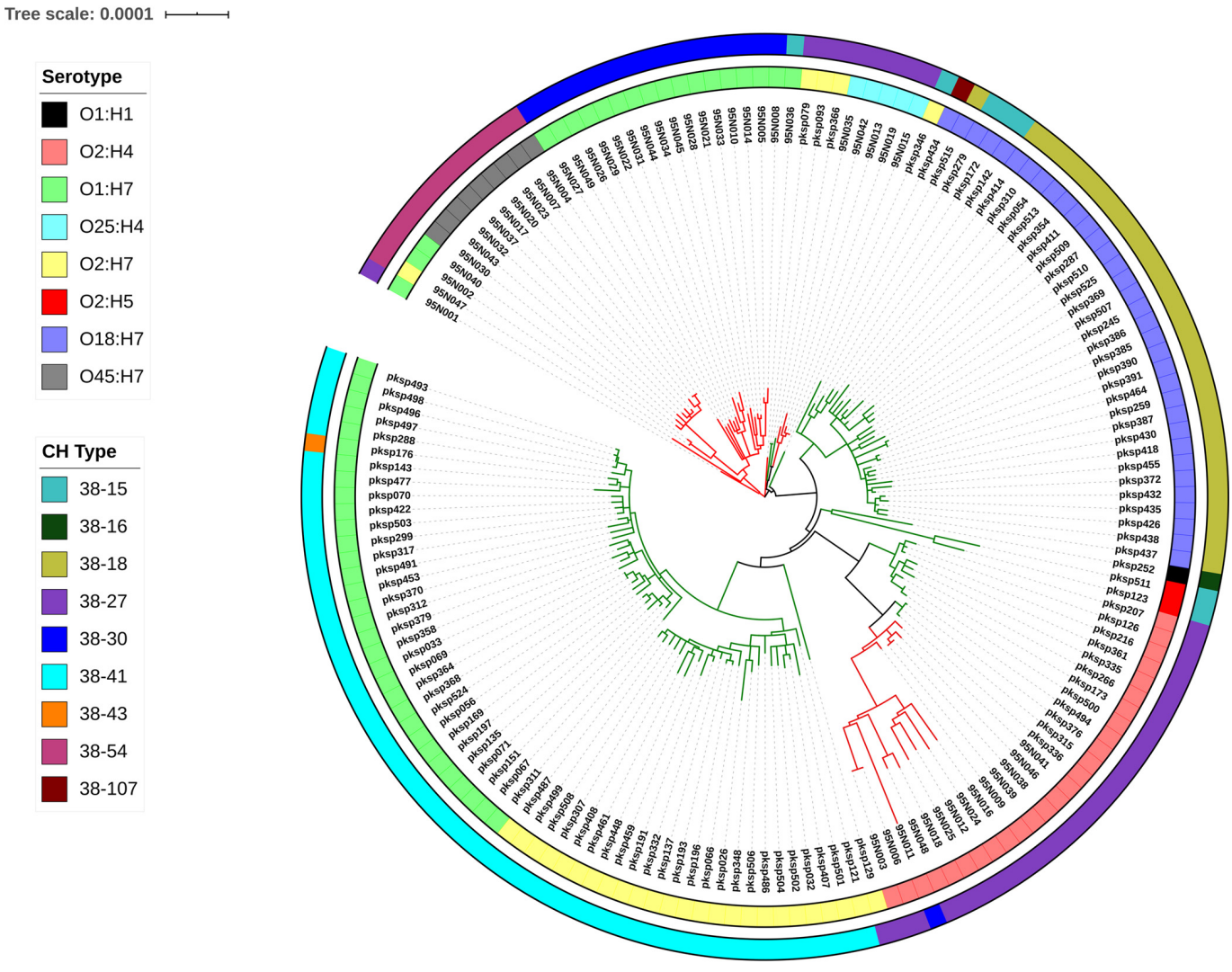


FIG 4 Maximum-likelihood intergenic region phylogeny of 159 ST95 isolates constructed using IQ-TREE and visualized using iTOL. Green and red branches represent *pks* island positives and negatives, respectively. The inner ring represents the serotypes of the genomes as identified by ECTyper, and the outer ring represents CH types of the genomes obtained from CHTyper.

***pks* island phylogeny.** The core genome (number of core genes = 2,579) phylogeny and the *pks* island sequence phylogeny of 247 genomes which contained the *pks* island in a single contig were compared to study the effect of the pattern of evolution of the island sequences with respect to the core genome and the subtype (sequence type and serotype). The core genome phylogeny of the 247 genomes constructed using IQ-TREE (30) showed a clading pattern reflective of the sequence type and serotype, with few exceptions (Fig. 6A). This core genome phylogeny (Fig. 6A) was compared with that of the phylogeny of *pks* island sequences derived from the 247 genomes (Fig. 6B) using Dendroscope (38) as shown in Fig. S3 in the supplemental material. Hierarchical BAPS clustering (31) of the island alignment provided 3 clusters at the first level, which are depicted using different clade colors (black, purple, and orange) in the phylogenetic tree (Fig. 6B). The clading pattern was in agreement with the obtained BAPS clusters. Cluster 1 (black clade) consisted of only 6 genomes, which formed a distinct clade compared to cluster 2 and cluster 3, which comprised the rest of the genomes analyzed. The islands did not show any clustering pattern reflective of the sequence type of their respective genomes in contrast to the core genome phylogeny, except for the *pks* island of genomes from ST998, which were found to cluster together. This lack of concordance with the core genome clustering pattern could indicate

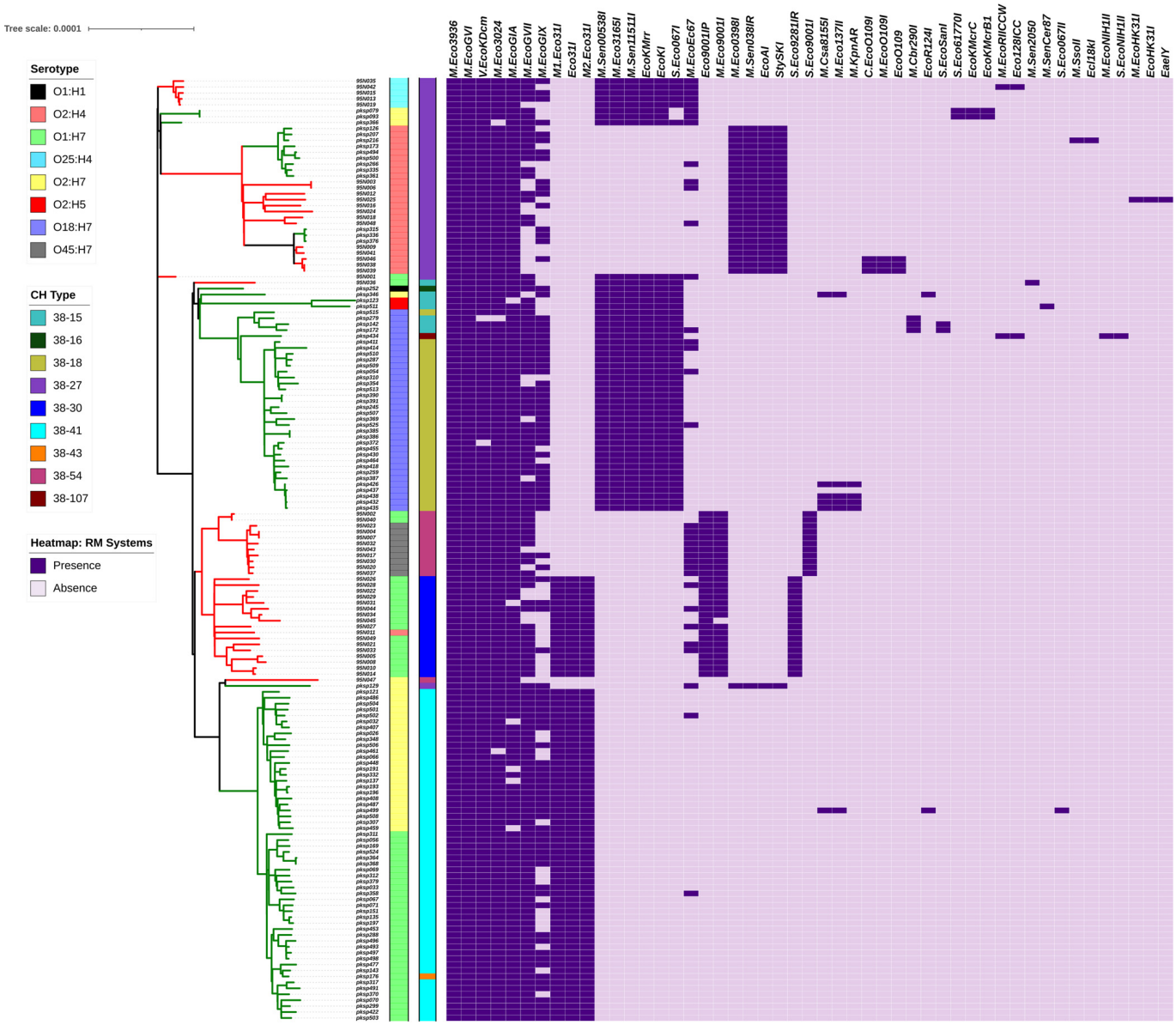


FIG 5 RM system prevalence pattern of ST95 genomes plotted as a heat map, along with core genome phylogeny with serotype and CH type labeled. The prevalence pattern showed accordance with the clading pattern of the phylogeny, as well as with the serotype distribution of the genomes.

that HGT could possibly be the mode of transfer of the island. Islands from ST12, ST73, and ST95 genomes were found to display an intermixed pattern, indicating the possibility of HGT of the island across sequence types (Fig. 6B). The bootstrap support values of the major clades of the core genome phylogeny of 247 *pks*-positive genomes (Fig. 6A) ranged from 98.6% to 100%, and that of the *pks* island phylogeny (Fig. 6B) ranged from 84.6% to 100%.

DISCUSSION

The colibactin-producing *pks* island found in certain members of *Enterobacteriaceae* is emerging as an important virulence marker in the progression of CRC, meningitis, and septicemia (9). Several studies have described the role of colibactin in CRC (19, 39–41), including the synergy between host cells and microbiota in CRC progression (41), making the genotoxin an important virulence factor that requires urgent attention owing to its clinical implications. The *pks* island shows wide distribution among neonatal *E. coli* K1 isolates and was observed to have a major role in the fully virulent

TABLE 3 Comparison of prevalence of selected RM systems among *pks*-positive and *pks*-negative genomes

RM system	Genes	Prevalence [no. (%)] among *pks* positives ($n = 530$)	Prevalence [no. (%)] among *pks* negatives ($n = 3,583$)
Eco900I/928I (TYPE-I-RM)	*Eco9001P/9281IP*	124 (23.4)	484 (13.5)
	M.Eco9001I/9281I	124 (23.4)	252 (7.03)
	S.Eco9001I/9281I	43 (8.11)	43 (1.2)
Eco.CFTI (TYPE-I RM)	*Eco.CFTIP*	257 (48.5)	761 (21.23)
	M.EcoCFTI	257 (48.5)	761 (21.23)
	S.EcoCFTI	156 (29.43)	7 (0.19)
Eco.CFTII (TYPE-III RM)	*Eco.CFTIIP*	159 (30)	4 (0.11)
	M.EcoCFTII	157 (29.62)	0 (0)

phenotype of the bacteria in a neonatal systemic infection model (12). In the present study, high-throughput phylogenomic comparison of *pks* island-harboring *E. coli* genomes from the in-house culture collection and publicly available ones from the NCBI were used to draw insights into the island's acquisition and evolution. The in-house genome collection ($n = 23$) was a part of a previous study from our group, where the isolates linked to a clinical setting from Pune, India, were subjected to epidemiological investigation and characterization of virulence and resistance attributes (26). Whole-genome-based virulome and resistome analysis revealed that the in-house *pks*-positive genomes possessed a high number of genes contributing to virulence (Fig. 1) (see Table S3 in the supplemental material). Genes conferring antimicrobial resistance prevalent in the *pks*-positive genomes mostly consisted of efflux pumps, and only a few specific antibiotic resistance determinants were observed (Fig. 2) (see Table S4 in the supplemental material). These findings were in line with the phenotypic observation of reduced antibiotic resistance and increased functional virulence characteristics displayed by the *pks*-positive isolates in our previous study (26) compared to the frequently observed multidrug-resistant *pks*-negative ExPEC clones obtained from Indian population (42–46). Our previous genomic studies on the *pks*-negative ExPEC collection displayed a higher prevalence of

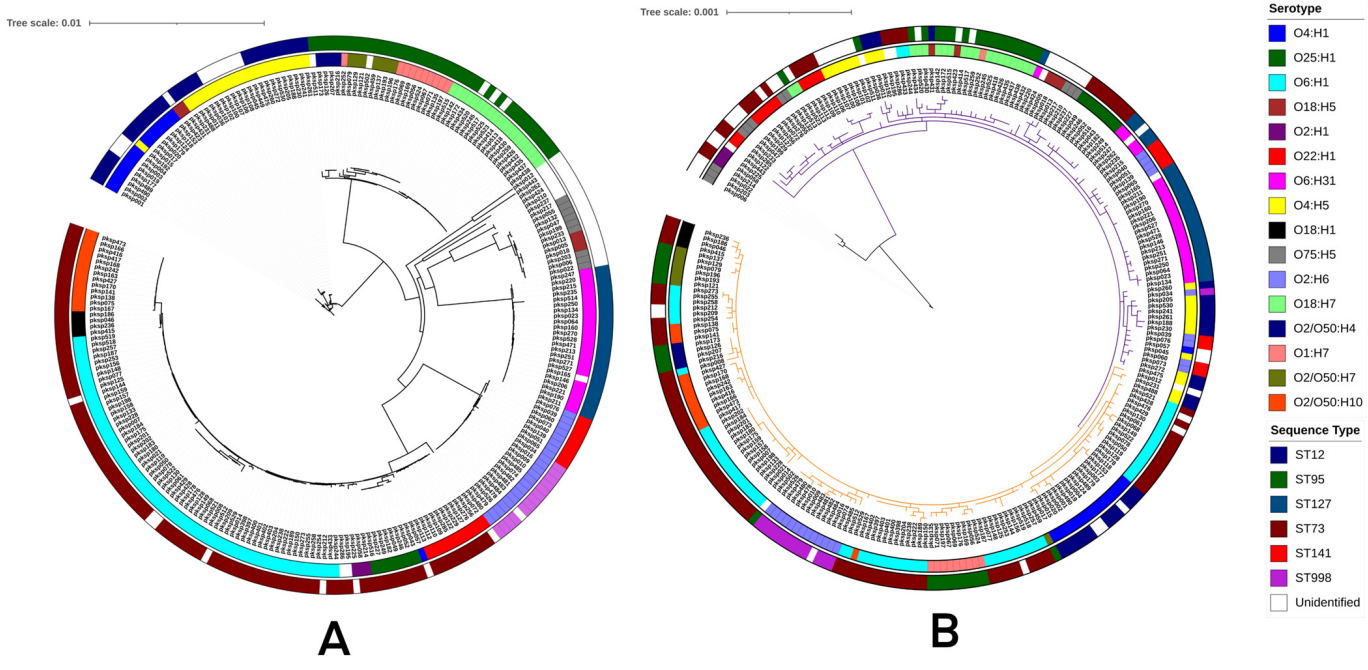


FIG 6 (A) Core genome phylogeny of 247 *pks*-positive genomes constructed using IQTree and visualized using iTOL. (B) Phylogeny of *pks* island sequences from 247 genomes constructed using IQ-TREE and visualized by iTOL. Clade colors depict the three BAPS clusters, i.e., cluster 1 (black), cluster 2 (purple), and cluster 3 (orange). In both panels A and B, the inner ring denotes serotypes and outer ring denotes sequence types of the genomes from which the island was derived; legends for the same have been provided in the figure.

specific antibiotic resistance genes and a relatively lower prevalence of virulence genes (45–47) compared to those in our current analysis of the *pks*-positive genomes. Notably, all of the *pks*-positive genomes harbored the *bacA* gene, which is involved in resistance against the antibiotic, bacitracin (48), and the gene(s) *pmrE*, *pmrC*, and *pmrF*, involved in the binding of polymyxin (49). The large virulence gene repertoire in *pks*-positive isolates is consistent with the previous report based on PCR-based observations on bacteremia isolates (50), implying its clinical significance. Adhesins and type VI secretion systems showed abundance, and there was an increased prevalence of genes belonging to different siderophore production systems (Fig. 1), in concordance with the phenotypic observations of siderophore production assay from our earlier study (26) and other reports, which indicate potential associations between the *pks* island and iron acquisition systems (51). A previous study reported that the *pks* island encoded peptidase ClbP is involved in the genotoxin activation as well as renders antimicrobial activity either through microcins (Mcc) biosynthesis or secretion independently, or in cooperation with glucosyltrasferase, thus reflecting the crucial co-selection of these islands in the evolution of pathogenic phylogroup B2 (16). In a recent study, microcin, salmochelin, and colibactin have also been indicated as a triad that could potentially provide a selective advantage for bacterial colonization in the rectal reservoir with minimal genetic cost (52). The abundance of the siderophore systems like yersiniabactin, enterobactin, salmochelin, and *chuASTUVWXY* genes, along with the *pks* island, could potentially play a role in the successful colonization and persistence of these isolates. Virulence factor profiling also showed an increased prevalence of the hemolysin system (*hly*) in *pks*-positive isolates (Fig. 1), the association of which was indicated previously as a risk factor for colorectal cancer (53).

The study was further expanded to screen 4,090 genomes of *E. coli* obtained from NCBI, out of which the *pks* island was detected in 507 genomes, in addition to the 23 in-house genomes. The 530 *pks*-positive genomes were further subjected to various *in silico* typing methods to identify distribution patterns among various *E. coli* subtypes (Table 1). All of the ST73 *E. coli* isolates were observed to harbor the *pks* island, in contrast to the most prevalent and highly successful ExPEC pandemic clone ST131 data set, which was notably completely *pks*-negative. It is also interesting to note that the prominent STs with *pks* island-positive genomes, i.e., ST73 and ST95, have been previously reported to show low antibiotic resistance (2, 54, 55), which, along with our previous study (26) and Comprehensive Antibiotic Resistance Database (CARD)-based genome analysis in the current study could indicate the association of the *pks* island with isolates having a reduced antimicrobial resistance profile. Phylogrouping showed a strong association of the *pks* island with the B2 phylogroup, in accordance with previous reports (15, 23, 50, 56, 57), as well as with our previous study (26).

ST95 is a successful ExPEC clonal complex that displays functional virulence properties of host adhesion, invasion, biofilm, and serum resistance (58), and it has clinical implications in urinary tract infection and newborn meningitis, while also being a predominant avian and companion animal pathogen (2). The ST95 data set was used as a model for studying the *pks* island, as it was the only sequence type which carried a comparable number of *pks*-positive and *pks*-negative genomes. A total of five clades were obtained (Fig. 3), which were comparable to a previous study about the analysis of STc95 genomes that identified 5 subgroups within the STc95 complex (59). Clades I, II, III, IV, and V of the core genome phylogeny in our study (Fig. 3) showed correspondence to subgroups C, E, B, D, and A, respectively, based on the similar serotype and *fimH* type (59) (clade III additionally carried O2:H5, O1:H7, O1:H1, and O2:H7 genomes in small numbers). The prevalence pattern of the *pks* island sequence was in line with the previously observed prevalence of the *clbB* gene in the same study (59). Differential clading patterns observed in ST95 core genome phylogeny with separate positive, negative, and mixed clusters indicate the potential role of the core genome in HGT and integration of the island (Fig. 3). Core intergenic region-based phylogeny showed a clading pattern more reflective of *pks* island carriage (Fig. 4). A previous study demonstrated that the patterns of polymorphism of the intergenic region *o454-*

nlpD displayed concordance with the phylogenetic background, as well as with some important virulence-associated genes in *E. coli* (60). Core intergenic region substitutions were previously described to show association with the acquisition of an accessory genome in ST131 *E. coli* (61), and the analysis of ST95 genomes with respect to the *pks* island exhibited a similar pattern. Although most of the clustering patterns of the ST95 core genome phylogeny reflected the serotypes, O1:H7 showed a peculiar distribution into different clades containing *pks*-positive and *pks*-negative genomes, and they also had a distinct *fimH* type (Fig. 4). Scoary (33) was used for pangenome-wide analysis of accessory genes, where the positive and mixed clusters were used as a combined data set (clades I, II, III, and V) to compare with genomes belonging to a completely *pks*-negative clade (clade IV) and the genes that showed differential enrichment among the groups listed in Table 2. It was interesting to note that the genes *idnODRKT* belonging to the subsidiary system for L-idonic acid catabolism, which may provide a metabolic advantage for colonization (62), were present in all genomes belonging to the *pks*-positive and mixed cluster, while being completely absent in the members of the exclusively *pks*-negative clade (clade IV). The type 1 restriction enzyme R protein *hsdR* and the DNA methylase *hsdM* were observed to be present only among the genomes belonging to the exclusively *pks*-negative clade (Table 2).

As RM systems are shown to be involved in the regulation of HGT and recombination (63), their prevalence was studied among *pks*-positive and *pks*-negative data sets as a preliminary analysis to determine their putative role in transfer or incompatibility of the acquisition of the *pks* island. A previous study has indicated the potential role of restriction modification systems in the acquisition of resistance plasmids in ST95 O1:H7 isolates (64). Since the *pks* island showed clade-specific distribution patterns within the ST95 core genome phylogenetic tree, the tree topology was compared with its RM system prevalence data as a model to study the RM system diversity and finer distribution pattern (Fig. 5). The analysis is limited to the RM systems in the curated Gold Standard Database of REBASE and their selected cognate restriction enzyme subunit counterparts of the systems. When overlaid with the core genome phylogeny, the topology of RM prevalence pattern showed relation to the subclades reflective of their serotypes (Fig. 5). This observation is similar to the results from a previous study describing the methyl transferase diversity among ST131 *E. coli* isolates in which the RM system profiles were observed to show relation to their phylogenetic clusters (65). Another study in *Burkholderia pseudomallei* showed the clade-specific complement of the RM system, which potentially led to the clade-specific patterns in the DNA methylome (66). The population structure of *Neisseria meningitidis* was also observed to coincide with its RM system distribution, suggesting a role of RM systems as a barrier in DNA exchange, driving the formation of distinct phylogenetic lineages (67). Similar sublineage correlations based on serovars and phylogenetic clading of genomes with identical RM profiles were observed in a previous study involving *Salmonella enterica* (68). Based on this evidence and our observations, we hypothesize that the RM system profiles of the isolates might have a potential role in shaping the phylogenetic lineages and guiding the DNA exchange, thus playing a role in the horizontal acquisition of the genomic island. Notably, in the analysis of the RM system profile in the entire *pks*-positive and *pks*-negative data sets, the type III RM system Eco.CFTII showed a higher prevalence in *pks*-positive genomes than in the *pks*-negative genomes (Table 3). However, the limitation of a small number of curated candidates available for RM system analysis is to be noted, and careful interpretation is mandated. Based on these preliminary observations from prevalence analysis of RM systems among *pks*-positive and *pks*-negative genomes, further studies on their probable role in the acquisition and maintenance of the mobile genetic elements will be required.

The phylogeny of the *pks* island, in contrast to core genome phylogeny, revealed its mixed distribution among various sequence types and serotypes (except in certain groups) indicative of a probable frequent HGT across the sequence types (Fig. 6; see also Fig. S3 in the supplemental material). This observation is in line with the evidence

from a previous study in which the comparison between phylogenetic trees of the core genome and the *pks* island sequences within the ECOR collection displayed different clustering patterns indicative of the transmission of the island to be horizontal and not vertical (25). This, along with other observations of prevalence patterns, demonstrated that certain sequence types of *E. coli*, such as ST73, ST95, and ST12, show increased capability to acquire the island, and frequent horizontal exchanges of the island could occur across these subtypes.

In conclusion, our study is perhaps the first one to perform large-scale, whole-genome-based investigations with respect to the distribution of the *pks* island(s) among different *E. coli* populations and the consequent evolutionary relationships. The preferential distribution pattern of the *pks* (encoded genotoxin)-harboring *E. coli* was studied using different computational methods of subtyping. These observations may be able to provide support to the diagnostic systems or health care modalities aimed at understanding the clinical implications of the potential genotoxic nature of *pks*-positive isolates. The *pks* island phylogeny indicated horizontal acquisition/transmission and the possibility of exchange between compatible *E. coli* subtypes. Investigation of the ST95 model data set revealed a higher prevalence of the *pks* island within specific serotypes and CH types, pointing at the role of HGT and finer evolution within a particular ST. The core genome and core intergenic region phylogeny were used to gain a comprehensive understanding of the clade-specific pattern of distribution of the island, which is otherwise a part of the accessory genome. Further studies on the potential role of RM systems in shaping the lineages and driving the acquisition of the island among compatible isolates needs to be performed at a higher resolution in order to gain interesting insights into the HGT and evolution of virulence in pathogenic *E. coli*.

MATERIALS AND METHODS

Ethics statement. All of the *E. coli* isolates that are newly unraveled here were originally isolated as part of our previous studies, as mentioned. Cultures and DNA preparations were handled as per standard biosafety guidelines for *E. coli* and within the ambit of available permissions.

Whole-genome sequencing, assembly, and annotation. Genomic DNA of 25 *pks*-positive (in-house) *E. coli* isolates (originally cultured and maintained by S.J. and her colleagues from Dr, D. Y. Patil University Hospital, Pune, India), which were characterized in our previous study (26), were isolated and purified of any RNA contamination using a Qiagen DNeasy blood and tissue kit (Qiagen, Germany) and sequenced using the Illumina MiSeq platform (69, 70). The paired-end reads were subjected to quality control using NGS QC Toolkit (71), trimmed using FastX-Trimmer (http://hannonlab.cshl.edu/fastx_toolkit/), and further assembled *de novo* using SPAdes Genome Assembler (v3.6.1) (72). Assembly statistics were obtained using QUAST (73). Two out of the 25 isolates were discarded from further analysis due to poor quality. The contigs were further ordered and scaffolded using C-L-Authenticator (74), using the *E. coli* ATCC 25922 complete genome as a reference, and the scaffolds were annotated using Prokka (75). Genome statistics were gleaned using Artemis (76), and sequence types of the isolates were identified using an *in silico* MLST pipeline using in-house scripts (47, 77) and the publicly available MLST pipeline (https://github.com/tseemann/mlst), which uses the PubMLST database (https://pubmlst.org/) (78). ECTyper (https://github.com/phac-nml/ecoli_serotyping) was used to perform *in silico* serotyping of the genomes.

Analysis of the genomes for the resistance and virulence determinants. Amino acid sequence files from annotated genomes were used to determine the resistance and virulence genes by performing BLASTp (79) against the Comprehensive Antibiotic Resistance Database (80) and Virulence Factor Database (28), respectively. A percentage identity of 70% and a query coverage of 75% were used as thresholds while analyzing the genomes for the presence of the respective genes. The heat plots depicting the presence-absence status of the genes were generated using the *gplots* (https://github.com/talgalili/gplots) package of R. The virulence and resistance profiles of 23 in-house *pks*-positive genomes were also compared with those of 23 in-house *pks*-negative genomes using the methodology mentioned above (accession IDs are listed in Table S6 in the supplemental material).

Whole-genome comparative analysis and visualization. The complete genome of *E. coli* IHE3034 (GenBank accession number CP001969.1) was used as the reference genome, and the assembled genomes of in-house isolates harboring the *pks* islands were compared to the reference *pks*-positive genomes using BLAST Ring Image Generator (BRIG) (27) to determine their genetic relatedness, with upper and lower identity thresholds of 70% and 50%, respectively. The annotation of phages in the reference genome was performed using the PHAST server (81), and the coordinates of the loci of the detected phages were plotted on the image. The coordinates of the *pks* island were also annotated in the BRIG image.

The trimmed, filtered reads of the genomes were mapped and aligned to the reference sequence of the *pks* island along with flanking regions obtained from NCBI (GenBank accession number AM229678.1) using SAMtools (82) and Bowtie 2 (83). The mapped reads were assembled *de novo* using SPAdes (72), and

the reconstructed island sequences were obtained and then further subjected to BRIG (27) using the complete *pks* island sequence as the reference to visualize the integrity of the island. The upper and lower identity thresholds used in the analysis were 70% and 50%, respectively.

Identification of *pks* island-containing genomes in the public domain. *E. coli* genomes were downloaded from the public database using in-house scripts, and the genomes with size greater than 4.8 Mb and with fewer than 200 contigs each, were used for the study. A total of 3,784 draft and 306 complete genomes were selected after curation as the final database for further downstream analysis. The complete *pks* island along with flanking regions of *E. coli* IHE3034 and its coding sequences were obtained from NCBI as the reference sequence of the genomic island (GenBank accession number AM229678.1). The genomes were screened for the presence of *pks* island genes obtained from the above references using BLASTn (79) with identity and query coverage thresholds of 85%.

Genome annotation, *in silico* MLST and phylogrouping. The genomes of both NCBI and in-house isolates were subjected to annotation using Prokka software (75). All the genomes were subjected to *in silico* phylogrouping and multilocus sequence typing (MLST) to determine the sequence types using an *in silico* MLST pipeline that harnessed in-house scripts (47, 77) and the MLST pipeline (https://github.com/tseemann/mlst), which uses the PubMLST database (https://pubmlst.org/) (78). ECTyper (https://github.com/phac-nml/ecoli_serotyping) was used to perform the *in silico* serotyping of the genomes.

ST95 pangenome analysis. ST95 was used as a model data set to study the distribution and evolutionary pattern of the *pks* island due to the availability of both *pks*-positive ($n = 110$) and *pks*-negative genomes ($n = 49$). The pangenome analysis of 159 genomes from ST95 after annotation using Prokka (75) was performed using Roary (84) with identity and *E* value cutoffs of 85% and 0.00001, respectively, for the determination of orthologous gene clusters. Genes which were shared by all the 159 isolates, which constitute the core genome, and the core genes were subjected to COG classification using eggNOG (29), and the COG groups were tabulated. The genomes were also analyzed using CHTyper (85) for the *in silico* determination of CH types based on *fumC* and *fimH* alleles.

ST95 core genome phylogeny. The core genes determined using Roary were subjected to nucleotide alignment using PRANK (86), and the resultant core genome alignment was further subjected to trimAl (87) (using -strict flag) for the trimming and refinement of the alignment. The alignment was also used as an input for hierBAPS (31) to perform hierarchical clustering based on sequence variations using Bayesian methods. The refined alignment was then subjected to IQ-TREE (30) with ModelFinder (88) to optimize the best nucleotide substitution model to construct a maximum-likelihood phylogenetic tree with 500 bootstrap replicates. The resultant core genome based maximum-likelihood phylogenetic tree was subjected to ClonalFrameML (89) to remove recombination regions and was visualized using interactive Tree Of Life (iTOL) (90). The branches were color coded according to the presence/absence of the *pks* island, and BAPS cluster, serogroup, and CH type information were also annotated in the tree using data strips. In addition, a core genome phylogenetic tree including an outgroup, ED1a (NCBI assembly number GCA_000026305.1), was also constructed using the methodology mentioned above. A pangenome-wide association study comparing the genomes belonging to *pks*-positive and mixed clades with genomes belonging to the exclusively *pks*-negative clade (as identified in core genome-based phylogeny) was performed using Scoary (33) with the help of the gene_presence_absence.csv output file of Roary (84). The *pks*-positive genomes ($n = 110$) and *pks*-negative genomes belonging to the mixed clade ($n = 20$) were grouped together and designated with trait value "1" and the *pks*-negative genomes forming the exclusively *pks*-negative clade ($n = 29$) were designated with trait value "0" in the Scoary (33) input. The prevalence of the differentially enriched genes between the two groups was determined across the *pks*-positive ($n = 530$) and *pks*-negative ($n = 3583$) genomes using BLASTn (79) with identity and query coverage thresholds of 85%.

ST95 IGR phylogeny. The GFF files that were derived from the annotation of 159 ST95 genomes using Prokka (75) and the gene presence-absence file obtained from pangenome analysis by Roary (84) were used to perform the intergenic region analysis using Piggy (35). The intergenic regions (IGRs) that were shared by all the genomes (core IGRs) were extracted and aligned using Prank (86), followed by trimming using trimAl (87) to refine the alignment by removing spurious and poorly aligned regions. IQ-TREE (30) was employed along with ModelFinder (88) (-MFP flag) for construction of IGR phylogeny with 500 bootstrap replicates, followed by ClonalFrameML (89) to produce a maximum-likelihood phylogeny of the core intergenic regions of ST95 genomes. The resultant phylogenetic tree was visualized using iTOL (90), with serogroup and CH type information of the isolates, which was previously obtained, labeled as data strips.

RM system analysis. The restriction modification (RM) gene profiling of the *E. coli* genomes was performed using the REBASE Gold Standard Database (36). The REBASE database was clustered using UCLUST (37) with an identity threshold of 90%. This curated database was used for the detection of RM systems in *E. coli* genomes. A BLASTn search was performed against genomes with identity and query coverage thresholds of 85%. In order to determine the pattern of distribution of RM systems among *pks*-positive and *pks*-negative genomes, BLAST analysis of curated RM systems database was performed against the data sets, namely, ST95 genomes ($n = 159$), *pks*-positive genomes ($n = 530$), and *pks*-negative genomes ($n = 3,583$). In cases where the genomes were observed to carry the modification and recognition subunits, the sequences of their cognate restriction enzymes obtained from REBASE were also separately analyzed, if they were not already included in the gold-standard database.

***pks* island phylogeny.** The *pks*-positive genomes were subjected to standalone BLAST analysis against *pks* island sequence, and the genomes with identity and query coverage of greater than 95% and 85%, respectively, for the *pks* island were used to obtain genome sequences harboring the *pks* island within a single contig. The core genome phylogeny of these selected genomes ($n = 247$) was

constructed per the methodology mentioned in the previous sections. The *pks* island sequences from these genomes were extracted from the locus information of BLAST outputs using the extract-align program from EMBOSS (http://emboss.sourceforge.net/apps/cvs/emboss/apps/extractalign.html) and in-house scripts to handle reverse complements. The island sequences were aligned using PRANK (86), and trimAl (used with -strict flag) (87) was used to refine the alignment. Bayesian analysis of population structure (BAPS) (31) clustering of the alignment was performed for the sequences, and IQ-TREE (30) with 1,000 bootstrap replicates was used for the construction of a maximum-likelihood phylogeny with ModelFinder enabled (88) and visualized using iTOL (90), along with annotations for sequence type and serotype information. The two phylogenetic trees were also compared using the "connect taxa" functionality of Dendroscope (v3.7.3) (38).

Data availability. The genome sequence data generated as a part of this study are deposited in NCBI under the BioProject identifier PRJNA667681. Accession numbers of the individual genomes described here are as follows: JADBJB000000000, JADBJA000000000, JADNRJ000000000, JADBIZ000000000, JADBIY000000000, JADBIX000000000, JADBIW000000000, JADBIV000000000, JADBIU000000000, JADBIT000000000, JADBIS000000000, JADBIR000000000, JADBIQ000000000, JADBIP000000000, JADBIO000000000, JADBIN000000000, JADBIM000000000, JADBIL000000000, JADBIK000000000, JADBIJ000000000, JADBII000000000, JADBIH000000000, and JADBIG000000000 (see Table S2 in the supplemental material for further details).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, TIF file, 2.9 MB.

FIG S2, TIF file, 1.2 MB.

FIG S3, TIF file, 1.2 MB.

TABLE S1, PDF file, 0.1 MB.

TABLE S2, PDF file, 0.1 MB.

TABLE S3, XLSX file, 0.1 MB.

TABLE S4, XLSX file, 0.02 MB.

TABLE S5, PDF file, 0.3 MB.

TABLE S6, PDF file, 0.1 MB.

TABLE S7, PDF file, 0.1 MB.

ACKNOWLEDGMENTS

This work represents part of the Ph.D. research of A.S., who is also a guarantor on this study for the purposes of isolates, sequences, accessions, workflows, and raw data. The study was part of the umbrella objectives of the Indo-German International Research Training Group (IRTG), *Internationales Graduiertenkolleg*—Functional Molecular Infection Epidemiology (GRK1673), a collaborative endeavor of the German Research Foundation (DFG) and the University of Hyderabad (India). We would like to thank Microsoft Corporation for the 'Microsoft Azure for Research' and 'AI for Earth' awards to N.A. and subsequent, on-demand Azure sponsorships for our analyses.

A.S. acknowledges the senior research fellowship (SRF) from the CSIR, India.

We acknowledge the valuable help and suggestions from Aditya Kumar Lankapalli, Arif Hussain, and Sumeet Tiwari. We also extend special thanks to the four referees for their critique and advice on our work and to Taane Clark (London School of Hygiene and Tropical Medicine) for his critical review, valuable suggestions on our bioinformatics analyses and comments on the manuscript.

REFERENCES

1. Croxen MA, Finlay BB. 2010. Molecular mechanisms of *Escherichia coli* pathogenicity. Nat Rev Microbiol 8:26–38. https://doi.org/10.1038/nrmicro2265.
2. Denamur E, Clermont O, Bonacorsi S, Gordon D. 2021. The population genetics of pathogenic *Escherichia coli*. Nat Rev Microbiol 19:37–54. https://doi.org/10.1038/s41579-020-0416-x.
3. Bliven KA, Maurelli AT. 2012. Antivirulence genes: insights into pathogen evolution through gene loss. Infect Immun 80:4061–4070. https://doi.org/10.1128/IAI.00740-12.
4. Clermont O, Bonacorsi S, Bingen E. 2000. Rapid and simple determination of the *Escherichia coli* phylogenetic group. Appl Environ Microbiol 66:4555–4558. https://doi.org/10.1128/aem.66.10.4555-4558.2000.
5. Ahmed N, Dobrindt U, Hacker J, Hasnain SE. 2008. Genomic fluidity and pathogenic bacteria: applications in diagnostics, epidemiology and intervention. Nat Rev Microbiol 6:387–394. https://doi.org/10.1038/nrmicro1889.
6. Hacker J, Kaper JB. 2000. Pathogenicity islands and the evolution of microbes. Annu Rev Microbiol 54:641–679. https://doi.org/10.1146/annurev.micro.54.1.641.
7. Dobrindt U, Hochhut B, Hentschel U, Hacker J. 2004. Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol 2:414–424. https://doi.org/10.1038/nrmicro884.
8. Groisman EA, Ochman H. 1996. Pathogenicity islands: bacterial evolution in quantum leaps. Cell 87:791–794. https://doi.org/10.1016/S0092-8674(00)81985-6.
9. Faïs T, Delmas J, Barnich N, Bonnet R, Dalmasso G. 2018. Colibactin: more

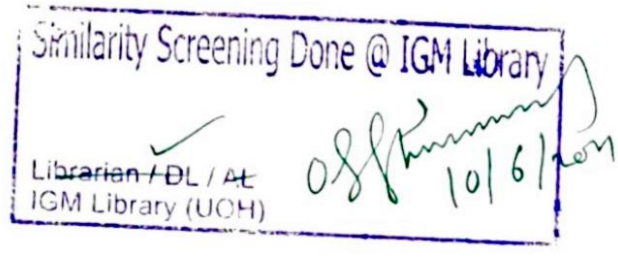
than a new bacterial toxin. Toxins (Basel) 10:151. https://doi.org/10.3390/toxins10040151.
10. Nougayrède J-P, Homburg S, Taieb F, Boury M, Brzuszkiewicz E, Gottschalk G, Buchrieser C, Hacker J, Dobrindt U, Oswald E. 2006. *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells. Science 313:848–851. https://doi.org/10.1126/science.1127059.
11. Cuevas-Ramos G, Petit CR, Marcq I, Boury M, Oswald E, Nougayrede J-P. 2010. *Escherichia coli* induces DNA damage *in vivo* and triggers genomic instability in mammalian cells. Proc Natl Acad Sci U S A 107:11537–11542. https://doi.org/10.1073/pnas.1001261107.
12. McCarthy AJ, Martin P, Cloup E, Stabler RA, Oswald E, Taylor PW. 2015. The genotoxin colibactin is a determinant of virulence in *Escherichia coli* K1 experimental neonatal systemic infection. Infect Immun 83:3704–3711. https://doi.org/10.1128/IAI.00716-15.
13. Micenková L, Beňová A, Frankovičová L, Bosák J, Vrba M, Ševčíková A, Kmet'ová M, Šmajs D. 2017. Human *Escherichia coli* isolates from hemocultures: septicemia linked to urogenital tract infections is caused by isolates harboring more virulence genes than bacteraemia linked to other conditions. Int J Med Microbiol 307:182–189. https://doi.org/10.1016/j.ijmm.2017.02.003.
14. Wallenstein A, Rehm N, Brinkmann M, Selle M, Bossuet-Greif N, Sauer D, Bunk B, Spröer C, Wami HT, Homburg S, Von Bünau R, König S, Nougayrède J-P, Overmann J, Oswald E, Müller R, Dobrindt U. 2020. ClbR is the key transcriptional activator of colibactin gene expression in *Escherichia coli*. mSphere 5:e00591-20. https://doi.org/10.1128/mSphere.00591-20.
15. Putze J, Hennequin C, Nougayrède JP, Zhang W, Homburg S, Karch H, Bringer MA, Fayolle C, Carniel E, Rabsch W, Oelschlaeger TA, Oswald E, Forestier C, Hacker J, Dobrindt U. 2009. Genetic structure and distribution of the colibactin genomic island among members of the family *Enterobacteriaceae*. Infect Immun 77:4696–4703. https://doi.org/10.1128/IAI.00522-09.
16. Massip C, Branchu P, Bossuet-Greif N, Chagneau CV, Gaillard D, Martin P, Boury M, Sécher T, Dubois D, Nougayrède JP, Oswald E. 2019. Deciphering the interplay between the genotoxic and probiotic activities of *Escherichia coli* Nissle 1917. PLoS Pathog 15:e1008029. https://doi.org/10.1371/journal.ppat.1008029.
17. Buc E, Dubois D, Sauvanet P, Raisch J, Delmas J, Darfeuille-Michaud A, Pezet D, Bonnet R. 2013. High prevalence of mucosa-associated *E. coli* producing cyclomodulin and genotoxin in colon cancer. PLoS One 8:e56964. https://doi.org/10.1371/journal.pone.0056964.
18. Cougnoux A, Dalmasso G, Martinez R, Buc E, Delmas J, Gibold L, Sauvanet P, Darcha C, Déchelotte P, Bonnet M, Pezet D, Wodrich H, Darfeuille-Michaud A, Bonnet R. 2014. Bacterial genotoxin colibactin promotes colon tumour growth by inducing a senescence-associated secretory phenotype. Gut 63:1932–1942. https://doi.org/10.1136/gutjnl-2013-305257.
19. Dejea CM, Fathi P, Craig JM, Boleij A, Taddese R, Geis AL, Wu X, DeStefano Shields CE, Hechenbleikner EM, Huso DL, Anders RA, Giardiello FM, Wick EC, Wang H, Wu S, Pardoll DM, Housseau F, Sears CL. 2018. Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria. Science 359:592–597. https://doi.org/10.1126/science.aah3648.
20. Marcq I, Martin P, Payros D, Cuevas-Ramos G, Boury M, Watrin C, Nougayrède JP, Olier M, Oswald E. 2014. The genotoxin colibactin exacerbates lymphopenia and decreases survival rate in mice infected with septicemic *Escherichia coli*. J Infect Dis 210:285–294. https://doi.org/10.1093/infdis/jiu071.
21. Secher T, Payros D, Brehin C, Boury M, Watrin C, Gillet M, Bernard-Cadenat I, Menard S, Theodorou V, Saoudi A, Olier M, Oswald E. 2015. Oral tolerance failure upon neonatal gut colonization with *Escherichia coli* producing the genotoxin colibactin. Infect Immun 83:2420–2429. https://doi.org/10.1128/IAI.00064-15.
22. Lu MC, Chen YT, Chiang MK, Wang YC, Hsiao PY, Huang YJ, Lin CT, Cheng CC, Liang CL, Lai YC. 2017. Colibactin contributes to the hypervirulence of *pks*$^+$ K1 CC23 *Klebsiella pneumoniae* in mouse meningitis infections. Front Cell Infect Microbiol 7:103. https://doi.org/10.3389/fcimb.2017.00103.
23. Sarshar M, Scribano D, Marazzato M, Ambrosi C, Aprea MR, Aleandri M, Pronio A, Longhi C, Nicoletti M, Zagaglia C, Palamara AT, Conte MP. 2017. Genetic diversity, phylogroup distribution and virulence gene profile of *pks* positive *Escherichia coli* colonizing human intestinal polyps. Microb Pathog 112:274–278. https://doi.org/10.1016/j.micpath.2017.10.009.
24. Morgan RN, Saleh SE, Farrag HA, Aboulwafa MM. 2019. Prevalence and pathologic effects of colibactin and cytotoxic necrotizing factor-1 (Cnf 1) in *Escherichia coli*: experimental and bioinformatics analyses. Gut Pathog 11:1–18. https://doi.org/10.1186/s13099-019-0304-y.
25. Messerer M, Fischer W, Schubert S. 2017. Investigation of horizontal gene transfer of pathogenicity islands in *Escherichia coli* using next-generation sequencing. PLoS One 12:e0179880. https://doi.org/10.1371/journal.pone.0179880.
26. Suresh A, Ranjan A, Jadhav S, Hussain A, Shaik S, Alam M, Baddam R, Wieler LH, Ahmed N. 2018. Molecular genetic and functional analysis of *pks*-harboring, extra-intestinal pathogenic *Escherichia coli* from India. Front Microbiol 9:2631. https://doi.org/10.3389/fmicb.2018.02631.
27. Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA. 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics 12:402. https://doi.org/10.1186/1471-2164-12-402.
28. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q. 2005. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res 33:D325–D328. https://doi.org/10.1093/nar/gki008.
29. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, Von Mering C, Bork P. 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44:D286–D293. https://doi.org/10.1093/nar/gkv1248.
30. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274. https://doi.org/10.1093/molbev/msu300.
31. Cheng L, Connor TR, Sirén J, Aanensen DM, Corander J. 2013. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. Mol Biol Evol 30:1224–1228. https://doi.org/10.1093/molbev/mst028.
32. Weissman SJ, Johnson JR, Tchesnokova V, Billig M, Dykhuizen D, Riddell K, Rogers P, Qin X, Butler-Wu S, Cookson BT, Fang FC, Scholes D, Chattopadhyay S, Sokurenko E. 2012. High-resolution two-locus clonal typing of extraintestinal pathogenic *Escherichia coli*. Appl Environ Microbiol 78:1353–1360. https://doi.org/10.1128/AEM.06663-11.
33. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. 2016. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol 17:238. https://doi.org/10.1186/s13059-016-1108-8.
34. Oren Y, Smith MB, Johns NI, Zeevi MK, Biran D, Ron EZ, Corander J, Wang HH, Alm EJ, Pupko T. 2014. Transfer of noncoding DNA drives regulatory rewiring in bacteria. Proc Natl Acad Sci U S A 111:16112–16117. https://doi.org/10.1073/pnas.1413272111.
35. Thorpe HA, Bayliss SC, Sheppard SK, Feil EJ. 2018. Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. Gigascience 7:1–11. https://doi.org/10.1093/gigascience/giy015.
36. Roberts RJ, Vincze T, Posfai J, Macelis D. 2015. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res 43:D298–D299. https://doi.org/10.1093/nar/gku1046.
37. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461. https://doi.org/10.1093/bioinformatics/btq461.
38. Huson DH, Scornavacca C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol 61:1061–1067. https://doi.org/10.1093/sysbio/sys062.
39. Arthur JC, Perez-Chanona E, Mühlbauer M, Tomkovich S, Uronis JM, Fan T-J, Campbell BJ, Abujamel T, Dogan B, Rogers AB, Rhodes JM, Stintzi A, Simpson KW, Hansen JJ, Keku TO, Fodor AA, Jobin C. 2012. Intestinal inflammation targets cancer-inducing activity of the microbiota. Science 338:120–123. https://doi.org/10.1126/science.1224820.
40. Lopès A, Billard E, Casse AH, Villéger R, Veziant J, Roche G, Carrier G, Sauvanet P, Briat A, Pagès F, Naimi S, Pezet D, Barnich N, Dumas B, Bonnet M. 2020. Colibactin-positive *Escherichia coli* induce a procarcinogenic immune environment leading to immunotherapy resistance in colorectal cancer. Int J Cancer 146:3147–3159. https://doi.org/10.1002/ijc.32920.
41. Chagneau CV, Garcie C, Bossuet-Greif N, Tronnet S, Brachmann AO, Piel J, Nougayrède J-P, Martin P, Oswald E. 2019. The polyamine spermidine modulates the production of the bacterial genotoxin colibactin. mSphere 4:e00414-19. https://doi.org/10.1128/mSphere.00414-19.
42. Hussain A, Ranjan A, Nanwar N, Babbar A, Jadhav S, Ahmed N. 2014. Genotypic and phenotypic profiles of *Escherichia coli* isolates belonging to clinical sequence type 131 (ST131), clinical non-ST131, and fecal non-ST131 lineages from India. Antimicrob Agents Chemother 58:7240–7249. https://doi.org/10.1128/AAC.03320-14.
43. Hussain A, Ewers C, Nandanwar N, Guenther S, Jadhav S, Wieler LH, Ahmed N. 2012. Multiresistant uropathogenic *Escherichia coli* from a region in India where urinary tract infections are endemic: genotypic and phenotypic characteristics of sequence type 131 isolates of the CTX-M-15

extended-spectrum-β-lactamase-producing lineage. Antimicrob Agents Chemother 56:6358–6365. https://doi.org/10.1128/AAC.01099-12.
44. Ranjan A, Shaik S, Hussain A, Nandanwar N, Semmler T, Jadhav S, Wieler LH, Ahmed N. 2015. Genomic and functional portrait of a highly virulent, CTX-M-15-producing H30-Rx subclone of *Escherichia coli* sequence type 131. Antimicrob Agents Chemother 9:6087–6095. https://doi.org/10.1128/AAC.01447-15.
45. Ranjan A, Shaik S, Mondal A, Nandanwar N, Hussain A, Semmler T, Kumar N, Tiwari S, Jadhav S, Wieler LH, Ahmed N. 2016. Molecular epidemiology and genome dynamics of New Delhi metallo-β-lactamase-producing extraintestinal pathogenic *Escherichia coli* strains from India. Antimicrob Agents Chemother 60:6795–6805. https://doi.org/10.1128/AAC.01345-16.
46. Ranjan A, Shaik S, Nandanwar N, Hussain A, Tiwari SK, Semmler T, Jadhav S, Wieler LH, Alam M, Colwell RR, Ahmed N. 2017. Comparative genomics of *Escherichia coli* isolated from skin and soft tissue and other extraintestinal infections. mBio 8:e01070-17. https://doi.org/10.1128/mBio.01070-17.
47. Shaik S, Ranjan A, Tiwari SK, Hussain A, Nandanwar N, Kumar N, Jadhav S, Semmler T, Baddam R, Islam MA, Alam M, Wieler LH, Watanabe H, Ahmed N. 2017. Comparative genomic analysis of globally dominant ST131 clone with other epidemiologically successful extraintestinal pathogenic *Escherichia coli* (ExPEC) lineages. mBio 8:e01596-17. https://doi.org/10.1128/mBio.01596-17.
48. El Ghachi M, Bouhss A, Blanot D, Mengin-Lecreulx D. 2004. The *bacA* gene of *Escherichia coli* encodes an undecaprenyl pyrophosphate phosphatase activity. J Biol Chem 279:30106–30113. https://doi.org/10.1074/jbc.M401701200.
49. Olaitan AO, Morand S, Rolain JM. 2014. Mechanisms of polymyxin resistance: acquired and intrinsic resistance in bacteria. Front Microbiol 5:643. https://doi.org/10.3389/fmicb.2014.00643.
50. Johnson JR, Johnston B, Kuskowski MA, Nougayrede JP, Oswald E. 2008. Molecular epidemiology and phylogenetic distribution of the *Escherichia coli pks* genomic island. J Clin Microbiol 46:3906–3911. https://doi.org/10.1128/JCM.00949-08.
51. Martin P, Tronnet S, Garcie C, Oswald E. 2017. Interplay between siderophores and colibactin genotoxin in *Escherichia coli*. IUBMB Life 69:435–441. https://doi.org/10.1002/iub.1612.
52. Massip C, Chagneau CV, Boury M, Oswald E. 2020. The synergistic triad between microcin, colibactin, and salmochelin gene clusters in uropathogenic *Escherichia coli*. Microbes Infect 22:144–147. https://doi.org/10.1016/j.micinf.2020.01.001.
53. Yoshikawa Y, Tsunematsu Y, Matsuzaki N, Hirayama Y, Higashiguchi F, Sato M, Iwashita Y, Miyoshi N, Mutoh M, Ishikawa H, Sugimura H, Wakabayashi K, Watanabe K. 2020. Characterization of colibactin-producing *Escherichia coli* isolated from Japanese patients with colorectal cancer. Jpn J Infect Dis 73:437–442. https://doi.org/10.7883/yoken.JJID.2020.066.
54. Bengtsson S, Naseer U, Sundsfjord A, Kahlmeter G, Sundqvist M. 2012. Sequence types and plasmid carriage of uropathogenic *Escherichia coli* devoid of phenotypically detectable resistance. J Antimicrob Chemother 67:69–73. https://doi.org/10.1093/jac/dkr421.
55. Roer L, Hansen F, Frølund Thomsen MC, Knudsen JD, Hansen DS, Wang M, Samulioniené J, Justesen US, Røder BL, Schumacher H, Østergaard C, Andersen LP, Dzajic E, Søndergaard TS, Stegger M, Hammerum AM, Hasman H. 2017. WGS-based surveillance of third-generation cephalosporin-resistant *Escherichia coli* from bloodstream infections in Denmark. J Antimicrob Chemother 72:1922–1929. https://doi.org/10.1093/jac/dkx092.
56. Dubois D, Delmas J, Cady A, Robin F, Sivignon A, Oswald E, Bonnet R. 2010. Cyclomodulins in urosepsis strains of *Escherichia coli*. J Clin Microbiol 48:2122–2129. https://doi.org/10.1128/JCM.02365-09.
57. Kohoutova D, Smajs D, Moravkova P, Cyrany J, Moravkova M, Forstlova M, Cihak M, Rejchrt S, Bures J. 2014. *Escherichia coli* strains of phylogenetic group B2 and D and bacteriocin production are associated with advanced colorectal neoplasia. BMC Infect Dis 14:733. https://doi.org/10.1186/s12879-014-0733-7.
58. Nandanwar N, Janssen T, Kühl M, Ahmed N, Ewers C, Wieler LH. 2014. Extraintestinal pathogenic *Escherichia coli* (ExPEC) of human and avian origin belonging to sequence type complex 95 (STC95) portray indistinguishable virulence features. Int J Med Microbiol 304:835–842. https://doi.org/10.1016/j.ijmm.2014.06.009.
59. Gordon DM, Geyik S, Clermont O, O'Brien CL, Huang S, Abayasekara C, Rajesh A, Kennedy K, Collignon P, Pavli P, Rodriguez C, Johnston BD, Johnson JR, Decousser J-W, Denamur E. 2017. Fine-scale structure analysis shows epidemic patterns of clonal complex 95, a cosmopolitan *Escherichia coli* lineage responsible for extraintestinal infection. mSphere 2:e00168-17. https://doi.org/10.1128/mSphere.00168-17.
60. Ewers C, Dematheis F, Singamaneni HD, Nandanwar N, Fruth A, Diehl I, Semmler T, Wieler LH. 2014. Correlation between the genomic *o454-nlpD* region polymorphisms, virulence gene equipment and phylogenetic group of extraintestinal *Escherichia coli* (ExPEC) enables pathotyping irrespective of host, disease and source of isolation. Gut Pathog 6:37. https://doi.org/10.1186/s13099-014-0037-x.
61. McNally A, Oren Y, Kelly D, Pascoe B, Dunn S, Sreecharan T, Vehkala M, Välimäki N, Prentice MB, Ashour A, Avram O, Pupko T, Dobrindt U, Literak I, Guenther S, Schaufler K, Wieler LH, Zhiyong Z, Sheppard SK, McInerney JO, Corander J. 2016. Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. PLoS Genet 12:e1006280. https://doi.org/10.1371/journal.pgen.1006280.
62. Bausch C, Peekhaus N, Utz C, Blais T, Murray E, Lowary T, Conway T. 1998. Sequence analysis of the GntII (subsidiary) system for gluconate metabolism reveals a novel pathway for L-idonic acid catabolism in *Escherichia coli*. J Bacteriol 180:3704–3710. https://doi.org/10.1128/JB.180.14.3704-3710.1998.
63. Oliveira PH, Touchon M, Rocha EPC. 2014. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. Nucleic Acids Res 42:10618–10631. https://doi.org/10.1093/nar/gku734.
64. Stephens CM, Adams-Sapper S, Sekhon M, Johnson JR, Riley LW. 2017. Genomic analysis of factors associated with low prevalence of antibiotic resistance in extraintestinal pathogenic *Escherichia coli* sequence type 95 strains. mSphere 2:e00390-16. https://doi.org/10.1128/mSphere.00390-16.
65. Forde BM, Phan MD, Gawthorne JA, Ashcroft MM, Stanton-Cook M, Sarkar S, Peters KM, Chan KG, Chong TM, Yin WF, Upton M, Schembri MA, Beatson SA. 2015. Lineage-specific methyltransferases define the methylome of the globally disseminated *Escherichia coli* ST131 clone. mBio 6:e01602-15. https://doi.org/10.1128/mBio.01602-15.
66. Nandi T, Holden MTG, Didelot X, Mehershahi K, Boddey JA, Beacham I, Peak I, Harting J, Baybayan P, Guo Y, Wang S, How LC, Sim B, Essex-Lopresti A, Sarkar-Tyson M, Nelson M, Smither S, Ong C, Aw LT, Hoon CH, Michell S, Studholme DJ, Titball R, Chen SL, Parkhill J, Tan P. 2015. *Burkholderia pseudomallei* sequencing identifies genomic clades with distinct recombination, accessory, and epigenetic profiles. Genome Res 25:129–141. https://doi.org/10.1101/gr.177543.114.
67. Budroni S, Siena E, Dunning Hotopp JC, Seib KL, Serruto D, Nofroni C, Comanducci M, Riley DR, Daugherty SC, Angiuoli SV, Covacci A, Pizza M, Rappuoli R, Moxon ER, Tettelin H, Medini D. 2011. *Neisseria meningitidis* is structured in clades associated with restriction modification systems that modulate homologous recombination. Proc Natl Acad Sci U S A 108:4494–4499. https://doi.org/10.1073/pnas.1019751108.
68. Roer L, Hendriksen RS, Leekitcharoenphon P, Lukjancenko O, Kaas RS, Hasman H, Aarestrup FM. 2016. Is the evolution of *Salmonella enterica* subsp. *enterica* linked to restriction-modification systems? mSystems 1:e00009-16. https://doi.org/10.1128/mSystems.00009-16.
69. Jadhav S, Hussain A, Devi S, Kumar A, Parveen S, Gandham N, Wieler LH, Ewers C, Ahmed N. 2011. Virulence characteristics and genetic affinities of multiple drug resistant uropathogenic *Escherichia coli* from a semi urban locality in India. PLoS One 6:e18063. https://doi.org/10.1371/journal.pone.0018063.
70. Hussain A, Shaik S, Ranjan A, Suresh A, Sarker N, Semmler T, Wieler LH, Alam M, Watanabe H, Chakravortty D, Ahmed N. 2019. Genomic and functional characterization of poultry *Escherichia coli* from India revealed diverse extended-spectrum β-lactamase-producing lineages with shared virulence profiles. Front Microbiol 10:2766. https://doi.org/10.3389/fmicb.2019.02766.
71. Patel RK, Jain M. 2012. NGS QC toolkit: a toolkit for quality control of next generation sequencing data. PLoS One 7:e30619. https://doi.org/10.1371/journal.pone.0030619.
72. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
73. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086.
74. Shaik S, Kumar N, Lankapalli AK, Tiwari SK, Baddam R, Ahmed N. 2016. Contig-Layout-Authenticator (CLA): a combinatorial approach to ordering and scaffolding of bacterial contigs for comparative genomics and molecular

epidemiology. PLoS One 11:e0155459. https://doi.org/10.1371/journal.pone.0155459.
75. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153.
76. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. 2000. Artemis: sequence visualization and annotation. Bioinformatics 16:944–945. https://doi.org/10.1093/bioinformatics/16.10.944.
77. Hussain A, Shaik S, Ranjan A, Nandanwar N, Tiwari SK, Majid M, Baddam R, Qureshi IA, Semmler T, Wieler LH, Islam MA, Chakravortty D, Ahmed N. 2017. Risk of transmission of antimicrobial resistant *Escherichia coli* from commercial broiler and free-range retail chicken in India. Front Microbiol 8:2120. https://doi.org/10.3389/fmicb.2017.02120.
78. Jolley KA, Maiden MCJ. 2010. BIGSdb: scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11:595. https://doi.org/10.1186/1471-2105-11-595.
79. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421.
80. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, Bhullar K, Canova MJ, De Pascale G, Ejim L, Kalan L, King AM, Koteva K, Morar M, Mulvey MR, O'Brien JS, Pawlowski AC, Piddock LJV, Spanogiannopoulos P, Sutherland AD, Tang I, Taylor PL, Thaker M, Wang W, Yan M, Yu T, Wright GD. 2013. The Comprehensive Antibiotic Resistance Database. Antimicrob Agents Chemother 57:3348–3357. https://doi.org/10.1128/AAC.00419-13.
81. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. 2011. PHAST: a fast phage search tool. Nucleic Acids Res 39:W347–W352. https://doi.org/10.1093/nar/gkr485.
82. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
83. Langmead B, Salzberg S. 2013. Fast-gapped read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923.
84. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691–3693. https://doi.org/10.1093/bioinformatics/btv421.
85. Roer L, Johannesen TB, Hansen F, Stegger M, Tchesnokova V, Sokurenko E, Garibay N, Allesøe R, Thomsen MCF, Lund O, Hasman H, Hammerum AM. 2018. CHTyper, a web tool for subtyping of extraintestinal pathogenic *Escherichia coli* based on the *fumC* and *fimH* alleles. J Clin Microbiol 56:e00063-18. https://doi.org/10.1128/JCM.00063-18.
86. Löytynoja A. 2014. Phylogeny-aware alignment with PRANK. Methods Mol Biol 2231:17–37. https://doi.org/10.1007/978-1-62703-646-7_10.
87. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973. https://doi.org/10.1093/bioinformatics/btp348.
88. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14:587–589. https://doi.org/10.1038/nmeth.4285.
89. Didelot X, Wilson DJ. 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol 11:e1004041. https://doi.org/10.1371/journal.pcbi.1004041.
90. Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 47:W256–W259. https://doi.org/10.1093/nar/gkz239.

Genomic and functional analysis of colibactin harboring Escherichia coli

by Arya Suresh



Genomic and functional analysis of colibactin harboring Escherichia coli

ORIGINALITY REPORT

14%	10%	12%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	edoc.rki.de Internet Source	7%
2	www.frontiersin.org Internet Source	1%
3	Arya Suresh, Amit Ranjan, Savita Jadhav, Arif Hussain, Sabiha Shaik, Munirul Alam, Ramani Baddam, Lothar H. Wieler, Niyaz Ahmed. "Molecular Genetic and Functional Analysis of pks-Harboring, Extra-Intestinal Pathogenic Escherichia coli From India", Frontiers in Microbiology, 2018 Publication	1%
4	Jolanta Sarowska, Bozena Futoma-Koloch, Agnieszka Jama-Kmiecik, Magdalena Frej-Madrzak et al. "Virulence factors, prevalence and potential transmission of extraintestinal pathogenic Escherichia coli isolated from different sources: recent reports", Gut Pathogens, 2019 Publication	1%

This is candidate's paper used as source.

Professor Dr. Niyaz Ahmed, PhD
Department of Biotechnology & Bioinformatics
University of Hyderabad, Hyderabad, India

This is candidate's paper.

Professor Dr. Niyaz Ahmed, PhD
Department of Biotechnology & Bioinformatics
University of Hyderabad, Hyderabad, India

Similarity Screening Done @ IGM Library
Sub ID: 1591786725
Sub date: 22-05-2021
Librarian / DLI / CL
IGM Library (UGM)

5	mdpi.com Internet Source	<1 %
6	Submitted to University of Liverpool Student Paper	<1 %
7	Jorge Carlos Navarro-Muñoz, Jérôme Collemare. "Evolutionary Histories of Type III Polyketide Synthases in Fungi", Frontiers in Microbiology, 2020 Publication	<1 %
8	irep.ntu.ac.uk Internet Source	<1 %
9	Ying-Tsong Chen, Yi-Chyi Lai, Mei-Chen Tan, Li-Yun Hsieh et al. "Prevalence and characteristics of pks genotoxin gene cluster-positive clinical Klebsiella pneumoniae isolates in Taiwan", Scientific Reports, 2017 Publication	<1 %
10	mbio.asm.org Internet Source	<1 %
11	pdfs.semanticscholar.org Internet Source	<1 %
12	Arif Hussain, Sabiha Shaik, Amit Ranjan, Nishant Nandanwar et al. "Risk of Transmission of Antimicrobial Resistant Escherichia coli from Commercial Broiler and	<1 %

Free-Range Retail Chicken in India", Frontiers in Microbiology, 2017
Publication

---

13 Zhang, Shuhong, Qingping Wu, Jumei Zhang, and Xuemei Zhu. "Occurrence and Characterization of Enteropathogenic Escherichia coli (EPEC) in Retail Ready-to-Eat Foods in China", Foodborne Pathogens and Disease, 2015.
Publication

<1 %

---

14 "Inflammation, Infection, and Microbiome in Cancers", Springer Science and Business Media LLC, 2021
Publication

<1 %

---

15 Pérez Carrascal, Olga M., David VanInsberghe, Soledad Juárez, Martin F. Polz, Pablo Vinuesa, and Víctor González. "Population Genomics of the Symbiotic Plasmids of Sympatric Nitrogen-Fixing Rhizobium Species Associated with Phaseolus vulgaris", Environmental Microbiology, 2016.
Publication

<1 %

---

16 Jessika Nowak, Cristina D. Cruz, Marcel Tempelaars, Tjakko Abee et al. "Persistent Listeria monocytogenes strains isolated from mussel production facilities form more biofilm but are not linked to specific genetic

<1 %

markers", International Journal of Food Microbiology, 2017
Publication

17	Sifuna Anthony Wawire, Oleg N. Reva, Thomas J. O'Brien, Wendy Figueroa, Victor Dinda, William A. Shivoga, Martin Welch. "Virulence and antimicrobial resistance genes are enriched in the plasmidome of clinical Escherichia coli isolates compared with wastewater isolates from western Kenya", Infection, Genetics and Evolution, 2021 Publication	<1 %
18	"Microbial Toxins", Springer Science and Business Media LLC, 2018 Publication	<1 %
19	Maria G. Spindola, Marcos P. V. Cunha, Luisa Z. Moreno, Cristina R. Amigo et al. " Genetic diversity, virulence genotype and antimicrobial resistance of uropathogenic (UPEC) isolated from sows ", Veterinary Quarterly, 2018 Publication	<1 %
20	aac.asm.org Internet Source	<1 %
21	seenthis.net Internet Source	<1 %
22	theses.gla.ac.uk Internet Source	

		<1 %
23	"Posters", FEMS Microbiology Letters, 2003 Publication	<1 %
24	Submitted to Central Queensland University Student Paper	<1 %
25	Danyu Chen, Wencheng Zou, Shengze Xie, Linghan Kong et al. " Serotype and Antimicrobial Resistance of Isolated from Feces of Wild Giant Pandas ( ) in Sichuan Province, China ", Journal of Wildlife Diseases, 2018 Publication	<1 %
26	abasy.ccg.unam.mx Internet Source	<1 %
27	www.mdpi.com Internet Source	<1 %
28	Nougayrede, J.P.. "Cyclomodulins: bacterial effectors that modulate the eukaryotic cell cycle", Trends in Microbiology, 200503 Publication	<1 %
29	Submitted to Stonehill College Student Paper	<1 %
30	www.researchgate.net Internet Source	<1 %

31	"Posters", Clinical Microbiology and Infection, 2009 Publication	<1 %
32	Arif Hussain, Sabiha Shaik, Amit Ranjan, Arya Suresh et al. "Genomic and Functional Characterization of Poultry Escherichia coli From India Revealed Diverse Extended-Spectrum β-Lactamase-Producing Lineages With Shared Virulence Profiles", Frontiers in Microbiology, 2019 Publication	<1 %
33	Ascel Samba-Louaka. "Bacterial cyclomodulin Cif blocks the host cell cycle by stabilizing the cyclin-dependent kinase inhibitors $p21^{waf1}$ and $p27^{kip1}$", Cellular Microbiology, 12/2008 Publication	<1 %
34	Oswald, E.. "Bacterial toxins that modulate host cell-cycle progression", Current Opinion in Microbiology, 200502 Publication	<1 %
35	Submitted to University of Bath Student Paper	<1 %
36	journals.plos.org Internet Source	<1 %
37	ueaeprints.uea.ac.uk Internet Source	<1 %

38	Secher, Thomas, Ascel Samba-Louaka, Eric Oswald, and Jean-Philippe NougayrÃ¨de. "Escherichia coli Producing Colibactin Triggers Premature and Transmissible Senescence in Mammalian Cells", PLoS ONE, 2013. Publication	<1 %
39	Harry L. T. Mobley. "Pathogenic Escherichia coli", Nature Reviews Microbiology, 02/2004 Publication	<1 %
40	Pham, P. H., Y. J. Huang, C. Chen, and N. C. Bols. "Corexit 9500 Inactivates Two Enveloped Viruses of Aquatic Animals but Enhances the Infectivity of a Nonenveloped Fish Virus", Applied and Environmental Microbiology, 2014. Publication	<1 %
41	SpringerBriefs in Food Health and Nutrition, 2015. Publication	<1 %
42	onlinelibrary.wiley.com Internet Source	<1 %
43	trepo.tuni.fi Internet Source	<1 %

Exclude quotes On Exclude matches < 14 words

Exclude bibliography On