High throughput computational genomics of different lineages of Extra-intestinal Pathogenic *Escherichia coli*

Thesis submitted to the University of Hyderabad
for the degree of

DOCTOR OF PHILOSOPHY

By

Shaik Sabiha
(Reg. No: 12LTPH04)



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

May, 2017

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 **"High throughput computational genomics of different lineages of Extra-intestinal Pathogenic *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. Dr. 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:

Signature:

Name: Shaik Sabiha

Reg. No.: 12LTPH04

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

Date: 17/03/2017

CERTIFICATE

This is to certify that the thesis entitled **"High throughput computational genomics of different lineages of Extra-intestinal Pathogenic *Escherichia coli*"** submitted by **Ms. Shaik Sabiha** bearing registration number **12LTPH04** in partial fulfillment of the requirements for award of **Doctor of Philosophy** in the School of Life Sciences is a bona fide 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 award of any degree or diploma.

Parts of this thesis have been:

A. Published in the following publications:

Shaik S, et al. 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.

B. Presented in the following conferences:

1. *Indo-German International Research Training Group (IRTG-GRK1673) Winter School on "Pathogen Biology and Genomics"* held from 23rd to 27th February, 2015 at University of Hyderabad, India (International).
2. *20th Scientific Meeting of Malaysian Society for Molecular Biology and Biotechnology* held from 26th -27th June 2013 at University of Malaya, Malaysia (International)

Further, the student has passed the following courses towards the fulfillment of course work required for the award of Ph.D

S.no	Course Code	Name	Credits	Pass/Fail
1	BT 801	Seminar 1	1	PASS
2	BT 802	Research Ethics & Management	2	PASS
3	BT 803	Biostatistics	2	PASS
4	BT 804	Analytical Techniques	3	PASS
5	BT 805	Lab Work	4	PASS

Prof. Dr. Niyaz Ahmed
Research supervisor
Dept. of Biotechnology and Bioinformatics

Head
Dept. of Biotechnology and Bioinformatics

Dean
School of Life Sciences

Dedicated to my beloved parents

Dr. Shaik Ahmed

Fathimunnisa Bibi

Acknowledgement

I am heavily indebted to my mentor Prof. Dr. Niyaz Ahmed for introducing me to the world of genomics. His mentorship and constant support has encouraged me to learn many new things as a part of my PhD. He has always been an inspiration which ultimately motivated us to take up challenges that could match up to his standards.

I would also like to thank the doctoral committee members Dr. Insaf Ahmed Qureshi and Dr. Vaibhav Vindal for their constant support and valuable suggestions during the entire thesis work. I would like to thank the present and past heads of Department of Biotechnology and Bioinformatics and present and past deans of School of Life sciences for use of the departmental/School facilities. I would also like to acknowledge the entire department and school's non-teaching staff for their help in doing all the official works throughout this period of work. Dr. Savita Jadhav is highly acknowledged for the provision of clinical samples from D. Y Patil hospital, Pune for the research work. Prof. Dr. L H Wieler and Mr. Torsten semmler are acknowledged for helping us with sequencing of the strains.

I would like to acknowledge DBT, CSIR for support through fellowship and University of Malaya-High Impact Research grant, IRTG GRK 1673 grant and DBT for funding my research and lab facilities. I would like to acknowledge the Bioinformatics Facility (BIF) for the use of servers and computational infrastructure. I would also like to acknowledge DBT, DST, UGC-SAP, DST-FIST, UPE, CSIR & DBT-CREB for funding Department and school facilities.

I am also thankful to the present and past members of Pathogen Biology for their lovely support and friendship. I have no words to express my gratitude for Dr. Narender Kumar who has helped in

understanding the concepts of NGS and genomics. My first objective would not have been possible without him. I am also thankful to Dr. Ramani Baddam for being a critic as well as a great friend. Her comments have always helped in the refinement of my work. Presence of my NGS team members Aditya, Kshitij, Sumeet, Arya and Kishore has always made the atmosphere in the computational lab livelier. Special thanks to Aditya for helping me out in troubleshooting the scripts. I would also like to thank Dr. Arif Hussain, Dr. Amit Ranjan and Dr. Nishant for their guidance and help in my other two objectives. I also thank Dr. Shivendra, Dr. Suhail, Dr. Ashutosh, Dr. Savita, Dr. Vidyullatha, Dr. Majid, Shankar, Priya, Shamsul, Naveen, Shveta and Majjid for their encouragement and lively discussions. Very special thanks to Akhila Bommakanti and Snigdha pola for being there through good as well as bad times.

I am thankful to God for giving me such encouraging and understanding parents. Its only because of the liberty they gave me I could choose to follow my passion. I and my brother could achieve so much in life only because of their sacrifices and hard work. I thank my brother for supporting and believing in me. And special thanks to Riaz and Daanish for making me more responsible and dedicating. Finally, I am grateful to God for providing this life.

Shaik Sabiha

Index

Abbreviations **i-ii**

Chapter1: Introduction and state of the art **1**

Chapter 2: Objective 1 **31**
Contig-Layout-Authenticator (CLA): Combinatorial approach for ordering and scaffolding of contigs

Chapter 3: Objective 2 **67**
To analyze genomes of pandemic ExPEC Sequence types- ST131, ST38, ST405 and ST648, to provide insight into the genetic level similarities/differences among these clones

Chapter 4: Objective 3 **121**
To analyze genomes of *Escherichia coli* isolates from Indian retail poultry and different human *E. coli* pathotypes in order to evaluate their risk in zoonotic transmission of antimicrobial resistance

Chapter 5: Summary and epilogue **139**

Publications **145**

Abbreviations

APEC	**Avian pathogenic *Escherichia coli***
CARD	**Comprehensive Antibiotic Resistance Database**
CDS	**Coding sequences**
CFs	**Colonization factors**
CLA	**Contig Layout Authenticator**
COG	**Clusters of Orthologous groups**
DAEC	**Diffusely adherent *Escherichia coli***
DEC	**Diarrheagenic *Escherichia coli***
E. coli	***Escherichia coli***
EAEC	**Enteroaggregative *Escherichia coli***
ECOR	***E. coli* reference strain collection**
EHEC	**Enterohaemorrhagic *Escherichia coli***
EIEC	**Enteroinvasive *Escherichia coli***
EPEC	**Enteropathogenic *Escherichia coli***
ESBL	**Extended spectrum of beta-lactamase**
ETEC	**Enterotoxigenic *Escherichia coli***
ExPEC	**Extraintestinal pathogenic *Escherichia coli***

GI tract	**Gastro-intestinal tract**
IS	**Insertion Sequence**
MCM	**Mauve contig mover**
MDR	**Multidrug resistance**
MGE	**Mobile genetic elements**
MLEE	**Multi-locus enzyme electrophoresis**
MLST	**Multi-locus strain typing**
MPEC	**Mammary associated *Escherichia coli***
NDM	**New Delhi metallo-beta-lactamase**
NGS	**Next generation sequencing**
NMEC	**Neonatal meningitis *Escherichia coli***
PCR	**Polymerase chain reaction**
PFGE	**Pulse field gel electrophoresis**
pMLST	**Plasmid Multi-locus strain typing**
SEPEC	**Sepsis associated pathogenic *Escherichia coli***
SSTI	**Skin and soft tissue infection**
ST	**Sequence type**
STEC	**Shigatoxigenic *Escherichia coli***
UPEC	**Uropathogenic *Escherichia coli***

UTI	**Urinary tract infection**
VF	**Virulence factors**
VFDB	**Virulence factor database**
WGS	**Whole genome sequencing**

CHAPTER 1

INTRODUCTION AND STATE OF ART

Escherichia coli

Escherichia coli (E. coli), most commonly found as a commensal in the gastrointestinal (GI) tracts of all warm-blooded and few cold-blooded organisms belongs to the family of *Enterobacteriaceae* (Blount, 2015; Conway & Cohen, 2015; Kaper, Nataro, & Mobley, 2004). *E. coli* is usually harbored in the gut of these organisms throughout their life after being acquired within a short span from birth (Kaper et al., 2004). It is known to restrict the invasion of pathogens by competing for nutrients and niche within the gut thus benefitting its host by being a commensal (Conway & Cohen, 2015; Sweeney et al., 1996). Other than the gut, it is also known to be ubiquitously present in the environmental water bodies, soil, algal surfaces, etc.(Ishii & Sadowsky, 2008). While these commensal forms are benefitting humans, there are other pathogenic forms that usually cause different types of diseases in humans and animals (Kaper et al., 2004). These pathogenic forms are hypothesized to have evolved from their commensal counterparts by gaining virulence and resistance gene repertoires in the form of mobile genetic elements through horizontal gene transfer (fig1)(Ahmed, Dobrindt, Hacker, & Hasnain, 2008; Tenaillon, Skurnik, Picard, & Denamur, 2010).

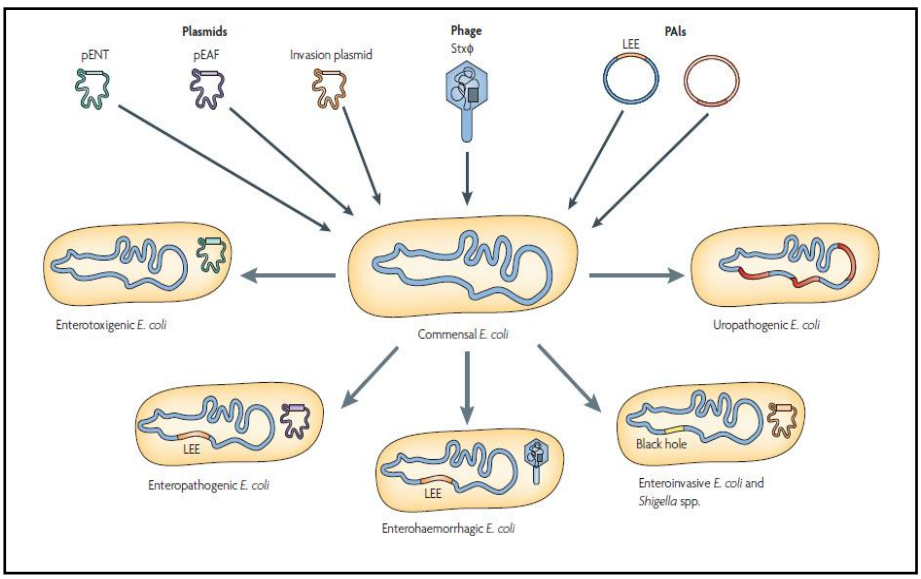


Fig1: Different mechanisms that have contributed to the evolution of several pathogenic forms from the commensal *E.coli* (Ahmed et al., 2008)

Pathogenic and non-pathogenic forms of *E. coli*

Based on their ability to cause infections in humans and other animal hosts, *E. coli* are broadly categorized into pathogenic and non-pathogenic forms. Harm-less non-pathogenic forms can be found in the environment and in the gut of warm blooded animals as commensals. Pathogenic forms can cause infections at different anatomical sites and based on this they are classified into diarrheagenic *E. coli* (DEC) and extraintestinal pathogenic *E. coli* (ExPEC) (Kaper et al., 2004; Nataro & Kaper, 1998). DEC pathotypes cause intestinal infections and are further classified into several pathotypes based on their preferential host colonization sites, virulence mechanisms, and the ensuing clinical symptoms and consequences (Gomes et al., 2016). These include enteropathogenic *E. coli* (EPEC), enterohemorrhagic (Shiga toxin-producing) *E. coli* (EHEC/STEC), enteroaggregative *E. coli* (EAEC), enterotoxigenic *E. coli* (ETEC), enteroinvasive *E. coli* (EIEC) and Diffusely Adherent *E. coli* (DAEC)(Clements, Young, Constantinou, & Frankel, 2012). Pathotypes such as uropathogenic *E. coli* (UPEC), neonatal meningitis-associated *E. coli* (NMEC), sepsis-causing *E. coli* (SEPEC), Avian Pathogenic *E. coli* (APEC), Mammary pathogenic *E. coli* (MPEC) and Skin and soft tissue infection associated (SSTI) *E. coli* are all different forms of ExPEC causing infections at extra-intestinal sites(Dale & Woodford, 2015; Ieka, Iddingsb, Oetkottc, Ohnsona, & Olana, 2005; Shpigel, Elazar, & Rosenshine, n.d.; Starc, Elers, Gubina, & Darja, 2009). Since infections caused by strains from these groups were not limited to one anatomical site, Russo and Johnson have defined EXPEC as any non-commensal *E. coli* that can cause any extra-intestinal diseases(Russo & Johnson, 2000).

Significance of ExPEC infections

ExPEC strains usually reside in the gut and cause infections after migrating to other parts such as urinary tract, blood or lungs (Smith, Fratamico, & Gunther, 2007). Since these strains are also carried by animals and poultry they exhibit greater zoonotic potential. ExPEC also includes APEC strains that cause illness in poultry. It has been reported that APEC strains share high genetic similarity and similar antibiotic and virulence gene patterns with ExPEC strains that cause diseases in poultry and humans. Other than this, it was observed that certain multi drug resistant clonal groups or lineages are usually responsible for these extra-intestinal infections in humans (Manges & Johnson, 2012; Smith et al., 2007).

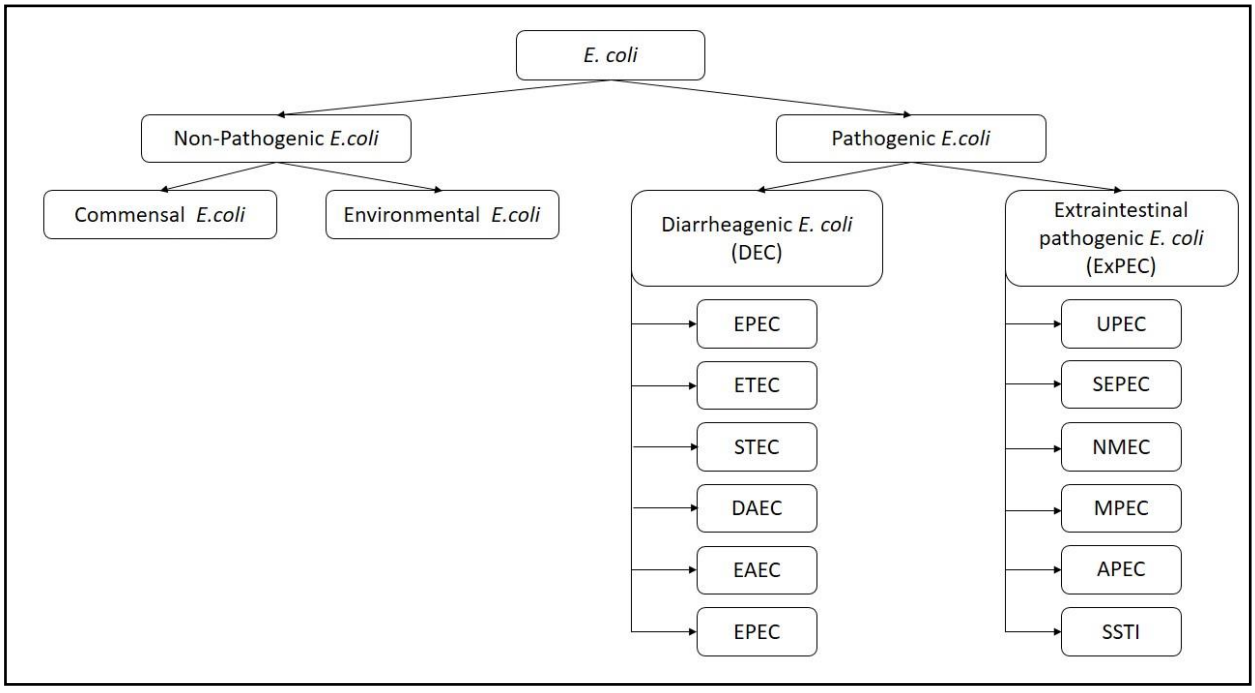


Fig 2: Schematic representation of *E.coli* pathotypes

Majority of the ExPEC are associated with community-acquired urinary tract infections (UTIs), pyelonephritis, prostatitis and catheter-associated UTIs and less frequently with skin and soft tissue infections, neonatal meningitis and hospital-acquired pneumonia(Dale & Woodford, 2015). If not treated properly, these infections can be life threatening. Usually

common among females, it is estimated that at least 40-50% of women and 5% of men suffer with UTI in their lifetime (Totsika et al., 2012). UTI can be non-symptomatic or symptomatic (Urethritis) and can further progress to complicated stages like cystitis, pyelonephritis and urosepsis. In developing countries like India, ExPEC infections are most common and this may be due to lack of proper sanitation facilities, awareness and high population; therefore requires significant attention.

Virulence potential in ExPEC

It was earlier reported that presence of certain pathogenic virulence factors such as *papA* and/or *papC, sfa/foc, afa/draBC, kpsM II* and *iutA* within ExPEC differentiated them from the commensal strains. Currently, many Virulence factors (VFs) are associated with ExPEC pathogenicity and they are classified into five major groups namely adhesins, toxins, iron acquisition systems, capsule production, protectins and invasins (Dale & Woodford, 2015). Genes from each of these groups have specific roles to play and they together contribute to the overall fitness of pathogenic strains. While adhesins help these bacteria to attach themselves to the surface of the epithelial cells, toxins secreted by them damage the host epithelial cells (Flores-Meireles, Walker, Caparon, & Hultgren, 2015). Siderophores and iron acquisition systems help in competing with the host for iron (Gao et al., 2012). Protectants and biofilms protect these organisms from the host serum and immune responses. Additionally, biofilms restrict drug accessibility (Beloin, Roux, & Ghigo, 2010). Table1 represents a brief description of virulence factors and genes associated with ExPEC.

Table 1: Virulence factors associated with ExPEC(Dale & Woodford, 2015)

Virulence factor	**gene**
Adhesins	
Adhesion siderophore	*iha*
Dr binding adhesins	*afa/draBC*
E. coli common pilus	*ecpA*
F1C fimbriae	*foc* gene cluster
Heat-resistant haemagglutinin	*hra*
M fimbriae	*bmaE*
N-acetyl D-glucosamine-specific fimbriae	*gaf*
P fimbriae	*papACEFG*
S fimbriae	*sfa/sfaS*
Type 1 fimbriae	*fimH*
Iron acquisition systems	
Aerobactin receptor	*iutA*
Peri-plasmic iron binding protein	*SitA*
Salmochelin receptor	*iroN*
Siderophore receptor	*ireA*
Yersiniabactin receptor	*fyuA*
Protectins and invasins	
Colicin V	*Cva*
Conjugal transfer surface exclusion protein	*traT*
Group 3 capsule	*kpsMT II*
Increased serum survival	*iss*
Invasion of brain endothelium	*ibeA*

K1/K2/K5 group 2 capsule variants	K1/K2/K5 genes
kpsM II group 2 capsule	*kpsM II*
Outer membrane protease T	*ompT*
Toxins	
α-haemolysin	*hylD*
Cytolethal distending toxin	*cdtB*
Cytotoxic necrotising factor	*Cnf1*
Enteroaggregative E. coli toxin	*astA*
Haemolysin A	*hylA*
Secreted autotransporter toxin	*sat*
Serine protease	*pic*
Vacuolating toxin	*vat*
Others	
β-glucoronidase	*uidA*
Colibactin synthesis	*Clb & clbB*
Uropathogenic-specific protein	*usp*
Flagellin variant	H7 *fliC*
Maltose and glucose-specific PTS transporter subunit IICB	*MalX*
Pathogenicity island marker	*malX*
D-serine deaminase	*DsdA*

Multi drug resistance in ExPEC

Multi drug resistance genes such as Extended spectrum β-lactamases (e.g., CTX-M-type), plasmid-mediated AMP C β-lactamases (e.g., CMY types) and carbapenemases are usually located on conjugative plasmids, transposases, insertion sequences and integrases and are acquired through horizontal gene transfer (Pitout, 2012). Treating ExPEC infections have

become complicated due to multi drug resistance in *E. coli* and this is because of the indiscriminate use of antibiotics such as cephalosporins, fluoroquinolones and trimethoprim-sulfamethoxazole (22). Treatment failures due to multidrug resistance are therefore leading to increased rates of hospitalization, morbidity, and mortality; and increased associated healthcare and societal costs (Poolman & Wacker, 2016). This is a grave concern in developing countries such as India where ~80% of infections are reported to be due to ESBL producers.

Evolution of successful lineages of ExPEC

Clonal evolution of certain *E. coli* strains had led to the worldwide dissemination of highly successful clones such as ST131, ST38, ST405 and ST648 (Pitout, 2012). ST131 with the best combination of virulence factors and multidrug resistance is the most studied clone all over the world (Petty et al., 2014). This clone which is usually associated with urinary tract as well as blood stream infections has been widely studied due to its pre-dominance over other pathogenic ExPEC clones(Hussain et al., 2012; Nicolas-Chanoine et al., 2008). This globally disseminated multidrug resistant clone belonging to phylogenetic group B2 has also been associated with the worldwide emergent CTX-M-15 enzyme (Schembri et al., 2015). Relatively higher virulence and metabolic capabilities along with resistance to wider classes of antibiotics are hypothesized to have contributed to the epidemiological success of this particular lineage (Hussain et al., 2014; Vimont et al., 2012).

E.coli typing methods

Subtyping *E.coli* or any other bacteria helps in discrimination beyond the species or sub-species levels and can be of great help while studying outbreaks (Fratamico et al., 2016). Many traditional as well advanced methodologies have been in use to define and understand

ExPEC lineages and their phylogeny (Dale & Woodford, 2015). The oldest approach (during 1970s) to delineate *E. coli* isolates was O, H and K antigen-based serotyping using which few ExPEC strains from humans and other animals were grouped (Bettelheim et al., 1976). During 1984, the *E. coli* reference strain collection (ECOR) comprising 72 isolates from human and other mammalian hosts were established by Ochman and Selander (Ochman & Selander, 1984). Phylogroups A, B1, B2, D and E were then used to categorize *E. coli* based on the Multi-locus enzyme electrophoresis (MLEE). This was followed by much faster and cost effective triplex polymerase chain reaction (PCR)-based method which in turn got replaced by more advanced multiplex PCR methods to assign the strains to these phylogroups (Clermont, Bonacorsi, Bingen, & Bonacorsi, 2000; Doumith, Day, Hope, Wain, & Woodford, 2012). Sequence based genotyping methods then superseded these traditional approaches to define in depth discrimination of *E. coli* strains. So far, MLST has been the most widely used technique where strains are categorized into different sequence types (STs) based on the allelic combinations of different housekeeping genes (Sullivan, Diggle, & Clarke, 2005; Van der Bij, Peirano, Pitondo-Silva, & Pitout, 2012). Analyzing ExPEC strains based on MLST can provide greater resolution by segregating them into different STs (Wirth et al., 2006). Currently three MLST schemes are available for *E. coli*, of which the Achtman scheme (http://mlst.warwick.ac.uk/mlst/dbs/Ecoli) is the most commonly used (Clermont, Gordon, & Denamur, 2015). For in-depth phylogenetic analysis of outbreak cases, much updated techniques such as pulsed-field gel electrophoresis (PFGE) and whole genome sequencing (WGS) can provide higher resolution when compared to MLST. Strengths and weaknesses of the above described methods are tabulated in Table 2.

Table 2: Strengths and weaknesses of ExPEC typing methods(Dale & Woodford, 2015)

Typing method	Strengths	Weaknesses
Phenotypic Multi-locus enzyme electrophoresis (MLEE)	• Majority of strains can be typed. • Easy to interpret results.	• Provides poor phylogenetic resolution compared with newer techniques.
Genetic Multiplex PCR-based techniques	• Allow for rapid determination of phylogroup. • Potential application in clinical setting (e.g. in predicting likely antibiogram or as screening tool in the setting outbreak • Relatively low cost.	• Only allow for detection of specific genomic targets as determined by primer mix.
Pulse-field gel electrophoresis (PFGE)	• Highly discriminative method with excellent phylogenetic resolution and epidemiological concordance. • Standardized systems available to aid interpretation and cross-comparison with other isolates, e.g. PulseNet. • Highly reproducible.	• Labor-intensive, time-consuming and technically demanding. • Results available in days. Resolution limited when DNA fragments of similar size. • Although standardization systems exist, an element of subjectivity remains.
Multi-locus sequence typing (MLST)	• Generated data highly portable compared with those produced by PFGE. • Internationally-standardized system that is highly reproducible	• Phylogenetic resolution is limited. • Expensive.
Whole genome sequencing (WGS)	• Provides the highest level of phylogenetic resolution and discriminatory power. • MLST can be determined from WGS data to allow for cross-comparison. • Lends itself to being automated.	• Requires significant resources, particularly bioinformatics expertise following initial sequencing.

Role of WGS in epidemiology based studies

With the recent advancements in the sequencing technologies, it is now possible to sequence a large number of strains at minimal cost and time and this has revolutionized the way epidemiological studies are performed (Kao, Haydon, Lycett, & Murcia, 2014). Whole genome sequencing can help in tracking the evolution and transmission patterns of bacterial pathogens (Koboldt, Steinberg, Larson, Wilson, & Mardis, 2013). A higher discrimination among the pathogens within the same species can throw more light on the spread of outbreaks involving MDR, nosocomial and pandemic infections (Fournier, Dubourg, & Raoult, 2014). Strain typing / bacterial typing can also help in understanding the dynamics behind the population structure of that species, but can sometimes be skewed because of high level of similarity among the house keeping genes. In contrast, Whole genome sequencing (WGS) can discriminate based on the whole genomic content, thus proving to be more useful in the epidemiological investigations of outbreaks (Parkhill, 2013). This was seen even in case of clonal organisms like *Mycobacterium tuberculosis* where other genotyping methods have failed (Harris et al., 2010). Pathogenic potential of bacteria can depend on the virulence it harbors and identification and characterization of these genes can be achieved with WGS. This can be done by either comparing the genomes of several strains or by performing BLAST search of the genomes of interest against the existing virulence factor databases (Karlin, 2001). WGS can also help in culturing the uncultivable pathogens by identifying the incomplete metabolic pathways and essential nutrients that are usually lacking in the standard culture media (Papin, Price, Wiback, Fell, & Palsson, 2003; Tyson & Banfield, 2005).

Next Generation Sequencing (NGS) platforms

Next generation sequencing was introduced to the world in 2001 by 454 sequencing(Ronaghi, 2001). This was followed by Illumina in 2008 and several others in the recent past(Bentley et al., 2008). A huge paradigm shift in whole genome sequencing was observed from then. Also with the continuous advancements in the sequencing technologies, most of the platforms are able to provide higher throughput data at minimal cost and time. The library preparation tasks which were much daunting initially are now being replaced with much faster and streamlined approaches with lesser manual interventions(Van Dijk, Auger, Jaszczyszyn, & Thermes, 2014). Currently many number of NGS platforms are available which differ in terms of output and fragment length and sequencing principle (Table 3)(Deurenberg et al., 2017).

Table 3: Comparison of different sequencing platforms(Deurenberg et al., 2017)

Company	Sequencing principle	Equipment	Output/run (Gb)	Maximum read length (bp)	Reads (x106)	Running time
Illumina	sequencing by synthesis	MiniSeq	0.6–7.5	2×150	25	4–24 h
Illumina		Miseq	0.3–15	2×300	25	5–55 h
Illumina		NextSeq	20–120	2×150	130/400	12–30 h
Illumina		HiSeq 3000	125–700	2×150	2500	<1–3.5 days
ThermoFisher	semi-conductor sequencing	Ion PGM™	0.03–2	200–400	0.4–5.5	2–7 h
ThermoFisher		Ion 5S™	0.6–15	200–400	3–80	2.5–4 h
ThermoFisher		Ion 5S™ XL	0.6–15	200–400	3–80	<24 h
Oxford Nanopore	ionic current sensing	MinION	21–42	230,000–300,000	2.2–4.4	1min–48 h
Pacific Biosciences[a]	SMRT (Single Molecule real time) technology	Sequel	0.75–1.25	>20,000	370,000	30 min–6 h
Pacific Biosciences[a]		RSII	0.5–1	>20,000	55,000	30 min–4 h

[a]The Pacific Biosciences data are per smart cell; both the Sequel and the RSII can run 1–16 smart cells in one run.

Rationale and objectives of the study

ExPEC infections have become a cause of concern worldwide due to the evolution of certain pandemic lineages which are difficult to treat. Emergence and rapid spread of MDR ExPEC strains are leading to a significant rise in the incidence rate of these infections all over the world.

In countries like India, infections due to *E. coli* are most commonly found due to the greater population density, sanitary and socio-economic reasons. These infections which were once easily treatable have now become difficult to treat due to the emergence of multidrug resistant lineages of these bacteria. Abuse of antibiotics in the medical as well as the food sectors is considered to be responsible for the evolution of MDR in these strains. This is all resulting in the increased rates of hospitalization, morbidity, and mortality; and increased associated healthcare and societal costs. In order to understand the mechanisms involved in the emergence of MDR ExPEC lineages, their genome architecture and the adaptation mechanism strategies have to be studied in detail.

Another major source of dissemination of MDR *E.coli* within environment and community could be through poultry. ExPEC, known for causing broad range of diseases in human also causes colibacillosis in chickens. Moreover, heavy doses of antibiotics used by the poultry industries while rearing these birds, are leading to the antibiotic resistant bacteria in their fecal flora. These MDR *E. coli* from poultry are often found to be zoonotic. Also, ExPEC isolates from human and animal sources share virulence genes and clonal backgrounds and the human health risk posed by such bacteria from poultry is still largely undefined. Therefore, the question of whether fecal isolates of poultry serving as a reservoir for extra intestinal and intestinal infections in humans remains unanswered.

Whole genome sequencing can address the above mentioned problems by providing greater resolution by comparing genomes from different backgrounds. Though it is much easier to sequence, the downstream job of reconstructing the genome from the millions of reads

generated is challenging. The usual approach involves assembling these reads into much larger contiguous fragments called contigs. These contigs can be either sorted and arranged with the help of a reference or linked into scaffolds based on the paired end information. A series of N's are placed at each of these links within the scaffolds which represents the missing information that needs to be filled. While the former approach might lose out some of the strain specific information, the latter fails in handling the repetitive regions. This problem has been tried to address in-house by designing a pipeline which follows a hybrid approach where the contigs are sorted according to a reference first followed by manually validating their sort-order based on the paired end information. This needed lot of manual intervention and was time consuming.

In order to address the above issues, the following objectives were framed.

1. To develop Contig-Layout-Authenticator (CLA), a combinatorial program to order and scaffold bacterial contigs *denovo* and its comparative evaluation with available ordering and scaffolding tools
2. To analyze genomes of pandemic ExPEC Sequence types- ST131, ST38, ST405 and ST648, for an insight into the genetic level similarities/differences among these clones
3. To analyze genomes of *Escherichia coli* isolates from Indian retail poultry and different human *E. coli* pathotypes to evaluate their risk in zoonotic transmission of antimicrobial resistance

References

Ahmed, N., Dobrindt, U., Hacker, J., & Hasnain, S. E. (2008). Genomic fluidity and pathogenic bacteria: Applications in diagnostics, epidemiology and intervention. *Nature Reviews Microbiology*, *6*(5), 387–394. https://doi.org/10.1038/nrmicro1889

Beloin, C., Roux, A., & Ghigo, J. (2010). Current Topics in Microbiology and Immunology Escherichia coli biofilms First Contacts with the Surface Approaching the Surface : The role of Motility in E . coli Adhesion to Surfaces, 249–289.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. https://doi.org/10.1038/nature07517

Bettelheim, K. A., Ismail, N., Shinebaum, R., Shooter, R. A., Moorhouse, E., & Farrell, W. (1976). The distribution of serotypes of Escherichia coli in cow-pats and other animal material compared with serotypes of E. coli isolated from human sources. *Journal of Hygiene*, *76*(3), 403–406. https://doi.org/10.1017/S0022172400055327

Blount, Z. D. (2015). The unexhausted potential of *E. coli*. *eLife*, *4*, 1–12. https://doi.org/10.7554/eLife.05826

Clements, A., Young, J. C., Constantinou, N., & Frankel, G. (2012). Infection strategies of enteric pathogenic Escherichia coli © 2012 Landes Bioscience . Do not distribute . © 2012 Landes Bioscience ., (April), 71–87.

Clermont, O., Bonacorsi, S., Bingen, E., & Bonacorsi, P. (2000). Rapid and Simple Determination of the Escherichia coli Phylogenetic Group Rapid and Simple Determination of the Escherichia coli Phylogenetic Group. *Applied and Environmental Microbiology*, *66*(10), 4555–4558. https://doi.org/10.1128/AEM.66.10.4555-4558.2000.Updated

Clermont, O., Gordon, D., & Denamur, E. (2015). Guide to the various phylogenetic classification schemes for Escherichia coli and the correspondence among schemes. *Microbiology (Reading, England)*, *161*(2015), 980–988. https://doi.org/10.1099/mic.0.000063

Conway, T., & Cohen, P. S. (2015). Commensal and Pathogenic Escherichia coli Metabolism in the Gut. *Microbiology Spectrum*, (3), 1–15. https://doi.org/10.1128/microbiolspec.MBP-0006-2014.Correspondence

Dale, A. P., & Woodford, N. (2015). Extra-intestinal pathogenic Escherichia coli (ExPEC): Disease , carriage and clones. *Journal of Infection*, (September), 1–12. https://doi.org/10.1016/j.jinf.2015.09.009

Deurenberg, R. H., Bathoorn, E., Chlebowicz, M. A., Couto, N., Ferdous, M., García-cobos, S., … Rossen, J. W. A. (2017). Application of next generation sequencing in clinical microbiology and infection prevention, *243*, 16–24.

Doumith, M., Day, M. J., Hope, R., Wain, J., & Woodford, N. (2012). Improved multiplex PCR strategy for rapid assignment of the four major Escherichia coli phylogenetic groups. *Journal of Clinical Microbiology*, *50*(9), 3108–3110. https://doi.org/10.1128/JCM.01468-12

Flores-Meireles, A., Walker, J., Caparon, M., & Hultgren, S. (2015). Urinary tract infections: epidemiology, mechanisms of infection and treatment options. *Nature Reviews Microbiology*, *13*(5), 269–284. https://doi.org/10.1038/nrmicro3432.Urinary

Fournier, P.-E., Dubourg, G., & Raoult, D. (2014). Clinical detection and characterization of bacterial pathogens in the genomics era. *Genome Medicine*, *6*(11), 114. https://doi.org/10.1186/s13073-014-0114-2

Fratamico, P. M., DebRoy, C., Liu, Y., Needleman, D. S., Baranzoni, G. M., & Feng, P.

(2016). Advances in molecular serotyping and subtyping of Escherichia coli. *Frontiers in Microbiology*, *7*(MAY), 1–8. https://doi.org/10.3389/fmicb.2016.00644

Gao, Q., Wang, X., Xu, H., Xu, Y., Ling, J., Zhang, D., … Liu, X. (2012). Roles of iron acquisition systems in virulence of extraintestinal pathogenic Escherichia coli: salmochelin and aerobactin contribute more to virulence than heme in a chicken infection model. *BMC Microbiology*, *12*(1), 143. https://doi.org/10.1186/1471-2180-12-143

Gomes, T. A. T., Elias, W. P., Scaletsky, I. C. A., Guth, B. E. C., Rodrigues, J. F., Piazza, R. M. F., … Martinez, M. B. (2016). Diarrheagenic Escherichia coli. *Brazilian Journal of Microbiology*, *47*, 3–30. https://doi.org/10.1016/j.bjm.2016.10.015

Harris, S. R., Feil, E. J., Holden, M. T. G., Quail, M. A., Nickerson, E. K., Chantratita, N., … Bentley, S. D. (2010). Evolution of MRSA During Hospital Transmission and Intercontinental Spread. *Science*, *327*(5964), 469–474. https://doi.org/10.1126/science.1182395

Hussain, A., Ewers, C., Nandanwar, N., Guenther, S., Jadhav, S., Wieler, L. H., & 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-??- lactamase-producing lineage. *Antimicrobial Agents and Chemotherapy*, *56*(12), 6358–6365. https://doi.org/10.1128/AAC.01099-12

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. *Antimicrobial Agents and Chemotherapy*, *58*(12), 7240–7249. https://doi.org/10.1128/AAC.03320-14

Ieka, K. E. R. O., Iddingsb, C. W. G., Oetkottc, C. D., Ohnsona, T. J. J., & Olana, L. K. N. (2005). Original article Characterizing the APEC pathotype **, *36*, 241–256. https://doi.org/10.1051/vetres

Ishii, S., & Sadowsky, M. J. (2008). Escherichia coli in the Environment: Implications for Water Quality and Human Health. *Microbes and Environments / JSME*, *23*(2), 101–108. https://doi.org/10.1264/jsme2.23.101

Kao, R. R., Haydon, D. T., Lycett, S. J., & Murcia, P. R. (2014). Supersize me: How whole-genome sequencing and big data are transforming epidemiology. *Trends in Microbiology*, *22*(5), 282–291. https://doi.org/10.1016/j.tim.2014.02.011

Kaper, J. B., Nataro, J. P., & Mobley, H. L. T. (2004). Pathogenic Escherichia coli. *Nature Reviews Microbiology*, *2*(2), 123–140. https://doi.org/10.1038/nrmicro818

Karlin, S. (2001). Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends in Microbiology*. https://doi.org/10.1016/S0966-842X(01)02079-0

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., & Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, *155*(1), 27–38. https://doi.org/10.1016/j.cell.2013.09.006

Manges, A. R., & Johnson, J. R. (2012). Food-borne origins of escherichia coli causing extraintestinal infections. *Clinical Infectious Diseases*, *55*(5), 712–719. https://doi.org/10.1093/cid/cis502

Nataro, J., & Kaper, J. (1998). Diarrheagenic Escherichia coli. *Clinical Microbiology Review*, *11*(1), 142.

Nicolas-Chanoine, M. H., Blanco, J., Leflon-Guibout, V., Demarty, R., Alonso, M. P.,

Caniça, M. M., … Johnson, J. R. (2008). Intercontinental emergence of Escherichia coli clone O25:H4-ST131 producing CTX-M-15. *Journal of Antimicrobial Chemotherapy*, *61*(2), 273–281. https://doi.org/10.1093/jac/dkm464

Ochman, H., & Selander, R. K. (1984). Standard reference strains of Escherichia coli from natural populations. *Journal of Bacteriology*, *157*(2), 690–693.

Papin, J. A., Price, N. D., Wiback, S. J., Fell, D. A., & Palsson, B. O. (2003). Metabolic pathways in the post-genome era. *Trends in Biochemical Sciences*. https://doi.org/10.1016/S0968-0004(03)00064-1

Parkhill, J. (2013). What has high-throughput sequencing ever done for us? *Nature Reviews. Microbiology*, *11*(10), 664–5. https://doi.org/10.1038/nrmicro3112

Petty, N. K., Ben Zakour, N. L., Stanton-Cook, M., Skippington, E., Totsika, M., Forde, B. M., … Beatson, S. a. (2014). Global dissemination of a multidrug resistant Escherichia coli clone. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(15), 5694–9. https://doi.org/10.1073/pnas.1322678111

Pitout, J. D. D. (2012). Extraintestinal pathogenic Escherichia coli: A combination of virulence with antibiotic resistance. *Frontiers in Microbiology*, *3*(JAN), 1–7. https://doi.org/10.3389/fmicb.2012.00009

Poolman, J. T., & Wacker, M. (2016). Extraintestinal Pathogenic Escherichia coli , a Common Human Pathogen : Challenges for Vaccine Development and Progress in the Field, *213*. https://doi.org/10.1093/infdis/jiv429

Ronaghi, M. (2001). Pyrosequencing sheds light on DNA sequencing. *Genome Research*. https://doi.org/10.1101/gr.11.1.3

Russo, T. A., & Johnson, J. R. (2000). Proposal for a New Inclusive Designation for

Extraintestinal Pathogenic Isolates of Escherichia coli: ExPEC. *The Journal of Infectious Diseases*, *181*, 1753–4. https://doi.org/10.1086/315418

Schembri, M. A., Zakour, N. L. Ben, Phan, M.-D., Forde, B. M., Stanton-Cook, M., & Beatson, S. A. (2015). Molecular Characterization of the Multidrug Resistant Escherichia coli ST131 Clone. *Pathogens (Basel, Switzerland)*, *4*(3), 422–30. https://doi.org/10.3390/pathogens4030422

Shpigel, N. Y., Elazar, S., & Rosenshine, I. (n.d.). Mammary pathogenic Escherichia coli. https://doi.org/10.1016/j.mib.2008.01.004

Smith, J. L., Fratamico, P. M., & Gunther, N. W. (2007). Extraintestinal Pathogenic *Escherichia coli*. *Foodborne Pathogens and Disease*, *4*(2), 134–163. https://doi.org/10.1089/fpd.2007.0087

Starc, M., Elers, K., Gubina, M., & Darja, Z. (2009). Virulence Potential of Escherichia coli Isolates from Skin and Soft Tissue Infections ˇ iva Petkovs, *47*(6), 1811–1817. https://doi.org/10.1128/JCM.01421-08

Sullivan, C. B., Diggle, M. A., & Clarke, S. C. (2005). Multilocus Sequence Typing: Data Analysis in Clinical Microbiology and Public Health. *Molecular Biotechnology*, *29*(3), 245–254. https://doi.org/10.1385/MB:29:3:245

Sweeney, N. J., Klemm, P., McCormick, B. A., Moller-Nielsen, E., Utley, M., Schembri, M. A., … Cohen, P. S. (1996). The Escherichia coli K-12 gntP gene allows E. coli F-18 to occupy a distinct nutritional niche in the streptomycin-treated mouse large intestine. *Infection and Immunity*, *64*(9), 3497–3503.

Tenaillon, O., Skurnik, D., Picard, B., & Denamur, E. (2010). The population genetics of commensal Escherichia coli. *Nat.Rev.Microbiol.*, *8*(1740–1534 (Electronic)), 207–217. https://doi.org/10.1038/nrmicro2298

Totsika, M., Moriel, D. G., Idris, A., Rogers, B. A., Wurpel, D. J., Phan, M. D., … Schembri, M. A. (2012). Uropathogenic Escherichia coli mediated urinary tract infection. *Curr Drug Targets*, *13*(11), 1386–1399. https://doi.org/10.2174/138945012803530206

Tyson, G. W., & Banfield, J. F. (2005). Cultivating the uncultivated: A community genomics perspective. *Trends in Microbiology*. https://doi.org/10.1016/j.tim.2005.07.003

Van der Bij, A. K., Peirano, G., Pitondo-Silva, A., & Pitout, J. D. D. (2012). The presence of genes encoding for different virulence factors in clonally related Escherichia coli that produce CTX-Ms. *Diagnostic Microbiology and Infectious Disease*, *72*(4), 297–302. https://doi.org/10.1016/j.diagmicrobio.2011.12.011

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*. https://doi.org/10.1016/j.tig.2014.07.001

Vimont, S., Boyd, A., Bleibtreu, A., Bens, M., Goujon, J. M., Garry, L., … Vandewalle, A. (2012). The CTX-M-15-Producing Escherichia coli Clone O25b: H4-ST131 Has High Intestine Colonization and Urinary Tract Infection Abilities. *PLoS ONE*, *7*(9). https://doi.org/10.1371/journal.pone.0046547

Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L. H., … Achtman, M. (2006). Sex and virulence in Escherichia coli: An evolutionary perspective. *Molecular Microbiology*, *60*(5), 1136–1151. https://doi.org/10.1111/j.1365-2958.2006.05172.x

Ahmed, N., Dobrindt, U., Hacker, J., & Hasnain, S. E. (2008). Genomic fluidity and pathogenic bacteria: Applications in diagnostics, epidemiology and intervention. *Nature Reviews Microbiology*, *6*(5), 387–394. https://doi.org/10.1038/nrmicro1889

Beloin, C., Roux, A., & Ghigo, J. (2010). Current Topics in Microbiology and Immunology Escherichia coli biofilms First Contacts with the Surface Approaching the Surface : The

role of Motility in E . coli Adhesion to Surfaces, 249–289.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. https://doi.org/10.1038/nature07517

Bettelheim, K. A., Ismail, N., Shinebaum, R., Shooter, R. A., Moorhouse, E., & Farrell, W. (1976). The distribution of serotypes of Escherichia coli in cow-pats and other animal material compared with serotypes of E. coli isolated from human sources. *Journal of Hygiene*, *76*(3), 403–406. https://doi.org/10.1017/S0022172400055327

Blount, Z. D. (2015). The unexhausted potential of *E. coli*. *eLife*, *4*, 1–12. https://doi.org/10.7554/eLife.05826

Clements, A., Young, J. C., Constantinou, N., & Frankel, G. (2012). Infection strategies of enteric pathogenic Escherichia coli © 2012 Landes Bioscience . Do not distribute . © 2012 Landes Bioscience ., (April), 71–87.

Clermont, O., Bonacorsi, S., Bingen, E., & Bonacorsi, P. (2000). Rapid and Simple Determination of the Escherichia coli Phylogenetic Group Rapid and Simple Determination of the Escherichia coli Phylogenetic Group. *Applied and Environmental Microbiology*, *66*(10), 4555–4558. https://doi.org/10.1128/AEM.66.10.4555-4558.2000.Updated

Clermont, O., Gordon, D., & Denamur, E. (2015). Guide to the various phylogenetic classification schemes for Escherichia coli and the correspondence among schemes. *Microbiology (Reading, England)*, *161*(2015), 980–988. https://doi.org/10.1099/mic.0.000063

Conway, T., & Cohen, P. S. (2015). Commensal and Pathogenic Escherichia coli Metabolism in the Gut. *Microbiology Spectrum*, (3), 1–15.

https://doi.org/10.1128/microbiolspec.MBP-0006-2014.Correspondence

Dale, A. P., & Woodford, N. (2015). Extra-intestinal pathogenic Escherichia coli (ExPEC): Disease , carriage and clones. *Journal of Infection*, (September), 1–12. https://doi.org/10.1016/j.jinf.2015.09.009

Deurenberg, R. H., Bathoorn, E., Chlebowicz, M. A., Couto, N., Ferdous, M., García-cobos, S., … Rossen, J. W. A. (2017). Application of next generation sequencing in clinical microbiology and infection prevention, *243*, 16–24.

Doumith, M., Day, M. J., Hope, R., Wain, J., & Woodford, N. (2012). Improved multiplex PCR strategy for rapid assignment of the four major Escherichia coli phylogenetic groups. *Journal of Clinical Microbiology*, *50*(9), 3108–3110. https://doi.org/10.1128/JCM.01468-12

Flores-Meireles, A., Walker, J., Caparon, M., & Hultgren, S. (2015). Urinary tract infections: epidemiology, mechanisms of infection and treatment options. *Nature Reviews Microbiology*, *13*(5), 269–284. https://doi.org/10.1038/nrmicro3432.Urinary

Fournier, P.-E., Dubourg, G., & Raoult, D. (2014). Clinical detection and characterization of bacterial pathogens in the genomics era. *Genome Medicine*, *6*(11), 114. https://doi.org/10.1186/s13073-014-0114-2

Fratamico, P. M., DebRoy, C., Liu, Y., Needleman, D. S., Baranzoni, G. M., & Feng, P. (2016). Advances in molecular serotyping and subtyping of Escherichia coli. *Frontiers in Microbiology*, *7*(MAY), 1–8. https://doi.org/10.3389/fmicb.2016.00644

Gao, Q., Wang, X., Xu, H., Xu, Y., Ling, J., Zhang, D., … Liu, X. (2012). Roles of iron acquisition systems in virulence of extraintestinal pathogenic Escherichia coli: salmochelin and aerobactin contribute more to virulence than heme in a chicken infection model. *BMC Microbiology*, *12*(1), 143. https://doi.org/10.1186/1471-2180-12-

Gomes, T. A. T., Elias, W. P., Scaletsky, I. C. A., Guth, B. E. C., Rodrigues, J. F., Piazza, R. M. F., … Martinez, M. B. (2016). Diarrheagenic Escherichia coli. *Brazilian Journal of Microbiology*, *47*, 3–30. https://doi.org/10.1016/j.bjm.2016.10.015

Harris, S. R., Feil, E. J., Holden, M. T. G., Quail, M. A., Nickerson, E. K., Chantratita, N., … Bentley, S. D. (2010). Evolution of MRSA During Hospital Transmission and Intercontinental Spread. *Science*, *327*(5964), 469–474. https://doi.org/10.1126/science.1182395

Hussain, A., Ewers, C., Nandanwar, N., Guenther, S., Jadhav, S., Wieler, L. H., & 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-??- lactamase-producing lineage. *Antimicrobial Agents and Chemotherapy*, *56*(12), 6358–6365. https://doi.org/10.1128/AAC.01099-12

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. *Antimicrobial Agents and Chemotherapy*, *58*(12), 7240–7249. https://doi.org/10.1128/AAC.03320-14

Ieka, K. E. R. O., Iddingsb, C. W. G., Oetkottc, C. D., Ohnsona, T. J. J., & Olana, L. K. N. (2005). Original article Characterizing the APEC pathotype **, *36*, 241–256. https://doi.org/10.1051/vetres

Ishii, S., & Sadowsky, M. J. (2008). Escherichia coli in the Environment: Implications for Water Quality and Human Health. *Microbes and Environments / JSME*, *23*(2), 101–108. https://doi.org/10.1264/jsme2.23.101

Kao, R. R., Haydon, D. T., Lycett, S. J., & Murcia, P. R. (2014). Supersize me: How whole-genome sequencing and big data are transforming epidemiology. *Trends in Microbiology*, *22*(5), 282–291. https://doi.org/10.1016/j.tim.2014.02.011

Kaper, J. B., Nataro, J. P., & Mobley, H. L. T. (2004). Pathogenic Escherichia coli. *Nature Reviews Microbiology*, *2*(2), 123–140. https://doi.org/10.1038/nrmicro818

Karlin, S. (2001). Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends in Microbiology*. https://doi.org/10.1016/S0966-842X(01)02079-0

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., & Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, *155*(1), 27–38. https://doi.org/10.1016/j.cell.2013.09.006

Manges, A. R., & Johnson, J. R. (2012). Food-borne origins of escherichia coli causing extraintestinal infections. *Clinical Infectious Diseases*, *55*(5), 712–719. https://doi.org/10.1093/cid/cis502

Nataro, J., & Kaper, J. (1998). Diarrheagenic Escherichia coli. *Clinical Microbiology Review*, *11*(1), 142.

Nicolas-Chanoine, M. H., Blanco, J., Leflon-Guibout, V., Demarty, R., Alonso, M. P., Caniça, M. M., … Johnson, J. R. (2008). Intercontinental emergence of Escherichia coli clone O25:H4-ST131 producing CTX-M-15. *Journal of Antimicrobial Chemotherapy*, *61*(2), 273–281. https://doi.org/10.1093/jac/dkm464

Ochman, H., & Selander, R. K. (1984). Standard reference strains of Escherichia coli from natural populations. *Journal of Bacteriology*, *157*(2), 690–693.

Papin, J. A., Price, N. D., Wiback, S. J., Fell, D. A., & Palsson, B. O. (2003). Metabolic

pathways in the post-genome era. *Trends in Biochemical Sciences*. https://doi.org/10.1016/S0968-0004(03)00064-1

Parkhill, J. (2013). What has high-throughput sequencing ever done for us? *Nature Reviews. Microbiology*, *11*(10), 664–5. https://doi.org/10.1038/nrmicro3112

Petty, N. K., Ben Zakour, N. L., Stanton-Cook, M., Skippington, E., Totsika, M., Forde, B. M., … Beatson, S. a. (2014). Global dissemination of a multidrug resistant Escherichia coli clone. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(15), 5694–9. https://doi.org/10.1073/pnas.1322678111

Pitout, J. D. D. (2012). Extraintestinal pathogenic Escherichia coli: A combination of virulence with antibiotic resistance. *Frontiers in Microbiology*, *3*(JAN), 1–7. https://doi.org/10.3389/fmicb.2012.00009

Poolman, J. T., & Wacker, M. (2016). Extraintestinal Pathogenic Escherichia coli , a Common Human Pathogen : Challenges for Vaccine Development and Progress in the Field, *213*. https://doi.org/10.1093/infdis/jiv429

Ronaghi, M. (2001). Pyrosequencing sheds light on DNA sequencing. *Genome Research*. https://doi.org/10.1101/gr.11.1.3

Russo, T. A., & Johnson, J. R. (2000). Proposal for a New Inclusive Designation for Extraintestinal Pathogenic Isolates of Escherichia coli: ExPEC. *The Journal of Infectious Diseases*, *181*, 1753–4. https://doi.org/10.1086/315418

Schembri, M. A., Zakour, N. L. Ben, Phan, M.-D., Forde, B. M., Stanton-Cook, M., & Beatson, S. A. (2015). Molecular Characterization of the Multidrug Resistant Escherichia coli ST131 Clone. *Pathogens (Basel, Switzerland)*, *4*(3), 422–30. https://doi.org/10.3390/pathogens4030422

Shpigel, N. Y., Elazar, S., & Rosenshine, I. (n.d.). Mammary pathogenic Escherichia coli. https://doi.org/10.1016/j.mib.2008.01.004

Smith, J. L., Fratamico, P. M., & Gunther, N. W. (2007). Extraintestinal Pathogenic *Escherichia coli*. *Foodborne Pathogens and Disease*, *4*(2), 134–163. https://doi.org/10.1089/fpd.2007.0087

Starc, M., Elers, K., Gubina, M., & Darja, Z. (2009). Virulence Potential of Escherichia coli Isolates from Skin and Soft Tissue Infections ˇ iva Petkovs, *47*(6), 1811–1817. https://doi.org/10.1128/JCM.01421-08

Sullivan, C. B., Diggle, M. A., & Clarke, S. C. (2005). Multilocus Sequence Typing: Data Analysis in Clinical Microbiology and Public Health. *Molecular Biotechnology*, *29*(3), 245–254. https://doi.org/10.1385/MB:29:3:245

Sweeney, N. J., Klemm, P., McCormick, B. A., Moller-Nielsen, E., Utley, M., Schembri, M. A., … Cohen, P. S. (1996). The Escherichia coli K-12 gntP gene allows E. coli F-18 to occupy a distinct nutritional niche in the streptomycin-treated mouse large intestine. *Infection and Immunity*, *64*(9), 3497–3503.

Tenaillon, O., Skurnik, D., Picard, B., & Denamur, E. (2010). The population genetics of commensal Escherichia coli. *Nat.Rev.Microbiol.*, *8*(1740–1534 (Electronic)), 207–217. https://doi.org/10.1038/nrmicro2298

Totsika, M., Moriel, D. G., Idris, A., Rogers, B. A., Wurpel, D. J., Phan, M. D., … Schembri, M. A. (2012). Uropathogenic Escherichia coli mediated urinary tract infection. *Curr Drug Targets*, *13*(11), 1386–1399. https://doi.org/10.2174/138945012803530206

Tyson, G. W., & Banfield, J. F. (2005). Cultivating the uncultivated: A community genomics perspective. *Trends in Microbiology*. https://doi.org/10.1016/j.tim.2005.07.003

Van der Bij, A. K., Peirano, G., Pitondo-Silva, A., & Pitout, J. D. D. (2012). The presence of genes encoding for different virulence factors in clonally related Escherichia coli that produce CTX-Ms. *Diagnostic Microbiology and Infectious Disease*, *72*(4), 297–302. https://doi.org/10.1016/j.diagmicrobio.2011.12.011

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*. https://doi.org/10.1016/j.tig.2014.07.001

Vimont, S., Boyd, A., Bleibtreu, A., Bens, M., Goujon, J. M., Garry, L., … Vandewalle, A. (2012). The CTX-M-15-Producing Escherichia coli Clone O25b: H4-ST131 Has High Intestine Colonization and Urinary Tract Infection Abilities. *PLoS ONE*, *7*(9). https://doi.org/10.1371/journal.pone.0046547

Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L. H., … Achtman, M. (2006). Sex and virulence in Escherichia coli: An evolutionary perspective. *Molecular Microbiology*, *60*(5), 1136–1151. https://doi.org/10.1111/j.1365-2958.2006.05172.x

CHAPTER 2: (Objective 1)

Contig-Layout-Authenticator (CLA): Combinatorial approach for ordering and scaffolding of contigs

This part is published as:

Contig-Layout-Authenticator (CLA): A Combinatorial Approach to Ordering and Scaffolding of Bacterial Contigs for Comparative Genomics and Molecular Epidemiology

Sabiha Shaik, Narender Kumar, Aditya K. Lankapalli, Sumeet K. Tiwari, Ramani Baddam, Niyaz Ahmed

PLoS One. 2016 Jun 1;11(6):e0155459. doi: 10.1371/journal.pone.0155459

Introduction

Emergence of newer platforms for whole genome sequencing has driven a revolution in the field of comparative genomics and epidemiological tracking of organisms (Ahmed 2009; Loman et al. 2012; Schuster 2008). These platforms are essentially an adaptation of shotgun sequencing producing higher throughput and being cost-effective when compared to traditional Sanger sequencing. Though there are a variety of platforms available such as Illumina, PacBio, Roche etc., they broadly involve fragmentation of the genomic DNA into small pieces or fragments followed by reading of bases in these fragments to generate read data (Quail et al. 2012). In addition, some platforms like Illumina also provide paired-end sequencing option which essentially means that the generated fragments are read from both the ends providing useful information for downstream processing. The challenge is then to reconstruct the genome out of these millions of reads generated from the sequencers. This has fostered development of a myriad of assembly tools with distinct algorithms to consolidate and harness the information spread across millions of sequence reads. Generally, these sequence reads are assembled into larger contiguous fragments called contigs, utilizing de novo algorithm based tools such as Velvet (Zerbino & Birney 2008), SSAKE (Warren et al. 2007), etc. Though this reduces complexity, contigs formed thereafter still require a lot of downstream analysis such as sorting of these contigs to their real biological order and resolution of repetitive elements before a high quality draft or complete genome is constructed.

The assembled contigs could be utilized in various ways to reconstruct the genome. One way is to directly order the contigs according to a chosen reference to build a chromosomal assembly using tools such as ABACAS (Assefa et al. 2009), Ragout (Kolmogorov et al. 2014), Mauve Contig Mover (MCM) (Rissman et al. 2009), CONTIGuator (Galardini et al. 2011), etc. The reference based approaches are known to be very efficient in analyzing specific

mutations/variations among highly similar organisms since the possibilities of finding novel genomic insertions are scarce. But, reference based approaches have limitations in terms of masking important biological co-ordinates specific to the genome under study such as repeats, duplications and prophage elements (Edwards & Holt 2013). These approaches may thus limit the resolution of final genome sequence, especially in highly repetitive and recombining organisms such as *Helicobacter pylori* (*H. pylori*) (Doolittle & Zhaxybayeva 2009).

The other approach is to construct scaffolds consisting of contigs joined together based on paired-end read information (Pop et al. 2004). While a majority of sequence read assemblers such as Velvet, SOAPdenovo2 (Luo et al. 2012), SOPRA (Dayarian et al. 2010), SGA(Simpson & Durbin 2012) etc. come with an additional option of scaffolding; exclusive stand-alone scaffolding tools such as Bambus, Bambus2 (Koren et al. 2011), SSPACE (Boetzer et al. 2011) etc. are also available. However, these scaffolds must be ordered to construct the draft genome which could be difficult to achieve. Therefore, in such cases consisting of multiple scaffolds, ordering and orientation is limited within scaffolds. Moreover, the approach poses difficulties in handling repetitive contigs resulting in mis-joins/misassemblies within scaffolds (Hunt et al. 2014). Though recent developments like PacBio sequencing helps address these limitations by virtue of long reads they generate, these platforms would take time to become mainstream technologies like Illumina due to the sequencing cost involved.

Addressing the drawbacks of both scaffolding as well as reference based whole genome ordering tools mainly for contigs from Illumina paired-end sequencing; we have developed Contig-Layout-Authenticator (CLA) which not only scaffolds the contigs but also tries to attain proper ordering of these scaffolds at whole genome level. In other words, the contigs are initially ordered at the whole genome level similar to that of ordering tools like ABACAS and Ragout based on a reference but in later stages the order is further refined. Thus ordered

contigs are further scaffolded only when supported by the paired-end information. Therefore, multiple scaffolds indicate that there is still missing genome information between two consecutive scaffolds. Thus CLA carries out the functions of ordering and scaffolding tools in a combinatorial approach yielding better draft genome sequences. The pipeline of CLA is also efficient in identifying probable contamination contigs as well as misassemblies within scaffolds, thus assisting users in improving the quality of the draft genome. The comparative analysis of CLA with other widely used ordering and scaffolding tools revealed its enhanced performance. Therefore, we believe CLA would serve as an efficient and user friendly tool for sorting and scaffolding of contigs to achieve better draft genome.

Results

As CLA performs both ordering at the whole genome level followed by scaffolding, its efficiency was validated at three different levels. Firstly, it was compared with the widely used ordering tools such as ABACAS, MCM, CONTIGuator and Ragout. Secondly, the scaffolding ability of CLA was compared with Bambus2, SSPACE, SOPRA, SOAPdenovo2, SGA and reference based scaffolder MeDuSa (Bosi et al. 2015). Finally to test the hybrid approach efficiency, scaffolds from SSPACE, SOPRA, SOAPdenovo2 and SGA were further ordered with Ragout and the final results were compared against CLA. All the in-house scripts and the command lines that were used for validation are provided along with the tool.

CLA vs. Reference based sorting tools

All the tools under this category such as ABACAS (v1.3.1), MCM (Mauve-2.3.1), CONTIGuator (v2.7.3) and Ragout (v1.1) were run with default parameters and their outputs were compared with CLA using QUAST(v3.1) (Gurevich et al. 2013) tool to avoid

manual errors. Ragout and MeDuSa were run using multiple reference genomes (Tables S1 and 2). The output generated by all the above mentioned tools using simulated and real sequence datasets were evaluated for parameters such as misassemblies in the chromosome assembly, number of contigs (>1kbp) unassigned in the final order, size of largest contig unassigned, genome fraction obtained and number of large repetitive contigs (contig size greater than 500bp) correctly placed. QUAST was used to identify number of misassemblies and to calculate the genome fractions with the original genome as reference for simulated datasets. While in real dataset, due to the unavailability of the original genome, a closely related genome was used as reference for QUAST. Therefore, the statistics generated in case of real dataset could be biased towards the reference genome and may not exactly depict the efficiency of the tools as shown with simulated dataset. In-house perl scripts were used to calculate the number of large repetitive contigs present in the original genome, as well as number of them placed by each of these tools.

Misassemblies in the chromosome assembly

A chromosomal assembly or draft genome constructed from reference based ordering tools comprised of ordered contigs merged together after placing N's in-between. Tools like Ragout, CONTIGuator and ABACAS usually provide the final ordered contigs in the form of a single scaffold or chromosome assembly whereas MCM only provides ordered contigs in multi-fasta format. In CLA, contigs are merged into scaffolds (placing N's in-between) only if supported by sufficient paired-end information. An in-house script was used to generate a chromosome assembly for the outputs of MCM and CLA. In order to identify the errors such as misassemblies encountered by each of these tools and compare their performance we used QUAST. The misassemblies observed in the chromosomal assemblies were further categorized into inversions, relocations and translocations by QUAST. As inferred from Table 1, CLA out competed other tools by producing least number of

misassemblies in the simulated datasets of all the organisms except *H. pylori*. In organisms like *Bartonella quintana* (*B. quintana)*, *Caulobacter crescentus* (*C. crescentus)* CLA showed only two inversions without any relocations (Table 2). After CLA, Ragout and CONTIGuator performed fairly well followed by ABACAS. MCM was observed to be the only tool with highest number of misplacements in all organisms which might be due to placement of unassigned contigs at the end of the order.

Table 1: Comparative statistics of CLA and reference based alignment tools generated using simulated dataset

	Genome*	Tool	# Misassemblies in Chromosome assembly	Unassigned Contigs in final order >1kbp	Largest contig unassigned	Genome Fraction	True Repeat places filled**
1.	***B. quintana*** #contigs:**47** #Total repeat pos: **19** #Misassmblies in input contigs: **0**	CLA	2	0	850	99.68	15
		Ragout	5	3	3347	99.299	15
		ABACAS	12	1	1104	98.648	9
		MCM	26	0	0	98.767	9
		CONTIGuator	4	7	10343	97.401	0
2.	***C. jejuni*** #contigs: **33** #Total repeat pos: **12** #Misassmblies in input contigs: **0**	CLA	4	0	425	99.728	10
		Ragout	5	4	4566	98.807	10
		ABACAS	9	7	481308	56.49	3
		MCM	21	0	0	98.82	5
		CONTIGuator	8	6	6071	97.71	1
3.	***C. crescentus*** #contigs:**49** #Total repeat pos: **28** #Misassmblies in input contigs: **0**	CLA	2	0	793	99.901	24
		Ragout	8	0	371	99.923	28
		ABACAS	11	0	371	99.355	9
		MCM	18	0	0	99.355	9
		CONTIGuator	4	6	5440	98.94	0
4.	***H. influenzae*** #contigs: **43** #Total repeat pos: **19** #Misassmblies in input contigs: **1**	CLA	6	0	606	98.963	9
		Ragout	12	0	375	98.559	19
		ABACAS	15	4	345463	70.201	3
		MCM	22	0	0	98.187	5
		CONTIGuator	11	3	3503	97.719	1
5.	***H. pylori*** #contigs: **46** #Total repeat pos: **20** #Misassmblies in input contigs: **0**	CLA	7	0	960	99.672	17
		Ragout	4	1	5487	98.97	14
		ABACAS	17	6	170928	82.357	6
		MCM	23	0	0	99.133	10
		CONTIGuator	8	4	4051	98.135	0

6.	***R. etli*** #contigs: **30** #Total repeat pos: **21** #Misassmblies in input contigs: **0**	CLA	3	0	601	99.756	19
		Ragout	5	1	1052	99.406	13
		ABACAS	5	7	798023	71.125	4
		MCM	24	0	0	99.221	8
		CONTIGuator	10	8	4331	98.761	0
7.	***S.* Typhi** #contigs: **67** #Total repeat pos: **24** #Misassmblies in input contigs: **0**	CLA	8	0	520	99.141	11
		Ragout	55	2	3835	98.873	12
		ABACAS	13	0	474	98.682	7
		MCM	32	0	0	98.698	7
		CONTIGuator	13	5	6299	98.311	1
8.	***T. pallidum*** #contigs: **22** #Total repeat pos: **17** #Misassmblies in input contigs: **0**	CLA	5	0	980	99.124	9
		Ragout	7	1	1617	99.432	17
		ABACAS	10	0	0	98.768	7
		MCM	15	0	0	98.769	7
		CONTIGuator	6	3	3283	97.95	1

*All the genomes were simulated with a read length of 100bp and insert size of 400bp. ** Total number of positions filled by the repetitive contigs in the final output. For example if tools have placed contig A at 2 out of 4 places and contig B at 1 out of 6 places, then total number of repeat positions in the original genome would be 10 and true repeats placed by the tool would be 3.

Table 2: Misassembly details of CLA and reference based ordering tools in simulated dataset. Table listing out number of relocations, translocations and inversions amounted to the total number of misassemblies.

	Genome	Tool	#relocations	#translocations	#inversions
1.	***B. quintana*** #contigs:**47** #Total repeat pos: **19** #Misassmblies in input contigs: **0**	CLA	0	0	2
		Ragout	5	0	0
		ABACAS	12	0	0
		MCM	26	0	0
		CONTIGuator	4	0	0
2.	***C. jejuni*** #contigs: **33** #Total repeat pos: **12** #Misassmblies in input contigs: **0**	CLA	4	0	0
		Ragout	4	0	1
		ABACAS	9	0	0
		MCM	21	0	0
		CONTIGuator	8	0	0
3.	***C. crescentus*** #contigs:**49** #Total repeat pos: **28** #Misassmblies in input contigs: **0**	CLA	0	0	2
		Ragout	8	0	0
		ABACAS	11	0	0
		MCM	18	0	0
		CONTIGuator	4	0	0
4.	***H. influenza*** #contigs: **43** #Total repeat pos: **19** #Misassmblies in input contigs: **1**	CLA	6	0	0
		Ragout	12	0	0
		ABACAS	15	0	0
		MCM	18	0	4
		CONTIGuator	9	0	2

5.	***H. pylori*** #contigs: **46** #Total repeat pos: **20** #Misassmblies in input contigs: **0**	CLA	5	0	2
		Ragout	2	0	2
		ABACAS	15	0	2
		MCM	21	0	2
		CONTIGuator	7	0	1
6.	***R. etli*** #contigs: **30** #Total repeat pos: **21** #Misassmblies in input contigs: **0**	CLA	3	0	0
		Ragout	5	0	0
		ABACAS	5	0	0
		MCM	23	0	1
		CONTIGuator	10	0	0
7.	***S.* Typhi** #contigs: **67** #Total repeat pos: **24** #Misassmblies in input contigs: **0**	CLA	8	0	0
		Ragout	55	0	0
		ABACAS	13	0	0
		MCM	32	0	0
		CONTIGuator	13	0	0
8.	***T. pallidum*** #contigs: **22** #Total repeat pos: **17** #Misassmblies in input contigs: **0**	CLA	5	0	0
		Ragout	6	0	1
		ABACAS	10	0	0
		MCM	14	0	1
		CONTIGuator	6	0	0

Since the original genome was unavailable for the real datasets, we utilized a closely related genome as reference. Because this modification would definitely favor the reference based tools, the performance of CLA varied from genome to genome as revealed in the Table 3. CLA gave out least number of misassemblies in all except *Campylobacter jejuni (C. jejuni)* as inferred from Table 3. Moreover, the performance of CLA was comparable to others exhibiting the robustness of the combinatorial approach used by CLA.

Table 3: Comparative statistics of CLA and reference based alignment tools generated using real dataset

	Genome	**Tool**	**# Misasse-mblies in Chromosome assembly**	**Unassign-ed Contigs in final order >1kbp**	**Largest contig unassigned**	**Genome Fraction**	**True Repeat places filled****
1.	***C. jejuni*** #contigs: **72** Readlength: **100**	CLA	19	0	965	90.553	3
		Ragout	18	8	32459	91.261	11
		ABACAS	22	8	32459	90.584	4

	Insert: **300** #Repeats: **11** #Misassmblies in input contigs: **7**	MCM	39	0	0	90.76	4
		CONTIGuator	20	8	32459	90.051	0
2.	***E. coli*** #contigs:**252** Readlength:**151** Insert:**300** #Repeats: **51** #Misassmblies in input contigs: **98**	CLA	231	0	1000	92.245	15
		Ragout	242	45	36453	91.65	36
		ABACAS	203	49	36453	91.264	10
		MCM	262	0	0	92.054	10
		CONTIGuator	176	46	36453	91.508	0
3.	***H. influenzae*** #contigs:**50** Readlength:**100** Insert:**200** #Repeats: **19** #Misassmblies in input contigs: **25**	CLA	32	0	774	88.984	8
		Ragout	33	8	5468	89.293	15
		ABACAS	24	11	208372	66.488	4
		MCM	39	0	0	88.194	4
		CONTIGuator	33	10	6440	87.842	0
4.	***M. tuberculosis*** #contigs:**241** Readlength:**101** Insert:**300** #Repeats: **43** #Misassmblies in input contigs: **6**	CLA	23	0	978	98.491	24
		Ragout	32	9	5385	98.375	38
		ABACAS	37	5	6881	98.441	13
		MCM	113	0	0	98.559	14
		CONTIGuator	24	28	5385	97.11	0
5.	***S.* Typhi** #contigs:**119** Readlength:**100** Insert:**200** #Repeats: **37** #Misassmblies in input contigs: **2**	CLA	31	0	978	98.75	13
		Ragout	32	4	93926	99.067	30
		ABACAS	23	4	93926	98.047	11
		MCM	65	0	0	98.812	11
		CONTIGuator	17	10	93926	98.328	0

** Total number of positions filled by the repetitive contigs in the final output.

Contigs unassigned in the final chromosome assembly

Majority of the reference based ordering tools often exclude contigs (which fail to align properly with reference genome) from the final order. These unassigned contigs are either provided as separate files or put towards the end in their output. The scenario becomes more pertinent in case of highly recombining organisms as the reference genome architecture might differ significantly from the target genome. These unassigned contigs could contain valuable strain specific information encoding novel elements, virulence genes etc. Since CLA utilizes paired-end information in addition to reference genome, it is able to place maximum

number of contigs in the chromosomal assembly retaining even potentially novel strain specific elements. As observed from the comparative statistics mentioned in the Table 1, CLA showed consistent results by not having unassigned any contig with size greater than 1kbp in all the organisms. On the other hand, CONTIGuator left more number of contigs with size greater than 1kbp unassigned to final order. It was also observed that a contig as long as 10kbp in *B. quintana,* was not considered in the final order. Though ABACAS performed considerably well, in some cases such as *C. jejuni*, *Haemophilus influenzae* (*H. influenzae)*, *H. pylori* and *Rhizobium etli (R. etli)* it ended up excluding contigs of lengths 481kbp, 345kbp, 170kbp and 798kbp, respectively. MCM was observed to assign all the contigs in the final assembly as it places the unaligned contigs towards the end of chromosomal assembly.

A similar pattern was also observed in real dataset (Table 3). While CLA and MCM have considered all the contigs that were greater than 1kbp in the final order contigs as large as 32kbp, 36kbp and 93kbp remained unassigned in the final order by Ragout, ABACAS and CONTIGuator in *C. jejuni, Escherichia coli* (*E. coli)* and *Salmonella* Typhi (*S.* Typhi) respectively. The number of contigs >1kbp that were not considered by these tools were also observed to be very high in real dataset.

Genome Fraction

The genome fraction obtained after the chromosome assembly of all the strains in simulated and real sequence datasets was calculated using QUAST. In almost all cases in simulated datasets, CLA and Ragout were observed to generate higher genome fraction when compared to others. While majority of tools achieved a genome fraction greater than 97%, genome fraction from ABACAS was found to be only 56.49%, 70.201%, 82.351% and 71.125% in *C. jejuni*, *H. influenzae*, *H. pylori* and *R. etli,* respectively (Table 1). In real datasets,

the genome fraction obtained by all tools was observed to be above 85% except in *H. influenza*, where ABACAS was observed to generate only 66.488%.

Resolving large repetitive contigs

Large repetitive contigs were defined as the contigs of length >500 bases having more than one coordinate in the original genome. To validate the placement of large repeats contigs greater than 500 bases were considered for this study. An in-house script was used to calculate the total number of positions for all repetitive contigs based on the BLASTn (Altschul et al. 1990) alignment against reference genome. BLAST version 2.2.29+ was used and the best hit with >90% identity and >90% query coverage was considered as a valid hit to identify the position coordinates. Then, the output of CLA and all ordering tools were scanned using another in house perl script to evaluate the total number of positions placed by all repetitive contigs.

The ultimate aim of any genome sequencing project is to harness as much information as possible by formulation of better assemblies. Given this, resolution of repeats becomes an unavoidable parameter to obtain near perfect genomes. Since all the tools mentioned above are designed to aid researchers for achieving better assemblies these tools were also compared for their performance in resolution of repeats. Except CLA and Ragout, other tools under this category were designed only to hint at the probable repetitive contigs and therefore required further manual intervention. As both CLA and Ragout handled repeats to some extent, they performed much better than the rest tools in all the genomes. CLA and Ragout could more or less fill similar number of true repeating positions in majority of cases. While CLA could correctly fill 19 out of 21 positions (Table 1) in *R. etli*, Ragout could fill only 13. Similarly, Ragout performed better in *Treponema pallidum* (*T. pallidum)* by filling all 17 positions whereas only 9 of them could be placed by CLA. Both CLA and Ragout have shown similar performance in highly recombining organisms like *B. quintana and C. jejuni*, by

filling 15 out of 19 positions and 10 out of 12 positions respectively. MCM and CONTIGuator were seen to be placing very few repeats in correct order and copy number. CLA and Ragout performed superior than other tools in handling the repetitive contigs placing them in correct positions followed by ABACAS.

In real dataset, number of repetitive positions was calculated with respect to the closely related genome used as reference which also served as input to these tools due to the unavailability of the original genome. Ragout was observed to perform well in all cases followed by CLA as inferred from Table 3. Therefore, it can be surmised that CLA and Ragout is comparable for this parameter in the chromosomal assembly process.

CLA vs. Scaffolding tools

Scaffolding further reduces the number of sequence fragments by joining consecutive contigs and helps in the improvement of assembly. This could be achieved with the help of paired-end information or cues from assembly graph, but was also attempted with the help of multiple reference genomes in the case of MeDuSa.

CLA which utilizes a combination of reference alignment and paired-end information for scaffolding was compared against recently published MeDuSa (v3) and scaffolders such as SSPACE (v2.0), SOPRA (v1.4.6), SOPAdenovo2 (v2.04-r240), SGA and BAMBUS2 (amos-3.1.0). For the simulated dataset, QUAST was used to deduce the assembly statistics from output files of each tool keeping original genome as reference. In the real dataset, a closely related strain was used as reference genome. Parameters such as number of scaffolds formed, number of misassemblies within the scaffolds, number of misassembled scaffolds and size of genome fraction were used to compare the performance of CLA with others.

It was observed that CLA performed better under this category owing to its combinatorial approach. The comparative statistics for simulated and real datasets are mentioned in Tables

4, 5 and 6, respectively. The lesser number of scaffolds with minimum number of misassemblies clearly depict the efficiency and accuracy of CLA over other scaffolding tools. For example, in *R. etli* of the simulated dataset, CLA could combine all 35 contigs in to just two scaffolds whereas MeDuSa, Bambus2, SSPACE with extension, SSPACE without extension, SOPRA, SOAPdenovo2 and SGA generated 10, 23, 24, 28, 30, 24 and 30 scaffolds, respectively (Table 4). CLA's performance was consistent for all the genomes under comparison except in *H. pylori* where number of misassemblies slightly increased to 7. In *C. jejuni*, 33 contigs were combined into 6 scaffolds without any misassemblies by CLA. In *B. quintana*, *C. crescentus* and *S*. Typhi only 2 inversions were found within scaffolds from CLA (Table 5). Number of misassemblies from CLA was also observed to be relatively few when compared with MeDuSa even in real dataset (Table 6).

Table 4: Comparative statistics of CLA and Scaffolding tools generated using simulated dataset

	Genome*	**Tool**	**# Scaffolds**	**# Misassemb-lies within scaffolds**	**# Misassembled Scaffolds**	**Genome Fraction**
1	***B. quintana*** #contigs:**47** #Misassmblies in input contigs: **0**	CLA	3	2	1	99.68
		MeDuSa	16	11	7	98.751
		Bambus2	32	3	1	98.639
		SSPACE (no extension)	41	1	1	98.763
		SSPACE (extension)	38	1	1	98.822
		SOPRA	43	1	1	98.751
		SOAPdenovo2	35	4	3	98.704
		SGA	45	0	0	98.743
2	***C. jejuni*** #contigs:**33** #Misassmblies in input contigs: **0**	CLA	6	0	0	99.728
		MeDuSa	10	18	1	98.756
		Bambus2	22	4	2	98.736
		SSPACE (no extension)	28	0	0	98.781
		SSPACE (extension)	24	2	2	98.829
		SOPRA	32	0	0	98.737
		SOAPdenovo2	28	3	1	98.792
		SGA	32	0	0	98.737
3	***C. crescentus*** #contigs:**49** #Misassmblies in input contigs: 0	CLA	4	2	1	99.901
		MeDuSa	12	24	7	99.356
		Bambus2	31	13	3	99.332
		SSPACE (no extension)	44	0	0	99.347
		SSPACE (extension)	38	3	3	99.363
		SOPRA	46	0	0	99.34
		SOAPdenovo2	46	0	0	99.34

		SGA	46	0	0	99.34
4	***H. influenzae*** #contigs:**43** #Misassmblies in input contigs: **1**	CLA	5	2	1	98.963
		MeDuSa	6	16	2	98.18
		Bambus2	31	6	3	98.158
		SSPACE (no extension)	38	2	2	98.003
		SSPACE (extension)	35	2	2	98.033
		SOPRA	26	1	1	98.156
		SOAPdenovo2	40	1	1	98.144
		SGA	27	1	1	98.156
5	***H. pylori*** #contigs:**46** #Misassmblies in input contigs: **0**	CLA	2	7	2	99.672
		MeDuSa	7	15	3	99.137
		Bambus2	25	10	6	99.103
		SSPACE (no extension)	39	0	0	99.101
		SSPACE (extension)	36	0	0	99.151
		SOPRA	42	0	0	99.092
		SOAPdenovo2	37	1	1	99.105
		SGA	44	0	0	99.078
6	***R. etli*** #contigs:**30** #Misassmblies in input contigs: **0**	CLA	2	2	1	99.811
		MeDuSa	10	13	1	99.204
		Bambus2	23	1	1	99.071
		SSPACE (no extension)	28	0	0	99.211
		SSPACE (extension)	24	2	2	99.225
		SOPRA	30	0	0	99.204
		SOAPdenovo2	24	4	1	99.215
		SGA	30	0	0	99.204
7	***S.* Typhi** #contigs:**67** #Misassmblies in input contigs: **0**	CLA	7	2	1	99.149
		MeDuSa	3	30	1	98.695
		Bambus2	47	11	6	98.666
		SSPACE (no extension)	51	1	1	98.679
		SSPACE (extension)	53	0	0	98.717
		SOPRA	63	0	0	98.647
		SOAPdenovo2	56	4	3	98.657
		SGA	66	0	0	98.644
8	***T. pallidum*** #contigs:**22** #Misassmblies in input contigs: **0**	CLA	3	3	2	99.124
		MeDuSa	1	13	1	98.765
		Bambus2	16	2	1	98.571
		SSPACE (no extension)	13	1	1	98.755
		SSPACE (extension)	14	2	1	98.865
		SOPRA	18	1	1	98.759
		SOAPdenovo2	14	2	2	98.747
		SGA	21	0	0	98.731

*All the genomes were simulated with a read length of 100bp and insert size of 400bp

Table 5: Misassembly details of CLA and Scaffolding tools in simulated dataset. Table listing out number of relocations, translocations and inversions which amounted to the total number of misassemblies

	Genome*	Tool	#relocations	#translocations	#inversions
1.	***B. quintana*** #contigs:**47** #Misassmblies in input contigs: **0**	CLA	0	0	2
		MeDuSa	10	0	1
		Bambus2	3	0	0
		SSPACE (no extension)	1	0	0
		SSPACE (extension)	1	0	0
		SOPRA	1	0	0
		SOAPdenovo2	4	0	0
		SGA	0	0	0
2.	***C. jejuni*** #contigs:**33** #Misassmblies in input contigs: **0**	CLA	0	0	0
		MeDuSa	18	0	0
		Bambus2	4	0	0
		SSPACE (no extension)	0	0	0
		SSPACE (extension)	2	0	0
		SOPRA	0	0	0
		SOAPdenovo2	3	0	0
		SGA	0	0	0
3.	***C. crescentus*** #contigs:**49** #Misassmblies in input contigs: **0**	CLA	0	0	2
		MeDuSa	22	0	2
		Bambus2	13	0	0
		SSPACE (no extension)	0	0	0
		SSPACE (extension)	3	0	0
		SOPRA	0	0	0
		SOAPdenovo2	0	0	0
		SGA	0	0	0
4.	***H. influenzae*** #contigs:**43** #Misassmblies in input contigs: **1**	CLA	2	0	0
		MeDuSa	16	0	0
		Bambus2	6	0	0
		SSPACE (no extension)	2	0	0
		SSPACE (extension)	2	0	0
		SOPRA	1	0	0
		SOAPdenovo2	1	0	0
		SGA	1	0	0
5.	***H. pylori*** #contigs:**45** #Misassmblies in input contigs: **0**	CLA	5	0	2
		MeDuSa	15	0	0
		Bambus2	9	0	1
		SSPACE (no extension)	0	0	0
		SSPACE (extension)	0	0	0
		SOPRA	0	0	0
		SOAPdenovo2	1	0	0
		SGA	0	0	0
6.	***R. etli*** #contigs:**46** #Misassmblies in input contigs: **0**	CLA	2	0	0
		MeDuSa	13	0	0
		Bambus2	1	0	0
		SSPACE (no extension)	0	0	0
		SSPACE (extension)	2	0	0
		SOPRA	0	0	0
		SOAPdenovo2	4	0	0
		SGA	0	0	0
		CLA	0	0	2
7.	***S.* Typhi**	MeDuSa	30	0	0
		Bambus2	11	0	0

	#contigs:**67** #Misassmblies in input contigs: **0**	SSPACE(no extension)	1	0	0
		SSPACE (extension)	0	0	0
		SOPRA	0	0	0
		SOAPdenovo2	4	0	0
		SGA	0	0	0
8.	***T. pallidum*** #contigs:**22** #Misassmblies in input contigs: **0**	CLA	3	0	0
		MeDuSa	13	0	0
		Bambus2	2	0	0
		SSPACE(no extension)	1	0	0
		SSPACE (extension)	2	0	0
		SOPRA	1	0	0
		SOAPdenovo2	2	0	0
		SGA	0	0	0

*All the genomes were simulated with a read length of 100bp and insert size of 400bp

Table 6: Comparative statistics of CLA and Scaffolding tools generated using real dataset

	Genome	**Tool**	**# Scaffolds**	**# Misassemb-lies within scaffolds**	**# Misassemb-led Scaffolds**	**Genome Fraction**
1.	***C. jejuni*** #contigs: **72** Readlength: **100** Insert: **300** #Misassmblies in input contigs: **7**	CLA	19	9	5	90.551
		MeDuSa	10	33	1	90.764
		Bambus2	45	16	9	90.692
		SSPACE(no extension)	50	10	8	90.589
		SSPACE (extension)	51	11	7	90.612
		SOPRA	44	9	6	90.685
		SOAPdenovo2	52	9	6	90.626
		SGA	54	9	5	90.613
2.	***E. coli*** #contigs:**252** Readlength:**151** Insert:**300** #Misassmblies in input contigs: **98**	CLA	67	165	37	92.31
		MeDuSa	74	223	11	92.047
		Bambus2	174	139	54	91.998
		SSPACE(no extension)	241	99	60	91.967
		SSPACE (extension)	249	104	63	92.132
		SOPRA	212	117	56	91.995
		SOAPdenovo2	206	103	56	91.935
		SGA	233	100	56	91.948
3.	***H. influenzae*** #contigs:**50** Readlength:**100** Insert:**200** #Misassmblies in input contigs: **25**	CLA	6	29	4	88.984
		MeDuSa	12	45	4	87.571
		Bambus2	43	25	10	87.562
		SSPACE(no extension)	32	26	11	87.533
		SSPACE (extension)	33	28	11	87.604
		SOPRA	24	27	6	88.184
		SOAPdenovo2	29	28	9	87.595
		SGA	34	28	4	88.148
4.	***M. tuberculosis*** #contigs:**241** Readlength:**101** Insert:**300** #Misassmblies in	CLA	44	10	5	98.52
		MeDuSa	9	93	3	98.549
		Bambus2	139	59	24	98.324
		SSPACE(no extension)	153	11	10	98.432
		SSPACE (extension)	144	14	11	98.534

	input contigs: **6**	SOPRA	119	10	8	98.455
		SOAPdenovo2	108	12	12	98.365
		SGA	139	9	9	98.388
5.	***S.* Typhi** #contigs:**119** Readlength:**100** Insert:**200** #Misassmblies in input contigs: **2**	CLA	40	5	3	98.771
		MeDuSa	22	47	6	98.796
		Bambus2	80	19	10	98.771
		SSPACE(no extension)	85	6	5	98.803
		SSPACE (extension)	91	2	2	98.855
		SOPRA	80	8	4	98.712
		SOAPdenovo2	89	3	3	98.709
		SGA	119	2	2	98.717

Though MeDuSa generated lesser number of scaffolds in few cases of both real and simulated datasets, number of misassemblies was significantly higher in all the organisms. While misassemblies in case of SSPACE, SOPRA and SOAPdenovo2 were more or less similar to that of CLA, these tools gave higher number of scaffolds. Considering the trade-off between number of scaffolds generated and number of misassemblies, CLA showed best performance with least number of scaffolds having minimum misassemblies and highest genome fraction.

CLA vs. Combination of Reference and Scaffolding tools

CLA being able to perform ordering followed by scaffolding, its performance was also compared by ordering already generated scaffolds. Ragout was used to order scaffolds from SSPACE with extension, SOPRA, SOAPdenovo2 and SGA and the final results were compared to that of CLA. The comparative statistics are listed out in Table 7.

CLA performed better with least number of misassemblies in majority of cases like *B. quintana*, *C. crescentus*, *R. etli*, *S.* Typhi and *T. pallidum*. Ragout and SGA together gave out least number of misassemblies in *C. jejuni*, *H. influenzae* and *H. pylori* followed by CLA. Amongst the combinations used, Ragout and SGA gave out good results and was found to be similar to that of CLA while Ragout and SSPACE with extension did not seem to be an ideal combination.

Table 7: Comparative statistics of CLA and Ragout ordering of scaffolds using simulated dataset

	Genome*	Tool	# of contigs and scaffolds in the input file	# Misassemblies in Chromosome assembly	Unassigned Contigs from final order >1kbp	Largest contig unassigend	Genome Fraction
1.	***B. quintana*** #Misassmblies in input contigs: **0**	CLA	47	2	0	850	99.68
		SSPACE_ext/Ragout	38	15	2	1155	99.242
		SOPRA/Ragout	43	7	6	3347	98.296
		SOAPdenovo2/Ragout	35	9	3	11040	98.644
		SGA/Ragout	45	2	0	739	99.153
2.	***C. jejuni*** #Misassmblies in input contigs: **0**	CLA	33	4	0	425	99.728
		SSPACE_ext/Ragout	24	5	5	15071	97.389
		SOPRA/Ragout	32	5	4	4566	98.807
		SOAPdenovo2/Ragout	28	4	3	1954	99.094
		SGA/Ragout	32	3	4	4566	98.807
3.	***C. crescentus*** #Misassmblies in input contigs: **0**	CLA	49	2	0	793	99.901
		SSPACE_ext/Ragout	38	32	0	371	99.386
		SOPRA/Ragout	46	8	0	371	99.896
		SOAPdenovo2/Ragout	46	6	0	371	99.923
		SGA/Ragout	46	6	0	371	99.923
4.	***H. influenza*** #Misassmblies in input contigs: **1**	CLA	43	6	0	606	98.963
		SSPACE_ext/Ragout	35	11	1	1088	98.136
		SOPRA/Ragout	26	11	0	572	99.013
		SOAPdenovo2/Ragout	40	12	0	375	98.517
		SGA/Ragout	27	3	1	2461	98.247
5.	***H. pylori*** #Misassmblies in input contigs: **0**	CLA	46	7	0	960	99.672
		SSPACE_ext/Ragout	36	12	1	6406	98.657
		SOPRA/Ragout	42	4	1	5487	98.965
		SOAPdenovo2/Ragout	37	6	1	10376	98.746
		SGA/Ragout	44	4	1	5487	98.97
6.	***R. etli*** #Misassmblies in input contigs: **0**	CLA	30	3	0	601	99.756
		SSPACE_ext/Ragout	24	8	5	3972720	89.983
		SOPRA/Ragout	30	5	1	1052	99.406
		SOAPdenovo2/Ragout	24	7	4	3831	99.102
		SGA/Ragout	30	4	1	1052	99.501
7.	***S.* Typhi** #Misassmblies in input contigs: **0**	CLA	67	8	0	520	99.141
		SSPACE_ext/Ragout	53	61	1	5127	98.625
		SOPRA/Ragout	63	9	0	519	99.236
		SOAPdenovo2/Ragout	56	32	0	283	99.104
		SGA/Ragout	66	54	3	3835	99.142
8.	***T. pallidum*** #Misassmblies in input contigs: **0**	CLA	22	5	0	980	99.124
		SSPACE_ext/Ragout	14	9	1	1696	98.711
		SOPRA/Ragout	18	6	1	1617	99.535
		SOAPdenovo2/Ragout	14	10	0	244	99.513
		SGA/Ragout	21	8	1	1617	99.537

*All the genomes were simulated with a read length of 100bp and insert size of 400bp

Additional features of CLA

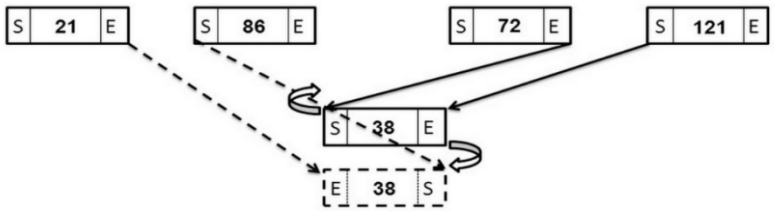
Handling intra-contig repeating segments

Small repetitive sequences which form only part of a contig also create problems during the assembly. Some of these elements consist of Insertion Sequence (IS) elements and transposable elements. CLA can split and precisely place certain segments of contigs which usually encode IS and transposable elements and display connections at different places given sufficient availability of paired-end information. Such segments within the contigs are extracted based on the read alignment positions and are placed according to its connections in the map-file ([b] in Fig 1). CLA in simulated data of *S*. Typhi, has placed a 500-600bp intra-contig repeat at about 24 places. This segment was found to be a transposable element that was present at about 26 positions in the original genome of *S*. Typhi TY2. The positioning of these small repeats according to CLA and their original position in the reference are mentioned in the Table 8. The performance of CLA in terms of placing small repeats appeared to vary and was dependent on the sequencing quality and coverage depth obtained. Therefore, CLA appears to be an efficient tool in handling even small repetitive elements in comparison to other tools which could not address these issues.

Contigs unrelated to the chromosomal assembly

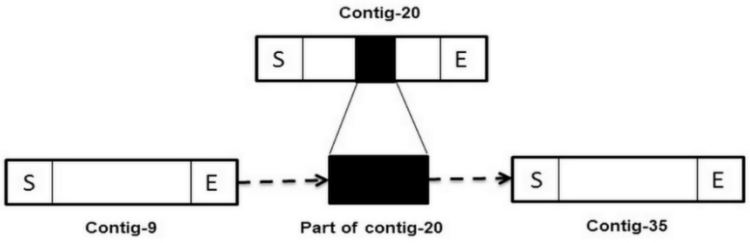
Possible contamination of DNA samples during sequencing and inaccurate de-multiplexing of the read data may lead to contamination of the reads potentially leading to the generation of un-related contigs. Such contigs might incur significant problems in the downstream analysis. In other cases, plasmids which are not a part of chromosomal sequences also pose problems in attaining accurate genome assembly. Therefore, CLA attempts to filter out such contigs and separates them from the main chromosomal assembly. Contigs with BLASTn identity (against reference) less than 5% and no paired-end link information are tagged as

(a)



Contig	Start	End
Contig_38_length_2410	Contig_86_length_1557-start;Contig_66_length_106077-mid;Contig_71_length_161135-mid;Contig_121_length_7537-mid;Contig_116_length_351813-mid;Contig_72_length_15210-end;Contig_8_length_70087-end;Contig_104_length_2525-end;Contig_65_length_7732-end;Contig_64_length_13919-end	Contig_63_length_6416-start;Contig_80_length_2674-start;Contig_121_length_7537-mid;Contig_34_length_34583-end;Contig_36_length_5124-end;Contig_121_length_7537-start;Contig_21_length_3050-end

(b)



CONTIG-NAME	START-REGION	END-REGION
Contig_9_length_56254	;Contig_45_length_4501-end	Contig_116_length_351813-mid;Contig_20_length_72457-mid;
Contig_35_length_32888	Contig_20_length_72457-mid;	-;-;-

(c)



CONTIG-NAME	START-REGION	END-REGION
Contig_95_length_1218	Contig_82_length_4595-end;-;Contig_27_length_46672-end	Contig_99_length_43852-start;Contig_82_length_4595-start

Fig 1. Pictorial representation of various scenarios where map-file can be used to resolve repetitive contigs
(a) A normal case scenario where a repetitive contig-38 is placed at two different positions based on its connections from the map-file.
(b) An example of an intra-contig repeating segment, where mid-region of contig-20 is connecting two contigs contig-9 and contig-35.
(c) Example of a Tandem repeat, where whole contig-95 has connections at both start and end pointing to another contig-82.

Table 8: BLAST output of intra-contig repeat from CLA result in simulated *S.* Typhi Ty2 data against the original Ty2 genome. The positions tabulated depict that around a 600bp intra-contig repeat is present at 26 different locations in the original genome.

	Intra contig repeat segment	Reference	%identity	Start position in reference	End position in reference
1	Contig_48_L_3835[1557-2173]	AE014613.1	100	3536792	3536176
2	Contig_48_L_3835[1557-2173]	AE014613.1	93.98	1898116	1898729
3	Contig_48_L_3835[1557-2173]	AE014613.1	94.11	2323287	2323896
4	Contig_48_L_3835[1557-2173]	AE014613.1	94.11	2534267	2534876
5	Contig_48_L_3835[1557-2173]	AE014613.1	94.1	528845	528237
6	Contig_48_L_3835[1557-2173]	AE014613.1	94.09	1802022	1802630
7	Contig_48_L_3835[1557-2173]	AE014613.1	94.1	2999401	3000009
8	Contig_48_L_3835[1557-2173]	AE014613.1	94.1	3108592	3107984
9	Contig_48_L_3835[1557-2173]	AE014613.1	94.08	1189885	1189279
10	Contig_48_L_3835[1557-2173]	AE014613.1	93.69	1745293	1745908
11	Contig_48_L_3835[1557-2173]	AE014613.1	93.93	3077484	3078093
12	Contig_48_L_3835[1557-2173]	AE014613.1	93.93	674093	673485
13	Contig_48_L_3835[1557-2173]	AE014613.1	93.93	1221320	1220712
14	Contig_48_L_3835[1557-2173]	AE014613.1	94.07	1265695	1265090
15	Contig_48_L_3835[1557-2173]	AE014613.1	93.93	1451127	1450519
16	Contig_48_L_3835[1557-2173]	AE014613.1	93.93	1647569	1646961
17	Contig_48_L_3835[1557-2173]	AE014613.1	93.93	1894909	1895517
18	Contig_48_L_3835[1557-2173]	AE014613.1	93.93	2116469	2117077
19	Contig_48_L_3835[1557-2173]	AE014613.1	93.93	3012263	3012871
20	Contig_48_L_3835[1557-2173]	AE014613.1	93.93	4604899	4605507
21	Contig_48_L_3835[1557-2173]	AE014613.1	93.78	1613193	1613802
22	Contig_48_L_3835[1557-2173]	AE014613.1	93.77	271634	271026
23	Contig_48_L_3835[1557-2173]	AE014613.1	93.77	3085076	3085684
24	Contig_48_L_3835[1557-2173]	AE014613.1	93.62	279969	280578
25	Contig_48_L_3835[1557-2173]	AE014613.1	93.44	2473149	2473757
26	Contig_48_L_3835[1557-2173]	AE014613.1	93.44	465283	464676

unrelated contigs. Such contigs were effectively excluded out by CLA in real datasets of *E. coli*, *H. influenzae*, *H. pylori*, *Mycobacterium tuberculosis* (*M. tuberculosis)* and *S*. Typhi. In *S*. Typhi one of the contigs tagged as unrelated were found to be plasmid contigs thus preventing false scaffolding with the chromosomal contigs. To examine the performance of CLA in detection of possible contamination, few random contigs from other genomes were introduced into the simulated data. CLA could detect these contigs as contamination from different source

and correctly labelled them as possibly unrelated contigs. Contigs tagged as unrelated could be cross-checked by the user for its use in the final assembly.

Information about probable misassemblies

Yet another advantage of CLA comes in the form of additional information about probable misassemblies occurring due to swapping contigs with same connections, which is given out in a log file to flag the issue(s).

Discussion

Genome assembly could be a challenging task especially for prokaryotes given the underlying plasticity of their genomes as dictated by discrete evolutionary events and bottlenecks that shape adaptation dynamics and lifestyles of bacterial organisms in different ecosystems (Ahmed 2009). Consequently, prokaryotic genomes are often replete with signatures reminiscent of various genetic rearrangements occurring due to frequent insertion, deletion and substitution events as well as enriched with multiple homopolymeric tracts (arising out of replication errors), repeat motifs of different composition and lengths, insertion sequences, prophages and genomic islands etc (Ahmed et al. 2008). All these plastic regions pose serious difficulties in assembling a genome mainly because of sequence redundancy that they bring in the form of multiple alleles, palindromes, inverted repeats and tandem duplications. Most of the available tools either rely on reference to order contigs at the whole genome level, or scaffold them based on the read data. While using a reference genome could result in omission/exclusion of certain strain specific elements (Galardini et al. 2011), the scaffolding struggles to resolve repetitive regions and limits ordering within the scaffolds (Hunt et al. 2014). Given these practical difficulties, we developed Contig-Layout-Authenticator (CLA), which combines the benefits of these individual approaches to

minimize the errors while generating a draft genome. CLA uses a reference at the beginning to create a preliminary sort order which then undergoes extensive validation based on paired-end read information to resolve repetitive elements and re-sorting of the contigs. The sorted contigs are only scaffolded based on the available read-pair information. Therefore, though ordering is tried to attain at the whole genome level, contigs are linked into scaffolds only when supported by paired-end information indicating their connectivity. Therefore, it is easier for the users to fill in the information between the scaffolds using further downstream processing in order to achieve a complete genome.

The existing reference based tools though efficient in sorting the contigs at the whole genome level, were observed to remove contigs from the final chromosome assembly which failed to align with the reference properly (Table 1, 2). These excluded contigs may lead to loss of significant genome information. For example, BLAST analysis of such excluded contigs from *H. influenzae* genome identified several metabolic genes which were otherwise discarded. On the other hand, the scaffolding of contigs is performed only based on paired-end information or in the case of MeDuSa using multiple reference genomes. In case of final assembly/genome obtained with multiple scaffolds, ordering is limited within the scaffolds and also repetitive regions with their misleading connections sometimes lead to intra-scaffold misassemblies. From our validation study, it was observed that CLA could address all these issues better than the compared tools (Table 1-8). Since CLA utilizes both reference and paired-end information, it performed better in retaining maximum number of contigs in the final output without compromising on accuracy. It also utilizes read-pair information to place some of the repetitive elements. The overall performance of CLA was found to be much better than the existing reference based ordering tools as well as scaffolding tools with minimum number of misassemblies. In all our case studies and comparative analyses, CLA was seen to be misplacing contigs in just two cases: 1) firstly, when there was insufficient paired-end information. 2) Secondly, when two contigs had same flanking contig

connections at both ends leading to their mis-joining within the scaffold. To avoid such scenarios, CLA also lists out contigs with probable swapping in the log file to alert the users of probable misassemblies within scaffolds.

The performance with the real and simulated datasets pointed out the capability of CLA in not only handling data from monomorphic bacteria such as *S*. Typhi but also highly diverse ones as *H. pylori*. The higher abundance of transposases in bacteria and important biological roles implicated for them in previous studies underlines the need to address them carefully during genome assembly (Aziz et al. 2010). CLA was observed to be efficient in this aspect and was able to resolve 24 out of 26 transposable elements in *S*. Typhi.

Though manual curation is inevitable for completing a genome, CLA leaves very less scope for manual intervention and also provides all the required blueprints to complete further re-construction of the genome. Therefore, we believe that in the light of existing difficulties regarding the genome assembly, CLA would be a significant step forward in improving the genome assembly pipeline with user friendly approach and efficient data usage.

Materials and Methods

Real and simulated data-sets

Eight complete bacterial genomes with varied genomic characteristics (Table 9) were downloaded from NCBI. The paired reads were simulated for each of the genomes using GemSIM (v.1.6) (McElroy et al. 2012) using Illumina error model. The genome characteristics of these genomes along with their accession numbers are provided in Table 9. The 100 bases long read-pairs were simulated with an insert length of 400 (+/- 20) along with genome coverage of 100X. The reads were filtered using NGS QCToolkit (v.2.3) (Patel et al. 2012) to remove bad quality reads.

These filtered reads were then assembled into contigs using Velvet *de novo* assembler (1.2.08) with an optimal k-mer was chosen with the help of VelvetOptimiser (2.2.5) (http://bioinformatics.net.au/software.velvetoptimiser.shtml). The assembled contigs were used as input for validating CLA with other tools using QUAST. For the real dataset, five paired-reads dataset with different organismal background from SRA dataset were considered. Information about these datasets is provided in the Table 10. Filtering followed by the assembly of reads was carried out in a similar manner as described above.

Table 9: Genome characteristics and information of the strains utilized for simulation along with accession numbers of the strains and their genomic characteristics that were used for the simulated dataset

S. no	Organism	GenBank ID	(G+C)%	Feature	Genome Size In Bp##	Reference genomes** GenBank ID
1	***Bartonella quintana*** **str. Toulouse**	BX897700.1	38.8	Clonal	1581384	CP003784.1 BX897699.1* CP001562.1*
2	***Campylobacter jejuni*** **NCTC 11168**	AL111168.1	30.55	Highly recombining	1641481	HG428754.1 CP003871.3* CP000025.1*
3	***Caulobacter crescentus*** **CB15**	AE005673.1	67.21	High (G+C)%	4016947	CP001340.1 CP002008.1* CP000927.1*
4	***Haemophilus influenzae*** **Rd KW20**	L42023.1	38.15	Highly recombining	1830023	FQ312006.1 CP005967.1* CP002277.1*
5	***Helicobacter pylori*** **J99**	AE001439.1	39.19	Highly recombining	1643831	CP001217.1 CP000012.1* AE000511.1*
6	***Rhizobium etli*** **CFN 42**	CP000133.1	61.27	Clonal, High (G+C)%	4381608	CP005950.1 CP001074.1* CP006986.1*
7	***Salmonella*** **Typhi Ty2**	AE014613.1	52.05	Clonal	4791950	CP003278.1 CP007559.1* AL513382.1*
8	***Treponema pallidum*** **str.Fribourg-Blanc**	CP003902.1	52.78	Highly clonal	1140253	CP003115.1 CP002375.1* CP002374.1*

** Reference strain used for their respective simulated data while validating the tools. *Additional reference genomes used for running MeDuSa and Ragout

Table 10: Strains of real dataset with accession numbers and reference genomes used for the study

S.no	Bacteria of real dataset	SRA Accession numbers	Reference genomes**
1.	***Campylobacter jejuni***	SRR1999732	CP006006.1 HG428754.1* CP003871.3* CP000025.1*
2.	***Escherichia coli***	SRR1788079	CP006636.1 HG941718.1* CP006784.1* CP000819.1*
3.	***Haemophilus influenzae***	DRR015774	CP005967.1 FQ312006.1* CP007471.1* CP002277.1*
4.	***Mycobacterium tuberculosis***	ERR845304	CP001664.1 HE663067.1* AP012340.1* CP009480.1*
5.	***Salmonella enterica***	ERR351254	AE006468.1 CP012151.1* AL513382.1* CP012091.1*

** Reference strain used for CLA,reference based ordering tools and for generating QUAST statistics. * Additional reference strains used for MeDuSa and Ragout. * Additional reference genomes used for running MeDuSa and Ragout

Pipeline of CLA

The tool runs in the following stages.

Sort order Creation

The detailed schematic of the tool is explained in the Fig 2. Contigs, paired-end sequence reads and a suitable reference genome sequence (preferably same species of the genome being assembled) serve as input for the tool. The process starts with the exclusion of contigs that are less than 200bp. A preliminary sort order is created after alignment of the contigs with the provided reference genome using BLASTn. Contigs with a BLASTn best hit of less than a threshold identity value (for contigs of size<1kbp: 50% identity; size>1-10kbp: 25% identity; size>10kbp: 10% identity) are placed at the end of the order. The individual contigs

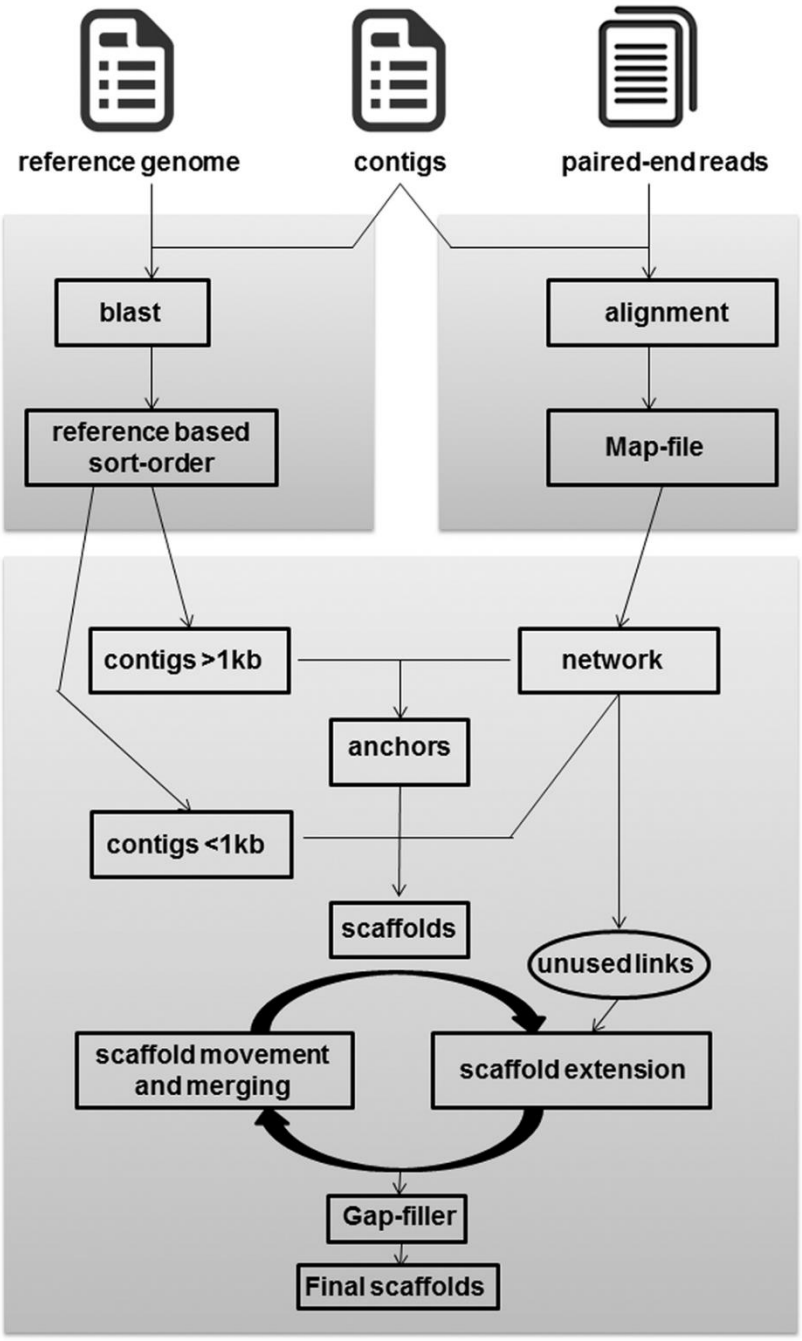


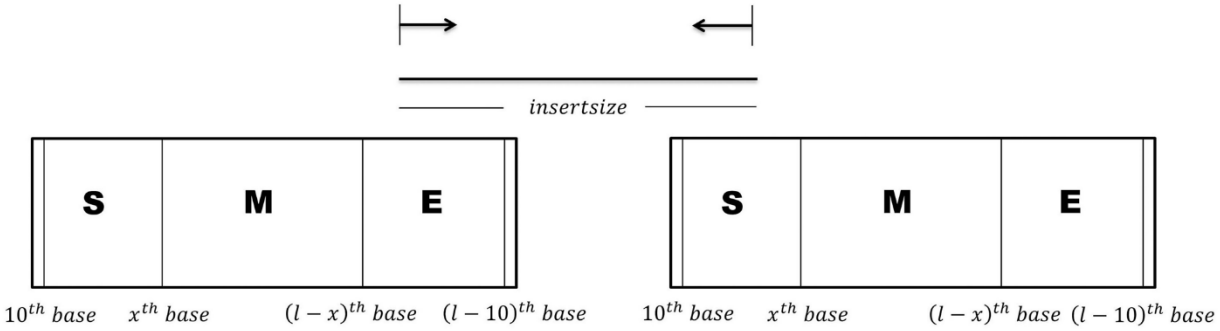
Fig 2. Schematic overview of the Contig-Layout-Authenticator pipeline. The schema of CLA pipeline is divided into three stages. In the first stage, a reference based order is derived followed by the second stage where connections between the individual contigs are extracted based on the alignment information. The final stage makes use of the information from the first two stages to decide the final order followed by scaffolding and gap filling.

are arranged according to the sort order with consequent reverse complementation of contigs wherever required in accordance with BLASTn output.

Extracting the connections

The paired reads are then mapped to the sorted contigs (obtained above) using BWA (Li & Durbin 2009) and their information regarding the alignment position and orientation is extracted. Each of the contigs is theoretically divided into start, mid and end regions as depicted in the Fig 3 and the sequence read having its pair in any other contig constitutes a potential link. The information of all the possible links (formed by the read pairs) at the start, mid and end regions of each of the contig is then tabulated in a map file. Since the read pairs are derived from a single insert, these links strongly suggest the proximity of these two contigs being together in the real genome. For example, as depicted in the Fig 3, if two contigs are in proximity then the reads aligned at the end of one contig would have their linking pairs aligned at the start of the following contig. The map file contains four columns representing the contig ID, start, end and mid region of the contigs respectively. To account for the erroneous assembly at the extreme ends of the contigs, first and last ten bases from the start and the end of each contig respectively, are not considered for calculating the links. A minimum of 10 read pair links are considered to calculate the proximity of two contigs to be together in the genome (valid connection between contigs). Therefore, the map file generation is instrumental in validation of the reference sorted contigs and also helps in resolving some of the issues such as repeats, duplications and inversions. Some examples regarding the resolutions of repetitive regions and possible duplications are detailed in the Fig 1

.



$$x = insertsize + (insertsize * 0.2)$$

$$l = contig\ length$$

$$start(S) = \left[10^{th}\ base\right]\ to\ [x^{th}\ base]$$

$$mid(M) = \left[x^{th}\ base\right]\ to\ [(l-x)^{th}\ base]$$

$$end(E) = \left[(l-x)^{th}\ base\right] to\ [(l-10)^{th}\ base]$$

Fig 3: Defining the start and end of a contig based on the insert size. Connectivity between two consecutive contigs is decided by the paired reads from a single insert. To find such connections between contigs, each contig is theoretically divided into start, mid and end regions based on the insert size.

Sort-order validation

Contigs that have less than 5% BLASTn identity against reference (as inferred from the earlier BLASTn output) and lacking suitable connections in the map-file are flagged as un-related and are removed from the downstream analysis. These might represent contigs formed due to contaminated reads or sometimes plasmid contigs as observed from the results of the real data set.

A network file is then created from the map-file, where start/end of a contig is represented as vertex/node and their respective connections as edges. Contigs that are less than 1kbp in size are initially excluded from the sort order. The tool then scans the map-file to validate the

reference based sort order looking for connections between the neighboring contigs (contigs >1kbp). In cases where there is no direct connection (no links connecting the end of the former and start of the later contig) between two consecutive contigs from the reference based sort order, the tool then looks for any intermediary contigs linking these two contigs based on the link information from the map-file. Each path is a representation of one or more contigs and a shortest path defined by one or more contigs connecting these two contigs based on the link information from the map-file are found using *Floyd Warshall* algorithm (Floyd 1962). Only paths that include contigs greater than 1kbp in size are considered for at this stage. The connecting contig is either copied or moved from its position based on the connections in the map-file. All the contigs at the end of this stage are referred to as anchors.

Connecting the anchors

The link information of the excluded contigs of size less than 1kbp are now used to connect the defined anchors again using *Floyd Warshall* algorithm as discussed in the previous section. All such connected anchors are then merged into scaffolds after placing N's between them. To prevent any false positive placements, the sort-order is scanned at every step based on the number of connections from the map-file.

Ordering the scaffolds

An iterative process of scaffold merging and extension is performed based on the unused connections and leftover contigs from the network file. After a search for inter scaffold connections, if two scaffolds are found to be connected, the one smaller in size is moved and oriented in accordance with the larger scaffold. Once merged, the map-file is scanned for any further unused connections at the ends of the newly extended scaffolds. Merging is again performed and the process is repeated until either all the connections are exhausted or no proper connections could be found.

A final sort order is then created and the contigs are merged into scaffolds followed by their ordering all in accordance with the new sort order. The gaps between the scaffolds are then closed by using GapFiller (Boetzer & Pirovano 2012). Even though the order is defined, all the contigs are not merged into a single pseudo-genome and are left at the stage of scaffolds because of lack of valid connections deciphered from the map-file. Thus, inter-scaffold gaps indicate the existence of gaps in the sequencing data. FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) is used by the tool for easy handling and manipulation of the intermediate files.

Tool availability

Project name: Contig-Layout-Authenticator

Project homepage: https://sourceforge.net/projects/c-l-authenticator/files

Supporting OS: Linux, Mac

Programming Language: Perl

References

Ahmed, N., 2009. A flood of microbial genomes - Do we need more? *PLoS ONE*, 4(6).

Ahmed, N. et al., 2008. Genomic fluidity and pathogenic bacteria: Applications in diagnostics, epidemiology and intervention. *Nature Reviews Microbiology*, 6(5), pp.387–394. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-42349093581&partnerID=40&md5=c11539ccfcccdaa6287e41a210202b72.

Altschul, S.F. et al., 1990. Basic local alignment search tool. *Journal of molecular biology*, 215(3), pp.403–10. Available at: http://www.sciencedirect.com/science/article/pii/S0022283605803602.

Assefa, S. et al., 2009. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics*, 25(15), pp.1968–1969. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19497936.

Aziz, R.K., Breitbart, M. & Edwards, R.A., 2010. Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res*, 38(13), pp.4207–4217. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20215432.

Boetzer, M. et al., 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, 27(4), pp.578–579. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21149342.

Boetzer, M. & Pirovano, W., 2012. Toward almost closed genomes with GapFiller. *Genome Biol*, 13(6), p.R56. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22731987.

Bosi, E. et al., 2015. MeDuSa: a multi-draft based scaffolder. *Bioinformatics*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25810435.

Dayarian, A., Michael, T.P. & Sengupta, A.M., 2010. SOPRA: Scaffolding algorithm for paired reads via statistical optimization. *BMC Bioinformatics*, 11, p.345. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/20576136.

Doolittle, W.F. & Zhaxybayeva, O., 2009. On the origin of prokaryotic species. *Genome Res*, 19(5), pp.744–756. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19411599.

Edwards, D.J. & Holt, K.E., 2013. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microb Inform Exp*, 3(1), p.2. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23575213.

Floyd, R.W., 1962. Algorithm 97: Shortest path. *Communications of the ACM*, 5(6), p.345. Available at: http://dl.acm.org/citation.cfm?id=367766.368168.

Galardini, M. et al., 2011. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. *Source Code Biol Med*, 6, p.11. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21693004.

Gurevich, A. et al., 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8), pp.1072–1075. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23422339.

Hunt, M. et al., 2014. A comprehensive evaluation of assembly scaffolding tools. *Genome Biol*, 15(3), p.R42. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24581555.

Kolmogorov, M. et al., 2014. Ragout-a reference-assisted assembly tool for bacterial genomes. *Bioinformatics*, 30(12), pp.i302-9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24931998.

Koren, S., Treangen, T.J. & Pop, M., 2011. Bambus 2: scaffolding metagenomes. *Bioinformatics*, 27(21), pp.2964–2971. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21926123.

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler

transform. *Bioinformatics*, 25(14), pp.1754–1760.

Loman, N.J. et al., 2012. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat Rev Microbiol*, 10(9), pp.599–606. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22864262.

Luo, R. et al., 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, 1(1), p.18. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23587118.

McElroy, K.E., Luciani, F. & Thomas, T., 2012. GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC Genomics*, 13, p.74. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22336055.

Patel, R.K. et al., 2012. NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data Z. Liu, ed. *PLoS ONE*, 7(2), p.e30619.

Pop, M., Kosack, D.S. & Salzberg, S.L., 2004. Hierarchical scaffolding with Bambus. *Genome Res*, 14(1), pp.149–159.

Quail, M.A. et al., 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 13, p.341. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22827831.

Rissman, A.I. et al., 2009. Reordering contigs of draft genomes using the Mauve aligner. *Bioinformatics*, 25(16), pp.2071–2073. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19515959.

Schuster, S.C., 2008. Next-generation sequencing transforms today's biology. *Nat Methods*, 5(1), pp.16–18. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18165802.

Simpson, J.T. & Durbin, R., 2012. Efficient de novo assembly of large genomes using

compressed data structures. *Genome Res*, 22(3), pp.549–556. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22156294.

Warren, R.L. et al., 2007. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, 23(4), pp.500–501. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17158514.

Zerbino, D.R. & Birney, E., 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5), pp.821–829.

CHAPTER 3: (Objective 2)

To analyze genomes of pandemic ExPEC Sequence types- ST131, ST38, ST405 and ST648, to provide insight into the genetic level similarities/differences among these clones

---

This chapter is in communication as:

Genomic analysis of dominant clonal lineages of extra-intestinal pathogenic *Escherichia coli*

INTRODUCTION

Extra-intestinal pathogenic *Escherichia coli* (ExPEC) strains are a versatile variant of commensal *E. coli* demonstrating a complex genome plasticity and phylogeny(Kaper et al. 2004). ExPEC strains can cause a range of disease syndromes, from uncomplicated urinary tract infections to life-threatening septicemia in humans (Kaper et al. 2004). Control of infections caused by ExPEC has been troublesome due to the emergence of antimicrobial resistance particularly the extended-spectrum-β-lactamase (ESBL)-producing *E. coli* strains which are linked with the increased rates of morbidity and mortality (Tumbarello et al. 2007). Association of these multi drug resistant (MDR) strains with high rate of community as well as hospital acquired infections can be largely attributed to the evolution of a number of different clones within the species with enhanced metabolic repertoires along with increased levels of virulence and antibiotic resistance (Dobrindt 2005).

To understand the evolution of such clones, several classification schemes have been suggested, of which multi locus sequence typing (MLST) is usually considered to be the gold standard (Sullivan et al. 2005; Van der Bij et al. 2012). ESBL-producing *E. coli* ST131 was identified among thousands of sequence types (STs) or clones defined by MLST based on the variations in seven housekeeping genes of *E. coli* (Peirano & Pitout 2010; Wirth et al. 2006). Compared to other pathogenic ExPEC clones, ST131 has been widely studied due to its predominant association with urinary tract as well as blood stream infections worldwide (Nicolas-Chanoine et al. 2008; Hussain et al. 2012). This globally disseminated multidrug resistant clone belonging to phylogenetic group B2 has also been associated with the worldwide emergence of a particular ESBL, CTX-M-15(Schembri et al. 2015). Relatively higher virulence and metabolic capabilities along with resistance to wider classes of antibiotics can be attributed to the epidemiological persistence of this particular lineage (Hussain et al. 2014; Vimont et al. 2012).

Recent reports suggest that a few other sequence types such as ST38, ST405 and ST648 were also associated with global dissemination of CTX-M-producing *E. coli* (Pitout 2012). *E. coli* ST38 carrying CTX-M, NDM-1 and OXA-48 genes were earlier reported from several countries such as Japan, Netherlands, Korea and Tanzania (Suzuki et al. 2009; Van Der Bij et al. 2011; Kim et al. 2011; Mshana et al. 2011; Poirel et al. 2011; Yamamoto et al. 2011; Ranjan et al. 2016). While CTX-M positive ST405 and ST648 were also reported from different parts of the world (Coque et al. 2008; Jones et al. 2008; Mihaila et al. 2010; Smet et al. 2010; Cortés et al. 2010; Zong & Yu 2010; Van Der Bij et al. 2011), NDM-producing ST405 and ST648 were reported from UK (Hornsey et al. 2011; Mushtaq et al. 2011). Comparative *in vitro* analysis of antibiotic resistance and virulence determinants of strains belonging to ST38, ST131, ST405 and ST648 reported the presence of certain virulence factors such as *sat*, *iutA*, *malX*, *usp* and *ompT* which were associated with greater adaptability, competitiveness and colonization capabilities (Peirano et al. 2013; Johnson et al. 2010; Van der Bij et al. 2012).

Whole genome based comparative analysis provides insights into the variation in the genetic architecture within or among different species (Ahmed 2009). While many ST131 whole genome sequences are already reported and analyzed worldwide, no such studies have yet been performed with ST38, ST405 and ST648 strains. In this study, we reported whole genome sequencing of five ExPEC strains belonging to these four STs which were previously obtained from western India. In addition, whole genome based comparative analysis of 139 strains belonging to ESBL-producing ST38, ST405 and ST648 with ST131 from public as well as in-house databases was performed. Findings from this study give an insight into the genetic level differences among these clones which could be the potential basis for their global predominance over each other.

RESULTS

Strain information and genome characteristics of in-house strains

The assembly of NA023, NA081, NA090, NA101 and NA112 strains comprised of 176, 223, 193, 175 and 168 contigs respectively, with an approximate genome size of 5.2 Mb. The average G + C content was ~50% and an average CDS (coding sequence) number of all the strains was found to be around ~5000 bases amounting to a coding percentage of ~87. The characteristics of the five genomes are presented in Table 1. While two strains NA101 and NA112 belonged to ST131, strains NA023, NA083 and NA090 belonged to ST648, ST405 and ST38, respectively.

Dataset collection for the comparative analysis

Apart from the five in-house strains, 134 strains belonging to the four STs were selected from the public databases. Thus, the dataset included 99 ST131 (from all three clades A, B and C), 12 ST38, 10 ST405 and 18 ST648 comprising of 139 strains from different isolation sources such as stool, patients suffering from UTI and bacteremia, etc. (Table 2). These 139 strains were used for the *in silico* comparative analysis and five in-house strains NA023, NA081, NA090, NA101 and NA112 were characterized by various functional assays.

Table1: Background data and Genome Characteristics of five in-house strains NA023, NA081, NA090, NA101 and NA112 belonging to four STs considered in this study.

Strain	Disease status	Collection date	ST	# Contigs	Genome size	G+C %	CDS	Coding %	# rRNAs	# tRNAs
NA023	Prostitis	12-03-2009	ST648	176	5193516	50.57	5077	86.8	5	73
NA081	UTI	10-08-2009	ST405	223	5242265	50.5	5060	86.6	7	68
NA090	Septicemia	09-02-2009	ST38	193	5371032	50.41	5268	86.2	20	87
NA101	Pyelonephritis	05-10-2009	ST131	175	5289993	50.72	5244	87.2	13	79
NA112	Pyrexia of unknown origin	03-11-2009	ST131	168	5150072	50.78	5085	87.3	6	67

Table2: Information about 139 strains belonging to ST38, ST405, ST648 and ST131 considered for the comparative genomic analysis in this study.

S.no	Strain Name	WGS project/ SRA Accession	Isolation source	ST
1	LAU-EC10	AYNJ01	gastrointestinal	ST38
2	IS1	CBWA01	Unknown	ST38
3	1-110-08_S4_C1	JHDK01	stool	ST38
4	1-392-07_S1_C1	JNPT01	stool	ST38
5	1-392-07_S1_C2	JNPU01	stool	ST38
6	6-175-07_S1_C2	JOMS01	stool	ST38
7	6-175-07_S1_C3	JORL01	stool	ST38
8	upec-221	JSKR01	patient with urinary tract infection	ST38
9	upec-133	JSNS01	patient with urinary tract infection	ST38
10	blood-10-180	JSPS01	blood from patient with bacteremia	ST38
11	blood-10-0183	JSQQ01	blood from patient with bacteremia	ST38
12	NA090	MVIO01	Septicemia	ST38
13	LAU-EC5	AYOG01	Gastrointestinal	ST405
14	LAU-EC4	AYOP01	Gastrointestinal	ST405
15	upec-211	JSLB01	patient with urinary tract infection	ST405
16	blood-90544	JSPE01	blood from patient with bacteremia	ST405
17	blood-90543	JSPF01	blood from patient with bacteremia	ST405
18	blood-9-0292	JSPI01	blood from patient with bacteremia	ST405
19	blood-10-0554	JSQN01	blood from patient with bacteremia	ST405
20	blood-10-0541	JSQO01	blood from patient with bacteremia	ST405
21	blood-09-0464	JSRM01	blood from patient with bacteremia	ST405
22	NA081	JSXM01	patient with Urinary tract infection	ST405

23	BIDMC 19C	AXLI01	Urine	ST648
24	LAU-EC8	AYNH01	gastrointestinal	ST648
25	LAU-EC9	AYNI01	gastrointestinal	ST648
26	668	AYQW01	Unknown	ST648
27	IS5	CBWB01	Unknown	ST648
28	ECONIH1	CP009859.1	rectal swab	ST648
29	BIDMC 43a	JAPF01	Blood Culture	ST648
30	BIDMC 19B	JAPI01	Urine	ST648
31	BIDMC 19A	JAPJ01	Bronchoalveolar Lavage	ST648
32	BIDMC 17B	JAPK01	Peritoneal Fluid	ST648
33	BIDMC 17A	JAPL01	Blood Culture	ST648
34	BIDMC 3	JAPP01	Swab, Abdominal Incision	ST648
35	BIDMC 2B	JAPQ01	Blood Culture	ST648
36	BIDMC 6	JAPV01	Bile	ST648
37	LR09	JDVF01	wastewater treatment plant	ST648
38	BIDMC 82	JJMY01	urine	ST648
39	CR694	JTGI01	urine	ST648
40	NA023	JSXK01	Prostitis	ST648
41	SE15	AP009378.1	feces	ST131-A
42	S120EC	ERR161283	Blood	ST131-A
43	S26EC	ERR161245	Urine	ST131-A
44	S2EC	ERR161235	Urine	ST131-A
45	S31EC	ERR161300	Urine	ST131-A
46	S34EC	ERR161247	Urine	ST131-A
47	S37EC	ERR161302	Urine	ST131-A
48	S5EC	ERR161236	Urine	ST131-A
49	S94EC	ERR161257	Blood	ST131-A
50	HVM1147	ERR161318	Peritoneal fluid	ST131-B
51	HVM2044	ERR161323	Blood	ST131-B
52	HVM2289	ERR161325	Urine	ST131-B
53	HVM277	ERR161315	Urine	ST131-B
54	HVM52	ERR161308	Urine	ST131-B
55	S104EC	ERR161267	Urine	ST131-B

56	S105EC	ERR161268	Urine	ST131-B
57	S114EC	ERR161277	Urine	ST131-B
58	S128EC	ERR161291	Blood	ST131-B
59	S19EC	ERR161241	Urine	ST131-B
60	S21EC	ERR161242	Urine	ST131-B
61	S22EC	ERR161243	Urine	ST131-B
62	S24EC	ERR161244	Urine	ST131-B
63	S32EC	ERR161301	Urine	ST131-B
64	S6EC	ERR161299	Urine	ST131-B
65	S79EC	ERR161305	Urine	ST131-B
66	B36EC	ERR161254	Blood	ST131-C
67	HVM1299	ERR161320	Abdominal abscess	ST131-C
68	HVM1619	ERR161321	Surgical wound	ST131-C
69	HVM1997	ERR161322	Urine	ST131-C
70	HVM3017	ERR161328	Urine	ST131-C
71	HVM3189	ERR161329	Urine	ST131-C
72	HVM5	ERR161306	Urine	ST131-C
73	HVM826	ERR161316	Blood	ST131-C
74	HVM834	ERR161317	Urine	ST131-C
75	HVR2496	ERR161326	Blood	ST131-C
76	HVR83	ERR161311	Blood	ST131-C
77	IR18E	ERR458470	Unknown	ST131-C
78	IR49	ERR458471	Unknown	ST131-C
79	IR65	ERR458472	Unknown	ST131-C
80	IR68	ERR458473	Unknown	ST131-C
81	MS2481	ERR161252	Blood	ST131-C
82	MS2493	ERR161253	Blood	ST131-C
83	P146EC	ERR161313	Rectal swab	ST131-C
84	P189EC	ERR161314	Rectal swab	ST131-C
85	P50EC	ERR161307	Rectal swab	ST131-C
86	P53EC	ERR161309	Rectal swab	ST131-C
87	P56EC	ERR161310	Rectal swab	ST131-C
88	S100EC	ERR161263	Rectal swab	ST131-C

89	S101EC	ERR161264	Rectal swab	ST131-C
90	S102EC	ERR161265	Urine	ST131-C
91	S103EC	ERR161266	Urine	ST131-C
92	S107EC	ERR161270	Urine	ST131-C
93	S108EC	ERR161271	Blood	ST131-C
94	S109EC	ERR161272	Urine	ST131-C
95	S10EC	ERR161237	Urine	ST131-C
96	S110EC	ERR161273	Urine	ST131-C
97	S111EC	ERR161274	Urine	ST131-C
98	S112EC	ERR161275	Urine	ST131-C
99	S113EC	ERR161276	Rectal swab	ST131-C
100	S115EC	ERR161278	Urine	ST131-C
101	S116EC	ERR161279	Unknown	ST131-C
102	S117EC	ERR161280	Unknown	ST131-C
103	S118EC	ERR161281	Urine	ST131-C
104	S119EC	ERR161282	Blood	ST131-C
105	S11EC	ERR161238	Urine	ST131-C
106	S121EC	ERR161284	Urine	ST131-C
107	S122EC	ERR161285	Urine	ST131-C
108	S123EC	ERR161286	Urine	ST131-C
109	S124EC	ERR161287	Urine	ST131-C
110	S125EC	ERR161288	Surgical wound	ST131-C
111	S126EC	ERR161289	Urine	ST131-C
112	S127EC	ERR161290	Urine	ST131-C
113	S129EC	ERR161292	Urine	ST131-C
114	S12EC	ERR161239	Urine	ST131-C
115	S130EC	ERR161293	Urine	ST131-C
116	S131EC	ERR161294	Urine	ST131-C
117	S132EC	ERR161295	Urine	ST131-C
118	S133EC	ERR161296	Blood	ST131-C
119	S134EC	ERR161297	Urine	ST131-C
120	S135EC	ERR161298	Blood	ST131-C
121	S15EC	ERR161240	Urine	ST131-C

122	S1EC	ERR161234	Urine	ST131-C
123	S30EC	ERR161246	Urine	ST131-C
124	S39EC	ERR161248	Unknown	ST131-C
125	S43EC	ERR161249	Unknown	ST131-C
126	S47EC	ERR161250	Unknown	ST131-C
127	S53EC	ERR161251	Unknown	ST131-C
128	S65EC	ERR161303	Urine	ST131-C
129	S77EC	ERR161304	Urine	ST131-C
130	S92EC	ERR161255	Blood	ST131-C
131	S93EC	ERR161256	Blood	ST131-C
132	S95EC	ERR161258	Blood	ST131-C
133	NA097	JSXJ01	Urine sample from patient with Septicemia	ST131-C
134	NA101	JSXN01	Pyelonephritis	ST131-C
135	NA112	JSXO01	Pyrexia of unknown origin	ST131-C
136	NA114	CP002797.2	urine	ST131-C
137	JJ1886	CP006784.1	patient with fatal urosepsis	ST131-C
138	MNCRE44	CP010876.1	Sputum of sepsis patient	ST131-C
139	EC958	HG941718.1	urine	ST131-C

Whole genome phylogenomic analysis confirms clonality within the Sequence Types

Core genome based phylogenetic tree (Fig 1) showed four distinct clades representing each sequence type, emulating the MLST scheme depicting the clonality as well as distinctness at the whole genome level. Three clades A (orange), B (pink) and C (red) as earlier reported were also observed within the ST131 dataset(Petty et al. 2014). Clades of ST131 were observed to be more distant from the remaining STs and also clades representing ST38 (purple) and ST405 (blue) were relatively closer than ST648 (green) (Fig 1).

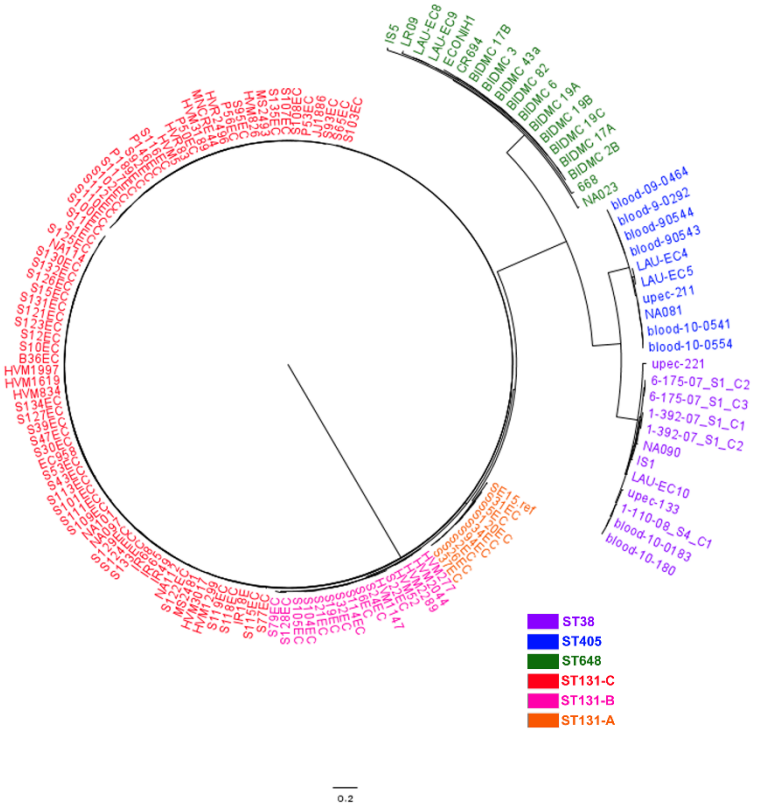


Fig 1: Whole genome phylogenetic tree depicting individual STs as separate clades. Clades A, B and C of ST131 were observed to be distant that the remaining ST38, ST405 and ST648.

ST's exhibit similar plasmid and resistance profiles

From the BLASTn (Altschul et al. 1990) analysis of 139 strains against the downloaded Plasmid Finder (Carattoli et al. 2014) database, it was observed that 98 out of 99 strains of ST131 harbored IncF plasmid. Among them, in-house strains NA097(Ranjan et al. 2015), NA101 and NA114(Avasthi et al. 2011) (earlier reported) harbored IncF plasmid with pMLST profile F2:A1:B- while NA112 has IncF with F4:A-:B- (Table 3). These two pMLST profiles were earlier reported to be most common among the ST131 strains(Phan et al. 2015). IncF was also observed to be the major class within ST38 and ST405, where 11 out of 12 strains of ST38 and all 10 strains of ST405 harbored IncF. In case of ST648, only 7 out of 18 were observed to have IncF. None of the strains from ST38, ST405 and ST648 had F2:A1: B- and F1:A2:B20 pMLST profiles. In addition to IncF, plasmid types IncHI2, IncI1,

IncN, IncP, IncQ, IncB/O/K/Z, IncX1, IncX3, IncX4 and IncY were observed in lower prevalence in ST131, ST38 and ST405. Few strains of ST648 strains harbored IncHI2 and IncN (Table 3).

BLASTp (Altschul et al. 1990) analysis of putative protein sequences of 139 strains against the CARD (McArthur et al. 2013) database (Fig 2A) showed that the resistance profiles of strains belonging to all four STs were nearly similar. A majority of the strains from ST131 (44/99), ST648 (6/18) and ST405 (8/10) carried different variants of CTX-M-15 gene, while only 2 out of 12 strains of ST38 carried this gene. Similarly, variants of TEM, SHV and OXA were also present in most of the strains from all four STs. Other resistance genes from different classes were also present in the majority of strains but no ST specific pattern was apparent. Five in-house strains NA023, NA081, NA090, NA101 and NA112 were tested *in vitro* for 18 different antimicrobials belonging to six different classes namely β-lactam antibiotics (penicillin and cephalosporin), fluoroquinolone, aminoglycoside, tetracycline, trimethoprin/sulphonamide and macrolide. The four strains NA101, NA023, NA081 and NA090 belonging to different sequence types ST131, ST648, ST405 and ST38 were resistant to 10, 12, 12 and 8 antibiotics, respectively. The other strain of ST131 NA112 was resistant to only six antibiotics. All five strains harbored CTX-M-15 gene. The complete resistome of strains is represented in the form of a heatmap (Fig 2B).

Table 3: Plasmid classes observed in each of the strains belonging to ST38, ST405, ST648 and ST131. A positive hit with identity >=95% and query coverage >=90% indicates presence of that particular class in the strain.

S.No	Strain	IncF	IncHI2	IncI1	IncN	IncP	IncQ	IncB /O/K/Z	IncX1	IncX3	IncX4	IncY
	ST38											
1	upec-221	-	-	-	-	P	P	-	-	-	-	-
2	6-175-07_S1_C3	[F1:A-:B-]	-	-	-	-	-	-	-	-	-	-
3	IS1	[F1:A-:B23]	-	-	-	-	-	-	-	-	-	-
4	6-175-07_S1_C2	[F1:A-:B33*]	-	-	-	-	-	-	-	-	-	-
5	1-392-07_S1_C1	[F1:A-:B33]	-	-	-	-	-	-	-	-	-	-
6	1-392-07_S1_C2	[F1:A-:B33]	-	-	-	-	-	-	-	-	-	-
7	NA090	[F1:A-:B33]	-	-	-	-	-	-	-	-	-	-
8	1-110-08_S4_C1	[F2:A-:B10]	-	-	-	-	-	-	-	-	-	-
9	LAU-EC10	[F29:A4*:B10]	-	-	-	-	-	-	-	-	-	-
10	upec-133	[F31:A4:B37*]	-	-	-	-	-	-	-	-	-	-
11	blood-10-180	[F51:A-:B10]	-	-	-	-	-	-	-	-	-	-
12	blood-10-0183	[F51:A-:B10]	-	-	-	-	-	-	-	-	-	-
	ST405											
1	upec-211	[F1:A1:B-]	-	P	-	-	-	P	-	-	-	-
2	blood-90543	[F1*:A1:B16]	-	-	-	-	-	P	-	-	-	-
3	blood-90544	[F1:A1:B16]	-	-	-	-	-	P	-	-	-	-

4	blood-9-0292	[F1:A1:B16]	-	-	-	-	-	-	-	-	-	-
5	blood-09-0464	[F1:A1:B16]	-	-	-	-	-	-	-	-	-	-
6	blood-10-0554	[F24:A-:B6]	-	-	-	-	-	-	-	-	-	-
7	blood-10-0541	[F24:A-:B6]	-	-	-	-	-	-	-	-	-	P
8	NA081	[F24:A-:B6]	-	-	-	-	-	-	-	-	-	P
9	LAU-EC4	[F36:A1:B10]	-	-	-	-	-	-	-	-	-	-
10	LAU-EC5	[F36:A1:B10]	-	-	-	-	-	-	-	-	-	-
	ST648											
1	BIDMC 19C	-	P	-	P	-	-	-	-	-	P	-
2	668	-	-	-	P	-	-	P	-	-	-	-
3	BIDMC 43a	-	P	-	-	-	-	-	-	-	-	-
4	BIDMC 19B	-	-	-	-	-	-	-	-	-	-	-
5	BIDMC 19A	-	-	-	-	-	-	-	-	-	-	-
6	BIDMC 17B	-	-	-	-	-	-	-	-	-	-	-
7	BIDMC 17A	-	P	-	P	-	-	-	-	-	-	-
8	BIDMC 3	-	P	-	P	-	-	-	-	-	-	-
9	BIDMC 2B	-	P	-	P	-	-	-	-	-	P	-
10	BIDMC 6	-	P	-	P	-	-	-	-	-	P	-
11	BIDMC 82	-	P	-	P	-	-	P	-	-	P	-

12	IS5	[F1:A1:B-]	-	-	-	-	-	P	-	-	-	-
13	ECONIH1	[F1:A1:B1]	-	-	P	-	-	-	-	-	-	-
14	LR09	[F1:A1:B1]	-	-	-	-	-	-	-	-	-	P
15	NA023	[F31:A4:B1]	-	-	-	-	-	-	-	-	-	-
16	CR694	[F36:A-:B32]	-	P	-	-	-	-	-	P	P	-
17	LAU-EC8	[F48:A1:B49]	-	-	-	-	-	-	-	-	-	-
18	LAU-EC9	[F48:A1:B49]	-	-	-	-	-	-	-	-	-	-
	ST131											
1	HVM5	-	-	-	-	-	-	-	-	-	-	-
2	S11EC	[F-:A1:B-]	-	-	-	-	-	P	-	-	-	-
3	S1EC	[F-:A1:B-]	-	-	-	-	-	-	-	-	-	-
4	HVR83	[F-:A2:B20]	-	P	-	-	-	-	P	-	P	-
5	S103EC	[F1:A-:B23]	-	-	-	-	-	-	-	-	-	-
6	S65EC	[F1:A-:B23]	-	-	-	-	-	-	-	-	-	-
7	S34EC	[F1:A1:B1]	-	-	-	-	P	-	-	-	-	-
8	S37EC	[F1:A1:B1]	-	-	-	-	-	-	-	-	-	-
9	S26EC	[F1:A1:B23]	-	-	-	-	-	-	-	-	-	-
10	S2EC	[F1:A1:B23]	-	-	-	-	-	-	-	-	-	-
11	S31EC	[F1:A1:B23]	-	-	-	-	-	-	-	-	-	-

12	S110EC	[F1:A2:B-]	-	-	-	-	-	-	-	-	-	-
13	S116EC	[F1:A2:B-]	-	-	-	-	-	-	-	-	-	-
14	P146EC	[F1:A2:B20*]	-	-	-	-	-	P	-	-	-	-
15	MNCRE44	[F1:A2:B20]	-	-	-	-	-	-	-	P	P	-
16	HVM3189	[F1:A2:B20]	-	-	-	-	-	-	-	-	-	-
17	HVM826	[F1:A2:B20]	-	-	-	-	-	-	-	-	-	-
18	HVR2496	[F1:A2:B20]	-	-	-	-	-	-	-	-	-	-
19	MS2493	[F1:A2:B20]	-	-	-	-	-	-	-	-	-	-
20	P189EC	[F1:A2:B20]	-	-	-	-	-	-	-	-	-	-
21	P50EC	[F1:A2:B20]	-	-	-	-	-	P	-	-	-	-
22	P53EC	[F1:A2:B20]	-	-	-	-	-	-	-	-	P	-
23	P56EC	[F1:A2:B20]	-	-	-	-	-	-	-	-	-	-
24	S100EC	[F1:A2:B20]	-	-	-	-	-	-	-	-	-	-
25	S102EC	[F1:A2:B20]	-	-	-	-	-	-	-	-	-	-
26	S107EC	[F1:A2:B20]	-	-	-	-	-	-	-	-	-	-
27	S108EC	[F1:A2:B20]	-	-	-	-	-	-	-	-	-	-
28	S112EC	[F1:A2:B20]	-	-	-	-	-	-	-	-	-	-
29	S120EC	[F1:A2:B20]	-	-	-	-	-	-	-	-	-	-
30	S135EC	[F1:A2:B20]	-	-	-	-	-	-	-	-	-	-

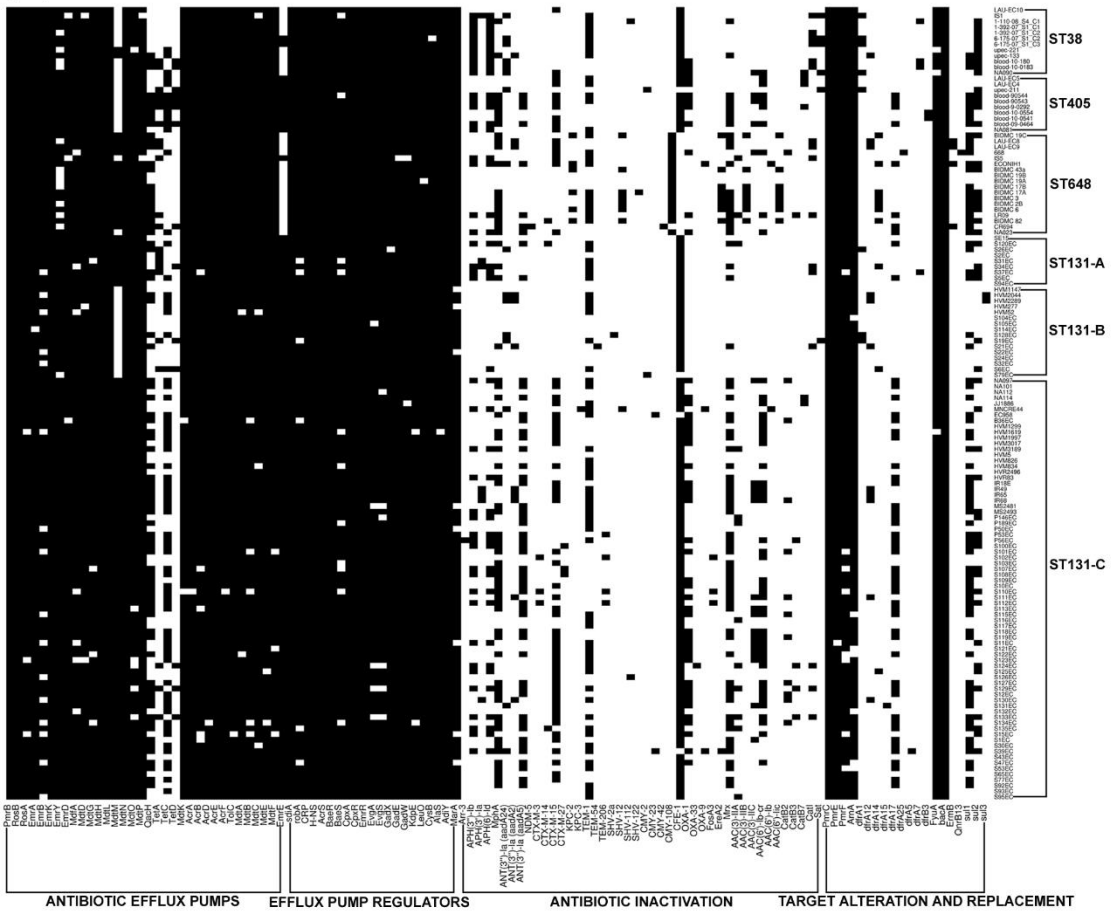
31	S92EC	[F1:A2:B20]	-	-	-	-	-	-	-	-	P	-
32	S95EC	[F1:A2:B20]	-	-	-	-	-	-	-	-	-	-
33	S117EC	[F10:A-:B-]	-	-	-	-	-	-	P	-	-	-
34	S131EC	[F16:A-:B1]	-	-	-	P	-	-	-	-	P	-
35	HVM2044	[F2:A-:B1]	-	P	-	-	-	-	-	-	-	-
36	HVM2289	[F2:A-:B1]	-	P	-	-	-	-	-	-	-	-
37	HVM277	[F2:A-:B1]	-	-	-	-	-	-	-	-	-	-
38	HVM52	[F2:A-:B1]	-	P	-	-	-	P	-	-	-	-
39	NA097	[F2:A1:B-]	-	-	-	-	-	-	-	-	-	-
40	NA101	[F2:A1:B-]	-	-	-	-	-	-	-	-	-	-
41	NA114	[F2:A1:B-]	-	-	-	-	-	-	-	-	-	-
42	JJ1886	[F2:A1:B-]	-	-	-	-	-	-	-	-	-	-
43	EC958	[F2:A1:B-]	-	-	-	-	-	-	-	-	-	-
44	B36EC	[F2:A1:B-]	-	-	-	-	-	-	-	-	-	-
45	HVM1619	[F2:A1:B-]	-	-	-	-	-	-	-	-	P	-
46	HVM1997	[F2:A1:B-]	-	-	-	-	-	-	-	-	-	-
47	HVM834	[F2:A1:B-]	-	-	-	-	-	-	P	-	P	-
48	S101EC	[F2:A1:B-]	-	-	-	-	-	-	-	-	-	-
49	S109EC	[F2:A1:B-]	-	-	-	-	-	P	-	-	-	-

50	S10EC	[F2:A1:B-]	-	-	-	-	-	-	-	-	-	-
51	S113EC	[F2:A1:B-]	-	-	-	-	-	-	-	-	-	P
52	S121EC	[F2:A1:B-]	-	-	-	-	-	-	-	-	-	-
53	S123EC	[F2:A1:B-]	-	-	-	-	-	-	-	-	-	-
54	S126EC	[F2:A1:B-]	-	-	-	-	-	-	-	-	-	-
55	S127EC	[F2:A1:B-]	-	-	-	-	-	-	-	-	P	-
56	S12EC	[F2:A1:B-]	-	-	-	-	-	-	-	-	-	-
57	S132EC	[F2:A1:B-]	-	-	-	-	-	-	-	-	-	-
58	S134EC	[F2:A1:B-]	-	-	-	-	-	-	-	-	-	-
59	S15EC	[F2:A1:B-]	-	-	-	-	-	-	P	-	-	-
60	S30EC	[F2:A1:B-]	-	-	-	-	-	-	-	-	-	-
61	S39EC	[F2:A1:B-]	-	-	P	-	-	-	-	-	P	-
62	S47EC	[F2:A1:B-]	-	-	-	-	-	-	-	-	-	-
63	S53EC	[F2:A1:B-]	-	-	-	-	-	-	-	-	-	-
64	S77EC	[F2:A1:B-]	-	-	-	-	-	-	-	-	-	-
65	S93EC	[F2:A1:B-]	-	-	-	-	-	-	-	-	-	-
66	S125EC	[F2:A6*:B1]	-	-	-	-	-	-	-	-	P	P
67	IR49	[F22:A1:B20]	-	-	-	-	-	-	-	-	-	-
68	IR65	[F22:A1:B20]	-	-	-	-	-	-	-	-	-	-

69	S111EC	[F22:A1:B20]	-	-	-	-	-	-	-	-	-	-
70	S122EC	[F22:A1:B20]	-	-	-	-	-	-	-	-	-	-
71	S130EC	[F22:A1:B20]	-	-	-	-	-	-	-	-	-	-
72	S128EC	[F24:A-:B40]	-	P	-	P	-	-	-	-	-	-
73	SE15	[F29:A-:B10]	-	-	-	-	-	-	-	-	-	-
74	HVM1147	[F29:A-:B10]	-	-	-	-	-	-	-	-	-	-
75	S104EC	[F29:A-:B10]	-	-	-	-	-	-	-	-	-	-
76	S105EC	[F29:A-:B10]	-	-	-	-	-	-	-	-	P	-
77	S114EC	[F29:A-:B10]	-	-	-	-	-	-	-	-	-	-
78	S21EC	[F29:A-:B10]	-	-	-	-	-	-	-	-	-	-
79	S22EC	[F29:A-:B10]	-	-	-	-	-	-	-	-	-	-
80	S24EC	[F29:A-:B10]	-	-	-	-	-	-	-	-	-	-
81	S5EC	[F29:A-:B10]	-	-	-	-	-	-	-	-	-	-
82	S94EC	[F29:A-:B10]	-	-	-	-	-	-	-	-	-	-
83	HVM3017	[F31:A4:B1]	-	-	-	-	-	-	-	-	-	-
84	IR18E	[F31:A4:B1]	-	-	-	-	-	-	-	-	-	-
85	IR68	[F36:A1:B20]	-	-	-	-	-	-	-	-	-	-
86	MS2481	[F36:A1:B20]	-	-	-	-	-	-	-	-	-	-
87	HVM1299	[F36:A4:B1]	-	-	-	-	-	-	-	-	-	-

88	S115EC	[F36:A4:B1]	-	-	-	-	-	-	-	-	-	-
89	S118EC	[F36:A4:B1]	-	-	-	-	-	-	-	-	-	-
90	S119EC	[F36:A4:B1]	-	-	-	-	-	-	-	-	-	-
91	NA112	[F4:A-:B-]	-	-	-	-	-	-	-	-	P	-
92	S79EC	[F4:A-:B1]	-	P	-	-	-	-	-	-	-	-
93	S19EC	[F4:A-:B10]	-	-	-	-	-	P	P	-	P	-
94	S124EC	[F48:A1:B49]	-	-	-	-	P	-	-	-	P	P
95	S129EC	[F48:A1:B49]	-	-	-	-	P	-	-	-	P	P
96	S133EC	[F48:A1:B49]	-	-	-	-	P	-	-	-	-	-
97	S43EC	[F51*:A-:B-]	-	-	-	-	-	-	-	-	-	-
98	S32EC	[F51:A-:B10]	-	-	-	-	-	-	-	-	-	-
99	S6EC	[F51:A-:B10]	-	-	P	-	-	-	-	-	-	-

(A)



(B)

	Antibiotics	Penicilinases			Cephalosporins					Fluoroquinolones				Trimethoprim/ sulphonamides	Aminoglycosides		Macrolides	Tetracyclines	
Sequence types	strains	AMP	PIT	CB	CTX	CEP	CPZ	CPM	CZ	NX	CIP	MO	GAT	COT	GEN	AK	NIT	DO	TE
ST131	112	■			■					■	■	■	■						
	101	■		■	■	■	■	■	■		■	■		■					
ST648	23	■		■	■	■	■	■	■	■	■	■		■					■
ST405	81	■	■	■	■	■	■	■	■	■	■	■	■						
ST38	90	■			■	■	■		■	■	■			■					

Fig 2A: Heat map depicting the 56 essential CARD resistance genes in our dataset containing 139 strains belonging to ST38, ST405, ST648 and ST131 (clades A, B, C). Black indicates a positive hit with identity >=70% and query coverage >=75%; Fig 2B: Representation of *in vitro* analysis of five in-house strains tested for 18 different antimicrobials belonging to six different classes.

ST131 exhibits relatively robust virulence profile

BLASTn analysis of the 139 strains against the virulence factor database (VFDB) is represented in the form of a heatmap (Fig 3), which emulates the results of previously reported *in vitro* virulence studies on these four STs (Van der Bij et al. 2012). Virulence factor such as *kpsM-III* group capsule was present in 4 out of 10 strains of ST405 and in a single strain of ST648 but was missing from all 99 strains of ST131. Also *hra* which was reported to be commonly present in ST38 was observed in 11 out of 12 strains and few strains from other STs. *ompT* which was earlier reported in the strains of ST131 and ST648 was observed to be present in 17/18 strains of ST648 and all 99 strains of ST131 but was missing in the remaining STs. While *usp* was exclusively present in all the 99 strains of ST131, *cnf1* and *iroN* were present only in 14 and 8 ST131 strains. These three virulence factors were completely missing from ST38, ST405 and ST648. Pathogenicity island marker (*malX*) was present in all 139 strains irrespective of the STs. Adhesin and invasion related genes such as *upaG*, *ehaG*, *yfaL*, *ibeB*, *ompA* were present in all 139 strains. But several autotransporters, serine proteases (*espP*, *epeA*, *pic*, EC55989_4660) and many UPEC-specific virulence genes were found to be completely absent among ST38, ST405 and ST648 strains, whereas they were present in the majority of ST131 strains. Few auto-transporter genes such as *ypjA*, *yejO*, ECIAI39_1904 were observed to be missing from ST131 but were present in the reaining STs (Fig 3).

The *in vitro* adhesion and invasion assays revealed that all five strains were able to adhere and invade the bladder epithelial cell line (T24). As compared to *E. coli* dh5-alpha (negative control), strains of all four sequence types were significantly adherent ($p<0.001$). There were certain strain specific differences in adherence capabilities (Fig 4A). Similarly, the strains were also significantly invasive as compared to negative control (Fig 4B). Thus, the strains demonstrated similar pathogenic capabilities in terms of adherence and invasion to epithelial cells, irrespective of STs.

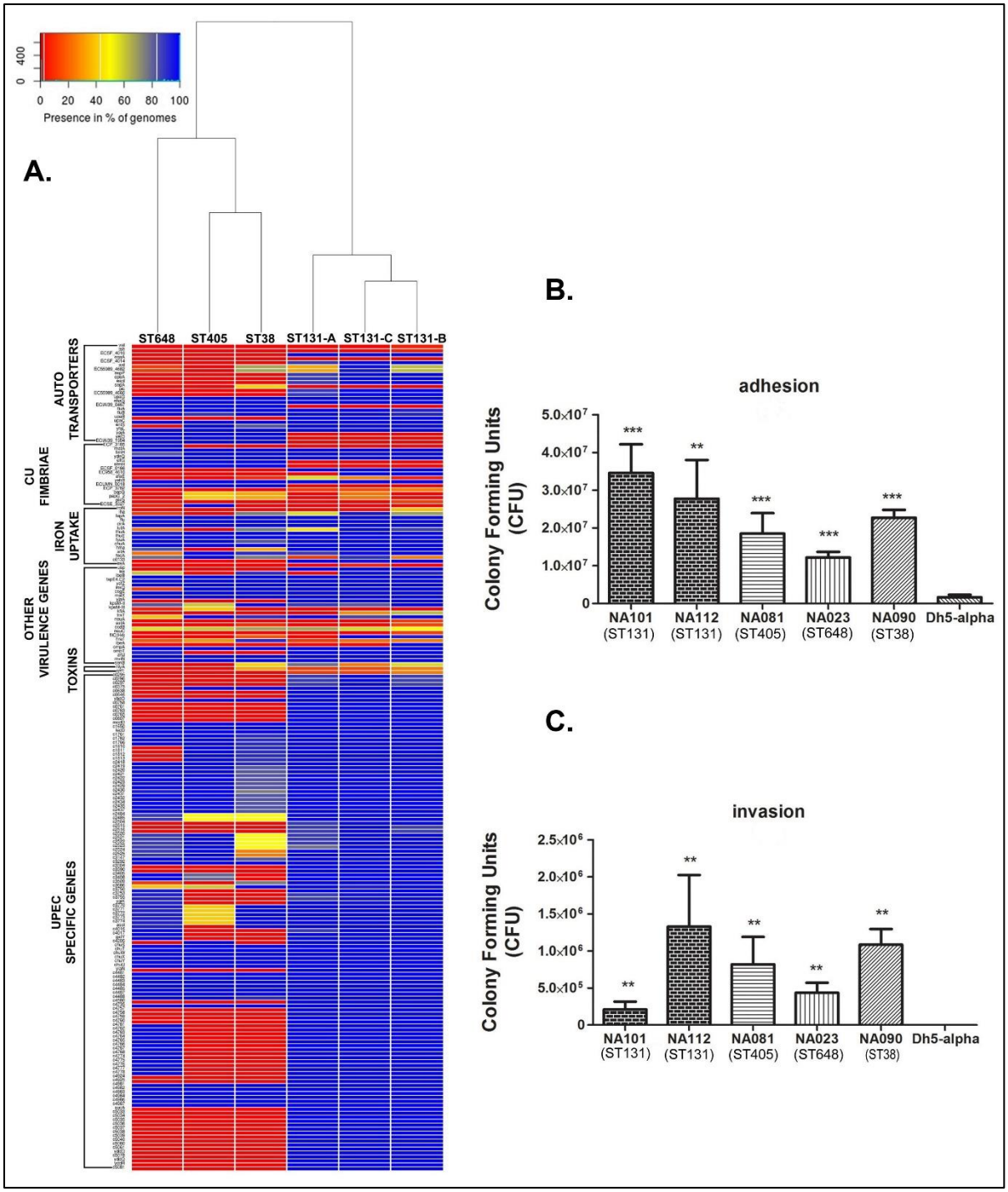


Fig 3a: Gradient Heat map depicting the frequency of virulence genes in ST38, ST405, ST648, ST131 (clades A, B and C). Frequency of each gene is calculated by the formula (Presence in no. of strains of ST / Total no. of strains in ST) * 100. Color ranging towards red and blue depict lower and greater frequencies respectively. Fig 3b,c: *In vitro* assays performed on five in-house strains NA023, NA081, NA090, NA101 and NA112 belonging to all four STs considered in this study. Fig 3B. Adhesion assay. Fig 3C. Invasion assay

Mobile genetic elements indeed define ST131

BRIG (Alikhan et al. 2011) analysis (Fig 5A) of mobile genetic elements revealed that *GI-leuX*, *GI-pheV* and *GI-thrW* were conserved in majority of the ST131 strains but were either missing partially or completely from non-ST131 strains. *GI-thrW* was observed to be missing in the in-house two ST131 strains NA101 and NA112. *GI-pheV* which did not follow any conservation pattern in case of ST131 strains was also completely or partially absent in all other STs. So was the case with the other mobile genetic elements. 31,448 bp High pathogenicity Island (HPI) from EC958 which is 99% identical to that of *Yersinia pestis* (Totsika et al. 2011) found to be the only mobile genetic element in all strains of ST131, ST405 and ST648; and 10 out of 12 strains of ST38.

A similar BRIG image (Fig 5B) analyzing MTases revealed that chromosome associated MTases such as M.EcoMDcm (EC958_2226), M.EcoMVI (EC958_3663) and M.EcoMDam (EC958_3778) were present in all the strains of the dataset. M.EcoMIII (EC958_0425), a Type I R-M (restriction modification system**)** which is a part of *GI-thrW* was missing from all strains of ST38, ST405 and ST648 and also in two ST131 strains (NA101 and NA112). Lack of *GI-thrW* in these strains as discussed above justifies the absence of this particular MTase. While M.EcoMIV (EC958_1101) encoded by prophage-2 was conserved in most of the strains of ST131 clade C strains, it was completely missing in the remaining strains. M.EcoMVII (EC958_4083) was observed to be present in only few strains of ST131 emulating the pattern of its associated mobile genetic element *GI-selC*. Other MTases such as M1.EcoMI /M2.EcoMI (EC958_0008/EC958_0009), M.EcoMII (EC958_0078), M.EcoMV (EC958_1545), M.EcoMVIII (pEC958_A0009), McrBC (EC958_0011 and EC958_0012) and Mrr (EC958_0079) were present in almost all the strains of clade C and few other strains.

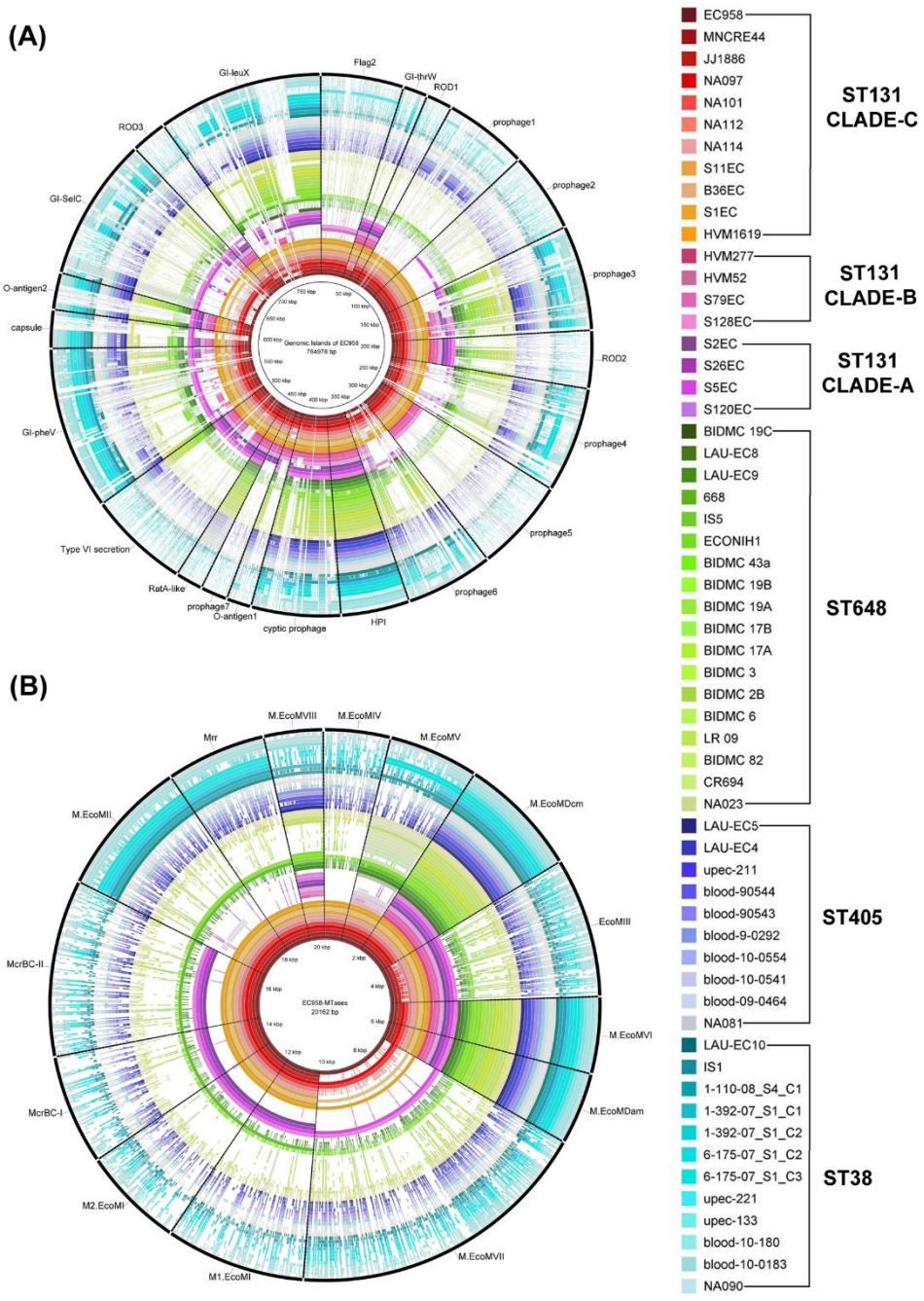


Fig 3: Mobile Genetic Elements and MTases in non-ST131 (ST38, ST405 and ST648) strains. Fig 3A. BRIG image showing the presence of mobile genetic elements reported in EC958 in our dataset; 3B. BRIG image with EC958 MTases as reference.

Comparison of Essential gene set of serum resistome

All 139 strains were analyzed for the presence of 56 essential genes belonging to the serum resistome of EC958 which were recently defined by TraDIS (Transposon directed insertion-site sequencing) (Phan et al. 2013). It was observed from the BLASTp analysis (Fig 6A) that few genes were missing (hit identity <70% and query coverage < 75%) from the complete dataset of ST38, ST405 and ST648 which include EC958_0460 (*hyxA*), EC958_0461(*hyxR*), EC958_1112, EC958_1114, EC958_2371, EC958_2373, EC958_4029 (*waaL*), EC958_4030 (*waaU /waaK*), EC958_4035 (*waaB/rfaB*), EC958_4032 (*waaY/rfaY*), EC958_4033 (*waaJ/rfaJ*) and EC958_4034 (*waaI/rfaI*) [fig 6b]. EC958_0858 (*tolA*) belonging to the Tol-Pal system of *E. coli* was observed to be only partially present in these strains (129 bp out of 436bp with identity >90%).

In vitro analysis of the in house strains NA101, NA112, NA090, NA023 and NA081 belonging to all four sequence types revealed that all strains were found to be resistant to the bactericidal activity of human serum and showed significant growth in 50% human serum in comparison to the negative control (Dh5-alpha) (Fig 6B). NA114 and NA097 were also earlier reported to be resistant to human serum (Avasthi et al. 2011; Ranjan et al. 2015).

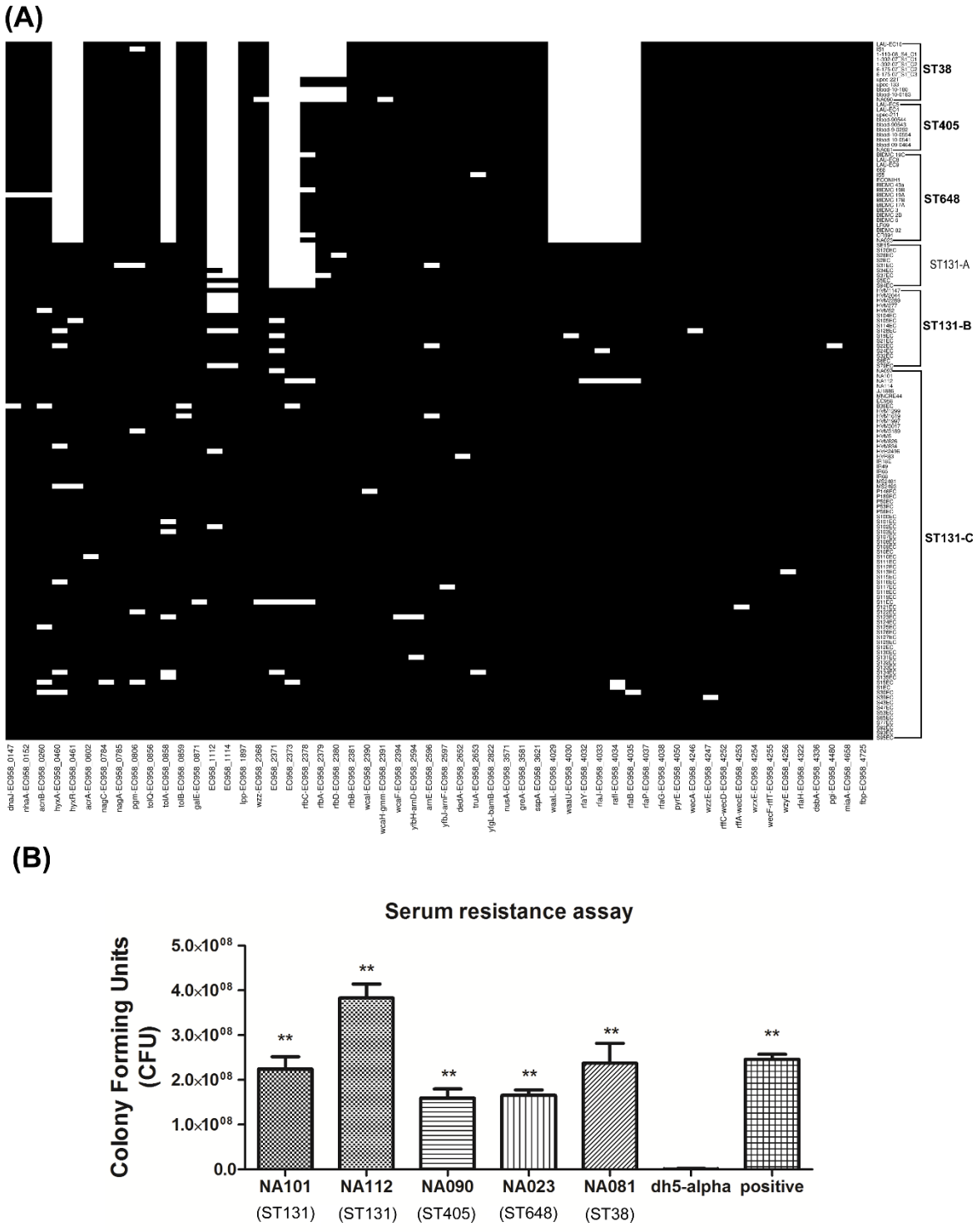


Fig 4: Serum Resistance Analysis.4A. Heat map depicting the 56 essential serum resistome genes in our dataset containing 139 strains belonging to ST38, ST405, ST648 and ST131 (clades A,B,C). Black indicates a positive hit with identity >=70% and query coverage >=75%; 4B. Serum resistance assay

Genes missing from ST131

Core genome analysis was performed to identify the core gene content separately for each sequence type; ST38, ST131, ST405 and ST648. About 5169, 8189, 5495 and 5851 orthologous groups were identified in case of ST38, ST131, ST405 and ST648, respectively. Of these, 4101, 4140, 4362 and 4279 belonged to their respective cores. The core protein sequences including paralogs were extracted from one representative strain of each ST and another round of core genome analysis was performed using only the core protein sequences of individual STs. This analysis resulted in 4328 orthologous groups, of which 3472 belonged to the common core and 856 groups belonged to the accessory content.

Accessory content of the four cores, excluding the paralog clusters was further analyzed to identify any proteins that were specifically present or absent in ST131. It was observed that no ST131 specific proteins were identified, however 224 proteins that were specifically absent in the core of ST131 were identified. Another round of BLASTp analysis of these 224 protein sequences against the complete dataset of 139 strains revealed that about 142 proteins (Table 4) were indeed absent in all 99 strains of ST131 but were present in the remaining 40 strains belonging to ST38, ST405 and ST648.

The functional annotation of these 142 sequences revealed that 99 proteins had a clusters of orthologous group (COG) (Koonin 2002) domain. These 99 proteins included membrane proteins, transcriptional regulators, transporters, toxins-antitoxins, chaperones, fimbrial proteins etc., belonging to a wide range of COG categories.

Table 4: Information about the list of proteins that are specifically missing in all 99 strains of ST131 while were present in remaining 40 strains belonging to ST38, ST405 and ST648. Presence indicates a positive BLASTp hit with identity >=70%.

S.no	UNIPROT ID	Function	COG ID	COG class	COG Category
1	V8KDQ1	Lipopolysaccharide 1,2-glucyltransferase	COG1442	[M] Cell wall/membrane/envelope biogenesis	Cellular Processes And Signaling
2	V8JY35	D-alanyl-D-alanine dipeptidase	COG2173	[M] Cell wall/membrane/envelope biogenesis	Cellular Processes And Signaling
3	V8KE30	Betaine/carnitine/choline BCCT transporter	COG1292	[M] Cell wall/membrane/envelope biogenesis	Cellular Processes And Signaling
4	V8KDC7	Beta-1,4-galactyltransferase	COG3306	[M] Cell wall/membrane/envelope biogenesis	Cellular Processes And Signaling
5	V8K6U2	Type-1 fimbrial protein subunit A	COG3539	[N] Cell motility	Cellular Processes And Signaling
6	V8K634	Fimbrial protein	COG3539	[N] Cell motility	Cellular Processes And Signaling
7	V8KUT9	Fimbrial protein	COG3539	[N] Cell motility	Cellular Processes And Signaling
8	V8KSH5	Fimbrial protein	COG3539	[N] Cell motility	Cellular Processes And Signaling
9	V8KT34	Fimbrial protein	COG3539	[N] Cell motility	Cellular Processes And Signaling
10	V8KUU3	Fimbrial protein	COG3539	[N] Cell motility	Cellular Processes And Signaling
11	V8K5N9	Adhesin	COG3539	[N] Cell motility	Cellular Processes And Signaling
12	V8K6T9	Membrane protein	COG0425	[O] Posttranslational modification, protein turnover, chaperones	Cellular Processes And Signaling

13	V8K087	Diguanylate cyclase yddV	COG2199	[T] Signal transduction mechanisms	Cellular Processes And Signaling
14	V8KAJ7	Antitoxin MazE	COG2336	[T] Signal transduction mechanisms	Cellular Processes And Signaling
15	V8K746	Uncharacterized protein	COG3521	[U] Intracellular trafficking, secretion, and vesicular transport	Cellular Processes And Signaling
16	V8K7E7	Type VI secretion protein	COG3518	[U] Intracellular trafficking, secretion, and vesicular transport	Cellular Processes And Signaling
17	V8K4Y2	Membrane protein	COG3455	[U] Intracellular trafficking, secretion, and vesicular transport	Cellular Processes And Signaling
18	V8K751	CAD protein	COG3515	[U] Intracellular trafficking, secretion, and vesicular transport	Cellular Processes And Signaling
19	V8K349	UPF0214 protein YfeW	COG1680	[V] Defense mechanisms	Cellular Processes And Signaling
20	V8KAZ7	Toxin mazF	COG2337	[V] Defense mechanisms	Cellular Processes And Signaling
21	V8KFX2	CRISPR-associated endonuclease Cas1	COG1518	[V] Defense mechanisms	Cellular Processes And Signaling
22	V8JSL0	Antitoxin YafN	COG2161	[V] Defense mechanisms	Cellular Processes And Signaling
23	V8K7W1	Molecular chaperone	COG3121	[W] Extracellular structures	Cellular Processes And Signaling
24	V8KUP8	Fimbrial protein	COG3121	[W] Extracellular structures	Cellular Processes And Signaling
25	V8KUP6	Fimbrial chaperone BcfG	COG3121	[W] Extracellular structures	Cellular Processes And Signaling
26	V8JKS1	XRE family transcriptional regulator	COG2944	[K] Transcription	Information Storage And Processing
27	V8KUZ3	Uncharacterized protein	COG2944	[K] Transcription	Information Storage And Processing
28	V8K3J8	Uncharacterized protein	COG4886	[K] Transcription	Information Storage And Processing

29	V8K7W6	Transcriptional regulator	COG2207	[K] Transcription	Information Storage And Processing
30	V8KFV9	Transcriptional regulator	COG1846	[K] Transcription	Information Storage And Processing
31	V8KG32	Transcriptional regulator	COG1609	[K] Transcription	Information Storage And Processing
32	V8KCQ7	LysR family transcriptional regulator	COG0583	[K] Transcription	Information Storage And Processing
33	V8KVE3	LysR family transcriptional regulator	COG0583	[K] Transcription	Information Storage And Processing
34	V8KFC2	LysR family transcriptional regulator	COG0583	[K] Transcription	Information Storage And Processing
35	V8K0Z8	HTH-type transcriptional regulator MurR	COG1737	[K] Transcription	Information Storage And Processing
36	V8KIN4	GntR family transcriptional regulator	COG2188	[K] Transcription	Information Storage And Processing
37	V8K2V1	Transpase	COG5464	[L] Replication, recombination and repair	Information Storage And Processing
38	V8JLS9	Transpase	COG5464	[L] Replication, recombination and repair	Information Storage And Processing
39	V8KCB7	ATP-dependent helicase	COG1201	[L] Replication, recombination and repair	Information Storage And Processing
40	V8K144	Hydrogenase-4 subunit E	COG4237	[C] Energy production and conversion	Metabolism
41	V8K394	Hydrogenase 3 membrane subunit	COG0650	[C] Energy production and conversion	Metabolism
42	V8KTL4	Glutamate symport transmembrane protein	COG1301	[C] Energy production and conversion	Metabolism
43	V8K3H5	Formate hydrogenlyase subunit	COG3260	[C] Energy production and conversion	Metabolism

		7			
44	V8K1L5	Formate hydrogenlyase subunit 2	COG1142	[C] Energy production and conversion	Metabolism
45	V8KF75	Dioxygenase subunit beta	COG1018	[C] Energy production and conversion	Metabolism
46	V8KMG1	C4-dicarboxylate ABC transporter	COG3069	[C] Energy production and conversion	Metabolism
47	V8K398	(4Fe-4S)-binding protein	COG1143	[C] Energy production and conversion	Metabolism
48	V8KT59	Transporter	COG0531	[E] Amino acid transport and metabolism	Metabolism
49	V8JYJ9	Peptide ABC transporter substrate-binding protein	COG0747	[E] Amino acid transport and metabolism	Metabolism
50	V8KV94	Methylaspartate ammonia-lyase	COG3799	[E] Amino acid transport and metabolism	Metabolism
51	V8KU99	Glutamate mutase epsilon subunit	COG4865	[E] Amino acid transport and metabolism	Metabolism
52	V8KFK2	Uncharacterized protein	COG0366	[G] Carbohydrate transport and metabolism	Metabolism
53	V8KJM8	PTS alpha-glucide transporter subunit IIBC	COG1263	[G] Carbohydrate transport and metabolism	Metabolism
54	V8K583	MFS transporter	COG2814	[G] Carbohydrate transport and metabolism	Metabolism
55	V8KJR6	6-phpho-alpha-glucidase	COG1486	[G] Carbohydrate transport and metabolism	Metabolism
56	V8KIC3	Probable UbiX-like flavin prenyltransferase	COG0163	[H] Coenzyme transport and metabolism	Metabolism
57	V8KI60	Phenolic acid decarboxylase subunit C	COG0043	[H] Coenzyme transport and metabolism	Metabolism

58	V8KV99	Cob(I)alamin adenolsyltransferase/cobinamide ATP-dependent adenolsyltransferase	COG2109	[H] Coenzyme transport and metabolism	Metabolism
59	V8KT09	Glutamate mutase sigma subunit	COG2185	[I] Lipid transport and metabolism	Metabolism
60	V8KFB6	3-phenylpropionate/cinnamic acid dioxygenase ferredoxin--NAD(+) reductase component	COG0446	[I] Lipid transport and metabolism	Metabolism
61	V8K8D7	Uncharacterized protein	COG1629	[P] Inorganic ion transport and metabolism	Metabolism
62	V8K4L2	Tautomerase PptA	COG1942	[Q] Secondary metabolites biosynthesis, transport and catabolism	Metabolism
63	V8KDZ8	3-phenylpropionate/cinnamic acid dioxygenase subunit beta	COG5517	[Q] Secondary metabolites biosynthesis, transport and catabolism	Metabolism
64	V8KGH6	Uncharacterized protein	COG3468	-	Multiple-Classes
65	V8K624	Type VI secretion system protein ImpI	COG3456	-	Multiple-Classes
66	V8K1M2	Transcriptional regulator	COG3604	-	Multiple-Classes
67	V8KD21	Tartrate dehydrogenase	COG0473	-	Multiple-Classes
68	V8JZ57	Peptide ABC transporter permease	COG0601	-	Multiple-Classes
69	V8K3H0	Oxidoreductase	COG0651	-	Multiple-Classes
70	V8K3F4	NADP-dependent oxidoreductase	COG2130	-	Multiple-Classes

71	V8K4J6	Methyl-accepting chemotaxis protein	COG0840	-	Multiple-Classes
72	V8KS91	Methyl-accepting chemotaxis protein	COG0840	-	Multiple-Classes
73	V8K1L8	Hydrogenase	COG0651	-	Multiple-Classes
74	V8K263	Hydantoin racemase	COG0651	-	Multiple-Classes
75	V8KEY2	Dioxygenase	COG4638	-	Multiple-Classes
76	V8K5Y2	Adhesin	COG5295	-	Multiple-Classes
77	V8KEC3	3-phenylpropionate/cinnamic acid dioxygenase subunit alpha	COG4638	-	Multiple-Classes
78	V8KGG9	3-phenylpropionate/cinnamic acid dioxygenase ferredoxin subunit	COG2146	-	Multiple-Classes
79	V8KG59	3-phenylpropionate-dihydrodiol/cinnamic acid-dihydrodiol dehydrogenase	COG1028	-	Multiple-Classes
80	V8KHE5	Invasion protein	COG0745	[K] Transcription, [T] Signal transduction mechanisms	Multiple-Classes
81	V8K6U7	Transcriptional regulator	COG2197	[T] Signal transduction mechanisms, [K] Transcription	Multiple-Classes
82	V8KL53	Uncharacterized protein	COG4804	[R] General function prediction only	Poorly Characterized
83	V8KH81	Multidrug ABC transporter ATP-binding protein	COG4178	[R] General function prediction only	Poorly Characterized

84	V8K6I9	Membrane protein	COG2391	[R] General function prediction only	Poorly Characterized
85	V8K9C1	Lyase	COG1413	[R] General function prediction only	Poorly Characterized
86	V8KIM2	Uncharacterized protein	COG5435	[S] Function unknown	Poorly Characterized
87	V8KB66	Uncharacterized protein	COG4453	[S] Function unknown	Poorly Characterized
88	V8KHN2	Uncharacterized protein	COG3111	[S] Function unknown	Poorly Characterized
89	V8KTZ8	Toxin RelE	COG4737	[S] Function unknown	Poorly Characterized
90	V8JV47	Membrane protein	COG3152	[S] Function unknown	Poorly Characterized
91	V8KN72	Uncharacterized protein	COG1357	[S] Function unknown	Poorly Characterized
92	V8KG33	Uncharacterized protein	-	-	-
93	V8KQH5	Uncharacterized protein	-	-	-
94	V8KAN6	Uncharacterized protein	-	-	-
95	V8KCJ2	Uncharacterized protein	-	-	-
96	V8KJU4	Uncharacterized protein	-	-	-
97	V8K5J4	Uncharacterized protein	-	-	-
98	V8KAC5	Uncharacterized protein	-	-	-
99	V8KU12	Uncharacterized protein	-	-	-
100	V8KSP2	Uncharacterized protein	-	-	-
101	V8K756	Uncharacterized protein	-	-	-
102	V8KVD3	Uncharacterized protein	-	-	-

103	V8KV88	Uncharacterized protein	-	-	-
104	V8JR73	Uncharacterized protein	-	-	-
105	V8JR03	Uncharacterized protein	-	-	-
106	V8K1W4	Uncharacterized protein	-	-	-
107	V8JVI1	Uncharacterized protein	-	-	-
108	V8KIP7	Uncharacterized protein	-	-	-
109	V8KJP2	Uncharacterized protein	-	-	-
110	V8KI19	Uncharacterized protein	-	-	-
111	V8K8B8	Uncharacterized protein	-	-	-
112	V8JX34	Uncharacterized protein	-	-	-
113	V8KHM0	Uncharacterized protein	-	-	-
114	V8K1A9	Uncharacterized protein	-	-	-
115	V8K832	Uncharacterized protein	-	-	-
116	V8KAI7	Uncharacterized protein	-	-	-
117	V8K997	Uncharacterized protein	-	-	-
118	V8KDD8	Uncharacterized protein	-	-	-
119	V8K7F9	Uncharacterized protein	-	-	-
120	V8JV31	Uncharacterized protein	-	-	-
121	V8K619	Uncharacterized protein	-	-	-

122	V8K629	Type VI secretion protein	-	-	-
123	V8KIH3	Type III effector protein	-	-	-
124	V8KHG4	Type III effector protein	-	-	-
125	V8KHR0	Type III effector	-	-	-
126	V8KJG9	Type III effector	-	-	-
127	V8KJQ2	Type III cell invasion protein SipB	-	-	-
128	V8JSN6	Toxin YafO	-	-	-
129	V8JLC1	Toxin RelE	-	-	-
130	V8KID8	SsRNA endonuclease	-	-	-
131	V8K633	Membrane protein	-	-	-
132	V8K1E3	Membrane protein	-	-	-
133	V8KDP0	Membrane protein	-	-	-
134	V8K1U4	Lipoprotein	-	-	-
135	V8KTK9	Glutamate mutase	-	-	-
136	V8K149	Formate hydrogenlyase maturation protein HycH	-	-	-
137	V8KFV6	DNA damage-indicible protein DnaD	-	-	-
138	V8K463	Deubiquitinase	-	-	-

139	V8KT14	Acyl-CoA synthetase	-	-	-
140	V8KCI2	Aankyrin	-	-	-
141	V8KH92	4-hydroxybenzoate decarboxylase	-	-	-
142	V8KVD8	3-methylaspartate ammonia-lyase	-	-	-

DISCUSSION

The enormous diversity of *E. coli* has made it the subject of innumerable comparative and phylogenetic studies in order to determine the mechanisms by which each of the subgroups/lineages has diversified and specialized. The recent emergence of the multi-drug resistant *E. coli* lineages (ST131, ST38, ST405 and ST648) as the globally dominant strains, isolated from extra intestinal infections provides a new and important opportunity to understand how *E. coli* evolves and diversifies(Pitout 2012). Although the genomic architecture of ST131 has been widely studied, this is a first report on a genome based comparison of ST131 strains together with three emerging ESBL producing ExPEC lineages; ST38, ST405 and ST648. Since *E. coli* is a very versatile organism, longitudinal studies on its genomics is very essential to understand the differences in the genomic coordinates and to understand its evolution into different lineages. A dataset of 5 in-house sequenced strains along with 134 strains from the public databases belonging to these four STs were analyzed.

Clonality within the STs as depicted from the phylogenetic analysis (Fig 1) suggests that these STs differ at the genetic level though they are all associated with ExPEC infections. Plasmid analysis has revealed that IncF plasmid is not just successful in ST131 as earlier reported but also were majorly found in the remaining STs considered in this study (Petty et al. 2014; Lanza et al. 2014; Nicolas-Chanoine et al. 2014) (Table 3). The pMLST profile F2:A1:B- and F1:A2:B20 which were commonly associated with IncF of ST131 strains (Petty et al. 2014) were not seen in any of the non-ST131 strains across the dataset. Together with IncF, an intermixture of other plasmid classes such as IncHI2, IncI1, IncN, IncP, IncQ, IncB/O/K/Z, IncX1, IncX3, IncX4 and IncY were found among all the STs to a lesser extent. Similar to plasmid profiles, the antimicrobial resistance profiles as inferred from both genomic analysis of 139 strains and antimicrobial susceptibility assays on 5 in-house strains revealed that overall antimicrobial resistance was common and differed insignificantly

between the four STs (Fig 2). Presence of ESBL genes; *bla*$_{CTX-M}$, *bla*$_{TEM}$, and *bla*$_{SHV}$ together with several other resistance genes belonging to different classes such as efflux pumps, pump regulators, target in activators etc. , support the MDR nature of these STs (Fig 2).

Presence of virulence genes such as *hra* (heat resistant agglutin), *kpsMIII* group 3 capsule, *ompT* (outer membrane protease T), *malx* (pathogenicity island marker) were found to be consistent with the previous reports on *in vitro* virulence profiles of these four STs (Van der Bij et al. 2012) (Fig 3A). Presence of genes related to adherence and invasion among all the strains indicated comparable adherence and invasion capabilities among four STs. This was further evident from adhesion and invasion assays on in-house strains (Fig 3B, C). Another important observation was the presence of several serine proteases, auto-transporters and UPEC specific virulence genes exclusively in the strains of ST131, which is an obvious reason for the predominance of ST131 strains in extra-intestinal infections than any other single sequence type.

Mobile genetic elements (MGEs) are usually associated with virulence factors (Totsika et al. 2011; Brian M Forde, M. Phan, et al. 2015). BRIG analysis of MGEs reported earlier in EC958(Totsika et al. 2011) revealed that most of these regions were missing either completely or partially in ST38, ST405 and ST648 (Fig 4A). Earlier *in silico* studies with ST131 strains (Petty et al. 2014) revealed that the presence of these regions were one of the major causes behind the diversity within the ST131 lineage. Also, MTases such as EcoMIII (EC958_0425), M.EcoMIV (EC958_1101), M.EcoMVII (EC958_4083), M1.EcoMI /M2.EcoMI (EC958_0008/EC958_0009) which are associated *GI-thrW*, *Phi-2*, *GI-selC* and *GI-leuX*, respectively were also found to be missing in strains lacking these MGEs (Fig 4B). Three of the above mentioned MTases are part of restriction modification systems (R-M systems) which are known to act as barriers for DNA exchange among inter as well as intra

species. It would be possible that these MTases have a prominent role in segregating and shaping up the gene pool of ST131(Brian M. Forde et al. 2015).

Serum resistance is one of the major survival mechanisms adapted by the pathogenic *E. coli* to survive in the bloodstream of the host. BLASTp analysis of the 139 strains against 56 essential serum resistance genes of EC958 have revealed that 12 of these genes were missing from all 40 strains belonging to the non-ST131 but present in ST131 strains (Fig 5A). Among these missing essential serum resistance genes, EC958_0460 (*lyxA*) and EC958_0461 (*lyxR*) were reported to have a role in regulating O-antigen chain length (Phan et al. 2013). Previous studies have also shown that EC958_0461 (*lyxR*) is linked to the suppression of nitrosoactive stress response as well as important for intracellular survival within macrophages (Bateman & Seed 2012). EC958_1112 is one of the four glucose transferases found in the O-antigen and the operons, EC958_4029 (*waaL*), EC958_4030 (*waaU /waaK*) and EC958_4035 (*waaB/rfaB*) were linked to the LPS core biosynthesis. Mutants of all these genes except for EC958_1114 were reported to be relatively more sensitive to the serum when compared to the wild types (Phan et al. 2013). However, all the five in-house strains representing the four STs exhibited similar resistance capabilities against the bactericidal activity of human serum, this may be due to the presence of different protectins in strains of other STs (Fig 5B).

OrthoMCL (Li et al. 2003) analysis of core proteins belonging to each of these four STs revealed that around 142 proteins were completely missing from all the 99 strains of ST131 (Table 4) but were present in all 40 strains from ST38, ST405 and ST648. COG functional annotation of these proteins revealed that they represent a wide variety of COG functional classes which include membrane proteins, transcriptional regulators, transcriptional chaperones, lyases, hydrogenases, etc. Also two toxin-antitoxin modules *mazE-mazF* and *yafO-yafN* were found to be missing from the ST131 strains. While *mazE-mazF* is known to

cause programmed cell death under stress conditions such as nutrient starvation, addition of antibiotics, etc.; *yafO-yafN* is involved in SOS response triggered by DNA damage (Yamaguchi & Inouye 2011). Another important protein diguanylate cyclase *yddV* involved in c-di-GMP signaling was also among these 142 genes that were absent in the ST131 strains analyzed. It was reported earlier that this protein was found to be either truncated or completely deleted in many ExPEC strains. It was hypothesized that losing this protein would negatively regulate c-di-GMP which may in turn enhance motility and also down express curli fibers to avoid local immune defenses (Povolotsky & Hengge 2016).

In summary, we provide novel genomic insights on population structure of *E. coli* with respect to four emerging sequence types; ST131, ST 38, ST 408, ST 648. Such a comparative analysis has not been reported yet. Our data suggests that although the *E. coli* strains belonging to four STs demonstrate similar functional propensities, they are evolving in a phylogenetically distinct manner from each other with no significant overlaps in genomic coordinates. Particularly, the emergence of ST131 lineage is marked by drastic reduction in the number of core genes belonging to different functional classes. Our genome sequence data opens up avenues of research to further strengthen our understanding on the evolution and interaction of different ExPEC lineages.

METHODOLOGY

Strain selection and Whole genome sequencing

Five strains NA023, NA081, NA090, NA101 and NA112 belonging to ST38, ST405, ST648 and ST131 from the collection as previously reported (Ranjan et al. 2016) were sequenced using Illumina Miseq sequencer with an insert size of 400 to 500 bp.

Assembly and Annotation of in-house strains

High quality reads obtained from NGS QC Toolkit (Patel & Jain 2012) were used for *denovo* assembly of contigs using Velvet (Zerbino & Birney 2008). Thus generated contigs were then ordered and scaffolded using C-L-Authenticator (Shaik et al. 2016) and the final draft genomes were obtained by merging these scaffolds using a series of N's. These draft genomes were then submitted to the RAST server (Aziz et al. 2008) for annotation, and the genome statistics from the resulting file were extracted using ARTEMIS (Rutherford et al. 2000). The Sequence Type (ST) of each of these strains was determined by submitting the *denovo* contigs to https://cge.cbs.dtu.dk/services/MLST/.

Data-set collection for Comparative genomics

Collection of 99 ST131 strains: Apart from the two strains from the current study, 97 whole-genome sequences including NA097, NA114, JJ1886, and EC958 from the NCBI and other ST131 genomes described by Petty et al (Petty et al. 2014) were considered for the study.

Collection of ST38, ST405 and ST648 strains: As there are not many studies performed at the genomic level of these ST's, an in-house pipeline was devised to extract the genomic data from NCBI. As on 08th Jun'15, genomic data (both complete and incomplete) belonging to 3,051 *E. coli* strains was downloaded and in case of strains with incomplete data, contigs >500 in number were eliminated. In case of each genome, in-house scripts were used to extract the sequences of the seven housekeeping genes (*fum, adk, gyr, icd, pur, mdh* and *rec*) and a ST number was assigned based on the combination of allelic numbers of these genes (supported by the database data downloaded from http://mlst.warwick.ac.uk/mlst/dbs/Ecoli/Downloads_HTML). Amongst all the ST's, 12 strains from ST38, 10 strains from ST405 and 18 strains from ST648 were separated out. The Sequence Type (ST) of each of these strains was further verified by submitting the contigs to https://cge.cbs.dtu.dk/services/MLST/. Information about all the strains used in this study is provided in the supplementary table 3.

Phylogenomic analysis

A core-genome-based phylogenetic tree of these 139 strains belonging to all four STs was constructed using Harvest (Treangen et al. 2014) and the resulting tree was visualized using the Figtree (Rambaut 2009).

Detection of Plasmid classes and Resistance profiles

In case of each of these strains, BLASTn of the contigs/ whole genome was performed against the database data downloaded from Plasmid Finder (Carattoli et al. 2014). A positive hit (90% query coverage and 95% identity) indicated the presence of a particular plasmid class in respective strains.

For all of the genomes, protein sequences were predicted using GeneMarkS (Besemer et al. 2001). Resistance profiles of each of these strains was obtained by performing a BLASTp analysis of the predicted protein sequences against the database of Comprehensive Antibiotic Resistance Database (CARD). Only hits with 70% identity and 75% query coverage were given the status of being present. A heat map was generated using R to depict the positive hits (represented in black) in each of these genomes.

Determination of Virulence profile

Virulence profiles of each of these strains was obtained by performing a BLASTn analysis of 139 genomes against the database of selected VFDB (Chen et al. 2005) virulence genes. Only hits with 70% identity and query coverage either >= 50% or >=800bp were given the status of being present. Calculation of the frequency of each of these virulence genes at the ST level was calculated using the formula (Presence in no. of strains of ST / Total no. of strains in ST)*100. Frequencies of all the virulence genes in each of these STs is represented in the form of a heat map using R.

Genomic Islands, MTases and Serum Resistome analysis

BRIG analysis (fig 3a) was performed with the dataset containing 12, 10 and 18 strains belonging to ST38, ST405 and ST648 using Mobile Genetic Elements (MGEs) and MTases reported in the ST131 representative strains EC958(Brian M Forde, M.-D. Phan, et al. 2015; Totsika et al. 2011) as reference. Only 20 strains from ST131 dataset belonging to all three clades A, B and C were selected for this analysis.

A BLASTp analysis was performed to detect the presence of 56 essential serum resistance genes that were earlier reported in EC958(Phan et al. 2013). Positive hits were defined and represented in the form of a heat map as discusses earlier in the resistance section.

Comparative Core-Accessory genome analysis

Due to the differences in the sample sizes of each Sequence Type, core genes of individual Sequence Types were determined first. Core orthologous clusters of individual STs were identified using OrthoMCL(Li et al. 2003). The parameters for deciding orthologs such as identity and e-value cutoff were set to 70% and 0.00001, respectively. The genes with less than 50 amino acids were excluded from the analysis. The clusters that contained orthologs in at least 90% of the sample size constituted the core. Genes belonging to the core Orthologous groups of ST38, ST131, ST405 and ST405 were extracted from one respective representative strains.

Core genes (including recent paralogs) from the orthologous clusters of the four STs were then considered for further inter-ST core-accessory genome analysis. To identify genes that were specifically present or missing in ST131, orthologous clusters of all four ST-core genes were once again determined using OrthoMCL. Orthologous clusters that had missing genes only from ST131 core were further analyzed at the complete dataset level using BLASTp as described above. Final list of genes that were present in all 40 strains belonging to ST38,

ST405 and ST648 but missing from all 99 strains of ST131 were identified. COG classification and their functional annotations of all the genes of interest were determined using batch CD-search(Marchler-Bauer & Bryant 2004) against the COG database.

In vitro phenotypic characterization of the strains

All the five representative strains of the four different sequence types were characterized for the phenotypic properties such as resistance to 18 different antibiotics, ESBL production, adhesion and invasion capabilities towards human bladder epithelial (T24) cell lines and ability to resist bactericidal activity of human serum. Resistance towards antimicrobials was determined by standard Kirby Bauer disk diffusion technique using the ICOSA UTI strips (Himedia, India) as per the guidelines of CLSI and ESBL production was tested by double disk synergy test (Hussain et al. 2014). The adhesion and invasion assay was performed as per the standard protocols described previously(Nandanwar et al. 2014). Briefly, T24 cells were grown to monolayer in 24 well plates and infected at MOI 10 with different strains in triplicates. After 3 hours of incubation at 37°C in CO_2 incubator the cells were washed thrice with 1X PBS and then lysed using 0.1% triton X-100. The lysates were collected and plated after serial dilutions. For invasion, after 3 hrs of incubation cells were additionally incubated for 1.5 hrs with medium containing 100ug/ml of gentamicin. After incubation cells were lysed as in adhesion step and plated. After incubation CFU was determined for all the strains. The serum resistance assay was performed in sterile 50% human serum (Pan Biotech, Germany) for three hours as per the protocols described elsewhere(Nandanwar et al. 2014). The growth in serum was observed in terms of CFU. The adhesion and invasion assay was performed three times while serum resistance assay was performed twice in triplicates. For serum resistance and adhesion assay Mann-Whitney test was employed to get p values against negative control (Dh5-alpha) while for invasion assay p values were obtained by employing Wilcoxon signed rank test against negative control.

REFERENCES

Ahmed, N., 2009. A flood of microbial genomes - Do we need more? *PLoS ONE*, 4(6).

Alikhan, N.F. et al., 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics*, 12, p.402.

Altschul, S.F. et al., 1990. Basic local alignment search tool. *J Mol Biol*, 215(3), pp.403–10.

Avasthi, T.S. et al., 2011. Genome of multidrug-resistant uropathogenic Escherichia coli strain NA114 from India. *Journal of Bacteriology*, 193(16), pp.4272–4273.

Aziz, R.K. et al., 2008. The RAST Server: rapid annotations using subsystems technology. *BMC genomics*, 9(1), p.75.

Bateman, S.L. & Seed, P.C., 2012. Epigenetic regulation of the nitrosative stress response and intracellular macrophage survival by extraintestinal pathogenic Escherichia coli. *Molecular Microbiology*, 83(5), pp.908–925.

Besemer, J., Lomsadze, A. & Borodovsky, M., 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic acids research*, 29(12), pp.2607–2618.

Van der Bij, A.K. et al., 2012. The presence of genes encoding for different virulence factors in clonally related Escherichia coli that produce CTX-Ms. *Diagnostic Microbiology and Infectious Disease*, 72(4), pp.297–302. Available at: http://dx.doi.org/10.1016/j.diagmicrobio.2011.12.011.

Van Der Bij, A.K. et al., 2011. Clinical and molecular characteristics of extended-spectrum-β-lactamase-producing Escherichia coli causing bacteremia in the Rotterdam Area, Netherlands. *Antimicrobial Agents and Chemotherapy*, 55(7), pp.3576–3578.

Carattoli, A. et al., 2014. In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrobial Agents and Chemotherapy*, 58(7), pp.3895–3903.

Chen, L. et al., 2005. VFDB: A reference database for bacterial virulence factors. *Nucleic Acids Research*, 33(DATABASE ISS.).

Coque, T.M. et al., 2008. Dissemination of clonally related Escherichia coli strains expressing extended-spectrum ??-lactamase CTX-M-15. *Emerging Infectious Diseases*, 14(2), pp.195–200.

Cortés, P. et al., 2010. Isolation and Characterization of Potentially Pathogenic Antimicrobial-Resistant Escherichia coli Strains from Chicken and Pig Farms in Spain. *Applied and Environmental Microbiology*, 76(9), pp.2799–2805.

Dobrindt, U., 2005. (Patho-)Genomics of Escherichia coli. *International Journal of Medical Microbiology*, 295(6–7), pp.357–371.

Forde, B.M. et al., 2015. Lineage-specific methyltransferases define the methylome of the globally disseminated escherichia coli ST131 clone. *mBio*, 6(6).

Forde, B.M., Phan, M., et al., 2015. Lineage-Specific Methyltransferases Define the Methylome of the Globally Disseminated Escherichia coli ST131 Clone. , 6(6), pp.1–12.

Forde, B.M., Phan, M.-D., et al., 2015. Lineage-Specific Methyltransferases Define the Methylome of the Globally Disseminated Escherichia coli ST131 Clone. *mBio*, 6(6), pp.e01602-15.

Hornsey, M., Phee, L. & Wareham, D.W., 2011. A novel variant, NDM-5, of the New Delhi metallo-??-lactamase in a multidrug-resistant Escherichia coli ST648 isolate recovered from a patient in the United Kingdom. *Antimicrobial Agents and Chemotherapy*, 55(12), pp.5952–5954.

Hussain, A. et al., 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. *Antimicrobial Agents and Chemotherapy*, 58(12), pp.7240–7249.

Hussain, A. et al., 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-??- lactamase-producing lineage. *Antimicrobial Agents and Chemotherapy*, 56(12), pp.6358–6365.

Johnson, J.R. et al., 2010. Escherichia coli sequence type ST131 as the major cause of serious multidrug-resistant E. coli infections in the United States. *Clinical Infectious Diseases*, 51(3),

pp.286–294. Available at: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=20572763&retmode=ref&cmd=prlinks%5Cnpapers3://publication/doi/10.1086/653932.

Jones, G.L. et al., 2008. Prevalence and distribution of plasmid-mediated quinolone resistance genes in clinical isolates of Escherichia coli lacking extended-spectrum beta-lactamases. *The Journal of antimicrobial chemotherapy*, 62(September), pp.1245–1251.

Kaper, J.B., Nataro, J.P. & Mobley, H.L.T., 2004. Pathogenic Escherichia coli. *Nature Reviews Microbiology*, 2(2), pp.123–140.

Kim, J. et al., 2011. Characterization of IncF plasmids carrying the blaCTX-M-14 gene in clinical isolates of Escherichia coli from Korea. *Journal of Antimicrobial Chemotherapy*, 66(6), pp.1263–1268.

Koonin, E. V, 2002. The Clusters of Orthologous Groups (COGs) Database: Phylogenetic Classification of Proteins from Complete Genomes. *Antenna*, pp.1–6. Available at: http://www.ncbi.nlm.nih.gov/books/NBK21090/.

Lanza, V.F. et al., 2014. Plasmid Flux in Escherichia coli ST131 Sublineages, Analyzed by Plasmid Constellation Network (PLACNET), a New Method for Plasmid Reconstruction from Whole Genome Sequences. *PLoS Genetics*, 10(12).

Li, L., Stoeckert, C.J. & Roos, D.S., 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9), pp.2178–2189.

Marchler-Bauer, A. & Bryant, S.H., 2004. CD-Search: Protein domain annotations on the fly. *Nucleic Acids Research*, 32(WEB SERVER ISS.).

McArthur, A.G. et al., 2013. The comprehensive antibiotic resistance database. *Antimicrobial Agents and Chemotherapy*, 57(7), pp.3348–3357.

Mihaila, L. et al., 2010. Probable intrafamily transmission of a highly virulent CTX-M-3-producing Escherichia coli belonging to the emerging phylogenetic subgroup D2 O102-ST405 clone.

Journal of Antimicrobial Chemotherapy, 65(7), pp.1537–1539.

Mshana, S.E. et al., 2011. Multiple ST clonal complexes, with a predominance of ST131, of Escherichia coli harbouring blaCTX-M-15 in a tertiary hospital in Tanzania. *Clinical Microbiology and Infection*, 17(8), pp.1279–1282.

Mushtaq, S. et al., 2011. Phylogenetic diversity of Escherichia coli strains producing NDM-type carbapenemases. *Journal of Antimicrobial Chemotherapy*, 66(9), pp.2002–2005.

Nandanwar, N. et al., 2014. Extraintestinal pathogenic Escherichia coli (ExPEC) of human and avian origin belonging to sequence type complex 95 (STC95) portray indistinguishable virulence features. *International Journal of Medical Microbiology*, 304(7), pp.835–842.

Nicolas-Chanoine, M.H. et al., 2008. Intercontinental emergence of Escherichia coli clone O25:H4-ST131 producing CTX-M-15. *Journal of Antimicrobial Chemotherapy*, 61(2), pp.273–281.

Nicolas-Chanoine, M.H., Bertrand, X. & Madec, J.Y., 2014. Escherichia coli st131, an intriguing clonal group. *Clinical Microbiology Reviews*, 27(3), pp.543–574.

Patel, R.K. & Jain, M., 2012. NGS QC toolkit: A toolkit for quality control of next generation sequencing data. *PLoS ONE*, 7(2).

Peirano, G. et al., 2013. Virulence potential and adherence properties of Escherichia coli that produce CTX-M and NDM ??-lactamases. *Journal of Medical Microbiology*, 62(PART4), pp.525–530.

Peirano, G. & Pitout, J.D.D., 2010. Molecular epidemiology of Escherichia coli producing CTX-M ??-lactamases: the worldwide emergence of clone ST131 O25:H4. *International Journal of Antimicrobial Agents*, 35(4), pp.316–321.

Petty, N.K. et al., 2014. Global dissemination of a multidrug resistant Escherichia coli clone. *Proceedings of the National Academy of Sciences of the United States of America*, 111(15), pp.5694–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3992628&tool=pmcentrez&rendertype=abstract.

Phan, M.D. et al., 2015. Molecular characterization of a multidrug resistance IncF plasmid from the globally disseminated escherichia coli ST131 clone. *PLoS ONE*, 10(4), pp.1–17.

Phan, M.D. et al., 2013. The Serum Resistome of a Globally Disseminated Multidrug Resistant Uropathogenic Escherichia coli Clone. *PLoS Genetics*, 9(10).

Pitout, J.D.D., 2012. Extraintestinal pathogenic Escherichia coli: A combination of virulence with antibiotic resistance. *Frontiers in Microbiology*, 3(JAN), pp.1–7.

Poirel, L. et al., 2011. Emergence of OXA-48-Producing Escherichia coli Clone ST38 in France. *Antimicrobial Agents and Chemotherapy*, 55(10), pp.4937–4938.

Povolotsky, T.L. & Hengge, R., 2016. Genome-based comparison of cyclic di-GMP signaling in pathogenic and commensal Escherichia coli strains. *Journal of Bacteriology*, 198(1), pp.111–126.

Rambaut, A., 2009. FigTree, a graphical viewer of phylogenetic trees. *Institute of Evolutionary Biology University of Edinburgh*.

Ranjan, A. et al., 2015. Genomic and functional portrait of a highly virulent, CTX-M-15-producing H30-Rx subclone of Escherichia coli sequence type 131. *Antimicrobial Agents and Chemotherapy*, 59(10), pp.6087–6095.

Ranjan, A. et al., 2016. Molecular epidemiology and genome dynamics of New Delhi metallo-beta-lactamase (NDM) producing extraintestinal pathogenic E. coli (ExPEC) strains from India. *Antimicrobial agents and chemotherapy*, (September). Available at: http://www.ncbi.nlm.nih.gov/pubmed/27600040.

Rutherford, K. et al., 2000. Artemis: sequence visualization and annotation. *Bioinformatics (Oxford, England)*, 16(10), pp.944–945.

Schembri, M.A. et al., 2015. Molecular Characterization of the Multidrug Resistant Escherichia coli ST131 Clone. *Pathogens (Basel, Switzerland)*, 4(3), pp.422–30. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26131613%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4584265.

Shaik, S. et al., 2016. Contig-Layout-Authenticator (CLA): A Combinatorial Approach to Ordering and Scaffolding of Bacterial Contigs for Comparative Genomics and Molecular Epidemiology. *Plos One*, 11(6), p.e0155459.

Smet, A. et al., 2010. Characterization of extended-spectrum beta-lactamases produced by Escherichia coli isolated from hospitalized and nonhospitalized patients: emergence of CTX-M-15-producing strains causing urinary tract infections. *Microbial drug resistance (Larchmont, N.Y.)*, 16(2), pp.129–134.

Sullivan, C.B., Diggle, M.A. & Clarke, S.C., 2005. Multilocus Sequence Typing: Data Analysis in Clinical Microbiology and Public Health. *Molecular Biotechnology*, 29(3), pp.245–254.

Suzuki, S. et al., 2009. Change in the prevalence of extended-spectrum-??-lactamase-producing Escherichia coli in Japan by clonal spread. *Journal of Antimicrobial Chemotherapy*, 63(1), pp.72–79.

Totsika, M. et al., 2011. Insights into a multidrug resistant Escherichia coli pathogen of the globally disseminated ST131 lineage: Genome analysis and virulence mechanisms. *PLoS ONE*, 6(10).

Treangen, T.J. et al., 2014. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome biology*, 15(11), p.524.

Tumbarello, M. et al., 2007. Predictors of mortality in patients with bloodstream infections caused by extended-spectrum-??-lactamase-producing Enterobacteriaceae: Importance of inadequate initial antimicrobial treatment. *Antimicrobial Agents and Chemotherapy*, 51(6), pp.1987–1994.

Vimont, S. et al., 2012. The CTX-M-15-Producing Escherichia coli Clone O25b: H4-ST131 Has High Intestine Colonization and Urinary Tract Infection Abilities. *PLoS ONE*, 7(9).

Wirth, T. et al., 2006. Sex and virulence in Escherichia coli: An evolutionary perspective. *Molecular Microbiology*, 60(5), pp.1136–1151.

Yamaguchi, Y. & Inouye, M., 2011. Regulation of growth and death in Escherichia coli by toxin-antitoxin systems. *Nat Rev Microbiol*, 9(11), pp.779–790. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21927020.

Yamamoto, T. et al., 2011. Emergence of NDM-1-positive capsulated Escherichia coli with high resistance to serum killing in Japan. *Journal of Infection and Chemotherapy*, 17(3), pp.435–439.

Zerbino, D.R. & Birney, E., 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5), pp.821–829.

Zong, Z. & Yu, R., 2010. Escherichia coli carrying the blaCTX-M-15 gene of ST648. *Journal of Medical Microbiology*, 59(12), pp.1536–1537.

CHAPTER 4: (Objective 3)

To analyze genomes of *Escherichia coli* isolates from Indian retail poultry and different human *E. coli* pathotypes in order to evaluate their risk in zoonotic transmission of antimicrobial resistance

---

This chapter is in communication as:

Risk of zoonotic transmission of antimicrobial resistant *Escherichia coli* from commercial broiler and free-range retail chicken in India

Introduction

The rapid global rise of *Escherichia coli* infections that are resistant to therapeutically important antimicrobials, including front-line drugs like cephalosporins and fluoroquinolone, is of serious concern, as it hampers treatment of infections leading to significant morbidity, mortality, medical costs as well as production losses in livestock (de Been et al. 2014). *E. coli* is responsible for infections that are amongst the most significant diseases in humans and animals, which can be nosocomial-acquired and or community-acquired (Jadhav et al. 2011). Being a part of endogenous flora, *E. coli* can easily acquire resistance to antimicrobials consumed by humans and animals (van den Bogaard et al. 2001).

Poultry are recognized as an important source for dissemination of antimicrobial resistant *E. coli* in community and environment (van den Bogaard et al. 2001). Pathogenic *E. coli* in poultry are a direct threat to both poultry industry and human health as they may result in hard-to-treat infections (Horn et al. 2012). Extraintestinal pathogenic *E. coli* (ExPEC), the causative agent of colibacillosis in chickens inflict severe losses due to morbidity and condemnations (Horn et al. 2012). ExPEC can also cause several extraintestinal diseases in humans, including urinary tract infections, neonatal meningitis, and sepsis (Hussain et al. 2012; Nandanwar et al. 2016). Apart from ExPEC, the poultry gut could also harbor other variants of intestinal pathogenic *E. coli* (Lutful Kabir 2010). Antimicrobial agents have been widely used in animal production in India, as recently suggested by a study carried out by Centre for Science and Environment, which detected high levels of antibiotic residues in broiler chicken samples from the national capital region of India, Delhi (www.cseindia.org). Such practices lead to high proportion of antibiotic resistant bacteria in their fecal flora. Consequently, meat at slaughtering operations can be extensively contaminated with fecal *E. coli* of poultry origin (Lutful Kabir 2010).

The ExPEC pathotype that is isolated from poultry with clinical signs of respiratory and systemic diseases is often termed as avian pathogenic *E. coli* (APEC) (Ewers et al. 2004). Although research has been mainly focused on infections caused by APEC pathotype, little is known about the reservoirs of these bacteria (Ewers et al. 2004). Also, many human and animal source ExPEC isolates share virulence genes and clonal backgrounds and the human health risk posed by such bacteria from poultry is still largely undefined (Dziva et al. 2013). Moreover, the pathotype APEC, itself is ill-defined (Collingwood et al. 2014). A study also suggests that one APEC strain showed high genome similarity to human enterotoxigenic *E. coli* (ETEC) (Dziva et al. 2013). Therefore, the question of whether fecal isolates of poultry serving as a reservoir for extraintestinal and intestinal infections in humans remains unanswered. Recent investigations compared *E. coli* from symptomatic poultry and human ExPEC by virulence genotyping, serotyping and *in vitro* assays (Mitchell et al. n.d.; Nandanwar et al. 2014). However the application of high-resolution, whole genome sequencing methods appears to be superior and is likely to provide accurate insights into the phylogenetic backgrounds of poultry *E. coli* and would highlight the similarities between isolates of different human pathotypes (de Been et al. 2014).

The presence of antimicrobial resistant bacteria in food has been attributed to the widespread use of antimicrobials in farming practices (Landers et al. 2012; Qumar et al. 2017). Currently few data are available regarding the contamination of retail foods with *E. coli*, especially those that are multi-resistant and pathogenic. Methods of livestock production differ in antibiotic usage practices, and this can influence the frequency of antibiotic-resistant bacteria on retail meats (Zhang et al. 2011). Consequently little is known about the frequency of antibiotic-resistant microorganisms in poultry that were raised by free-range farming. What is widely unknown, however, is the molecular characteristics of *E. coli* isolates from clinically healthy poultry at the whole genome level. Therefore, improved understanding of the origins and transmission of antimicrobial-resistant and pathogenic *E. coli* is required as an

important part of One Health strategy. In order to address these critical questions, we employed a wide range of genotyping and phenotyping methods to characterize *E. coli* recovered from poultry meats and carcasses, sampled from retail markets during 2015, in India and compared them with genomes of human *E. coli* pathotypes.

Results

Poultry E. coli isolates shared remarkable similarities with human and avian E. coli pathotypes

The general features of 10 in-house whole genome sequenced poultry *E. coli* isolates are shown in Table 1. We generated a phylogenetic tree of 10 in-house isolates to study their relationship with 40 publicly available genomes comprising of 20 human disease-associated *E. coli* (10 ExPEC and 10 *enteric* pathogens), 10 Avian pathogenic *E. coli* (APEC) and 10 *E. coli* genomes from healthy broiler chicken (Fig 1A). Despite genetic diversity, the phylogenetic tree was able to largely group isolates of different pathotypes with clade1 being dominant in enteric strains and clade 2 mainly comprised of ExPEC, APEC and poultry isolates. The out-group comprised of few APEC and poultry *E. coli* which may in fact be pure commensals in nature. Out of 10 poultry genomes analyzed, three were represented in clade 1, six genomes in clade 2 and one genome in the out-group. Overall, we observed that the poultry genomes demonstrated closer genetic associations with ExPEC than with human enteric pathogens.

Table 1: Genome features and assembly statistics of 10 whole genome sequenced poultry *E. coli* isolates

Genome	Accession Number	Isolation source	No. of Raw reads	Genome coverage (X)	Number of contigs >=500bp	N50 value	Sequence Type (ST)	G+C content (%)	Genome size (bp)	Coding %	No. of CDS	No. of rRNAs
NA1001	MKWV00000000	Broiler cecum	457192	21.9	232	79906	UNK	50.41	5230886	85.8	5198	10
NA1002	MKWW00000000	Broiler cecum	1330724	69.7	187	123923	ST117	50.72	5256107	86.3	5227	16
NA1003	MKWX00000000	Broiler cecum	1430266	73.8	257	86543	ST1640	50.84	5461579	86.1	5444	13
NA1004	MKWY00000000	Broiler cecum	1580522	80.3	167	165460	UNK	50.48	5291577	86.3	5229	18
NAEC1	MLAZ00000000	Broiler meat	2197830	60.8	138	163214	ST131	50.73	5152191	87.4	5037	13
NAEC2	MLBA00000000	Broiler cecum	1779900	61.7	222	88992	ST115	50.36	5507065	85.9	5398	7
NAEC3	MLBB00000000	Broiler meat	1446944	52.8	181	111651	UNK	50.47	5144964	85.9	5084	10
NAEC4	MLBC00000000	Broiler meat	1806644	62.1	276	94921	ST-115	50.37	5786021	85.4	5766	9
NAEC5	MLBD00000000	Broiler cecum	1189470	41.6	193	124307	ST-155	50.45	5195760	85.8	5025	14
NAEC6	MKWZ00000000	Broiler cecum	1377018	46.5	119	123677	ST-155	50.48	5095482	87	4948	12

UNK; Unknown sequence type

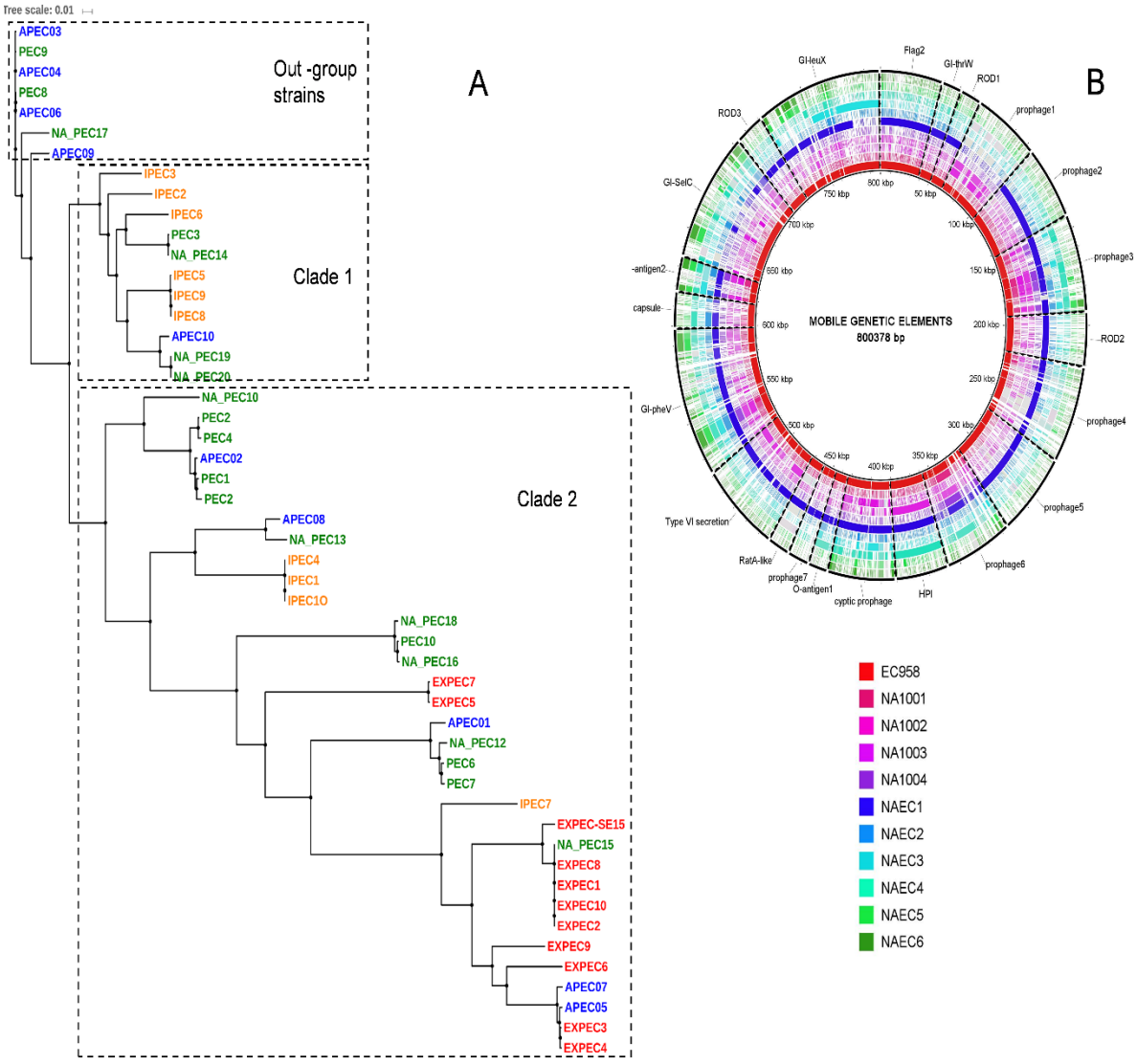


Figure 1: A. Phylogenetic tree of 50 *E. coli* strains: The core genome based consensus Maximum Likelihood phylogenetic tree of 10 in-house poultry *E. coli* isolates together with 40 other publicly available human and animal disease associated *E. coli* isolates generated after 1000 replicates/bootstraps using RAxML. The output of RAxML was visualized using iTOL tool (http://itol.embl.de/). Despite genetic diversity, the phylogenetic tree on the whole was able to diffusely group isolates belonging to different pathotypes. APEC (avian pathogenic *E. coli*), ExPEC (Extraintestinal pathogenic *E. coli*), IPEC (intestinal pathogenic *E. coli*), PEC (healthy poultry *E. coli*), NA_PEC (in-house poultry *E. coli*). B. Distribution of EC958 genomic islands (GIs) in 10 poultry *E. coli* isolates: the whole genome alignment of 10 Indian poultry *E. coli* isolates was generated using the concatenated sequence of mobile genetic elements of EC958 strain as reference in Blast Ring Image Generator (BRIG). Results show that EC958 GIs are well conserved in one ST131 poultry genome but partially or completely absent from the genomes of other nine strains.

BRIG analysis was done for 10 in-house poultry *E. coli* isolates to determine the status of 22 mobile genetic elements as reported in EC958 (Fig 1B). Out of 10 genomes, only one (NAEC1) poultry *E. coli* revealed presence of maximum number of complete mobile genetic elements, this is the same genome that was co-clustered with the ST131 clade. The genomes NA1002 (ST131) and NAEC4 (ST115) contained few complete mobile elements, while the rest (seven genomes) harbored only partial or no mobile elements.

In order to understand the genetic relationships of virulence and resistance of poultry *E. coli* in juxtaposition with the genetic landscapes of different human *E. coli* pathotypes, we compared virulence and resistance profiles of 10 sequenced poultry *E. coli* genotypes with 40 publicly *E. coli* genomes. Hierarchical clustering of resistance profiles (Fig 2B) revealed that the two sister clusters 1A and 1B represented isolates from all pathotypes including our five (50%) poultry genomes. However, cluster 2 was well marked and dominated by only poultry *E. coli* isolates that contained the rest of our five (50%) poultry genomes. The cluster obtained by the virulence gene profiles (Fig 2A) grouped strains into two categories; one composed of mixed strains (sister clusters 1A and 1B) and the other dominated with enteric isolates (cluster 2). All in-house poultry *E. coli* isolates were represented in the mixed clusters indicating closer genetic identity with other pathotypes. These observations reinforce the zoonotic potential of poultry *E. coli* in humans.

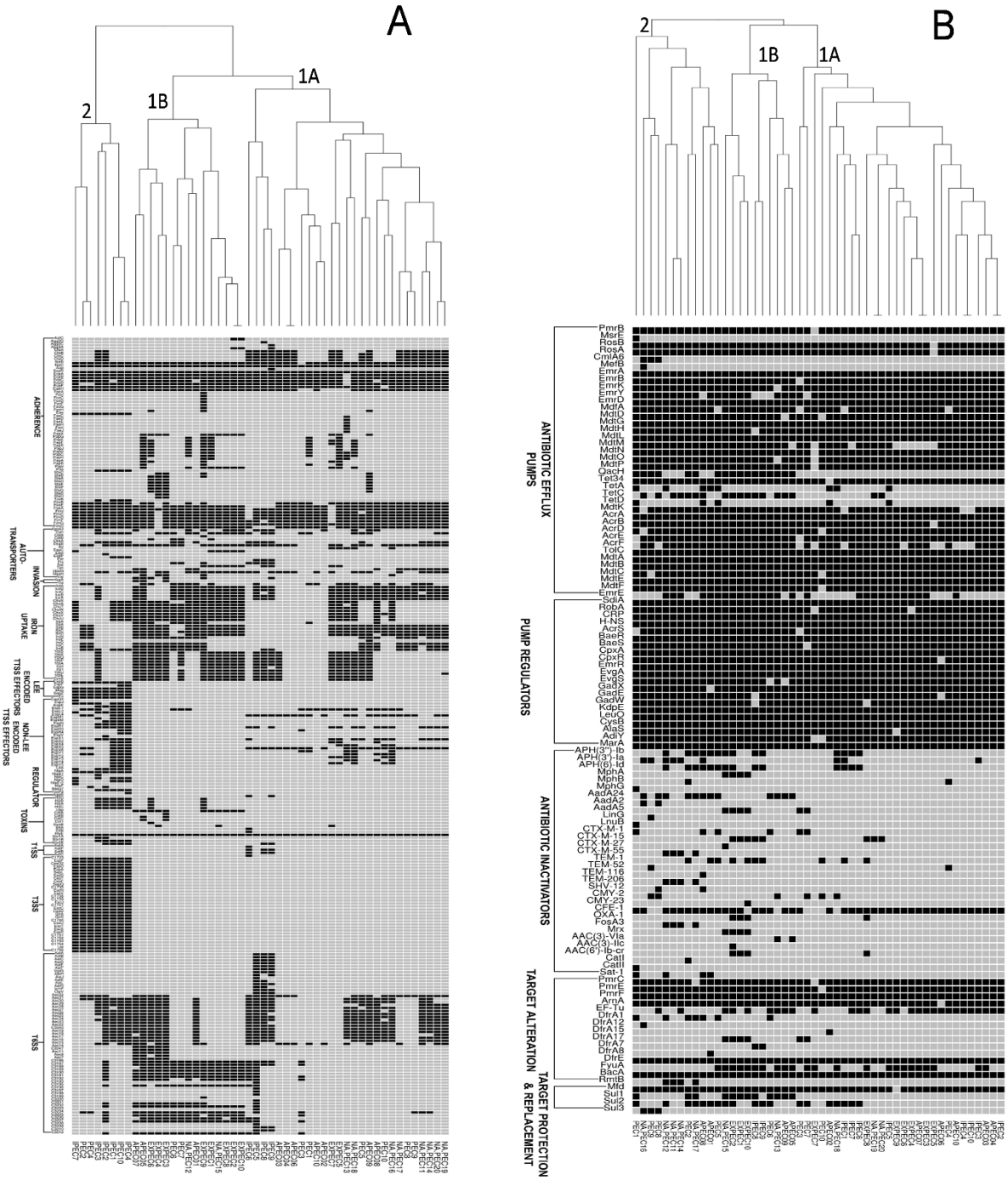


Figure 2: Gene cluster results for 50 *Escherichia coli* isolates: The presence (Black Square) and the absence (Grey Square) of virulence genes (A) and resistance genes (B) are represented in the image. Gene names are listed on the left. *E. coli* pathotypes are listed below the image. The cluster diagram showed that 10 sequenced poultry pan-genome virulence/resistance genes were more similar to extraintestinal pathogenic *E. coli* strains and poultry commensals compared to intestinal pathogenic *E. coli.* APEC (avian pathogenic *E. coli*), ExPEC (Extraintestinal pathogenic *E. coli*), IPEC (intestinal pathogenic *E. coli*), PEC (healthy poultry *E. coli*), NA_PEC (in-house poultry *E. coli*).

Discussion

In this study, we assessed *E. coli* isolates obtained from retail poultry (meat/carcass) from across the South and West central India (four states) by using WGS methods. Our findings provide evidence that a considerable fraction of raw retail poultry chicken were contaminated with antimicrobial-resistant and potentially pathogenic *E. coli*. This is particularly alarming for countries like India, given the high burden of diseases, emergence of resistance determinants, and the confluence of prevailing socio-economic, demographic and environmental factors (Dikid et al. 2013; Jadhav et al. 2011).

Pathogenic *E. coli* (ExPEC and diarrheagenic) is a leading cause of infections in both humans and poultry (EWERS et al. 2007). Antimicrobial usage in the poultry industry and an increasing trend of chicken meat consumption might contribute significantly to the emergence and dissemination of multiresistant and pathogenic *E. coli* variants, which could be a serious public health threat (Laxminarayan & Chaudhury 2016). Hence, constant surveillance of their epidemiology with respect to their genetic backgrounds, antimicrobial resistance patterns and specific virulence attributes is pertinent [21]. Our findings therefore have potential implications for public-health and antibiotic usage regulation (Johnson et al. 2005). Poultry industries uses antibiotics both for therapeutic purposes and for growth promotion (Singer & Hofacre 2006).

The phylogenetic tree, based on core genomes demonstrated that, the poultry *E. coli* isolates are not a homogeneous group, because some of the poultry *E. coli* isolates revealed similar genetic backgrounds with human ExPEC and enteric isolates, pointing to the role of chickens as reservoir of potentially pathogenic *E. coli*. This is also confirmed by the presence of two globally emerging pathogenic lineages of *E. coli*- ST131 (*H*30-Rx subclone) and ST117. The other commonly identified sequence types among the 10 poultry *E. coli* genomes

were -ST115, ST155 and ST1640 which are reported to be mostly associated with ESBL phenotypes (Börjesson et al. n.d.).

Comparison of genomic islands of EC958 with poultry *E. coli* genomes revealed that the genome of ST131 showed conservation of several complete genomic islands. However, two isolates one belonging to ST117 and another to ST115 also harbored few genomic islands. The conserved section of the genomic islands hint towards the evolution of ExPEC virulence among poultry *E. coli*. Virulence gene profiling revealed that the poultry *E. coli* are not homogenous group and showed mixed clustering. However, it provided better resolution between ExPEC and enteric pathotypes. Nonetheless, a majority of poultry *E. coli* isolates and human ExPEC were clustered together suggesting that they share remarkable similarities with human ExPEC pathotypes. The comparative genomic analysis suggests that the poultry *E. coli* isolates share closer genetic identity to human *E. coli* as regards to core genome phylogeny, antimicrobial genes and virulence gene content and thus represent potential zoonotic risks. The possibility that multi-drug resistant and or pathogenic *E. coli* could be transmitted from food sources like chicken meat raises serious public health concerns. Studies like this and more will provide data to the regulatory bodies in order to eliminate and prevent transmission of such organisms in food supply.

Material and methods

Specimen collection

Between February, 2015 to September, 2015, 22 poultry retail outlets were sampled from four cities each corresponding to three South Indian states and a West-central Indian state; Karnataka, Telangana, Andhra Pradesh and Maharashtra, respectively. A total of 75 poultry cecum, comprising broiler and free-range chicken (39 and 36, respectively) together with 45

raw meat, comprising broiler and free-range chicken samples (32 and 13, respectively) were obtained from retail poultry outlets. From each shop, multiple samples (different birds) were procured, resulting in a total of 120 samples. Samples arrived at the laboratory were stored at 4°C and processed within 24 hours.

Whole genome sequencing (WGS) and comparison with human E. coli pathotypes

Paired end sequencing of the 10 random ESBL producing poultry *E. coli* (7 cecum and 3 meat) isolates was carried out using Illumina MiSeq with an insert size of 400-500 bp. The accession numbers of 10 poultry *E. coli* genomes including their genomic features are represented in table 1. Paired end read data was filtered for high quality reads followed by *denovo* assembly using NGS QC Toolkit (v2.3.3) (Patel et al. 2012) and SPAdes Genome Assembler (v3.6.1) (Bankevich et al. 2012), respectively. Thus generated *denovo* contigs were ordered and scaffolded using Contig-Layout-Authenticator (Shaik et al. 2016; Avasthi et al. 2011). Final draft genomes were obtained by merging the scaffolds using a series of N's and then submitted to the RAST (Overbeek et al. 2014) server for annotation. The genome statistics were gleaned using ARTEMIS (Rutherford et al. 2000). The Sequence Type (ST) of each of these strains were determined by *in silico* MLST (Ranjan et al. 2015). Mobile genetic elements as reported in the ExPEC strain EC958 (Totsika et al. 2011) were used as reference to compare 10 poultry *E. coli* using BRIG (Alikhan et al. 2011). A core-genome-based phylogenetic tree of 50 strains; 10 in house poultry, 10 poultry from public sources, 10 APEC, 10 ExPEC and 10 enetric pathogens (Table S1) was constructed using Harvest (Treangen et al. 2014) and the resulting tree was visualized using Figtree. Virulence and Resistance profiles of these 50 strains were generated using the methods described in our earlier manuscript (Ranjan et al. 2016).

References

Alikhan, N.-F. et al., 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics*, 12(1), p.402.

Avasthi, T.S. et al., 2011. Genome of multidrug-resistant uropathogenic Escherichia coli strain NA114 from India. *Journal of bacteriology*, 193(16), pp.4272–3.

Bankevich, A. et al., 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19(5), pp.455–477.

de Been, M. et al., 2014. Dissemination of Cephalosporin Resistance Genes between Escherichia coli Strains from Farm Animals and Humans by Specific Plasmid Lineages. *PLoS Genetics*, 10(12).

van den Bogaard, A.E. et al., 2001. Antibiotic resistance of faecal Escherichia coli in poultry, poultry farmers and poultry slaughterers. *The Journal of antimicrobial chemotherapy*, 47(6), pp.763–71.

Börjesson, S. et al., Frequent Occurrence of Extended-Spectrum Beta-Lactamase-and Transferable AmpC Beta-Lactamase-Producing Escherichia coli on Domestic Chicken Meat in Sweden.

Chen, L.F. et al., 2014. Widespread dissemination of CTX-M-15 genotype extended-spectrum-β-lactamase-producing enterobacteriaceae among patients presenting to community hospitals in the southeastern United States. *Antimicrobial agents and chemotherapy*, 58(2), pp.1200–2.

Collingwood, C. et al., 2014. Is the Concept of Avian Pathogenic Escherichia coli as a Single Pathotype Fundamentally Flawed? *Frontiers in veterinary science*, 1, p.5.

Dikid, T. et al., 2013. Emerging & re-emerging infections in India: an overview. *The*

Indian journal of medical research, 138(1), pp.19–31.

Dziva, F. et al., 2013. Sequencing and functional annotation of avian pathogenic escherichia coli serogroup O78 strains reveal the evolution of E. coli lineages pathogenic for poultry via distinct mechanisms. *Infection and Immunity*, 81(3), pp.838–849.

Ewers, C. et al., 2004. Molecular epidemiology of avian pathogenic Escherichia coli (APEC) isolated from colisepticemia in poultry. *Veterinary Microbiology*, 104(1), pp.91–101.

EWERS, C. et al., 2007. Avian pathogenic, uropathogenic, and newborn meningitis-causing Escherichia coli: How closely related are they? *International Journal of Medical Microbiology*, 297(3), pp.163–176.

Firildak, G., Asan, A. & Goren, E., 2015. Chicken Carcasses Bacterial Concentration at Poultry Slaughtering Facilities. *Asian Journal of Biological Sciences*, 8(1), pp.16–29.

Franz, E. et al., 2015. Pathogenic Escherichia coli producing Extended-Spectrum β-Lactamases isolated from surface water and wastewater. *Scientific reports*, 5(SEPTEMBER), p.14372.

Horn, F. et al., 2012. Infections with avian pathogenic and fecal escherichia coli strains display similar lung histopathology and macrophage apoptosis. *PLoS ONE*, 7(7).

Hussain, A. et al., 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-??- lactamase-producing lineage. *Antimicrobial Agents and Chemotherapy*, 56(12), pp.6358–6365.

Jadhav, S. 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(3), p.e18063.

Johnson, J.R. et al., 2005. Antimicrobial-resistant and extraintestinal pathogenic Escherichia

coli in retail foods. *The Journal of infectious diseases*, 191(7), pp.1040–9.

Kobayashi, R.K.T. et al., 2011. EcoR Phylogenetic Analysis and Virulence Genotyping of Avian Pathogenic *Escherichia coli* Strains and *Escherichia coli* Isolates from Commercial Chicken Carcasses in Southern Brazil. *Foodborne Pathogens and Disease*, 8(5), pp.631–634.

Koga, V.L. et al., 2015. Comparison of Antibiotic Resistance and Virulence Factors among Escherichia coli Isolated from Conventional and Free-Range Poultry. *BioMed research international*, 2015, p.618752.

Landers, T.F. et al., 2012. A review of antibiotic use in food animals: perspective, policy, and potential. *Public health reports (Washington, D.C. : 1974)*, 127(1), pp.4–22.

Laxminarayan, R. & Chaudhury, R.R., 2016. Antibiotic Resistance in India: Drivers and Opportunities for Action. *PLoS medicine*, 13(3), p.e1001974.

Lutful Kabir, S.M., 2010. Avian colibacillosis and salmonellosis: a closer look at epidemiology, pathogenesis, diagnosis, control and public health concerns. *International journal of environmental research and public health*, 7(1), pp.89–114.

Mellata, M., 2013. Human and avian extraintestinal pathogenic Escherichia coli: infections, zoonotic risks, and antibiotic resistance trends. *Foodborne pathogens and disease*, 10(11), pp.916–32.

Mitchell, N.M. et al., Zoonotic Potential of Escherichia coli Isolates from Retail Chicken Meat Products and Eggs.

Nandanwar, N. et al., 2014. Extraintestinal pathogenic Escherichia coli (ExPEC) of human and avian origin belonging to sequence type complex 95 (STC95) portray indistinguishable virulence features. *International Journal of Medical Microbiology*, 304(7), pp.835–842.

Nandanwar, N. et al., 2016. Population structure and molecular epidemiology of human

clinical multi-drug resistant (MDR) Escherichia coli strains from Pune, India. *International Journal of Infectious Diseases*, 45, pp.343–344.

Overbeek, R. et al., 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic acids research*, 42(Database issue), pp.D206-14.

Patel, R.K. et al., 2012. NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data Z. Liu, ed. *PLoS ONE*, 7(2), p.e30619.

Qumar, 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. *Applied and environmental microbiology*, 83(1), pp.e02305-16.

Ranjan, A. et al., 2015. Genomic and functional portrait of a highly virulent, CTX-M-15-producing H30- Rx subclone of Escherichia coli sequence type (ST) 131. *Antimicrobial Agents and Chemotherapy*, p.AAC.01447-15.

Ranjan, A. et al., 2016. Molecular epidemiology and genome dynamics of New Delhi metallo-beta-lactamase (NDM) producing extraintestinal pathogenic E. coli (ExPEC) strains from India. *Antimicrobial agents and chemotherapy*, (September). Available at: http://www.ncbi.nlm.nih.gov/pubmed/27600040.

Rasschaert, G. et al., 2008. Contamination of carcasses with Salmonella during poultry slaughter. *Journal of food protection*, 71(1), pp.146–52.

Rodriguez-Siek, K.E. et al., 2005. Comparison of Escherichia coli isolates implicated in human urinary tract infection and avian colibacillosis. *Microbiology*, 151(6), pp.2097–2110.

Rutherford, K. et al., 2000. Artemis: sequence visualization and annotation. *Bioinformatics (Oxford, England)*, 16(10), pp.944–5.

Shaik, S. et al., 2016. Contig-Layout-Authenticator (CLA): A Combinatorial Approach to

Ordering and Scaffolding of Bacterial Contigs for Comparative Genomics and Molecular Epidemiology P. J. Biggs, ed. *PLOS ONE*, 11(6), p.e0155459.

Singer, R.S. & Hofacre, C.L., 2006. Potential Impacts of Antibiotic Use in Poultry Production. *Avian Diseases*, 50(2), pp.161–172.

Totsika, M. et al., 2011. Insights into a multidrug resistant Escherichia coli pathogen of the globally disseminated ST131 lineage: Genome analysis and virulence mechanisms. *PLoS ONE*, 6(10).

Treangen, T.J. et al., 2014. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome biology*, 15(11), p.524.

Zhang, J. et al., 2011. Contamination rates and antimicrobial resistance in Enterococcus spp., Escherichia coli, and Salmonella isolated from "no antibiotics added"-labeled chicken products. *Foodborne pathogens and disease*, 8(11), pp.1147–52.

CHAPTER 5

Summary and epilogue

ExPEC infections from *E. coli* have become a major source of concern worldwide mainly because of the dissemination of bacterial clones with greater virulence, drug resistance and zoonotic potential. This is majorly seen in developing countries such as India where there is a high disease burden due to the greater population density, sub-standard hygiene and socio-economic status. In order to understand the underlying genomic architecture and strategies adopted by ExPEC associated *E. coli* clones; several strains of *E. coli* were sequenced, assembled and analyzed as a part of this thesis.

Downstream protocol after sequencing whole genome included quality control where bad quality reads are excluded and then the good ones were assembled *denovo* into contigs. Further, ordering and scaffolding of these contigs in to a draft genome is equally challenging even in the case of prokaryotic genomes because of various genetic rearrangements occurring there in due to frequent insertion, deletion and substitution events, repeat motifs of different composition and lengths, insertion sequences, prophages and genomic islands etc. Most of the currently available tools either rely on a reference genome to order contigs at the whole genome level, or scaffold them based on the paired end read data. While using a reference genome could result in omission/exclusion of certain strain specific elements, the scaffolding methods struggle to resolve repetitive regions and limit ordering within the scaffolds. So, as a primary step, Contig-Layout-Authenticator (CLA) was developed in Perl which combines the benefits of individual approaches to minimize errors while generating a draft genome. The program runs in three stages. In the first stage, a reference based sort order is deduced by performing a BLAST analysis of the *de novo* contigs against the reference genome. The contigs are then sorted and reverse complemented with respect to the reference genome which marks the end of the stage one. In the second stage, a map file is created from alignment data obtained by aligning the paired end reads against the sorted and reverse

complemented contigs from stage one. Map file provides information about linking of start and end of all the contigs. Finally, in stage three, the reference based sort order from stage one is validated with the map file generated in the stage two. If the positions of any two consecutive contigs are not validated, it indicates that there must be missing information between them or they have to relocate to a different place. To do this, a network file was generated from the map file and then using all paths shortest path by Floyd Warshall was used to find the interlinking contigs and also relocating contigs if needed. This is followed by several iterations of shuffling contigs based on the network information. Finally, all the contigs whose contiguity is validated by the paired end information were only linked into scaffolds. Apart from ordering and scaffolding the contigs, this program also provides important information about the probable misassemblies and contaminations. After the development of Contig-Layout-Authenticator (CLA), its performance was well validated using WGS data from several bacterial species with different genetic backgrounds using simulated as well as real datasets.

Once a proper protocol for the assembly of the sequenced reads into draft genomes was standardized, whole genome based studies to compare different lineages of ExPEC *E. coli* were performed. Though ST131 lineage of *E. coli* has been widely studied, not many genome based studies of other recently emerging MDR lineages are reported. Genome based comparison of ST131 strains together with three emerging ESBL producing ExPEC lineages; ST38, ST405 and ST648 has been carried out to understand differences in their genomic coordinates. Analysis of 5 in-house sequenced strains along with 134 strains from the public databases belonging to these four STs has shown that ST131 harbored virulence genes mainly belonging to UPEC and a majority of these genes were missing from the other four STs. Also, the mobile genetic elements that were earlier reported to be associated with

ST131, were observed to be missing from ST38, ST405 and ST648. Though *in vivo* functional serum resistance assay has shown that all the strains were resistance to the bactericidal activity of the serum, *in silico* analysis revealed that few genes from 56 essential serum resistome genes from earlier studies were missing from ST38, ST405 and ST648. Core genome analysis of these 139 strains has revealed that a set of genes from different COG categories were missing from all 99 strains of ST131 but present in all the strains of other STs considered in this study. These findings have demonstrated that, although these STs have similar functional propensities they are evolving in a phylogenetically distinct manner from each other.

As a final objective of this dissertation, *E. coli* isolates obtained from retail poultry (meat/carcass) from across the South and West central India (four states) were analyzed by using WGS methods. As a part of this objective, 10 strains were sequenced, assembled and annotated. Further analysis of these 10 strains with 40 other strains from public databases has suggested that the poultry *E. coli* isolates share closer genetic identity to human *E. coli* as regards to core genome phylogeny, antimicrobial genes and virulence gene content and thus, represent potential zoonotic risks. The possibility that multi-drug resistant and or pathogenic *E. coli* could be transmitted from food sources such as chicken meat raises serious public health concerns.

In conclusion, as a part of this thesis work, a combinatorial program named Contig-Layout-Authenticator (CLA) was developed. This program is capable of ordering and scaffolding bacterial *de novo* contigs. It was also validated against the currently available reference based ordering as well as scaffolding tools with the help of both simulated as well as real datasets. Comparative study of *E. coli* strains belonging to ST131, ST38, ST405 and ST648 was

performed where five strains from these STs were sequenced, assembled and annotated. Core genome comparison of these four STs using five in-house and 134 publicly available strains have revealed 142 missing genes only from all the 99 strains of ST131. Another whole genome based study was conducted to study the zoonotic potential of strains isolated from Indian retail poultry where 10 in-house strains were sequenced, assembled and annotated and then their genomic co-ordinates were compared against 40 strains from different pathotypes available from public databases. Overall, the study potrays the patterns of evolution and possible zoonotic potential in different lineages of ESBL resistant ExPEC in addition to the development of a program to organize the bacterial *de-novo* contigs into a draft genome with higher accuracy and less time.

Publications

PLOS ONE

RESEARCH ARTICLE

Contig-Layout-Authenticator (CLA): A Combinatorial Approach to Ordering and Scaffolding of Bacterial Contigs for Comparative Genomics and Molecular Epidemiology

Sabiha Shaik, Narender Kumar, Aditya K. Lankapalli, Sumeet K. Tiwari, Ramani Baddam, Niyaz Ahmed*

Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad, Hyderabad, India

* niyaz.ahmed@uohyd.ac.in

CrossMark
click for updates

OPEN ACCESS

Citation: 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(6): e0155459. doi:10.1371/journal.pone.0155459

Editor: Patrick Jon Biggs, Massey University, NEW ZEALAND

Received: July 25, 2015

Accepted: April 29, 2016

Published: June 1, 2016

Copyright: © 2016 Shaik et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. Other than that the pipeline script is available from source forge, Project homepage: https://sourceforge.net/projects/c-l-authenticator/files.

Funding: The authors acknowledge funding support from the Department of Biotechnology, Government of India (BT/PR6921/MED/29/699/2013). NK would like to acknowledge Council for Scientific and Industrial Research, India for the award of Senior Research Fellowship. The authors also would like to

Abstract

A wide variety of genome sequencing platforms have emerged in the recent past. High-throughput platforms like Illumina and 454 are essentially adaptations of the shotgun approach generating millions of fragmented single or paired sequencing reads. To reconstruct whole genomes, the reads have to be assembled into contigs, which often require further downstream processing. The contigs can be directly ordered according to a reference, scaffolded based on paired read information, or assembled using a combination of the two approaches. While the reference-based approach appears to mask strain-specific information, scaffolding based on paired-end information suffers when repetitive elements longer than the size of the sequencing reads are present in the genome. Sequencing technologies that produce long reads can solve the problems associated with repetitive elements but are not necessarily easily available to researchers. The most common high-throughput technology currently used is the Illumina short read platform. To improve upon the shortcomings associated with the construction of draft genomes with Illumina paired-end sequencing, we developed Contig-Layout-Authenticator (CLA). The CLA pipeline can scaffold reference-sorted contigs based on paired reads, resulting in better assembled genomes. Moreover, CLA also hints at probable misassemblies and contaminations, for the users to cross-check before constructing the consensus draft. The CLA pipeline was designed and trained extensively on various bacterial genome datasets for the ordering and scaffolding of large repetitive contigs. The tool has been validated and compared favorably with other widely-used scaffolding and ordering tools using both simulated and real sequence datasets. CLA is a user friendly tool that requires a single command line input to generate ordered scaffolds.

acknowledge funding support from the HIR project of the University of Malaya (UM.C/625/1/HIR/MOHE/CHAN-02 - Molecular Genetics). NA is a Visiting Professor at the University of Malaya, Malaysia and is an Adjunct Professor of the Academy of Scientific and Innovative Research (AcSIR), India. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

The emergence of newer platforms for whole genome sequencing has driven a revolution in the field of comparative genomics and epidemiological tracking of microorganisms [1–3]. Some of these platforms are essentially an adaptation of shotgun sequencing producing higher throughput and being cost-effective when compared to traditional Sanger sequencing. Though there are a variety of platforms available, such as Illumina, 454, Ion Torrent etc., their utility mainly entails reading of bases from short DNA fragments to generate read data [4]. In addition, some platforms such as Illumina also provide a paired-end sequencing option which essentially means that the generated fragments are read from both the ends providing useful information for downstream processing. Although the advent of these platforms has revolutionized the way we analyze and compare genomes, the main challenge refers to reconstructing the complete or a maximum draft genome out of millions of reads generated by the sequencers. This has underpinned development of multiple assembly tools with similar or distinct algorithms to harness and consolidate the sequence read data into larger contiguous fragments called contigs. Some examples of such assembly tools based on *de novo* algorithms include Velvet [5], SSAKE [6], etc. Although many of the available assembly processes reduce complexity of the data by generating contigs, the latter still require a lot of downstream analysis such as sorting based on their real biological order and resolution of repetitive elements before a high quality draft or complete genome is constructed.

The assembled contigs could be incorporated in various ways to reconstruct the genome. One important approach refers to directly ordering the contigs according to a chosen reference to build a chromosomal assembly with the help of tools such as ABACAS [7], Ragout [8], Mauve Contig Mover (MCM) [9], CONTIGuator [10] and Scaffold_builder [11] etc. These reference based approaches are known to be very efficient in analyzing specific mutations/variations among highly similar organisms since the possibilities of finding novel genomic insertions are scarce in some clonal populations. However, these methods have limitations in identifying strain-specific genes as well as handling mobile genetic elements which usually contribute to high strain to strain variation [12]. These approaches could therefore affect the resolution of an assembled genome sequence, especially in case of highly diverse and recombining organisms such as *Helicobacter pylori* [13].

The other approach for genome assembly is based on constructing scaffolds consisting of contigs joined together based on paired-end read information [14]. While a majority of sequence read assemblers such as Velvet [5], SOAPdenovo2 [15], SOPRA [16] and SGA [17] etc. come with an additional option of scaffolding, exclusive and stand-alone scaffolding tools such as Bambus [14], Bambus2 [18] and SSPACE [19] are also available. These tools usually generate multiple scaffolds where only intra-scaffold ordering and orientation of contigs is observed. Therefore, further ordering of scaffolds is required to obtain a draft genome. Other than this, these tools also face difficulties during the scaffolding of the repetitive contigs which sometimes result in misjoining/misassembly of the scaffolds [20]. The emergence of sequencing platforms that no longer depend on genome fragmentation and attempt to read directly from the genomic DNA such as PacBio, Oxford Nanopore etc. has shown promise to overcome these limitations by virtue of the long reads they generate. However, these platforms would take some more time to become mainstream and to displace some of the popular yet affordable platforms such as Illumina in terms of availability and affordability in some parts of the world [21–23].

In order to address the limitations of both the reference and scaffolding based whole genome ordering tools mainly entailing Illumina paired-end sequencing, we developed Contig-Layout-Authenticator (CLA) which not only scaffolds the contigs but also attempts to achieve

proper ordering of the scaffolds at the whole genome level. In other words, the contigs are initially ordered at the whole genome level similar to the ordering achieved by ABACAS and Ragout, based on a reference, but the order is much more refined at the later stages. The ordered contigs are further scaffolded when supported by the paired-end information; this way, if multiple scaffolds are generated, then it clearly indicates that some genome information between two consecutive scaffolds still needs to be incorporated into the assembly. Thus, CLA carries out the functions of ordering and scaffolding tools in a combinatorial and efficient manner yielding better draft genomes. The pipeline of CLA is proficient and adept in identifying probable contaminating contigs as well as misassemblies within the scaffolds. It assists the users in improving quality of the draft genome in an informed and step by step manner. The comparative analysis of CLA with other widely used ordering and scaffolding tools revealed its enhanced performance. Therefore, we believe CLA would serve as an efficient and user friendly tool for sorting and scaffolding of contigs to achieve better draft genomes of prokaryotic organisms.

Results

As CLA performs both ordering at the whole genome level followed by scaffolding, its efficiency was validated at three different levels. Firstly, it was compared with the widely used ordering tools such as ABACAS, MCM, CONTIGuator and Ragout. Secondly, the scaffolding ability of CLA was compared with Bambus2, SSPACE, SOPRA, SOAPdenovo2, SGA and the reference-based scaffolder, MeDuSa [24]. Finally, to test the hybrid approach efficiency, scaffolds from SSPACE, SOPRA, SOAPdenovo2 and SGA were further ordered with Ragout and the final results were compared against CLA. All the in-house scripts (along with their command line history) that were used for validation, are provided along with the tool.

CLA versus reference based sorting tools

All the tools under this category such as ABACAS (v1.3.1), MCM (Mauve-2.3.1), CONTIGuator (v2.7.3) and Ragout (v1.1) were run with default parameters and their outputs were compared with CLA using the QUAST(v3.1) [25] tool to avoid manual errors. Ragout and MeDuSa were run using multiple reference genomes (S1 and S2 Tables). The outputs generated by all the above mentioned tools using simulated and real sequence datasets were evaluated for parameters such as misassemblies within the chromosome assembly, the number of contigs (>1kb) unassigned in the final order, the size of largest unassigned contig, the genome fraction obtained and the number of large repetitive contigs (contig size greater than 500bp) correctly placed. QUAST was used to identify number of misassemblies and to calculate the genome fractions with the original genome as reference for simulated datasets. While in the case of a real dataset, due to the unavailability of the original genome, a closely related genome was used as reference for QUAST. Therefore, the statistics generated in case of real dataset could be biased towards the reference genome and may not exactly depict the efficiency of the tools as shown with the simulated dataset. In-house Perl scripts were used to calculate the number of large repetitive contigs present in the original genome, as well as number of them placed by each of these tools.

Misassemblies in the chromosome assembly

A chromosomal assembly or draft genome constructed from reference based ordering tools comprised of ordered contigs merged together after placing 'Ns' in between. Tools such as Ragout, CONTIGuator and ABACAS usually provide the final ordered contigs in the form of a single scaffold or chromosome assembly, whereas, MCM only provides ordered contigs in

multi-FASTA format. In CLA, contigs are merged into scaffolds (by placing 'Ns' in between) only if supported by sufficient paired-end information. An in-house script was used to generate a chromosome assembly for the outputs of MCM and CLA. In order to identify the errors such as misassemblies encountered by each of the compared tools and judge their performance, we used QUAST. The misassemblies observed in the chromosomal assemblies were further linked to inversions, relocations and translocations by QUAST. Information about different types of misassemblies observed in case of simulated dataset(s) is given in S3 Table. As inferred from Table 1, CLA outcompeted other tools by producing least number of misassemblies in case of simulated datasets of all the test organisms except *H. pylori*. In case of organisms such as *Bartonella quintana*, and *Caulobacter crescentus*, CLA showed only two inversions without any relocations (S3 Table). After CLA, Ragout and CONTIGuator performed fairly well followed by ABACAS. MCM was observed to be the only tool with the highest number of misplacements in all organisms; this might be due to placement of unassigned contigs at the end of the order.

While dealing with real datasets, a reference genome required to act as an exact representative of the test strains was not available for some bacterial species/strains; we therefore chose a closely related genome as reference. Because this modification would definitely favor the reference based tools, the performance of CLA varied from genome to genome as revealed in the Table 2. CLA vended least number of misassemblies in all cases except *Campylobacter jejuni*, as inferred from Table 2. Moreover, the performance of CLA was comparable and even superior in some instances to other tools thus exhibiting the robustness of the combinatorial approach it used.

Contigs unassigned in the final chromosome assembly

Majority of the reference based ordering tools often exclude contigs (which fail to align properly with a reference genome) from the final order. These unassigned contigs are either provided as separate files or put towards the end in their output. The scenario becomes more pertinent in cases of highly recombining organisms as the reference genome architecture might differ significantly from the target genome. These unassigned contigs might encompass valuable strain specific information encoding novel elements and virulence genes etc. Since CLA utilizes paired-end information in addition to a reference genome, it is able to incorporate the maximum number of contigs in the chromosomal assembly and thereby retaining potentially novel, strain specific elements. As observed from the comparative statistics mentioned in Table 1, CLA showed consistent results by not having unassigned any contig with size greater than 1kb in all the organisms. On the other hand, CONTIGuator left more number of contigs with size greater than 1kb unassigned to the final order. It was also observed that a contig as long as 10kb in *B. quintana* was not considered in the final order generated by CONTIGuator. Though ABACAS performed considerably well, in some cases such as *C. jejuni*, *Haemophilus influenzae*, *H. pylori* and *Rhizobium etli*, it ended up excluding contigs of lengths 481kb, 345kb, 170kb and 798kb, respectively, in these organisms. MCM was observed to assign all the contigs in the final assembly as it places the unaligned contigs towards the end of the chromosomal assembly.

A similar pattern was observed with the real dataset (Table 2). While CLA and MCM have both incorporated in the final order all the contigs that were greater than 1kb, contigs as large as 32kb, 36kb and 93kb remained unassigned, respectively, in the final order, when Ragout, ABACAS and CONTIGuator were used to assemble genomes of *C. jejuni*, *Escherichia coli* and *Salmonella* Typhi. The number of contigs that were >1kb and were not considered by these tools were also observed to be very high in case of the real dataset.

Table 1. Comparative statistics of CLA and reference based alignment tools generated using simulated dataset.

	Genome*	Tool	# misassemblies in chromosome assembly	Unassigned contigs in final order >1kb	Largest contig unassigned	Genome fraction	True repeat places filled**
1.	***B. Quintana*** #contigs:**47** #total repeat positions: **19** #misassemblies in input contigs: **0**	CLA	2	0	850	99.68	15
		Ragout	5	3	3347	99.299	15
		ABACAS	12	1	1104	98.648	9
		MCM	26	0	0	98.767	9
		CONTIGuator	4	7	10343	97.401	0
2.	***C. jejuni*** #contigs: **33** #total repeat positions: **12** #misassemblies in input contigs: **0**	CLA	4	0	425	99.728	10
		Ragout	5	4	4566	98.807	10
		ABACAS	9	7	481308	56.49	3
		MCM	21	0	0	98.82	5
		CONTIGuator	8	6	6071	97.71	1
3.	***C. crescentus*** #contigs:**49** #total repeat positions: **28** #misassemblies in input contigs: **0**	CLA	2	0	793	99.901	24
		Ragout	8	0	371	99.923	28
		ABACAS	11	0	371	99.355	9
		MCM	18	0	0	99.355	9
		CONTIGuator	4	6	5440	98.94	0
4.	***H. influenzae*** #contigs: **43** #total repeat positions: **19** #misassemblies in input contigs: **1**	CLA	6	0	606	98.963	9
		Ragout	12	0	375	98.559	19
		ABACAS	15	4	345463	70.201	3
		MCM	22	0	0	98.187	5
		CONTIGuator	11	3	3503	97.719	1
5.	***H. pylori*** #contigs: **46** #total repeat positions: **20** #misassemblies in input contigs: **0**	CLA	7	0	960	99.672	17
		Ragout	4	1	5487	98.97	14
		ABACAS	17	6	170928	82.357	6
		MCM	23	0	0	99.133	10
		CONTIGuator	8	4	4051	98.135	0
6.	***R. etli*** #contigs: **30** #total repeat positions: **21** #misassemblies in input contigs: **0**	CLA	3	0	601	99.756	19
		Ragout	5	1	1052	99.406	13
		ABACAS	5	7	798023	71.125	4
		MCM	24	0	0	99.221	8
		CONTIGuator	10	8	4331	98.761	0
7.	***S.* Typhi** #contigs: **67** #total repeat positions: **24** #misassemblies in input contigs: **0**	CLA	8	0	520	99.141	11
		Ragout	55	2	3835	98.873	12
		ABACAS	13	0	474	98.682	7
		MCM	32	0	0	98.698	7
		CONTIGuator	13	5	6299	98.311	1
8.	***T. pallidum*** #contigs: **22** #total repeat positions: **17** #misassemblies in input contigs: **0**	CLA	5	0	980	99.124	9
		Ragout	7	1	1617	99.432	17
		ABACAS	10	0	0	98.768	7
		MCM	15	0	0	98.769	7
		CONTIGuator	6	3	3283	97.95	1

*All the genomes were simulated with a read length of 100bp and insert size of 400bp.

** Total number of positions filled by the repetitive contigs in the final output; for example, if tools have placed contig A at 2 out of 4 places and contig B at 1 out of 6 places, then total number of repeat positions in the original genome would be 10 and true repeats placed by the tool would be 3.

Number of

doi:10.1371/journal.pone.0155459.t001

Table 2. Comparative statistics of CLA and reference based alignment tools generated using real dataset.

	Genome	Tool	# misassemblies in chromosome assembly	Unassigned contigs in final order >1kb	Largest contig unassigned	Genome fraction	True repeat places filled**
1.	***C. jejuni*** #contigs: **72** Read length: **100** Insert: **300** #repeats: **11** #misassemblies in input contigs: **7**	CLA	19	0	965	90.553	3
		Ragout	18	8	32459	91.261	11
		ABACAS	22	8	32459	90.584	4
		MCM	39	0	0	90.76	4
		CONTIGuator	20	8	32459	90.051	0
2.	***E. coli*** #contigs: **252** Read length: **151** Insert: **300** #repeats: **51** #misassemblies in input contigs: **98**	CLA	231	0	1000	92.245	15
		Ragout	242	45	36453	91.65	36
		ABACAS	203	49	36453	91.264	10
		MCM	262	0	0	92.054	10
		CONTIGuator	176	46	36453	91.508	0
3.	***H. influenzae*** #contigs: **50** Read length: **100** Insert: **200** #repeats: **19** #misassemblies in input contigs: **25**	CLA	32	0	774	88.984	8
		Ragout	33	8	5468	89.293	15
		ABACAS	24	11	208372	66.488	4
		MCM	39	0	0	88.194	4
		CONTIGuator	33	10	6440	87.842	0
4.	***M. tuberculosis*** #contigs: **241** Read length: **101** Insert: **300** #repeats: **43** #misassemblies in input contigs: **6**	CLA	23	0	978	98.491	24
		Ragout	32	9	5385	98.375	38
		ABACAS	37	5	6881	98.441	13
		MCM	113	0	0	98.559	14
		CONTIGuator	24	28	5385	97.11	0
5.	***S.* Typhi** #contigs: **119** Read length: **100** Insert: **200** #repeats: **37** #misassemblies in input contigs: **2**	CLA	31	0	978	98.75	13
		Ragout	32	4	93926	99.067	30
		ABACAS	23	4	93926	98.047	11
		MCM	65	0	0	98.812	11
		CONTIGuator	17	10	93926	98.328	0

** Total number of positions filled by the repetitive contigs in the final output.

Number of

doi:10.1371/journal.pone.0155459.t002

Genome Fraction

The genome fraction obtained after the chromosome assembly of all the strains in simulated and real sequence datasets was calculated using QUAST. In almost all cases in simulated datasets, CLA and Ragout were observed to generate a higher genome fraction when compared to others. While the majority of tools achieved a genome fraction greater than 97%, the genome fraction from ABACAS was found to be 56.49%, 70.20%, 82.35% and 71.12% in *C. jejuni*, *H. influenzae*, *H. pylori* and *R. etli*, respectively (Table 1). In real datasets, the genome fraction obtained by all the tools was observed to be above 85% except in *H. influenzae*, where ABACAS was observed to generate a fraction only equal to 66.48%.

Resolving large repetitive contigs

Large repetitive contigs were defined as the contigs of length >500 bases having more than one coordinate in the original genome. To validate the placement, large repeat containing contigs

of size >500 bases were considered. An in-house script was used to calculate the total number of positions for all repetitive contigs based on a BLASTn [26] alignment against a reference genome. BLAST version 2.2.29+ was used and the best hit with >90% identity and >90% query coverage was considered as a valid hit to identify the position coordinates. Then, the output of CLA and all ordering tools were scanned using another in house Perl script to evaluate the total number of positions filled by repetitive region containing contigs.

The ultimate aim of any genome sequencing project is to harness as much information as possible by formulation of better assemblies. Given this, resolution of repeats becomes an unavoidable parameter to obtain near perfect genomes. Since all the tools mentioned above are designed to aid researchers in achieving better assemblies, these tools were also compared for their performance related to the resolution of repeats. Except CLA and Ragout, other tools under this category were designed only to hint at the probable repetitive contigs and therefore required further manual intervention. As both CLA and Ragout handled repeats to some extent, they performed much better than the other tools in all the genomes. CLA and Ragout could more or less fill a similar number of true repeating positions in the majority of cases. While CLA could correctly fill 19 out of 21 positions (Table 1) in *R. etli*, Ragout could fill only 13. Similarly, Ragout performed better in the *Treponema pallidum* (*T. pallidum)* assembly by filling all 17 positions whereas only 9 of them could be placed by CLA. Both CLA and Ragout have shown similar performance in highly recombining organisms like *B. quintana and C. jejuni*, by filling 15 out of 19 positions and 10 out of 12 positions respectively. MCM and CONTIGuator were observed to be placing very few repeats in the correct order and copy number. CLA and Ragout performed in a superior way to other tools in handling the repetitive contigs by placing them in correct positions, followed by ABACAS.

In a real dataset, the number of repetitive positions was calculated with respect to the closely related genome used as reference which also served as input to these tools due to the unavailability of the original genome. Ragout was observed to perform well in all cases followed by CLA as inferred from Table 2. Therefore, it can be surmised that CLA and Ragout are comparable for this parameter in the chromosomal assembly process.

CLA versus Scaffolding tools

Scaffolding further reduces the number of sequence fragments by joining consecutive contigs and helps in the improvement of assembly. This could be achieved with the help of paired-end information or cues from the assembly graph, but was also attempted with the help of multiple reference genomes in the case of MeDuSa.

CLA, which utilizes a combination of reference alignment and paired-end information for scaffolding was compared against the recently published MeDuSa (v3) and other scaffolders such as SSPACE (v2.0), SOPRA (v1.4.6), SOAP denovo2 (v2.04-r240), SGA and BAMBUS2 (amos-3.1.0). For the simulated dataset, QUAST was used to deduce the assembly statistics from output files of each tool keeping the original genome as reference. Whereas, for the real dataset, genome of a closely related strain was used as reference genome. Parameters such as number of scaffolds formed, number of misassemblies within the scaffolds, number of misassembled scaffolds and size of genome fraction were used to compare the performance of CLA with others.

It was observed that CLA performed better under this category owing to its combinatorial approach. The comparative statistics for simulated and real datasets are mentioned in Tables 3 and 4, respectively. The lesser number of scaffolds with minimum number of misassemblies clearly depict the efficiency and accuracy of CLA over other scaffolding tools. For example, in the *R. etli* genome of the simulated dataset, CLA could combine all 30 contigs into just two

Table 3. Comparative statistics of CLA and Scaffolding tools generated using simulated dataset.

	Genome*	Tool	# scaffolds	# misassemblies within scaffolds	# misassembled Scaffolds	Genome fraction
1.	***B. Quintana*** #contigs: **47** #misassemblies in input contigs: **0**	CLA	3	2	1	99.68
		MeDuSa	16	11	7	98.751
		Bambus2	32	3	1	98.639
		SSPACE (no extension)	41	1	1	98.763
		SSPACE (extension)	38	1	1	98.822
		SOPRA	43	1	1	98.751
		SOAPdenovo2	35	4	3	98.704
		SGA	45	0	0	98.743
2.	***C. jejuni*** #contigs: **33** #misassemblies in input contigs: **0**	CLA	6	0	0	99.728
		MeDuSa	10	18	1	98.756
		Bambus2	22	4	2	98.736
		SSPACE (no extension)	28	0	0	98.781
		SSPACE (extension)	24	2	2	98.829
		SOPRA	32	0	0	98.737
		SOAPdenovo2	28	3	1	98.792
		SGA	32	0	0	98.737
3.	***C. crescentus*** #contigs: **49** #misassemblies in input contigs: 0	CLA	4	2	1	99.901
		MeDuSa	12	24	7	99.356
		Bambus2	31	13	3	99.332
		SSPACE (no extension)	44	0	0	99.347
		SSPACE (extension)	38	3	3	99.363
		SOPRA	46	0	0	99.34
		SOAPdenovo2	46	0	0	99.34
		SGA	46	0	0	99.34
4.	***H. influenzae*** #contigs: **43** #misassemblies in input contigs: **1**	CLA	5	2	1	98.963
		MeDuSa	6	16	2	98.18
		Bambus2	31	6	3	98.158
		SSPACE (no extension)	38	2	2	98.003
		SSPACE (extension)	35	2	2	98.033
		SOPRA	26	1	1	98.156
		SOAPdenovo2	40	1	1	98.144
		SGA	27	1	1	98.156
5.	***H. pylori*** #contigs: **46** #misassemblies in input contigs: **0**	CLA	2	7	2	99.672
		MeDuSa	7	15	3	99.137
		Bambus2	25	10	6	99.103
		SSPACE (no extension)	39	0	0	99.101
		SSPACE (extension)	36	0	0	99.151
		SOPRA	42	0	0	99.092
		SOAPdenovo2	37	1	1	99.105
		SGA	44	0	0	99.078

(*Continued*)

Table 3. (*Continued*)

	Genome*	Tool	# scaffolds	# misassemblies within scaffolds	# misassembled Scaffolds	Genome fraction
6.	***R. etli*** #contigs: **30** #misassemblies in input contigs: **0**	CLA	2	2	1	99.811
		MeDuSa	10	13	1	99.204
		Bambus2	23	1	1	99.071
		SSPACE (no extension)	28	0	0	99.211
		SSPACE (extension)	24	2	2	99.225
		SOPRA	30	0	0	99.204
		SOAPdenovo2	24	4	1	99.215
		SGA	30	0	0	99.204
7.	***S.* Typhi** #contigs: **67** #misassemblies in input contigs: **0**	CLA	7	2	1	99.149
		MeDuSa	3	30	1	98.695
		Bambus2	47	11	6	98.666
		SSPACE (no extension)	51	1	1	98.679
		SSPACE (extension)	53	0	0	98.717
		SOPRA	63	0	0	98.647
		SOAPdenovo2	56	4	3	98.657
		SGA	66	0	0	98.644
8.	***T. pallidum*** #contigs: **22** #misassemblies in input contigs: **0**	CLA	3	3	2	99.124
		MeDuSa	1	13	1	98.765
		Bambus2	16	2	1	98.571
		SSPACE (no extension)	13	1	1	98.755
		SSPACE (extension)	14	2	1	98.865
		SOPRA	18	1	1	98.759
		SOAPdenovo2	14	2	2	98.747
		SGA	21	0	0	98.731

*All the genomes were simulated with a read length of 100bp and insert size of 400bp.

Number of

doi:10.1371/journal.pone.0155459.t003

scaffolds whereas MeDuSa, Bambus2, SSPACE with extension, SSPACE without extension, SOPRA, SOAPdenovo2 and SGA generated 10, 23, 24, 28, 30, 24 and 30 scaffolds, respectively. CLA's performance was consistent for all the genomes under comparison except in *H. pylori* where the number of misassemblies slightly increased to 7. In *C. jejuni*, 33 contigs were combined by CLA into 6 scaffolds without any misassemblies. In *B. quintana*, *C. crescentus* and *S.* Typhi, only 2 inversions were found within scaffolds from CLA (S4 Table). Numbers of misassemblies performed by CLA were relatively fewer when compared with MeDuSa even in the real dataset (Table 4).

Though MeDuSa generated a lesser number of scaffolds in a few cases of both the real and simulated datasets, the number of misassemblies was significantly higher in all the organisms. While misassemblies in case of SSPACE, SOPRA and SOAPdenovo2 were more or less similar to that of CLA, these tools gave a higher number of scaffolds. Considering the trade-off between number of scaffolds generated and the number of misassemblies, CLA showed the best performance with least number of scaffolds having the minimum misassemblies and the highest genome fraction recovered.

Table 4. Comparative statistics of CLA and Scaffolding tools generated using real dataset.

	Genome*	Tool	# scaffolds	# misassemblies within scaffolds	# misassembled scaffolds	Genome fraction
1.	***C. jejuni*** #contigs: **72** Read length: **100** Insert: **300** #misassemblies in input contigs: **7**	CLA	19	9	5	90.551
		MeDuSa	10	33	1	90.764
		Bambus2	45	16	9	90.692
		SSPACE (no extension)	50	10	8	90.589
		SSPACE (extension)	51	11	7	90.612
		SOPRA	44	9	6	90.685
		SOAPdenovo2	52	9	6	90.626
		SGA	54	9	5	90.613
2.	***E. coli*** #contigs: **252** Read length: **151** Insert: **300** #misassemblies in input contigs: **98**	CLA	67	165	37	92.31
		MeDuSa	74	223	11	92.047
		Bambus2	174	139	54	91.998
		SSPACE (no extension)	241	99	60	91.967
		SSPACE (extension)	249	104	63	92.132
		SOPRA	212	117	56	91.995
		SOAPdenovo2	206	103	56	91.935
		SGA	233	100	56	91.948
3.	***H. influenzae*** #contigs: **50** Read length: **100** Insert: **200** #misassemblies in input contigs: **25**	CLA	6	29	4	88.984
		MeDuSa	12	45	4	87.571
		Bambus2	43	25	10	87.562
		SSPACE (no extension)	32	26	11	87.533
		SSPACE (extension)	33	28	11	87.604
		SOPRA	24	27	6	88.184
		SOAPdenovo2	29	28	9	87.595
		SGA	34	28	4	88.148
4.	***M. tuberculosis*** #contigs: **241** Read length: **101** Insert: **300** #misassemblies in input contigs: **6**	CLA	44	10	5	98.52
		MeDuSa	9	93	3	98.549
		Bambus2	139	59	24	98.324
		SSPACE (no extension)	153	11	10	98.432
		SSPACE (extension)	144	14	11	98.534
		SOPRA	119	10	8	98.455
		SOAPdenovo2	108	12	12	98.365
		SGA	139	9	9	98.388
5.	***S.* Typhi** #contigs: **119** Read length:**100** Insert: **200** #misassemblies in input contigs: **2**	CLA	40	5	3	98.771
		MeDuSa	22	47	6	98.796
		Bambus2	80	19	10	98.771
		SSPACE (no extension)	85	6	5	98.803
		SSPACE (extension)	91	2	2	98.855
		SOPRA	80	8	4	98.712
		SOAPdenovo2	89	3	3	98.709
		SGA	119	2	2	98.717

* Real genome data used;

Number of

doi:10.1371/journal.pone.0155459.t004

CLA versus combination of reference based and scaffolding tools

CLA being able to perform ordering followed by scaffolding, its performance was also compared by ordering already generated scaffolds. Ragout was used to order scaffolds from SSPACE with extension, SOPRA, SOAPdenovo2 and SGA and the final results were compared to that of CLA. The comparative statistics are listed out in Table 5.

CLA performed better with the least number of misassemblies in the majority of cases like *B. quintana*, *C. crescentus*, *R. etli*, *S.* Typhi and *T. pallidum*. Ragout and SGA together generated the least number of misassemblies in *C. jejuni*, *H. influenzae* and *H. pylori* followed by CLA. Amongst the combinations used, Ragout and SGA gave good results similar to that of CLA while Ragout and SSPACE with extension did not seem to be an ideal and consistent combination.

Additional features of CLA

Handling intra-contig repeating segments. Small repetitive sequences which form only part of a contig also create problems during assembly. Some of these elements consist of insertion sequences (IS) and transposable elements. CLA could split and precisely place certain segments of contigs which comprised of IS and transposable elements and displayed connections at different places, given sufficient availability of paired-end information. Such segments within the contigs are extracted based on the read alignment positions and are placed according to their connections in the map-file ([b] in S1 Fig). CLA in simulated data of *S.* Typhi placed a 500-600bp intra-contig repeat at about 24 places. This segment was found to be a transposable element that was present at about 26 different positions in the original genome of *S.* Typhi TY2. The positioning of these small repeats according to CLA and their original position in the reference are mentioned in S5 Table. The performance of CLA in terms of placing small repeats appeared to vary and was dependent on the sequencing quality and coverage depth. Therefore, CLA appears to be an efficient tool in handling even small repetitive elements in comparison to other tools which could not address these issues.

Contigs unrelated to the chromosomal assembly. Possible contamination of DNA samples during sequencing and inaccurate de-multiplexing of read data may lead to contamination of the reads and potentially resulting in un-related contigs. Such contigs might cause significant problems during downstream analysis. In other cases, plasmids which are not a part of chromosomal sequences also pose problems in achieving accurate genome assembly. Given this, CLA was observed to filter out such contigs and separate them from the main chromosomal assembly. Contigs with BLASTn identity (against reference) of less than 5% and with no paired-end link information are tagged as unrelated contigs. Such contigs were effectively excluded out by CLA in real datasets of *E. coli*, *H. influenzae*, *H. pylori*, *Mycobacterium tuberculosis* and *S.* Typhi. In *S.* Typhi, one of the contigs tagged as unrelated was found to be a plasmid contig thus preventing false scaffolding with the chromosomal contigs. To examine the performance of CLA in the detection of possible contamination, a few random contigs from other genomes were introduced into simulated data. CLA could detect these contigs as contamination from a different source and correctly labelled them as possibly unrelated contigs. Contigs tagged as unrelated could be cross-checked by the user for its use in the final assembly.

Information about probable misassemblies. Another advantage of using CLA is that the user is provided with additional files containing information on the efficiency of read-pair mapping and the extent of possible misassembling. The same can be useful for advanced users in order to further improve their assembly. CLA flags this information by vending a log file to the user.

Table 5. Comparative statistics of CLA and Ragout ordering of scaffolds using simulated dataset.

	Genome*	Tool	# of contigs and scaffolds in the input file	# misassemblies in chromosome assembly	Unassigned contigs from final order >1kb	Largest contig unassigned	Genome fraction
1.	***B. quintana*** #misassemblies in input contigs: **0**	CLA	47	2	0	850	99.68
		SSPACE_ext/ Ragout	38	15	2	1155	99.242
		SOPRA/Ragout	43	7	6	3347	98.296
		SOAPdenovo2/ Ragout	35	9	3	11040	98.644
		SGA/Ragout	45	2	0	739	99.153
2.	***C. jejuni*** #misassemblies in input contigs: **0**	CLA	33	4	0	425	99.728
		SSPACE_ext/ Ragout	24	5	5	15071	97.389
		SOPRA/Ragout	32	5	4	4566	98.807
		SOAPdenovo2/ Ragout	28	4	3	1954	99.094
		SGA/Ragout	32	3	4	4566	98.807
3.	***C. crescentus*** #misassemblies in input contigs: **0**	CLA	49	2	0	793	99.901
		SSPACE_ext/ Ragout	38	32	0	371	99.386
		SOPRA/Ragout	46	8	0	371	99.896
		SOAPdenovo2/ Ragout	46	6	0	371	99.923
		SGA/Ragout	46	6	0	371	99.923
4.	***H. influenzae*** #misassemblies in input contigs: **1**	CLA	43	6	0	606	98.963
		SSPACE_ext/ Ragout	35	11	1	1088	98.136
		SOPRA/Ragout	26	11	0	572	99.013
		SOAPdenovo2/ Ragout	40	12	0	375	98.517
		SGA/Ragout	27	3	1	2461	98.247
5.	***H. pylori*** #misassemblies in input contigs: **0**	CLA	46	7	0	960	99.672
		SSPACE_ext/ Ragout	36	12	1	6406	98.657
		SOPRA/Ragout	42	4	1	5487	98.965
		SOAPdenovo2/ Ragout	37	6	1	10376	98.746
		SGA/Ragout	44	4	1	5487	98.97
6.	***R. etli*** #misassemblies in input contigs: **0**	CLA	30	3	0	601	99.756
		SSPACE_ext/ Ragout	24	8	5	3972720	89.983
		SOPRA/Ragout	30	5	1	1052	99.406
		SOAPdenovo2/ Ragout	24	7	4	3831	99.102
		SGA/Ragout	30	4	1	1052	99.501
7.	***S.* Typhi** #misassemblies in input contigs: **0**	CLA	67	8	0	520	99.141
		SSPACE_ext/ Ragout	53	61	1	5127	98.625
		SOPRA/Ragout	63	9	0	519	99.236
		SOAPdenovo2/ Ragout	56	32	0	283	99.104
		SGA/Ragout	66	54	3	3835	99.142

(*Continued*)

Table 5. (*Continued*)

	Genome*	Tool	# of contigs and scaffolds in the input file	# misassemblies in chromosome assembly	Unassigned contigs from final order >1kb	Largest contig unassigned	Genome fraction
8.	***T. pallidum*** #misassemblies in input contigs: **0**	CLA	22	5	0	980	99.124
		SSPACE_ext/ Ragout	14	9	1	1696	98.711
		SOPRA/Ragout	18	6	1	1617	99.535
		SOAPdenovo2/ Ragout	14	10	0	244	99.513
		SGA/Ragout	21	8	1	1617	99.537

*All the genomes were simulated with a read length of 100bp and insert size of 400bp.

Number of

doi:10.1371/journal.pone.0155459.t005

Discussion

Genome assembly could be a challenging task especially for prokaryotes. This is mainly due to the plasticity of prokaryotic genomes as dictated by discrete evolutionary events and bottlenecks that shape adaptation dynamics and lifestyles of bacterial organisms in different ecosystems [1]. Consequently, prokaryotic genomes are often replete with signatures reminiscent of various genetic rearrangements occurring due to frequent insertion, deletion and substitution events as well as enriched with multiple homopolymeric tracts (arising out of replication errors), repeat motifs of different composition and lengths, insertion sequences, prophages and genomic islands etc. [27]. All these plastic regions pose serious difficulties in assembling a genome mainly because of sequence redundancy that they bring in the form of multiple alleles, palindromes, inverted repeats and tandem duplications. Most of the available tools either rely on reference to order contigs at the whole genome level, or scaffold them based on the read data [7, 9, 10, 14, 16, 18]. While using a reference genome could result in omission/exclusion of certain strain specific elements [10], the scaffolding methods struggle to resolve repetitive regions and limit ordering within the scaffolds [20]. Given these practical difficulties, we developed CLA, which combines the benefits of individual approaches to minimize errors while generating a draft genome. CLA uses a reference at the beginning to create a preliminary sort order which then undergoes extensive validation based on paired-end read information to resolve repetitive elements and re-sorting of the contigs. The sorted contigs are only scaffolded based on the available read-pair information. Although ordering is tried to be attained at the whole genome level, contigs are linked into scaffolds only when supported by paired-end information indicating their connectivity. Hence, it is easier for the users to fill in the information between the scaffolds using further downstream processing in order to achieve a complete genome.

The existing reference based tools though efficient in sorting the contigs at the whole genome level, were observed to remove those contigs from the final chromosome assembly which failed to properly align with the reference (Tables 1 and 2). These excluded contigs may lead to loss of significant genome information. For example, BLAST analysis of such excluded contigs from the *H. influenzae* genome identified several genes encoding metabolic functions, which were otherwise discarded. On the other hand, the scaffolding of contigs is performed only based on paired-end information [14, 18, 19] or in the case of MeDuSa using multiple reference genomes [24]. In the case of final assembly/genome obtained with multiple scaffolds, ordering is limited within the scaffolds and also repetitive regions with their misleading

connections sometimes lead to intra-scaffold misassemblies. From our validation study, it was observed that CLA could address all these issues better than the compared tools (Tables 1–5). Since CLA utilizes both reference and paired-end information, it performed better in retaining maximum number of contigs in the final output without compromising on accuracy. It also utilizes read-pair information to place some of the repetitive elements. The overall performance of CLA was found to be much better than the existing reference based ordering tools as well as scaffolding tools with a minimum number of misassemblies. In all our case studies and comparative analyses, CLA was seen to be misplacing contigs in just two cases: 1) firstly, when there was insufficient paired-end information, 2) secondly, when two contigs had same flanking contig connections at both ends leading to their misjoining within the scaffold. To avoid such scenarios, CLA lists out contigs with probable swapping, in a log file to alert the users of probable misassemblies within scaffolds.

The performance with the real and simulated datasets pointed out the capability of CLA in not only handling data from monomorphic bacteria such as *S*. Typhi but also highly diverse ones as *H*. *pylori*. The higher abundance of transposases in bacteria and important biological roles proposed for them in previous studies underlines the need to handle them carefully during genome assembly [28]. CLA was observed to be efficient in this aspect and was able to resolve 24 out of 26 transposable elements in *S*. Typhi.

Though manual curation is inevitable for completing a genome, CLA leaves less scope for manual intervention and also provides all the required blueprints to complete further re-construction of the genome. Therefore, we believe that in the light of existing difficulties regarding the genome assembly, CLA would be a significant step forward in improving the genome assembly pipeline with a user friendly approach and efficient data usage.

Materials and Methods

Real and simulated datasets

Eight complete bacterial genomes with varied genome characteristics were downloaded from NCBI. The paired reads were simulated for each of the genomes with the help of GemSIM (v.1.6) [29] using its Illumina error model. The genome characteristics of these genomes along with their accession numbers are provided in S1 Table. The 100 bases long read-pairs were simulated with an insert length of 400(±20) along with genome coverage of 100X. The reads were filtered using NGS QC Toolkit (v.2.3) [30] to remove bad quality reads.

These filtered reads were then assembled into contigs using Velvet *de novo* assembler (1.2.08) with an optimal k-mer chosen with the help of VelvetOptimiser (2.2.5) (http://bioinformatics.net.au/software.velvetoptimiser.shtml). The assembled contigs were used as input for validating CLA with other tools using QUAST. For the real dataset, five paired-read datasets with different organismal background from SRA data were considered. Information about these datasets is provided in the S2 Table. Filtering followed by the assembly of reads was carried out in a similar manner as described above.

Pipeline of CLA

The tool was designed and developed to run in the following stages:

Sort order creation. The detailed schematic of CLA is explained in Fig 1. Contigs, paired-end sequence reads and a suitable reference genome sequence (preferably of the same species as the genome being assembled) serve as input for the tool. The process starts with the exclusion of contigs that are less than 200bp. A preliminary sort order is created after alignment of the contigs with the reference genome by using BLASTn. Contigs with a BLASTn best hit of less than a threshold identity value (for contigs of size <1kb: 50% identity; size>1-10kb: 25%

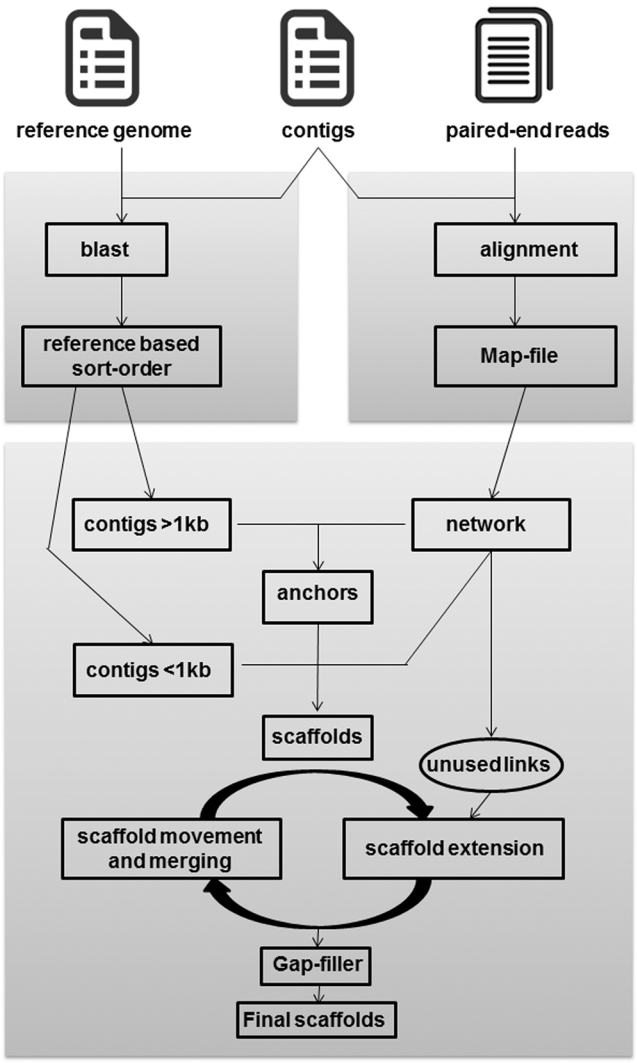
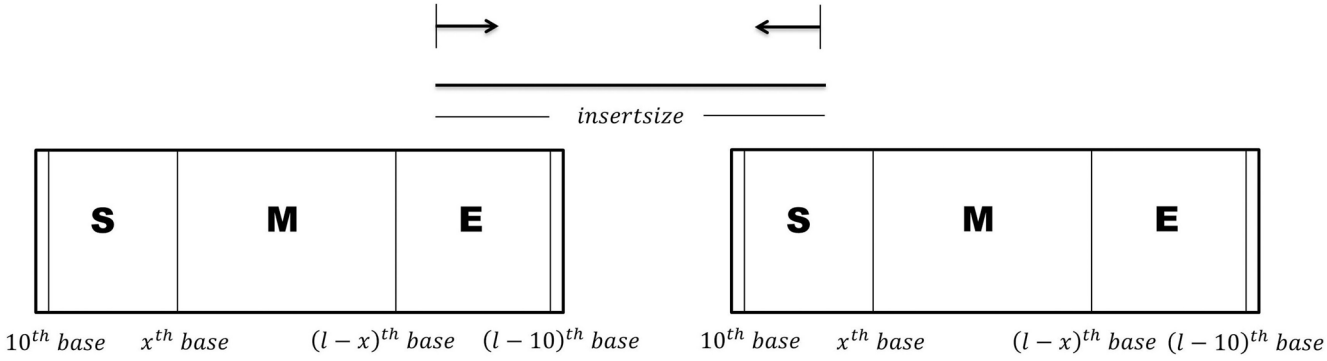


Fig 1. Schematic overview of CLA pipeline. The schema of CLA pipeline is divided into three stages. In the first stage, a reference based order is derived followed by the second stage where connections between the individual contigs are extracted based on alignment information. The final stage makes use of the information from the first two stages to decide the final order followed by scaffolding and gap filling.

doi:10.1371/journal.pone.0155459.g001

identity; size >10kb: 10% identity) are placed at the end of the order. The individual contigs are arranged according to the sort order with consequent reverse complementation of contigs wherever required in accordance with BLASTn output.

Extracting the connections. The paired reads are then mapped to the sorted contigs (obtained as above) using BWA [31] and their information regarding the alignment position and orientation is extracted. Each of the contigs is theoretically divided into start, mid and end regions as depicted in Fig 2 wherein the sequence read having its pair in any other contig constitutes a potential link. The information of all possible links (formed by the read pairs) at the start, mid and end regions of each of the contigs is then tabulated in a map file. Since the read pairs are derived from a single insert, these links strongly suggest the proximity or contiguity of these two contigs in the real genome. For example, as depicted in Fig 2, if two contigs are in proximity then the reads aligned at the end of one contig would have their linking pairs aligned



$$x = insertsize + (insertsize * 0.2)$$

$$l = contig\ length$$

$$start(S) = \left[10^{th}\ base\right]\ to\ [x^{th}\ base]$$

$$mid(M) = \left[x^{th}\ base\right]\ to\ [(l-x)^{th}\ base]$$

$$end(E) = \left[(l-x)^{th}\ base\right] to\ [(l-10)^{th}\ base]$$

Fig 2. Defining the start and end of a contig based on the insert size. Connectivity between two consecutive contigs is decided by the paired reads from a single insert. To find such connections between contigs, each contig is theoretically divided into start, mid and end regions based on the insert size.

doi:10.1371/journal.pone.0155459.g002

at the start of the following contig. The map file contains four columns representing the contig ID, start, end and mid region of the contigs, respectively. To avoid any erroneous assembly at the extreme ends of the contigs, the first and last ten bases from the start and the end of each of the contigs are not considered for calculating the links. A minimum of 10 read pair links are considered to calculate the proximity of two contigs in the genome (valid connection between contigs). Therefore, the map file generation is instrumental in validation of the reference sorted contigs and also helps in resolving some of the issues such as repeats, duplications and inversions. Some examples regarding the resolutions of repetitive regions and possible duplications are detailed in S1 Fig.

Sort-order validation. Contigs that have less than 5% BLASTn identity against reference (as inferred from the earlier BLASTn output) and lacking suitable connections in the map-file are flagged as un-related and are removed from the downstream analysis. These might represent contigs formed due to contaminating reads or sometimes due to plasmid contigs as observed from the results of the real datasets. A network file is then created from the map file, where start/end of a contig is represented as vertex/node and their respective connections as edges. Contigs that are less than 1kb in size are initially excluded from the sort order. The tool then scans the map file to validate the reference based sort order while looking for connections between the neighboring contigs (contigs >1kb). In cases where there is no direct connection (no links connecting the end of the former and start of the later contig) between two consecutive contigs from the reference based sort order, the tool then looks for any intermediary contigs linking these two contigs based on the link information from the map file. Each path is a representation of one or more contigs and a shortest path (defined by one or more contigs

connecting to other contigs based on the link information gleaned from the map file) was found using the Floyd Warshall algorithm [32]. Only paths that include contigs greater than 1kb in size are considered at this stage. The connecting contig is either copied or moved from its position based on the connections in the map file. All the contigs at the end of this stage are referred to as anchors.

Connecting the anchors. The link information of the excluded contigs of size less than 1kb were then used to connect the defined anchors using the Floyd Warshall algorithm as discussed in the previous section. All such connected anchors were then merged into scaffolds after placing 'Ns' between them. To prevent any false positive placements, the sort-order was scanned at every step based on the number of connections from the map file.

Ordering the scaffolds. An iterative process of scaffold merging and extension is performed by the CLA based on the unused connections and leftover contigs from the network file. After a search for inter scaffold connections, if two scaffolds are found to be connected, the one smaller in size is moved and oriented in accordance with the larger scaffold. Once merged, the map-file is scanned for any further unused connections at the ends of the newly extended scaffolds. Merging is again performed and the process is repeated until either all the connections are exhausted or no proper connections could be found. A final sort order is then created and the contigs are merged into scaffolds followed by their ordering, all in accordance with the new sort order. The gaps between the scaffolds are then closed by using GapFiller [33]. Even though the order is defined, all the contigs are not merged into a single pseudogenome and are left at the stage of scaffolds because of lack of valid connections available from the map file. Thus, inter-scaffold gaps indicate the existence of gaps at the level of the sequencing data itself. CLA uses the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) for easy handling and manipulation of the intermediate files.

Supporting Information

S1 Fig. Pictorial representation of various scenarios where map-file can be used to resolve repetitive contigs. (a) A normal case scenario where a repetitive contig 38 is placed at two different positions based on its connections from the map-file. (b) An example of an intra-contig repeating segment, where mid-region of contig 20 is connecting two contigs—contig 9 and contig 35. (c) Example of a tandem repeat, where the whole contig 95 has connections at both start and end pointing to another contig 82.
(TIF)

S1 Table. Genome characteristics and information of the strains utilized for simulation. Table with the accession numbers of the strains and their genomic characteristics that were used for the simulated dataset
(PDF)

S2 Table. Information about strains under real dataset. Table with the accession numbers of the strains and reference genomes used for the study
(PDF)

S3 Table. Misassembly details of CLA and reference based ordering tools in simulated dataset. Table listing out number of relocations, translocations and inversions which amounted to the total number of misassemblies
(PDF)

S4 Table. Misassembly details of CLA and Scaffolding tools in simulated dataset. Table listing out number of relocations, translocations and inversions which amounted to the total

number of misassemblies
(PDF)

S5 Table. BLAST output of intra-contig repeat from CLA result in simulated *S.* Typhi Ty2 data against the original Ty2 genome. The positions tabulated depict that a ~600bp intra-contig repeat is present at 26 different locations in the original genome
(PDF)

Acknowledgments

We would like to thank all the members of Pathogen Biology Laboratory for fruitful discussions and suggestions. NA is an Adjunct Professor of the Academy of Scientific and Innovative Research, India.

Author Contributions

Conceived and designed the experiments: NA SS NK. Performed the experiments: SS AKL NK SKT RB. Analyzed the data: SS AKL SKT RB. Contributed reagents/materials/analysis tools: NA. Wrote the paper: NA SS.

References

1. Ahmed N. A flood of microbial genomes-do we need more? PLoS One. 2009; 4(6):e5831. doi: 10.1371/journal.pone.0005831 PMID: 19513110
2. Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, Penn CW, et al. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat Rev Microbiol. 2012; 10(9):599–606. doi: 10.1038/nrmicro2850 PMID: 22864262
3. Schuster SC. Next-generation sequencing transforms today's biology. Nat Methods. 2008; 5(1):16–8. doi: 10.1038/nmeth1156 PMID: 18165802
4. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012; 13:341. doi: 10.1186/1471-2164-13-341 PMID: 22827831
5. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18(5):821–9. doi: 10.1101/gr.074492.107 PMID: 18349386
6. Warren RL, Sutton GG, Jones SJ, Holt RA. Assembling millions of short DNA sequences using SSAKE. Bioinformatics. 2007; 23(4):500–1. doi: 10.1093/bioinformatics/btl629 PMID: 17158514.
7. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics. 2009; 25(15):1968–9. doi: 10.1093/bioinformatics/btp347 PMID: 19497936
8. Kolmogorov M, Raney B, Paten B, Pham S. Ragout-a reference-assisted assembly tool for bacterial genomes. Bioinformatics. 2014; 30(12):i302–9. doi: 10.1093/bioinformatics/btu280 PMID: 24931998
9. Rissman AI, Mau B, Biehl BS, Darling AE, Glasner JD, Perna NT. Reordering contigs of draft genomes using the Mauve aligner. Bioinformatics. 2009; 25(16):2071–3. doi: 10.1093/bioinformatics/btp356 PMID: 19515959
10. Galardini M, Biondi EG, Bazzicalupo M, Mengoni A. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. Source Code Biol Med. 2011; 6:11. doi: 10.1186/1751-0473-6-11 PMID: 21693004
11. Silva GG, Dutilh BE, Matthews TD, Elkins K, Schmieder R, Dinsdale EA, et al. Combining de novo and reference-guided assembly with Scaffold_Builder. Source Code Biol Med. 2013; 8(1):23. doi: 10.1186/1751-0473-8-23 PMID: 24267787
12. Edwards DJ, Holt KE. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. Microb Inform Exp. 2013; 3(1):2. doi: 10.1186/2042-5783-3-2 PMID: 23575213
13. Doolittle WF, Zhaxybayeva O. On the origin of prokaryotic species. Genome Res. 2009; 19(5):744–56. doi: 10.1101/gr.086645.108 PMID: 19411599
14. Pop M, Kosack DS, Salzberg SL. Hierarchical scaffolding with Bambus. Genome Res. 2004; 14 (1):149–59. PMID: 14707177

15. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 2012; 1(1):18. doi: 10.1186/2047-217X-1-18 PMID: 23587118

16. Dayarian A, Michael TP, Sengupta AM. SOPRA: Scaffolding algorithm for paired reads via statistical optimization. BMC Bioinformatics. 2010; 11:345. doi: 10.1186/1471-2105-11-345 PMID: 20576136

17. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 2012; 22(3):549–56. doi: 10.1101/gr.126953.111 PMID: 22156294

18. Koren S, Treangen TJ, Pop M. Bambus 2: scaffolding metagenomes. Bioinformatics. 2011; 27 (21):2964–71. doi: 10.1093/bioinformatics/btr520 PMID: 21926123

19. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011; 27(4):578–9. doi: 10.1093/bioinformatics/btq683 PMID: 21149342

20. Hunt M, Newbold C, Berriman M, Otto TD. A comprehensive evaluation of assembly scaffolding tools. Genome Biol. 2014; 15(3):R42. doi: 10.1186/gb-2014-15-3-r42 PMID: 24581555

21. Karlsson E, Larkeryd A, Sjodin A, Forsman M, Stenberg P. Scaffolding of a bacterial genome using MinION nanopore sequencing. Sci Rep. 2015; 5:11996. doi: 10.1038/srep11996 PMID: 26149338

22. Mardis ER. Next-generation sequencing platforms. Annu Rev Anal Chem (Palo Alto Calif). 2013; 6:287–303.

23. Metzker ML. Sequencing technologies—the next generation. Nat Rev Genet. 2010; 11(1):31–46. doi: 10.1038/nrg2626 PMID: 19997069

24. Bosi E, Donati B, Galardini M, Brunetti S, Sagot MF, Lio P, et al. MeDuSa: a multi-draft based scaffolder. Bioinformatics. 2015; 31(15):2443–51. doi: 10.1093/bioinformatics/btv171 PMID: 25810435

25. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013; 29(8):1072–5. doi: 10.1093/bioinformatics/btt086 PMID: 23422339

26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215(3):403–10. PMID: 2231712

27. Ahmed N, Dobrindt U, Hacker J, Hasnain SE. Genomic fluidity and pathogenic bacteria: applications in diagnostics, epidemiology and intervention. Nat Rev Microbiol. 2008; 6(5):387–94. doi: 10.1038/nrmicro1889 PMID: 18392032

28. Aziz RK, Breitbart M, Edwards RA. Transposases are the most abundant, most ubiquitous genes in nature. Nucleic Acids Res. 2010; 38(13):4207–17. doi: 10.1093/nar/gkq140 PMID: 20215432

29. McElroy KE, Luciani F, Thomas T. GemSIM: general, error-model based simulator of next-generation sequencing data. BMC Genomics. 2012; 13:74. doi: 10.1186/1471-2164-13-74 PMID: 22336055

30. Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One. 2012; 7(2):e30619. doi: 10.1371/journal.pone.0030619 PMID: 22312429

31. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25(14):1754–60. doi: 10.1093/bioinformatics/btp324 PMID: 19451168

32. Floyd RW. Algorithm 97: Shortest path. Communications of the ACM. 1962; 5:345.

33. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. Genome Biol. 2012; 13(6): R56. doi: 10.1186/gb-2012-13-6-r56 PMID: 22731987

AMERICAN SOCIETY FOR MICROBIOLOGY | Antimicrobial Agents and Chemotherapy

Molecular Epidemiology and Genome Dynamics of New Delhi Metallo-β-Lactamase-Producing Extraintestinal Pathogenic *Escherichia coli* Strains from India

Amit Ranjan,[a] Sabiha Shaik,[a] Agnismita Mondal,[a] Nishant Nandanwar,[a] Arif Hussain,[a] Torsten Semmler,[b] Narender Kumar,[a] Sumeet K. Tiwari,[a] Savita Jadhav,[c] Lothar H. Wieler,[b] Niyaz Ahmed[a]

Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, University of Hyderabad, Gachibowli, Hyderabad, India[a]; Robert Koch Institute, Berlin, Germany[b]; Department of Microbiology, D. Y. Patil University, Pimpri, Pune, India[c]

The global dissemination and increasing incidence of carbapenem-resistant, Gram-negative organisms have resulted in acute public health concerns. Here, we present a retrospective multicenter study on molecular characterization of metallo-β-lactamase (MBL)-producing clinical *Escherichia coli* isolates recovered from extraintestinal infections in two hospitals in Pune, India. We screened a large sample size of 510 *E. coli* isolates for MBL production wherein we profiled their molecular determinants, antimicrobial resistance phenotypes, functional virulence properties, genomic features, and transmission dynamics. Approximately 8% of these isolates were MBL producers, the majority of which were of the NDM-1 (69%) type, followed by NDM-5 (19%), NDM-4 (5.5%), and NDM-7 (5.5%). MBL producers were resistant to all antibiotics tested except for colistin, fosfomycin, and chloramphenicol, which were effective to various extents. Plasmids were found to be an effective means of dissemination of NDM genes and other resistance traits. All MBL producers adhered to and invaded bladder epithelial (T24) cells and demonstrated significant serum resistance. Genomic analysis of MBL-producing *E. coli* isolates revealed higher resistance but a moderate virulence gene repertoire. A subset of NDM-1-positive *E. coli* isolates was identified as dominant sequence type 101 (ST101) while two strains belonging to ST167 and ST405 harbored NDM-5. A majority of MBL-producing *E. coli* strains revealed unique genotypes, suggesting that they were clonally unrelated. Overall, the coexistence of virulence and carbapenem resistance in clinical *E. coli* isolates is of serious concern. Moreover, the emergence of NDM-1 among the globally dominant *E. coli* ST101 isolates warrants stringent surveillance and control measures.

Infections with Gram-negative multidrug-resistant (MDR) pathogens, particularly those caused by β-lactamase-producing *Escherichia coli*, constitute a major reason for the "global SOS" in public health (1). Carbapenems are often considered to be the last-resort antibiotics available for treating infections caused by extended-spectrum-β-lactamase (ESBL)- or AmpC-producing bacteria (2) and are categorized in three major types: the KPC types, the metallo-β-lactamases (MBLs), and the oxacillinases (3). The major geographically widespread MBLs consist of imipenemase (IMP), Verona integron-encoded metallo-β-lactamase (VIM), and New Delhi metallo-β-lactamase (NDM); NDM has attracted significant worldwide attention as the NDM gene has been detected on different transferable plasmid types (both on typeable and untypeable) and has also been shown to be chromosomally integrated in other bacterial species (4, 5).

NDM belongs to the amber class B β-lactamase. Unlike other classes of β-lactamases (A, C, and D), NDM requires zinc ions to catalyze the hydrolysis of β-lactam antibiotics, and due to this, MBL activity can be inhibited in the presence of metal chelating agents such as EDTA (6). NDM-producing bacteria are mostly multidrug resistant and show resistance toward sulfonamides, fluoroquinolones, aminoglycosides, and macrolides but are susceptible to only a few antimicrobials, such as aztreonam, tigecycline, and colistin (7, 8). Recent reports suggest that this susceptibility is expected to be short-lived (1). Moreover, NDM enzymes are generally not inhibited by any of the known and available β-lactamase inhibitors such as sulbactam, clavulanate, and tazobactam (9).

NDM-producing bacteria have been implicated in both hospital- and community-acquired infections and have been recovered from several infection sites, including those associated with urinary tract infections, pneumonia, septicemia, and wound infections as well as from device-associated infections and companion animals (10–13). Although the exact geographical origin and the precise time for the emergence of the bla_{NDM} gene are unknown, the first NDM-1-positive isolate was recovered in 2009 from a Swedish patient who previously (2007) had an episode of hospitalization in New Delhi, India (14). Later, in the year 2010, Kumarasamy et al. provided evidence that NDM-1-positive *Enterobacteriaceae* are widespread in India, Pakistan, and the United Kingdom (15). Recent reports further indicate that several variants of NDM (NDM-1 to NDM-14)-producing *Enterobacteriaceae* are currently spreading worldwide, including in the United States, Netherlands, Australia, Canada, France, and Oman (16, 17) (http://www.lahey.org/studies). Most of these case reports/epidemiological studies provide evidence of connections of pa-

Received 23 June 2016 **Returned for modification** 21 July 2016
Accepted 25 August 2016
Accepted manuscript posted online 6 September 2016
Citation Ranjan A, Shaik S, Mondal A, Nandanwar N, Hussain A, Semmler T, Kumar N, Tiwari SK, 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. doi:10.1128/AAC.01345-16.
Address correspondence to Niyaz Ahmed, ahmed.nizi@gmail.com.
Copyright © 2016, American Society for Microbiology. All Rights Reserved.

tients to the Indian subcontinent in the form of migration/travel visits or medical treatment received there. However, certain reports indicate infection sources outside this region with patients without any obvious connection to this area (18). Nevertheless, the Indian subcontinent particularly offers several risk factors for the emergence and spread of NDM-producing superbugs, which include poor sanitation, easy access to broad-spectrum antibiotics over the counter without proper prescriptions, increased medical tourism for health care, and lack of stringent antibiotic policies.

Members of *Enterobacteriaceae*, particularly *E. coli*, are often disseminated globally through distinct and successful lineages, such as sequence types (ST) 131, 95, 10, etc. (19–22). Our previous studies have put forward strong evidence for the widespread occurrence of ST131 strains among extraintestinal infections in India and have also reported the existence of NDM-1-positive ST131 *E. coli*, thus representing a serious therapeutic challenge (23–27). This, together with the implication of NDM-producing strains globally, warrants serious attention of health care specialists and epidemiologists with respect to surveillance and resistance tracking in India and elsewhere so as to monitor threats locally and globally. Also, the data on the virulence traits associated with NDM-producing *E. coli* strains are currently scarce. With the advent of high-resolution, whole-genome sequencing methods, it has now become easy to decipher the genomes of the NDM-producing isolates and understand the underlying resistance and virulence mechanisms. The use of comparative genomics in infection tracking has enabled a comprehensive understanding of epidemiology, resistomes, virulomes, and the genetic background of NDM-producing MDR *E. coli* strains.

In the present study, we sought to determine the prevalence and phenotypes of MBL-producing strains from among a large number of *E. coli* isolates collected chronologically for 6 years (2009 to 2014) from a major tertiary hospital and to determine the antibiotic susceptibility patterns by taking into account the allelic diversity of bla_{NDM} genes and the plasmid replicon types. Further, we determined the virulence traits of NDM-producing *E. coli* by various phenotypic assays in conjunction with whole-genome sequences of five MDR NDM-producing *E. coli* strains and consequent analyses of resistomes, virulomes, and genetic affinities as important parameters required to portray a correct epidemiological scenario. We believe that such exhaustive analyses potentially provide for a clear and solid epidemiological baseline to guide an effective antibiotic policy in a country of endemicity such as India.

MATERIALS AND METHODS

Bacterial isolates. A total of 510 clinical isolates of *E. coli* were included in this study. The isolates were collected consecutively between January 2009 and December 2014 from in-patients and outpatients treated at two major tertiary care hospitals in Pune, India. The collection included clinical isolates from urine ($n = 393$), pus ($n = 88$), and other sources ($n = 29$). All the isolates were obtained as pure cultures and were cryopreserved at −80°C in 20% glycerol. The isolates were revived on LB agar plates and in LB broth prior to experiments. Collection of isolates was in accordance with the guidelines/approvals of Institutional Biosafety Committees at participating institutes/centers, where appropriate.

Detection of metallo-β-lactamase production and its molecular characterization. A total of 510 *E. coli* isolates were screened for MBL production using meropenem (with and without EDTA) with the help of Ezy MIC Strips (HiMedia; India), as per the manufacturer's instructions. All positive *E. coli* strains were also screened for the presence of bla_{NDM-1} by a gene-specific PCR (28). *E. coli* strains confirmed as MBL producers were further analyzed by high-fidelity PCR (with the use of Q5 DNA polymerase from NEB) using flanking primers of the NDM gene, as described earlier (29). Custom sequencing of the resulting amplicons was performed, and the nucleotide and translated protein sequences were analyzed using MEGA6 (30). NDM types were determined in accordance with the sequences available in the NCBI database(s).

Testing for antimicrobial susceptibility and its molecular determinants. Antimicrobial sensitivity of 39 strains against 23 different antibiotics was performed on Mueller-Hinton agar (MHA) plates by the Kirby-Bauer disc diffusion method using antibiotic discs (HiMedia) and Icosa urinary tract infection (UTI) rings (HiMedia), according to the manufacturer's protocols. Two in-house strains that were consistently resistant and susceptible to most of the antimicrobial agents were employed as positive and negative controls, respectively. Results were interpreted according to the criteria of the Clinical and Laboratory Standards Institute. For colistin, a zone of <11 mm was considered resistant. The MIC values of four frontline antibiotics, namely, tetracycline, co-trimoxazole, gentamicin, and ciprofloxacin, were determined by using HiComb MIC strips (HiMedia, India) as per the manufacturer's protocol. MIC values were determined and compared among a subset of 10 MBL-producing strains, 10 non-ESBL-producing strains, and 7 meropenem-resistant transconjugants. All MBL-producing *E. coli* strains were analyzed for molecular determinants of antibiotic resistance by PCR. DNA template was prepared using a standard heat lysis method. Thereafter, gene-specific primer sets were used to screen for the following ESBL genes: bla_{TEM}, bla_{SHV}, $bla_{CTX-M-15}$, bla_{OXA-48}, bla_{KPC}, and bla_{NDM-1} (28, 31). Other antibiotic resistance genes, such as those conferring resistance to tetracycline [*tet*(A) and *tet*(C)], sulfonamides (*sul1* and *sul2*), streptomycin (*strA*), and trimethoprim (*dfr*) and including the quinolone resistance determinants (carried on plasmids) such as *aac(6′)-Ib*, *qnrB*, and *qnrS*, were amplified as per standard PCR protocols (32, 33). The amplicons were run on 1% agarose gels, and the images were documented using a gel documentation system (Major Sciences, USA). The correlation between the phenotypic and different genotypic determinants of resistance was calculated with the help of observed phenotypic resistance and the frequency of the genotypic determinant probed.

Conjugation, mating experiments, and plasmid analysis. Conjugal transfer of resistance genes was performed by broth mating or filter mating experiments for a subset of nine carbapenem-resistant *E. coli* strains using the sodium azide-resistant recipient strain *E. coli* K-12 J53 (34). Transconjugants were selected on selection plates supplemented with a combination of meropenem (4 μg/ml) and sodium azide (50 μg/ml). In addition, other antibiotics and sodium azide combinations were used to select transconjugants. However, as meropenem-containing transconjugants were the focus of this study, the same were subjected to further analyses. Pure cultures of the transconjugants were preserved in 20% glycerol at −80°C. Antibiotic susceptibility and MICs were compared between the parental strains and their respective transconjugants. Plasmid classes were determined using PCR-based replicon typing (35).

Phenotypic assays. A total of nine random isolates were analyzed for functional virulence traits as described below. Nine MBL-producing *E. coli* strains were tested for their adhesion and invasion efficiency. Human bladder epithelial cell lines (T24) were cultured and maintained in RPMI 1640 medium (Invitrogen, USA) with 10% fetal bovine serum (FBS). Cells were seeded in 24-well plates and grown until a monolayer was obtained. Infection was done at a multiplicity of infection (MOI) of 10. After 3 h of incubation at 37°C with 5% CO_2, cells were washed three times with 1× phosphate-buffered saline (PBS), followed by lysis with 1 ml of 0.1% Triton X-100. Cell lysates were serially diluted, plated on LB agar plates, and incubated overnight at 37°C. For invasion assays, after 3 h of incubation, the cells were washed and further incubated for 1.5 h with 1 ml of RPMI medium containing 100 μg/ml of fosfomycin. After incubation, cells were washed three times with 1× PBS and processed in a way similar to that described above. Assays were performed three times with three technical replicates.

Serum resistance was determined for nine MBL-producing *E. coli* isolates by *in vitro* time-kill experiments as described earlier (22). Briefly, 5 μl of overnight-grown culture was added to 495 μl of LB broth and incubated for 1 h at 37°C in a shaker incubator. The culture was pelleted and resuspended in 1 ml of 1× PBS, and then 30 μl of this bacterial suspension was added to 96-well microtiter plate wells, each containing 270 μl of 50% sterile human serum (Pan Biotech, Germany) in 1× PBS. Thirty microliters of serum was then collected (0-h count), serially diluted, and plated on LB agar plates for CFU enumeration. The plate was then incubated for 3 h at 37°C. After the incubation time (3-h count), 30 μl of sample was pipetted for serial dilution and plated as described above. Growth was measured in terms of CFU counts. Samples were considered serum resistant if the 3-h CFU count equaled or exceeded the 0-h CFU count of the same sample. The serum assay was performed twice in triplicates.

Total siderophore production was tested by streaking all 39 MBL-positive *E. coli* samples on chrome azurol S (CAS) agar plates (blue medium). Color changes (haloes) from yellow to orange on the CAS agar medium plates, after overnight incubation at 37°C, were recorded. No color change in the medium corresponded to no siderophore production by the culture.

ERIC-PCR-based strain typing. Enterobacterial intergenic repetitive element sequence (ERIC)-PCR-based fingerprint analysis was performed as described previously (36). Strain-specific DNA banding profiles obtained on 1.5% agarose gel images were compared using BioNumerics software (version 7.1; Applied Maths, Belgium). After the bands were scored on dice similarity indices based on the unweighted-pair group method using average linkages (UPGMA; arithmetic mean), a dendrogram was generated in order to analyze similarity/diversity within carbapenem-resistant *E. coli* strains. The definition of ERIC clusters was based on a Dice similarity index.

WGS, assembly, annotation, and genomic analysis. Whole-genome sequencing (WGS) of five selected strains was carried out on an Illumina MiSeq system with an insert size of 400 to 500 bp. The paired-end sequence reads were filtered and trimmed using the NGS QC Toolkit (version 2.3.3) (37). These high-quality reads were assembled into contigs using SPAdes Genome Assembler (version 3.6.1) (38). The contigs were then ordered and scaffolded using an in-house workbench, Contig-Layout-Authenticator (39), and the resulting ordered scaffolds were joined by placing Ns between them to construct draft genomes. These were then submitted to the RAST (40) server for annotation, and the genome statistics were extracted using ARTEMIS (41). For all of the genomes, genes were predicted using GeneMarkS (42). Amino acid sequences of the products of the predicted genes from all strains were compared using blastp (43) against the database of *E. coli* virulence genes from the Virulence Factors Database (VFDB) (44). Only genes with 60% identity and 85% query coverage were considered present, and in this way the presence-absence status of each virulence gene in each of these strains was represented in the form of a heat map using R. A similar heat map was generated for putative resistance-related genes after comparison with the Comprehensive Antibiotic Resistance Database (CARD) (45), using blastp. Insertion sequence (IS) elements and putative phage sequences in the genome were identified using IS Finder (46) and PHAST (47), respectively. The sequence type (ST) of each of these strains was determined by submitting the contigs to the server of the Centre for Genomic Epidemiology (CGE) (https://cge.cbs.dtu.dk/services/MLST/).

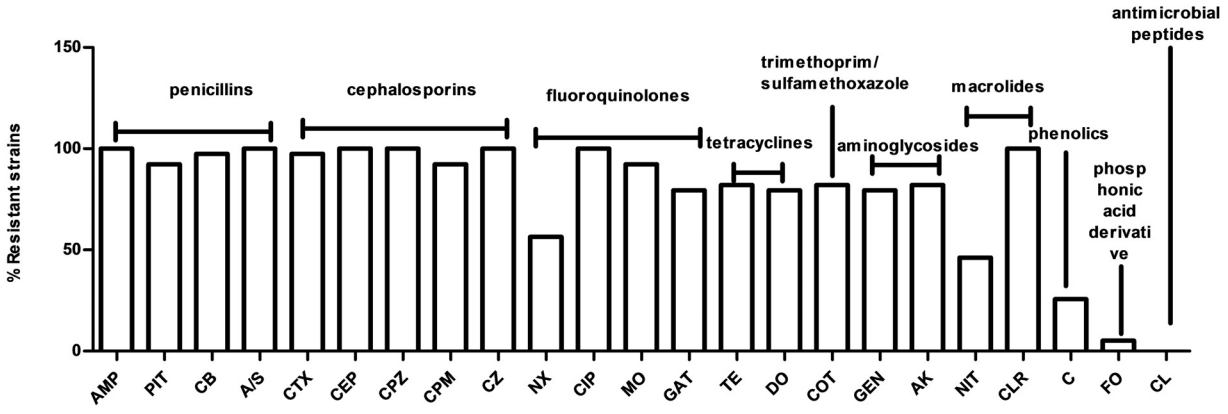
Statistical analysis. All statistical analyses were performed using GraphPad Prism (version 5.01). For adhesion and serum resistance, a nonparametric Mann-Whitney U test was performed while for the invasion assay Wilcoxon matched-pair tests were performed, and *P* values were determined.

Accession number(s). The results of this whole-genome shotgun project have been deposited in the DDBJ/ENA/GenBank databases under the accession numbers LXFJ00000000 (NA703), LXFK00000000 (NA724), LXEZ00000000 (NA086), LXFA00000000 (NA099), and LWTZ00000000 (NA084). The versions described in this paper are LXFJ01000000, LXFK01000000, LXEZ01000000, LXFA01000000, and LWTZ01000000, respectively.

RESULTS

Prevalence of MBL producers was low with a high abundance of the NDM-1 genotype. Of the 510 clinical *E. coli* isolates, 39 (7.6%) were found to be MBL producers. These 39 MBL-producing *E. coli* isolates were deemed to be of potential interest and were used further for in depth analyses, as described above. All of the 39 MBL-producing *E. coli* isolates were also found to be ESBL producers phenotypically. Most of the MBL producers (25/39) were from urine isolates, followed by those from pus (11/39) and other sources (3/39). The 39 MBL producers as identified at the phenotypic level were screened for the presence of major carbapenemase genes, including bla_{NDM}, bla_{KPC}, and bla_{OXA-48}. With the exception of the bla_{KPC} gene, which was not detected in any of the MBL producers, 36 (92%) MBL producers were bla_{NDM} positive, and 6 (15%) of them were also bla_{OXA-48} positive. Sequence analysis of the full-length NDM genes from 36 MBL producers revealed that the majority of them carried the bla_{NDM-1} gene (25 isolates), followed by bla_{NDM-5} (7 isolates). Further, the gene bla_{NDM-4} was present in two isolates, and the gene bla_{NDM-7} was harbored by two other MBL producers.

All MBL-producing *E. coli* isolates were identified to be MDR and pan-drug resistant. Of the 23 antimicrobial agents representing 10 different classes, our MBL-producing *E. coli* isolates showed an average resistance to nearly 18 antibiotics, with maximum resistance to 21 and minimum resistance to 10 antimicrobial agents. The frequencies of susceptibility/resistance are indicated in Fig. 1. Among all MBL-positive isolates, highest resistance was observed against the β-lactam antibiotic class while highest sensitivity (100%) was observed to the peptide antibiotic class, i.e., colistin. Each of the MBL-producing *E. coli* isolates was found to be at least 90% resistant to all the β-lactam and β-lactam/β-lactamase inhibitor combinations used, including ampicillin, carbenicillin, cefotaxime, cephalothin, cefoperazone, cefepime, cefazolin, piperacillin-tazobactam, and ampicillin-sulbactam. In contrast, each of the MBL-producing *E. coli* isolates was found to be sensitive to colistin. All MBL-positive isolates were highly susceptible to fosfomycin and chloramphenicol (Fig. 1). Other resistance patterns were as follows: 80% of isolates were resistant to each of the tetracycline and aminoglycoside classes and 82% were resistant to the trimethoprim/sulfonamide class. Notably, all MBL-positive isolates (100%) were identified to be MDR as they were resistant to three or more classes of antimicrobials. Moreover, all the MBL-producing *E. coli* isolates were found to be pan-drug resistant, meaning that they were resistant to as many as seven antimicrobial agents. The MIC values of MBL-ESBL-producing *E. coli* and non-ESBL *E. coli* isolates varied widely for different antibiotics used in the study (Table 1). Overall, the MBL-ESBL-producing *E. coli* isolates revealed higher MICs of all four frontline antimicrobial substances than non-ESBL strains. That the MIC values of the four antimicrobials for the meropenem transconjugants were higher than those of the recipient strain J53 *E. coli* indicated horizontal transmission of respective resistance genes. All 39 ESBL-MBL producers were negative for bla_{SHV} and bla_{KPC}. However, 31 isolates harbored the $bla_{CTX-M-15}$ gene, and 28 isolates harbored bla_{TEM}. The correlation among different resistance genotypes (presence or absence of resistance genes) and phenotypes (resistance or susceptibility) was found to be high for aminoglycosides (94%



Antibiotic class	Cephalosporins (beta lactams)						Fluoroquinolones			Tetracyclines		Sulphonamides-tri-methoprime			Amino glycosides
Resistance genotype	*ctx-m-15*	*tem*	*shv*	*ndm1*	*kpc*	*oxo48*	*aac6'*	*qnrB*	*qnrS*	*tetA*	*tetC*	*sul1*	*sul2*	*dfr*	*strA*
Number (% of positive strains)	31(79)	28(72)	0(0)	36(92)	0(0)	6(15)	26(67)	8(21)	7(18)	11(28)	22(56)	38(97)	32(82)	4(10)	30(77)

FIG 1 (Top) Antibiogram of MBL-producing carbapenem-resistant *E. coli* strains as tested by the disc diffusion method for the following 23 different antibiotics: ampicillin (Amp), piperacillin-tazobactam (PIT), carbenicillin (CB), ampicillin/sulbactam (A/S), cefotaxime (CTX), cephalothin (CEP), cefoperazone (CPZ), cefepime (CPM), cefazolin (CZ), nalidixic acid (NX), ciprofloxacin (CIP), moxifloxacin (MO), gatifloxacin (GAT), tetracycline (TE), doxycycline HCl (DO), co-trimoxazole (COT), gentamicin (GEN), amikacin (AK), nitrofurantoin (NIT), clarithromycin (CLR), chloramphenicol (C), fosfomycin (FO), and colistin (CL). All MBL-producing isolates were detected to be pan-drug resistant. (Bottom) Identification of resistance determinants among 39 MBL-producing *E. coli* isolates by gene-specific PCRs.

agreement) and sulfonamides (84% agreement), whereas the agreement between resistance to tetracycline and the presence of *tet*(A) and *tet*(C) was 69% and that for fluoroquinolone was (67%). The overall prevalences of all the major resistance genes in 39 MBL-producing *E. coli* isolates are presented in Fig. 1.

Metallo-β-lactamase resistance was naturally transferable via plasmids. NDM genes are basically located on plasmids. In order to determine the capability of these MBL-positive *E. coli* strains to disseminate NDM genes horizontally, we carried out conjugation mating experiments with the sodium azide-resistant recipient strain *E. coli* K-12 J53. Out of the nine representative MBL-positive strains employed, we were able to obtain seven meropenem-resistant transconjugants alongside several other transconjugants. However, only meropenem transconjugants were further analyzed for this study. The transconjugants were analyzed for their phenotypic drug resistance to four different classes of antibiotics, along with MBL and ESBL detection. All transconjugants were identified to be MBL and ESBL positive, with coresistance varying widely for different antibiotics used in the study. The detailed results are shown in Fig. 2. Transconjugants were confirmed to be carrying the NDM gene by gene-specific PCR. PCR-based replicon typing revealed the presence of primarily two plasmid types, the FIA and FIB plasmids, among transconjugants. These observations reflect the role of transmissible plasmids in dissemination of NDM-based carbapenem resistance.

MBL-producing *E. coli* isolates were moderately virulent. Phenotypic virulence properties of the MBL-positive *E. coli* isolates were assayed using adhesion, invasion, serum bactericidal, and siderophore production assays. The two former assays were carried out on nine randomly selected strains, and siderophore production was determined for all 39 strains. Adhesive and invasive virulence properties are important features during intestinal as well as extraintestinal infections. In order to evaluate this, we quantitatively assayed the pathogen's ability to adhere and invade the urinary bladder epithelial cells. Compared to the negative-control strain *E. coli* DH5α, all nine MBL-positive *E. coli* isolates showed significant adhesion to T24 cells with various efficiencies, as illustrated in Fig. 3a. Among the nine MBL-positive isolates tested, the NA084 isolate demonstrated highest adherence capability toward T24 cells. T24 cells were invaded by all investigated isolates, in contrast to the absence of invasion by the negative-control strain *E. coli* DH5α. Numbers of intracellular bacteria are depicted in Fig. 3b.

Resistance to serum is quite essential for the pathogens to survive and cause invasive infections; we found all nine tested isolates to be resistant (100%) to serum bactericidal activity compared to results with the control DH5α *E. coli*. Bacterial survival in serum over a period of 3 h is depicted in Fig. 3c. Though the virulence genes tested were not prevalent in these isolates, bacterial resistance toward the bactericidal effect of serum could result from individual or combined effects of capsular polysaccharide, surface proteins, and toxin secretions.

Siderophore production, as visualized by an iron mobilization zone around the bacterial colonies with orange haloes on the chrome azurol S agar, was similarly prevalent among 36 (92%) of

TABLE 1 MIC values of carbapenem-resistant strains, meropenem transconjugants, a recipient strain, and non-ESBL non-carbapenem-resistant *E. coli* isolates from the present study

Strain group and no.	MIC (μg/ml)[b]			
	Gentamicin	Ciprofloxacin	Co-trimoxazole	Tetracycline
Carbapenem-resistant strains ($n = 17$)[a]				
NA084	>240	>240	>240	3
NA644	>240	>240	>240	5
NA708	>240	60	>240	3
NA086	>240	>240	>240	2
NA086 T	>240	0.01	0.05	0.1
NA099	>240	>240	>240	0.1
NA099T	0.01	0.1	0.01	0.01
NA316	>240	>240	>240	3
NA316T	0.1	ND	0.05	0.1
NA635	>240	120	>240	5
NA635T	24	ND	>240	3
NA703	>240	>240	>240	0.001
NA703T	16	ND	0.05	0.001
NA724	>240	120	>240	3
NA724T	30	0.01	0.05	0.1
NA805	60	10	>240	0.01
NA805T	ND	1	0.05	0.01
Recipient strain ($n = 1$)				
K-12 J53	0.001	0.008	0.05	0.01
Non-ESBL non-carbapenem-resistant strains ($n = 10$)				
NA433	0.1	60	>240	10
NA010	0.1	0.008	30	0.1
NA295	0.01	0.01	120	0.1
NA053	0.25	1	>240	2
NA057	0.1	30	>240	0.1
NA120	0.01	0.1	>240	0.5
NA140	0.1	0.01	>240	30
NA150	0.1	0.25	2	0.01
NA203	0.01	0.01	0.1	0.01
NA274	0.1	10	>240	30

[a] The total includes seven meropenem transconjugants (T suffix).
[b] ND, not determined.

the 39 MBL-producing *E. coli* isolates. The role of iron in the virulence mechanism of human and animal pathogens is well established (48). Siderophore production has been shown to be more frequently associated with extraintestinal pathogenic *E. coli* (ExPEC).

Whole-genome analysis reveals high resistance and moderate virulence of MBL strains. Out of the 39 MBL-positive strains, whole-genome sequencing of five representative strains (three NDM-1 and two NDM-5) was carried out. Their genome charac-

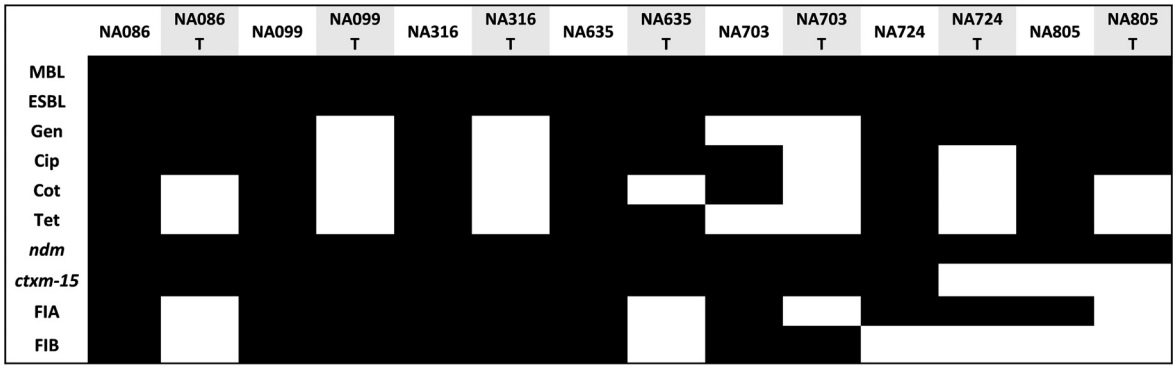


FIG 2 *In vitro* conjugation experiment and comparative analysis between clinical strains and their respective transconjugants revealed transfer of resistances via plasmids. The T suffix designates transconjugants. Gen, gentamicin; Cip, ciprofloxacin; Cot, co-trimoxazole; Tet, tetracycline. *ndm* and *ctxm-15* represent bla_{NDM} and $bla_{CTX\text{-}M\text{-}15}$ genes, respectively. FIA and FIB are the plasmid replicon types. Black and white boxes indicate presence and absence of corresponding traits, respectively.

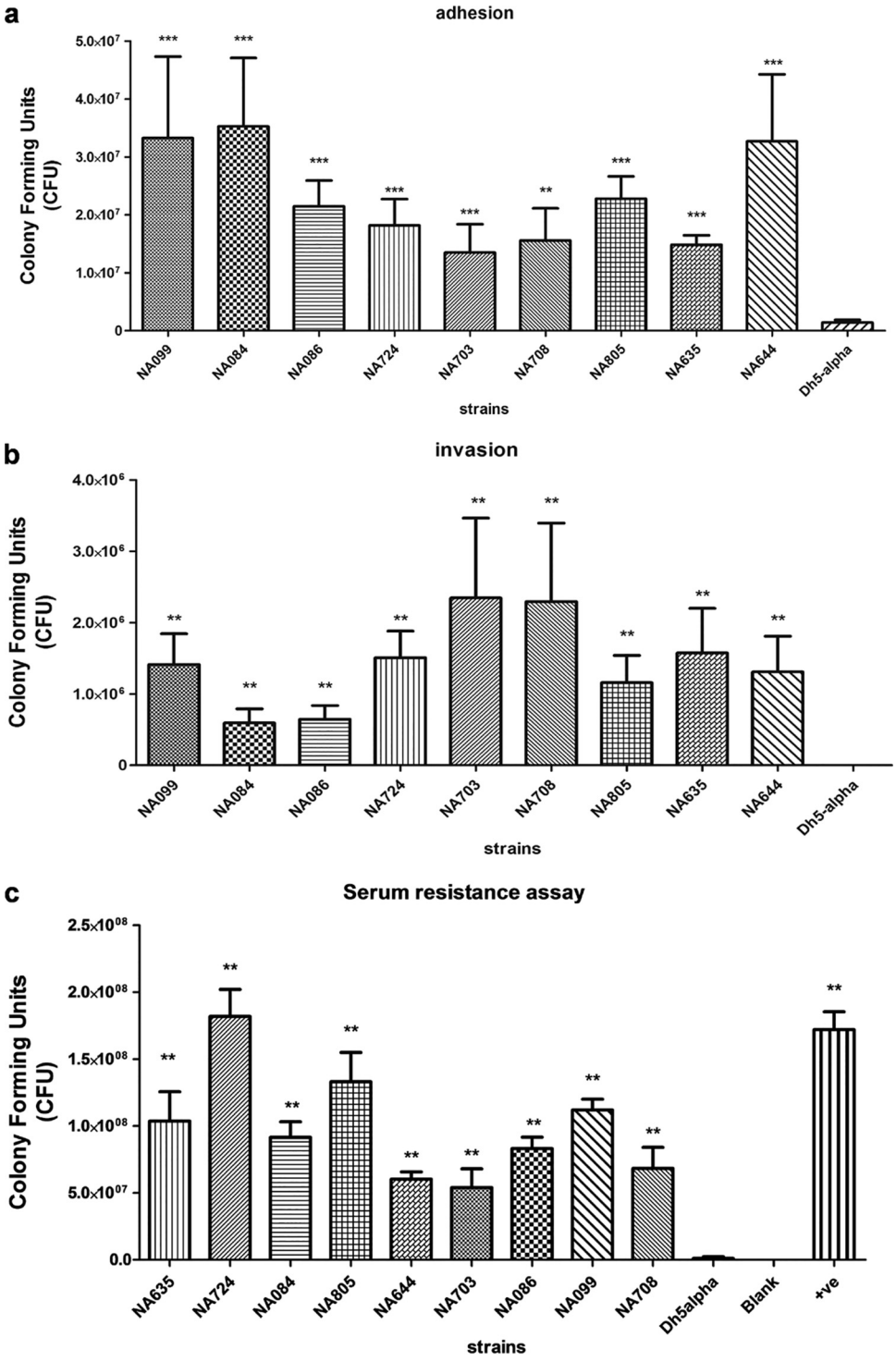


FIG 3 A subset of nine MBL-producing *E. coli* isolates was tested for three virulence-associated phenotypes: adhesion to human bladder epithelial cells (T24) at an MOI of 10 (a), invasion of human bladder epithelial cells (T24) (b), and resistance to 50% human serum for a 3-h time period (c). Adhesion and invasion assays were repeated three times in triplicates while the serum resistance assay was repeated twice in triplicates. All nine MBL-producing *E. coli* isolates demonstrated ExPEC-associated phenotypes. **, $P < 0.01$; ***, $P < 0.001$, for results compared to those with the control (DH5α).

teristics are represented in Table 2. The three NDM-1 strains belonged to ST101 while the two NDM-5 strains were characterized as ST167 and ST405. The whole-genome-based resistome reflected the high prevalence of resistance genes of all categories within all the strains, similar to the phenotypic drug resistance profile (Fig. 4). The virulome of these strains reflected moderate virulence wherein all the strains inherited genes belonging to adhesins, autotransporters, type three secretion systems

TABLE 2 The genome statistics and sequence types of five whole-genome-sequenced NDM-harboring strains comprising three NDM-1 and two NDM-5 strains

Strain	No. of raw reads	Genome coverage from filtered reads (×)	No. of contigs	ST	Genome size (no. of bases)	G+C content (%)	No. of CDS[a]	% coding capacity	No. of rRNAs	NDM type	Specimen source
NA084	1,741,426	73.25	109	101	5,147,001	50.59	5,000	86.8	13	NDM-1	Urine
NA086	1,643,674	63.13	115	101	5,147,281	50.59	4,994	86.9	13	NDM-1	Urine
NA099	2,406,806	104.68	155	101	5,197,495	50.57	5,063	85.8	17	NDM-1	Urine
NA703	2,401,628	92.94	211	167	5,263,381	50.66	5,186	86.0	16	NDM-5	Pus
NA724	2,784,910	118.36	166	405	5,424,442	50.55	5,273	87.0	9	NDM-5	Pus

[a] CDS, coding sequences.

(TTSS), and type 6 secretion systems (T6SS). NA099 harbors the least number of virulence genes while NA724 carries the maximum number of genes. The details are shown in Fig. 4. All the strains showed different phage contents and variable IS elements. The upstream sequence analysis of these five NDM strains showed insertion elements such as IS*Aba125* upstream of NDM-1 and IS*5* and truncated IS*Aba125* upstream of NDM-5 sequences.

NDM strains reveal high genomic diversity. PCR-based

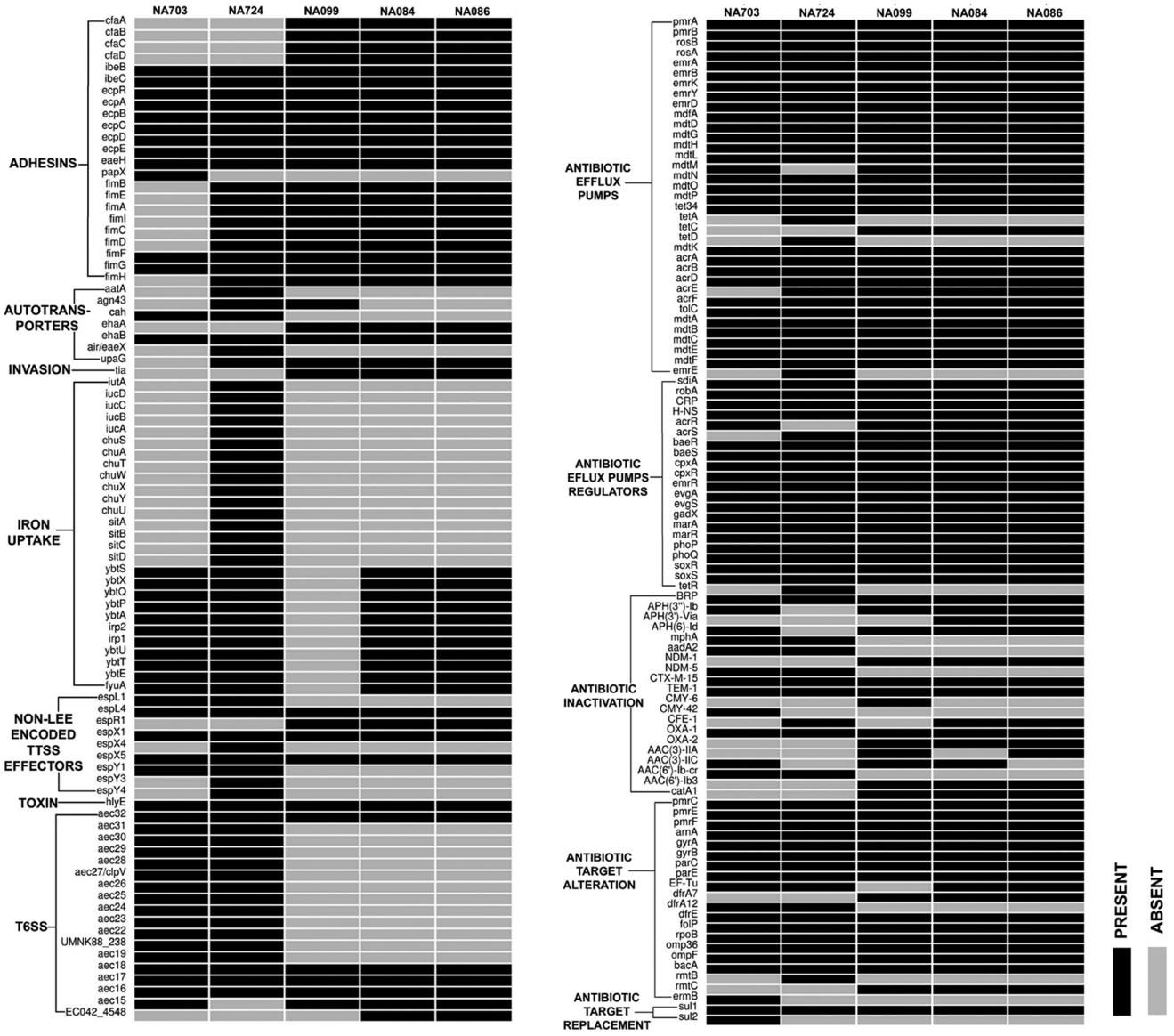


FIG 4 Whole-genome-based identification of virulence (left) and resistance (right) determinants of five MBL-producing *E. coli* isolates.

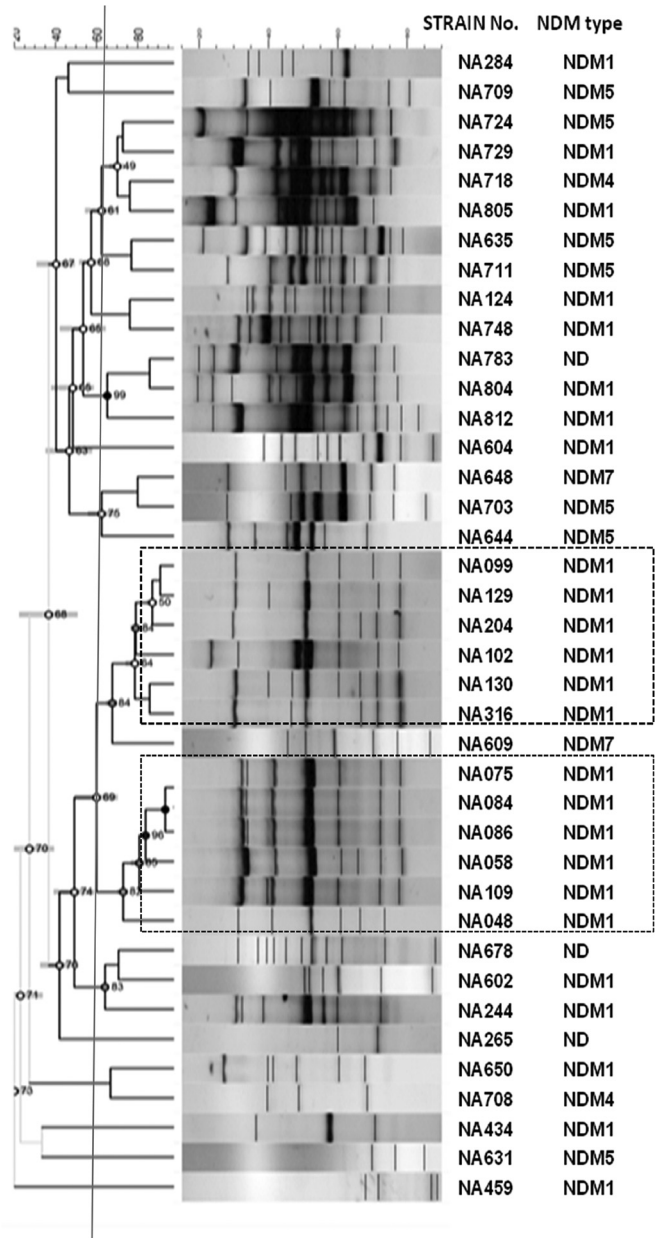


FIG 5 Dendrogram based on the phylogenetic analysis of 39 MBL-producing carbapenem-resistant clinical *E. coli* strains using ERIC-PCR banding analysis by BioNumerics, version 7.5 (Applied Maths, Belgium). Dotted boxes indicate clusters into which the majority of NDM-1 profiles segregated. ND, not determined.

genotyping was used to analyze the genetic relationships of the organisms. The ERIC-PCR-based dendrogram of these strains (Fig. 5) revealed high diversity among them. However, when strains were grouped at a lower threshold of ~65% identity, two diffusely clonal clades were observed. As expected, all the strains in this cluster harbored NDM-1 genes. Further, the *in silico* multilocus sequence typing (MLST) analysis of whole-genome sequences identified all three strains to be ST101. Overall, this cluster might include clonally related ST101 strains which are strongly associated with the NDM genotype. No such clusters were obtained for any other NDM variants. The above observations also support the idea that plasmids are the important means for NDM dissemination.

DISCUSSION

The emerging NDM, an acquired class B carbapenemase from among *Enterobacteriaceae*, has become a major public health concern worldwide also due to its medical and economic impact (49), and infections with NDM-positive pathogens pose a major health care problem, particularly for high-burden countries such as India. Previous studies from our group have shown an endemic *E. coli* infection burden and the presence of clonal group ST131 isolates with NDM-1 (23–25, 50). This study was aimed at detection, comprehensive characterization, and whole-genome sequence analysis of MBL-producing MDR organisms that can lead to successful infection. This study would likely be useful in guiding the formulation of control strategies involving antimicrobial stewardship and public health interventions in order to control the escalation of life-threatening MDR bacterial infections involving ExPEC organisms.

In general, production of MBL in *Enterobacteriaceae* shows an increasing prevalence pattern. However, the prevalence of carbapenem resistance among 510 *E. coli* isolates analyzed from 2009 to 2014 was found to be 7.6%, which was within the range of previously reported carbapenem resistance from India (51, 52); however, this prevalence might be variable in different geographical locations and different hospital settings. The dissemination of NDM carbapenemases involves several bla_{NDM} gene variants that are associated with different plasmid types among several Gram-negative species. A total of 14 NDM types have been reported globally, which were due to one or more point mutations (http://www.lahey.org/studies). In our study, we identified four NDM variants responsible for MBL production, with NDM-1 being the dominant variant, while other variants comprising NDM-4, NDM-5, and NDM-7 were observed with various frequencies. Previous reports from India have shown the presence of the NDM-6 type in addition to the above-mentioned NDM types (52). These four variants did not demonstrate any significant difference in MIC values for meropenem, nor did they show any specific association with different antimicrobial resistances. The carbapenems are considered the antibiotics of last resort for treatment of MDR bacterial infections; hence, bacteria which are resistant to these agents are often termed superbugs as they possess a resistance phenotype against different classes of broad-spectrum antibiotics. The antibiogram of these 39 isolates (Fig. 1) reflected a similar scenario wherein all MBL-producing strains (100%) revealed pan-drug resistance phenotypes as they were completely resistant to almost all the empirically used antibiotics. Nevertheless, all the strains were susceptible to colistin, and a majority of them were susceptible to fosfomycin. Polymyxins are the antibiotics of last resort to treat infections caused by MBL-producing *Enterobacteriaceae* (53), and as combination therapies are being advocated for treatment of MDR pathogens (54), the dual combination of colistin and fosfomycin could be the best option for effective treatment of carbapenem-resistant infections, which also minimizes the chances of resistance development.

The spread of antibiotic resistance genes is mediated by horizontal gene transfer (HGT) between bacteria (55). As multidrug-resistant plasmids serve as the most efficient means for dissemination of resistance traits via HGT, we also observed that the NDM genes and other resistances were transferred (via plasmids)

to recipient strains, as confirmed genotypically and phenotypically by our conjugation experiments (Fig. 2). ERIC-PCR-based clusters revealed the presence of NDM genes in *E. coli* isolates of mostly unique genotypes, suggesting that they are clonally unrelated and thereby have similar propensities to acquire the NDM genes (Fig. 5). Recent reports have shown a greater association of the NDM-1 genotype with sequence type 101, which is consistent with our findings as we also observed three NDM-1-bearing *E. coli* isolates belonging to ST101 (56, 57). We identified NDM-5 to be present in *E. coli* isolates belonging to ST405 and ST167. These sequence types were also linked to NDM dissemination worldwide (58–60). Further, we found that the NDM-1 sequences were preceded by the IS*Aba125* insertion element which belongs to the IS*30* family while NDM-5 has an IS*5* insertion element upstream of its sequence, followed by sequences of IS*Aba125*. IS*Aba125* is known to be located upstream of NDM in both *Acinetobacter* and *E. coli*, similar to our observations. It has also been reported that the IS*5* element is followed by a truncated IS*Aba125* sequence upstream of the NDM gene (61, 62).

The virulence phenotypes such as adhesion to and invasion of epithelial cells play an important role in initiation and establishment of infection. The virulence gene complement as detected at the whole-genome level (Fig. 4) and the adhesion/invasion (Fig. 3a and b) capabilities of the strains toward human urinary bladder epithelial cells demonstrate their pathogenic potentials and associated risks during infection as these strains are resistant to all empirically used antibiotics. Peirano et al. have also shown the ability of CTX-M- and NDM-positive strains to adhere to Caco2 and HepG2 cell lines (63). The ability to resist the bactericidal effect of serum complement as determined by serum resistance assays (Fig. 3c) points toward the enhanced survival advantages of these strains at different extraintestinal infection sites.

In summary, our findings demonstrated that the human extraintestinal pathogenic MBL-producing *E. coli* strains were associated with promiscuous plasmid types, resulting in a diverse range of clones harboring different variants of the NDM gene. Furthermore, whole-genome sequencing and genetic fingerprinting suggest that bla_{NDM-1} may also disseminate via dominant clones such as ST101. The genomes newly sequenced herein would be an important addition to the existing uropathogenic *E. coli* (UPEC)/ExPEC genome archives and would be helpful in future studies involving comparative genomics and molecular epidemiology. Given that the emergence of newer variants of NDM-1 in countries of endemicity and their association with different ExPEC infections represent an alarming problem, large-scale surveillance and molecular genotyping studies are needed in order to develop strategies to circumvent the emergence of dominant clones potentially leading to institutional outbreaks and nosocomial epidemics of the future.

ACKNOWLEDGMENTS

We acknowledge funding received from the Department of Biotechnology, Government of India [BT/HRD/NBA/34/01/2011(ix)], and support from the Indo-German International Research Training Group, Internationales Graduiertenkolleg (GRK1673), Functional Molecular Infection Epidemiology, an initiative of the German Research Foundation and the University of Hyderabad (India), for which N.A. and L.H.W. served as speakers. N.A. is an Adjunct Professor at the Academy of Scientific and Innovative Research, India. A.R. acknowledges the Indian Council of Medical Research for the award of a Junior Research Fellowship.

We thank Ramani Baddam for her help in discussing the data and their interpretation and help with the analysis of plasmid sequences. The Department of Biotechnology and Bioinformatics is supported by DST-FIST and UGC-SAP-DRS (level 1) programs.

FUNDING INFORMATION

This work, including the efforts of Lothar H. Wieler and Niyaz Ahmed, was funded by Deutsche Forschungsgemeinschaft (DFG) (GRK1673). This work, including the efforts of Niyaz Ahmed, was funded by Department of Biotechnology, Ministry of Science and Technology (DBT) (BT/HRD/NBA/34/01/2011(ix)).

REFERENCES

1. **Bonomo RA.** 2011. New Delhi metallo-beta-lactamase and multidrug resistance: a global SOS? Clin Infect Dis **52:**485–487. http://dx.doi.org/10.1093/cid/ciq179.
2. **Fair RJ, Tor Y.** 2014. Antibiotics and bacterial resistance in the 21st century. Perspect Medicin Chem **6:**25–64. http://dx.doi.org/10.4137/PMC.S14459.
3. **Queenan AM, Bush K.** 2007. Carbapenemases: the versatile beta-lactamases. Clin Microbiol Rev **20:**440–458. http://dx.doi.org/10.1128/CMR.00001-07.
4. **Poirel L, Dortet L, Bernabeu S, Nordmann P.** 2011. Genetic features of bla_{NDM-1}-positive *Enterobacteriaceae*. Antimicrob Agents Chemother **55:**5403–5407. http://dx.doi.org/10.1128/AAC.00585-11.
5. **Hu H, Hu Y, Pan Y, Liang H, Wang H, Wang X, Hao Q, Yang X, Yang X, Xiao X, Luan C, Yang Y, Cui Y, Yang R, Gao GF, Song Y, Zhu B.** 2012. Novel plasmid and its variant harboring both a bla_{NDM-1} gene and type IV secretion system in clinical isolates of *Acinetobacter lwoffii*. Antimicrob Agents Chemother **56:**1698–1702. http://dx.doi.org/10.1128/AAC.06199-11.
6. **Bush K, Jacoby GA.** 2010. Updated functional classification of beta-lactamases. Antimicrob Agents Chemother **54:**969–976. http://dx.doi.org/10.1128/AAC.01009-09.
7. **Peirano G, Schreckenberger PC, Pitout JD.** 2011. Characteristics of NDM-1-producing *Escherichia coli* isolates that belong to the successful and virulent clone ST131. Antimicrob Agents Chemother **55:**2986–2988. http://dx.doi.org/10.1128/AAC.01763-10.
8. **Tada T, Miyoshi-Akiyama T, Dahal RK, Sah MK, Ohara H, Kirikae T, Pokhrel BM.** 2013. NDM-8 metallo-beta-lactamase in a multidrug-resistant *Escherichia coli* strain isolated in Nepal. Antimicrob Agents Chemother **57:**2394–2396. http://dx.doi.org/10.1128/AAC.02553-12.
9. **Watkins RR, Papp-Wallace KM, Drawz SM, Bonomo RA.** 2013. Novel β-lactamase inhibitors: a therapeutic hope against the scourge of multidrug resistance. Front Microbiol **4:**392. http://dx.doi.org/10.3389/fmicb.2013.00392.
10. **Seija V, Medina Presentado JC, Bado I, Papa Ezdra R, Batista N, Gutierrez C, Guirado M, Vidal M, Nin M, Vignoli R.** 2015. Sepsis caused by New Delhi metallo-beta-lactamase (bla_{NDM-1}) and *qnr*D-producing *Morganella morganii*, treated successfully with fosfomycin and meropenem: case report and literature review. Int J Infect Dis **30:**20–26. http://dx.doi.org/10.1016/j.ijid.2014.09.010.
11. **Bogaerts P, Bouchahrouf W, de Castro RR, Deplano A, Berhin C, Pierard D, Denis O, Glupczynski Y.** 2011. Emergence of NDM-1-producing *Enterobacteriaceae* in Belgium. Antimicrob Agents Chemother **55:**3036–3038. http://dx.doi.org/10.1128/AAC.00049-11.
12. **Shaheen BW, Nayak R, Boothe DM.** 2013. Emergence of a New Delhi metallo-β-lactamase (NDM-1)-encoding gene in clinical *Escherichia coli* isolates recovered from companion animals in the United States. Antimicrob Agents Chemother **57:**2902–2903. http://dx.doi.org/10.1128/AAC.02028-12.
13. **Gupta N, Limbago BM, Patel JB, Kallen AJ.** 2011. Carbapenem-resistant *Enterobacteriaceae*: epidemiology and prevention. Clin Infect Dis **53:**60–67. http://dx.doi.org/10.1093/cid/cir202.
14. **Yong D, Toleman MA, Giske CG, Cho HS, Sundman K, Lee K, Walsh TR.** 2009. Characterization of a new metallo-β-lactamase gene, bla_{NDM-1}, and a novel erythromycin esterase gene carried on a unique genetic structure in *Klebsiella pneumoniae* sequence type 14 from India. Antimicrob Agents Chemother **53:**5046–5054. http://dx.doi.org/10.1128/AAC.00774-09.
15. **Kumarasamy KK, Toleman MA, Walsh TR, Bagaria J, Butt F, Balakrishnan R, Chaudhary U, Doumith M, Giske CG, Irfan S, Krish-**

nan P, Kumar AV, Maharjan S, Mushtaq S, Noorie T, Paterson DL, Pearson A, Perry C, Pike R, Rao B, Ray U, Sarma JB, Sharma M, Sheridan E, Thirunarayan MA, Turton J, Upadhyay S, Warner M, Welfare W, Livermore DM, Woodford N. 2010. Emergence of a new antibiotic resistance mechanism in India, Pakistan, and the UK: a molecular, biological, and epidemiological study. Lancet Infect Dis **10:**597–602. http://dx.doi.org/10.1016/S1473-3099(10)70143-2.

16. **Dortet L, Poirel L, Nordmann P.** 2014. Worldwide dissemination of the NDM-type carbapenemases in Gram-negative bacteria. Biomed Res Int **2014:**249856. http://dx.doi.org/10.1155/2014/249856.
17. **Zou D, Huang Y, Zhao X, Liu W, Dong D, Li H, Wang X, Huang S, Wei X, Yan X, Yang Z, Tong Y, Huang L, Yuan J.** 2015. A novel New Delhi metallo-β-lactamase variant, NDM-14, isolated in a Chinese Hospital possesses increased enzymatic activity against carbapenems. Antimicrob Agents Chemother **59:**2450–2453. http://dx.doi.org/10.1128/AAC.05168-14.
18. **Poirel L, Revathi G, Bernabeu S, Nordmann P.** 2011. Detection of NDM-1-producing *Klebsiella pneumoniae* in Kenya. Antimicrob Agents Chemother **55:**934–936. http://dx.doi.org/10.1128/AAC.01247-10.
19. **Petty NK, Ben Zakour NL, Stanton-Cook M, Skippington E, Totsika M, Forde BM, Phan MD, Gomes Moriel D, Peters KM, Davies M, Rogers BA, Dougan G, Rodriguez-Bano J, Pascual A, Pitout JD, Upton M, Paterson DL, Walsh TR, Schembri MA, Beatson SA.** 2014. Global dissemination of a multidrug resistant *Escherichia coli* clone. Proc Natl Acad Sci U S A **111:**5694–5699. http://dx.doi.org/10.1073/pnas.1322678111.
20. **Aibinu I, Odugbemi T, Koenig W, Ghebremedhin B.** 2012. Sequence type ST131 and ST10 complex (ST617) predominant among CTX-M-15-producing *Escherichia coli* isolates from Nigeria. Clin Microbiol Infect **18:**E49–E51. http://dx.doi.org/10.1111/j.1469-0691.2011.03730.x.
21. **Lau SH, Reddy S, Cheesbrough J, Bolton FJ, Willshaw G, Cheasty T, Fox AJ, Upton M.** 2008. Major uropathogenic *Escherichia coli* strain isolated in the northwest of England identified by multilocus sequence typing. J Clin Microbiol **46:**1076–1080. http://dx.doi.org/10.1128/JCM.02065-07.
22. **Nandanwar N, Janssen T, Kuhl 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. http://dx.doi.org/10.1016/j.ijmm.2014.06.009.
23. **Hussain A, Ranjan A, Nandanwar 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. http://dx.doi.org/10.1128/AAC.03320-14.
24. **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 **59:**6087–6095. http://dx.doi.org/10.1128/AAC.01447-15.
25. **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. http://dx.doi.org/10.1128/AAC.01099-12.
26. **Avasthi TS, Kumar N, Baddam R, Hussain A, Nandanwar N, Jadhav S, Ahmed N.** 2011. Genome of multidrug-resistant uropathogenic *Escherichia coli* strain NA114 from India. J Bacteriol **193:**4272–4273. http://dx.doi.org/10.1128/JB.05413-11.
27. **Nandanwar N, Hussain A, Ranjan A, Jadhav S, Ahmed N.** 2016. Population structure and molecular epidemiology of human clinical multidrug resistant (MDR) *Escherichia coli* strains from Pune, India. Int J Infect Dis **45 (Suppl 1):**343–344.
28. **Pfeifer Y, Wilharm G, Zander E, Wichelhaus TA, Gottig S, Hunfeld KP, Seifert H, Witte W, Higgins PG.** 2011. Molecular characterization of bla_{NDM-1} in an *Acinetobacter baumannii* strain isolated in Germany in 2007. J Antimicrob Chemother **66:**1998–2001. http://dx.doi.org/10.1093/jac/dkr256.
29. **Rasheed JK, Kitchel B, Zhu W, Anderson KF, Clark NC, Ferraro MJ, Savard P, Humphries RM, Kallen AJ, Limbago BM.** 2013. New Delhi metallo-beta-lactamase-producing *Enterobacteriaceae*, United States. Emerg Infect Dis **19:**870–878. http://dx.doi.org/10.3201/eid1906.121515.
30. **Tamura K, Stecher G, Peterson D, Filipski A, Kumar S.** 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol **30:**2725–2729. http://dx.doi.org/10.1093/molbev/mst197.
31. **Grobner S, Linke D, Schutz W, Fladerer C, Madlung J, Autenrieth IB, Witte W, Pfeifer Y.** 2009. Emergence of carbapenem-non-susceptible extended-spectrum beta-lactamase-producing *Klebsiella pneumoniae* isolates at the university hospital of Tubingen, Germany. J Med Microbiol **58:**912–922. http://dx.doi.org/10.1099/jmm.0.005850-0.
32. **Robicsek A, Strahilevitz J, Sahm DF, Jacoby GA, Hooper DC.** 2006. *qnr* prevalence in ceftazidime-resistant *Enterobacteriaceae* isolates from the United States. Antimicrob Agents Chemother **50:**2872–2874. http://dx.doi.org/10.1128/AAC.01647-05.
33. **Park CH, Robicsek A, Jacoby GA, Sahm D, Hooper DC.** 2006. Prevalence in the United States of *aac*(6′)-*lb-cr* encoding a ciprofloxacin-modifying enzyme. Antimicrob Agents Chemother **50:**3953–3955. http://dx.doi.org/10.1128/AAC.00915-06.
34. **Martinez-Martinez L, Pascual A, Jacoby GA.** 1998. Quinolone resistance from a transferable plasmid. Lancet **351:**797–799. http://dx.doi.org/10.1016/S0140-6736(97)07322-4.
35. **Carattoli A, Bertini A, Villa L, Falbo V, Hopkins KL, Threlfall EJ.** 2005. Identification of plasmids by PCR-based replicon typing. J Microbiol Methods **63:**219–228. http://dx.doi.org/10.1016/j.mimet.2005.03.018.
36. **Versalovic J, Koeuth T, Lupski JR.** 1991. Distribution of repetitive DNA sequences in eubacteria and application to fingerprinting of bacterial genomes. Nucleic Acids Res **19:**6823–6831. http://dx.doi.org/10.1093/nar/19.24.6823.
37. **Patel RK, Jain M.** 2012. NGS QC toolkit: a toolkit for quality control of next generation sequencing data. PLoS One **7:**e30619. http://dx.doi.org/10.1371/journal.pone.0030619.
38. **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. http://dx.doi.org/10.1089/cmb.2012.0021.
39. **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. http://dx.doi.org/10.1371/journal.pone.0155459.
40. **Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R.** 2014. The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). Nucleic Acids Res **42:**D206–D214. http://dx.doi.org/10.1093/nar/gkt1226.
41. **Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B.** 2000. Artemis: sequence visualization and annotation. Bioinformatics **16:**944–945. http://dx.doi.org/10.1093/bioinformatics/16.10.944.
42. **Besemer J, Lomsadze A, Borodovsky M.** 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res **29:**2607–2618.
43. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. J Mol Biol **215:**403–410. http://dx.doi.org/10.1016/S0022-2836(05)80360-2.
44. **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.
45. **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 LJ, 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. http://dx.doi.org/10.1128/AAC.00419-13.
46. **Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M.** 2006. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res **34:**D32–D36. http://dx.doi.org/10.1093/nar/gkj014.
47. **Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS.** 2011. PHAST: a fast phage search tool. Nucleic Acids Res **39:**W347–W352. http://dx.doi.org/10.1093/nar/gkr485.
48. **Gao Q, Wang X, Xu H, Xu Y, Ling J, Zhang D, Gao S, Liu X.** 2012. Roles

of iron acquisition systems in virulence of extraintestinal pathogenic *Escherichia coli*: salmochelin and aerobactin contribute more to virulence than heme in a chicken infection model. BMC Microbiol **12:**143. http://dx.doi.org/10.1186/1471-2180-12-143.

49. **Russo TA, Johnson JR.** 2003. Medical and economic impact of extraintestinal infections due to *Escherichia coli*: focus on an increasingly important endemic problem. Microbes Infect **5:**449–456. http://dx.doi.org/10.1016/S1286-4579(03)00049-2.
50. **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. http://dx.doi.org/10.1371/journal.pone.0018063.
51. **Bora A, Ahmed GU, Hazarika NK, Prasad KN, Shukla SK, Randhawa V, Sarma JB.** 2013. Incidence of bla_{NDM-1} gene in *Escherichia coli* isolates at a tertiary care referral hospital in Northeast India. Indian J Med Microbiol **31:**250–256. http://dx.doi.org/10.4103/0255-0857.115628.
52. **Rahman M, Shukla SK, Prasad KN, Ovejero CM, Pati BK, Tripathi A, Singh A, Srivastava AK, Gonzalez-Zorn B.** 2014. Prevalence and molecular characterisation of New Delhi metallo-beta-lactamases NDM-1, NDM-5, NDM-6 and NDM-7 in multidrug-resistant *Enterobacteriaceae* from India. Int J Antimicrob Agents **44:**30–37. http://dx.doi.org/10.1016/j.ijantimicag.2014.03.003.
53. **Morrill HJ, Pogue JM, Kaye KS, LaPlante KL.** 2015. Treatment options for carbapenem-resistant *Enterobacteriaceae* infections. Open Forum Infect Dis **2:**ofv050. http://dx.doi.org/10.1093/ofid/ofv050.
54. **Albur MS, Noel A, Bowker K, MacGowan A.** 2015. The combination of colistin and fosfomycin is synergistic against NDM-1-producing *Enterobacteriaceae* in *in vitro* pharmacokinetic/pharmacodynamic model experiments. Int J Antimicrob Agents **46:**560–567. http://dx.doi.org/10.1016/j.ijantimicag.2015.07.019.
55. **Dzidic S, Bedekovic V.** 2003. Horizontal gene transfer-emerging multidrug resistance in hospital bacteria. Acta Pharmacol Sin **24:**519–526.
56. **Toleman MA, Bugert JJ, Nizam SA.** 2015. Extensively drug-resistant New Delhi metallo-beta-lactamase-encoding bacteria in the environment, Dhaka, Bangladesh, 2012. Emerg Infect Dis **21:**1027–1030. http://dx.doi.org/10.3201/eid2106.141578.
57. **Yoo JS, Kim HM, Koo HS, Yang JW, Yoo JI, Kim HS, Park HK, Lee YS.** 2013. Nosocomial transmission of NDM-1-producing *Escherichia coli* ST101 in a Korean hospital. J Antimicrob Chemother **68:**2170–2172. http://dx.doi.org/10.1093/jac/dkt126.
58. **Zhang LP, Xue WC, Meng DY.** 2016. First report of New Delhi metallo-beta-lactamase 5 (NDM-5)-producing *Escherichia coli* from blood cultures of three leukemia patients. Int J Infect Dis **42:**45–46. http://dx.doi.org/10.1016/j.ijid.2015.10.006.
59. **Yang P, Xie Y, Feng P, Zong Z.** 2014. bla_{NDM-5} carried by an IncX3 plasmid in *Escherichia coli* sequence type 167. Antimicrob Agents Chemother **58:**7548–7552. http://dx.doi.org/10.1128/AAC.03911-14.
60. **Mushtaq S, Irfan S, Sarma JB, Doumith M, Pike R, Pitout J, Livermore DM, Woodford N.** 2011. Phylogenetic diversity of *Escherichia coli* strains producing NDM-type carbapenemases. J Antimicrob Chemother **66:**2002–2005. http://dx.doi.org/10.1093/jac/dkr226.
61. **Sole M, Pitart C, Roca I, Fabrega A, Salvador P, Munoz L, Oliveira I, Gascon J, Marco F, Vila J.** 2011. First description of an *Escherichia coli* strain producing NDM-1 carbapenemase in Spain. Antimicrob Agents Chemother **55:**4402–4404. http://dx.doi.org/10.1128/AAC.00642-11.
62. **Krishnaraju M, Kamatchi C, Jha AK, Devasena N, Vennila R, Sumathi G, Vaidyanathan R.** 2015. Complete sequencing of an IncX3 plasmid carrying bla_{NDM-5} allele reveals an early stage in the dissemination of the bla_{NDM} gene. Indian J Med Microbiol **33:**30–38. http://dx.doi.org/10.4103/0255-0857.148373.
63. **Peirano G, Mulvey GL, Armstrong GD, Pitout JD.** 2013. Virulence potential and adherence properties of *Escherichia coli* that produce CTX-M and NDM beta-lactamases. J Med Microbiol **62:**525–530. http://dx.doi.org/10.1099/jmm.0.048983-0.

AAC Journals.ASM.org

Genomic and Functional Portrait of a Highly Virulent, CTX-M-15-Producing *H*30-Rx Subclone of *Escherichia coli* Sequence Type 131

Amit Ranjan,[a] Sabiha Shaik,[a] Arif Hussain,[a] Nishant Nandanwar,[a] Torsten Semmler,[b,d] Savita Jadhav,[c] Lothar H. Wieler,[b,d] Niyaz Ahmed[a]

Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, University of Hyderabad, Gachibowli, India[a]; Centre for Infection Medicine, Institute of Microbiology and Epizootics, Freie Universität Berlin, Berlin, Germany[b]; Department of Microbiology, D. Y. Patil University, Pimpri, India[c]; Robert Koch Institute, Berlin, Germany[d]

***Escherichia coli* sequence type 131 (ST131) is a pandemic clone associated with multidrug-resistant, extraintestinal infections, attributable to the presence of the CTX-M-15 extended-spectrum β-lactamase gene and mutations entailing fluoroquinolone resistance. Studies on subclones within *E. coli* ST131 are critically required for targeting and implementation of successful control efforts. Our study comprehensively analyzed the genomic and functional attributes of the *H*30-Rx subclonal strains NA097 and NA114, belonging to the ST131 lineage. We carried out whole-genome sequencing, comparative analysis, phenotypic virulence assays, and profiling of the antibacterial responses of THP1 cells infected with these subclones. Phylogenomic analysis suggested that the strains were clonal in nature and confined entirely to a single clade. Comparative genomic analysis revealed that the virulence and resistance repertoires were comparable among the *H*30-Rx ST131 strains except for the commensal ST131 strain SE15. Similarly, seven phage-specific regions were found to be strongly associated with the *H*30-Rx strains but were largely absent in the genome of SE15. Phenotypic analysis confirmed the virulence and resistance similarities between the two strains. However, NA097 was found to be more robust than NA114 in terms of virulence gene carriage (*dra* operon), invasion ability ($P < 0.05$), and antimicrobial resistance (streptomycin resistance). RT^2 gene expression profiling revealed generic upregulation of key proinflammatory responses in THP1 cells, irrespective of ST131 lineage status. In conclusion, our study provides comprehensive, genome-inferred insights into the biology and immunological properties of ST131 strains and suggests clonal diversification of genomic and phenotypic features within the *H*30-Rx subclone of *E. coli* ST131.**

Extraintestinal infections due to *Escherichia coli* cause significant morbidity and death and increased health care costs (1, 2). Management of such infections is difficult due to the increasing abundance and wide spectrum of antimicrobial resistance, ranging from fluoroquinolones to extended-spectrum cephalosporins and to the "last resort" antimicrobials (e.g., carbapenems) (3).

Many multidrug-resistant (MDR) bacterial strains are now being reported, especially those belonging to clones capable of rapid global dissemination (4, 5). *H*30 sequence type 131 (ST131) is one predominant *E. coli* clone or lineage that is responsible for large proportions of community- and hospital-acquired urinary tract infections (UTIs) and bloodstream infections throughout the world (6–8). The *E. coli H*30-Rx subclone has emerged with strong associations with fluoroquinolone resistance and CTX-M-15 extended-spectrum β-lactamase (ESBL) production (9); *H*30 denotes the presence of allele 30 for *fimH* (type 1 fimbrial adhesin gene), R indicates its association with fluoroquinolone resistance, and x conveys its carriage of $bla_{\text{CTX-M-15}}$. Clonal expansion has now become a dominant mechanism in the proliferation of antimicrobial resistance in ST131 (10). Alarmingly, ST131 strains harboring New Delhi metallo-β-lactamase 1 (NDM-1), which confers carbapenem resistance, have also been reported among *E. coli* strains in particular and *Enterobacteriaceae* in general (11). ST131 *E. coli* strains not only are prevalent in human populations but also demonstrate interspecies transmission (12–14).

The epidemiological success of *E. coli* ST131 has been attributed to a wide variety of factors, including enhanced virulence and metabolic capabilities, the ability to acquire extensive and newer resistance traits, and better survivability in intestinal and extraintestinal sites (15, 16). However, the exact reasons for the rapid emergence and successful spread of ST131 clones remain largely undefined. Whole-genome sequencing (WGS) has revolutionized the study of bacterial evolution. In order to understand the recent evolutionary history of the most important *E. coli* strains in circulation, it is necessary to analyze their diversity in terms of the mutations, recombinations, and gene gains and losses in the population, including their profiles of bacteriophage contents (17).

Although ST131 is recognized as a pandemic clone involved in serious extraintestinal pathogenic *Escherichia coli* (ExPEC) infections, posing grave public health risks (15), it has received meager attention in countries such as India, despite tremendous host diversity, poor sanitation, and widespread abuse of antibiotics with an unclear antibiotic policy. Despite ongoing research on the epidemiological significance of ST131, little is known regarding the host responses and immunological profiles during infection for members of this lineage.

Received 19 June 2015 **Returned for modification** 28 June 2015
Accepted 12 July 2015

Accepted manuscript posted online 20 July 2015

Citation 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 *H*30-Rx subclone of *Escherichia coli* sequence type 131. Antimicrob Agents Chemother 59:6087–6095. doi:10.1128/AAC.01447-15.

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

Supplemental material for this article may be found at http://dx.doi.org/10.1128/AAC.01447-15.

Copyright © 2015, American Society for Microbiology. All Rights Reserved.

doi:10.1128/AAC.01447-15

To gain insight into its pathogenicity and to identify distinct traits associated with the high-risk *H*30-Rx subclone, we sought to investigate the whole-genome sequence of one of the representative ST131 strains (NA097) and to perform comparative genomic analysis using several reference genomes; we characterized the ST131 strain NA097 because of its functional and genomic attributes, including antimicrobial resistance and pathogenicity. Additionally, we performed profiling of innate immune responses to *E. coli* infections in differentiated THP1 macrophages with an adherent and invasive ST131 *E. coli* strain (NA097). The results of this study support the idea that expansion of the *H*30-Rx sublineage has contributed substantially to the dissemination of ST131 *E. coli*, and they provide a snapshot of the innate immune signatures of the host against ST131 infections.

MATERIALS AND METHODS

ST131 strains and genome sequencing. The *E. coli* strain NA097 was isolated from a urine sample from a 29-year-old male patient with septicemia who presented to the D. Y. Patil Hospital (Pimpri, India) in 2009. Strain NA114 was isolated in 2009 and was sequenced and reported previously by our group (18). Strain NA097 was characterized with routine biochemical techniques and was stored at −80°C in Luria-Bertani (LB) broth supplemented with 20% glycerol. High-quality genomic DNA was obtained using the Qiagen DNeasy blood and tissue kit (Qiagen, Germany). Whole-genome sequencing (WGS) was carried out with an Illumina MiSeq sequencer, which generated 1.3 million paired-end reads of 251 bp, with an insert size of 400 to 500 bp. After filtering and trimming of the paired-end reads using the NGS QC Toolkit (http://www.nipgr.res.in/ngsqctoolkit.html), the high-quality reads amounted to genome coverage of approximately 34 times. The reads were then assembled into contigs using the Velvet *de novo* assembler (19), run with an optimum hash length of 77. The contigs were then screened for the possible presence of plasmids based on their alignment with the online *E. coli* sequence databases, with parameters such as 70% identity and query coverage, using BLASTn (20). Only chromosomal contigs were ordered and scaffolded using an in-house workbench, C-L-Authenticator, to avoid misscaffolding due to plasmid-related contigs. The generated scaffolds and plasmid contigs were used for further analysis. The draft genome was then submitted to the RAST server (21) for annotation, and the genome statistics from the resulting file were extracted using ARTEMIS (22). Gene prediction was carried out using GeneMarkS (23). The numbers of tRNAs and rRNAs were identified using tRNAscan-SE (24) and RNAmmer (25), respectively.

Genomic data and comparative genomic analysis. A total of 107 whole-genome sequences, including those of ST131 strains NA097, NA114, JJ1886, and EC958 from the NCBI and other genomes described by Petty et al. (4), were used to construct a core-genome-based phylogenetic tree using Harvest (26), with the complete genome of the SE15 strain taken as a reference. The resulting tree was visualized using the SplitsTree program (http://www.splitstree.org). BRIG (27) was used to visualize the genome variations within clade C, including NA097 along with other reference ST131 genomes from the NCBI. For all of the genomes, predicted protein sequences from GeneMarkS were searched against the database of *E. coli* virulence genes downloaded from the Virulence Factors Database (VFDB) (28) by using BLASTp. Only hits with 80% identity and 70% query coverage were considered positive hits. A comparative table was created based on the presence or absence of these virulence genes in all of the genomes compared. A similar table was generated for putative resistance-related genes after comparison with the Comprehensive Antibiotic Resistance Database (CARD) (29), using BLASTp. PhiSpy (30) and PHAST (31) were used to identify the putative phage sequences in the genome. Insertion sequence (IS) elements and genomic islands were identified using IS finder (32) and IslandViewer (33), respectively. Predicted protein sequences were also searched against the Conserved Domains Database (CDD) (34) using NCBI Batch CD-Search, for functional categorization. In-house scripts were used to extract the best hits and corresponding cluster of orthologous groups (COG) classes from the results.

Molecular characterization and susceptibility testing. *E. coli* phylogenetic groups (A, B1, B2, and D) were determined with a well-established PCR method, as described previously (35). The isolates were preliminarily assessed for ST131 status with an allele-specific PCR assay for *pabB* gene detection (36). Antimicrobial susceptibility was examined with the standard Kirby-Bauer method, using an Icosa UTI-1 antibiotic ring (HiMedia, Mumbai, India), for 20 different antibiotics. ESBL production was detected using the double-disc synergy test. The results of both tests were interpreted according to Clinical and Laboratory Standards Institute (CLSI) criteria. Metallo-β-lactamase (MBL) production was screened with meropenem and meropenem-EDTA Etest strips (HiMedia, Mumbai, India), according to the manufacturer's directions. Detection of ESBL ($bla_{\text{CTX-M-15}}$, bla_{TEM}, and bla_{SHV}) and MBL ($bla_{\text{NDM-1}}$) genes was carried out by PCR (15). Plasmids were determined using Kado lysis, and replicon types were identified using five multiplex and three single PCR assays for the presence of a total of 18 plasmid types (37).

Phenotypic characterization of strains. Serum resistance assays were performed in 50% human serum (Pan Biotech) as described previously (38), with some modifications. Briefly, 5 μl of overnight bacterial culture was added to 495 μl of fresh LB medium and grown for 1 h. After centrifugation, the bacteria were resuspended in 1 ml of sterile phosphate-buffered saline (PBS), and 30 μl of culture with 270 μl of 50% human serum was added in triplicate to 96-well microtiter plates. An aliquot of 30 μl of each culture was serially diluted and plated (0-h count), and the rest was incubated for 3 h at 37°C. After incubation, the bacteria were serially diluted and plated (3-h count). Counting was performed after an overnight incubation, and CFU were calculated. The initial CFU value was subtracted from the final count in order to obtain growth in serum, and data were plotted on a graph. Hemagglutination assays were carried out with freshly isolated red blood cells (RBCs) from human blood, and results were interpreted as described elsewhere (39). Biotyping of strains for carbohydrate utilization of 35 sugar sources was performed using the Hi-Carbo kit (HiMedia, Mumbai, India), according to the manufacturer's protocol. Biofilm formation was examined in a total of nine replicates in 96-well microtiter plates with M63 medium, after 48 h, as described previously (8).

Epithelial cell adhesion and invasion assays. Adhesion and invasion assays were performed as described previously (38), with some modifications. T24 and ACHN cells (human bladder and kidney epithelial cell lines) were grown in RPMI 1640 medium and Ham's F-12 medium (Invitrogen) supplemented with 10% and 15% fetal bovine serum (FBS) (Invitrogen), respectively. Cells were grown in 5% CO_2 at 37°C. Cells were seeded into 24-well plates 48 h prior to cell infection assays. Infection was performed at a multiplicity of infection (MOI) of 10. Following 3 h of incubation at 37°C, the monolayer was washed three times using 1× PBS (Gibco) and subsequently lysed with 0.1% Triton X-100 in 1× PBS for 10 min at 37°C. The serially diluted lysates were plated on LB agar plates for determination of CFU of adherent bacteria. For invasion assays, the cells were incubated for 3 h and the monolayer was washed three times with 1× PBS and incubated for 1.5 h with 1 ml of fresh Ham's F-12 medium containing 100 μg/ml gentamicin. After incubation, the cells were lysed with 0.1% Triton X-100 in PBS for 10 min at 37°C. Lysates were serially diluted and CFU were determined as in the adhesion assays described above.

Infection of cultured THP1 cells and RT^2 profiling of antibacterial responses. Monocyte cells were cultured and maintained at 37°C in RPMI 1640 medium (Invitrogen) supplemented with 10% FBS (Invitrogen), in 5% CO_2. Prior to infection, about 1×10^6 cells were seeded into 35-mm-diameter petri dishes (Corning) and stimulated for 36 h using phorbol myristate acetate (PMA). After stimulation, cells were incubated for 12 h in antibiotic-free medium. Subsequently, *E. coli* strains (NA097 and DH5α) were grown overnight, followed by secondary growth to an optical

density at 600 nm of 0.95. *E. coli* was harvested and added to the differentiated THP1 cells at an MOI of 10. Similarly, another set of 1×10^6 cells was treated with 1 μg lipopolysaccharide (LPS). Infected cells were incubated for 3 h. Total RNA was isolated using a GenElute kit (Sigma), and cDNA was synthesized from 1 μg of DNase (Sigma)-treated RNA, using a first-strand cDNA synthesis kit (Qiagen, Germany). Real-time PCR was performed on cDNA using the RT^2 Profiler PCR array system (Qiagen), according to the manufacturer's instructions, in an Eppendorf RealPlex 4 system. The experiments were performed in triplicate, and the data were analyzed using the online RT^2 Profiler array data analysis tool (version 3.5) from Qiagen.

Statistical analysis. Statistical analyses for biofilm assays, serum bactericidal assays, adhesion assays, and invasion assays were carried out using the nonparametric Mann-Whitney *U* test, using GraphPad Prism 5.01. For calculations with quantitative real-time (qRT)-PCR data, custom online software was used as directed in the PAHS-148Z RT^2 Profiler manual (Qiagen), and calculations were performed according to the manufacturer's instructions. *P* values of <0.05 were considered to be significant.

Nucleotide sequence accession numbers. This whole-genome shotgun project has been deposited in DDBJ/EMBL/GenBank under accession number JSXJ00000000. The version described in this paper is version JSXJ01000000.

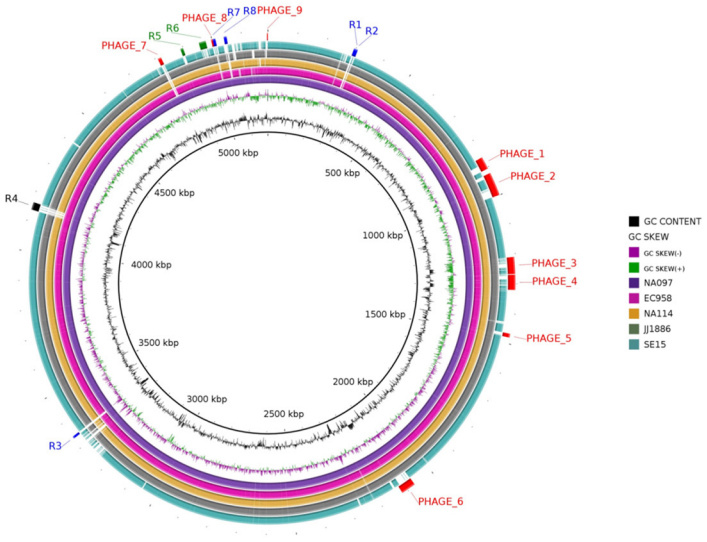
RESULTS

Genome statistics and comparative genomic analysis. The genome of NA097 comprised 187 contigs, constituting an approximate size of 5.1 Mb. The genome was predicted to encode ~5,164 protein coding sequences, with an average coding sequence length of 885 bp, amounting to a coding capacity of 88%. The genome revealed 51.83% G+C, 73 tRNAs, and 3 rRNA genes. The functional annotation of the NA097 genome using the COG database revealed a majority of genes related to metabolism (37%), followed by cellular processing and signaling (21%) and information storage and processing (17%). A fraction of the genes (14%) belonged to multiple classes, while 10% remained poorly characterized (see Fig. S2 in the supplemental material). *In silico* multilocus sequence typing (MLST) analysis using the MLST database (http://mlst.ucc.ie/mlst/mlst/dbs/Ecoli) showed that this organism belongs to sequence type 131. We identified 13 families of insertion sequence (IS) elements within the NA097 genome, including at least one copy each of IS*21*, IS*66*, IS*6*, IS*1*, IS*30*, Tn*3*, IS*1380*, IS*110*, IS*3*, ISL*3*, IS*4*, IS*200*, and IS*As1*.

Whole-genome sequence-based phylogenetic analysis of ST131 strains together with ST95 strains revealed three different clades (clade A, clade B, and clade C), as described in Fig. S1 in the supplemental material. The strains NA097 and NA114 clustered within clade C, which predominantly included strains of *H*30 and *H*30-Rx subclones of ST131 *E. coli*. The outcomes of uropathogenic *E. coli* (UPEC)-specific virulence and resistance gene profiling for five ST131 genomes are summarized in Tables S1 and S2 in the supplemental material. The virulence and resistance repertoires were comparable among the ST131 strains except for SE15, which had few virulence and fewer resistance genes. NA097 primarily revealed the *dra* operon, in contrast to the pyelonephritis-associated pili (*pap*) operon in NA114. The strains had similar resistance patterns except that NA097 was also positive for the *strA*, *strB*, and *aac3′IIC* genes. The genomes of EC958 and NA097 contained both bla_{TEM} and $bla_{CTX\text{-}M\text{-}15}$ genes; SE15 was negative for both, and the rest of the genomes had one of the two β-lactamase genes. To elucidate the variations among genomes that clustered in clade C, we subjected four well-characterized reference ST131 genomes, i.e., EC958, JJ1886, NA114, and SE15, with NA097 as a reference, to comparative analysis using BRIG. The most common variation observed for SE15 (a commensal ST131 strain) and the four *H*30-Rx subclonal ST131 strains was the absence of seven phage-associated regions in the SE15 genome (Fig. 1). Three major genomic islands, namely, GI-*selC*, GI-*pheV*, and GI-*leuX*, were also found in the NA097 genome, along with several other genomic islands with sizes ranging from 0.5 kb to 30 kb, as predicted by IslandViewer. The genome of NA097 exclusively contained the lysogenic phage P4 and other genomic regions encoding functions such as autoinduction, capsular polysaccharide synthesis, quorum sensing, aminoglycoside resistance, and others (labeled in blue in Fig. 1). A detailed description of different genomic regions and phages is also provided in Fig. 1. The genome of NA097 shared similarity with the recently published EC958 *E. coli* ST131 genome sequence, which contains several mobile genetic elements (MGEs) that would account for increased fitness or pathogenicity. Overall, the genomes of four *H*30-Rx strains were found to be highly clonal in nature, despite their distinct geographical and temporal origins.

Molecular characterization and susceptibility testing. The NA097 strain was positive for typical characteristics of clone ST131, i.e., it belonged to phylogroup B2 and was PCR positive for the *pabB* allele as well as CTX-M-15 and TEM β-lactamase genes. Phenotypically, NA097 was determined to be an ESBL producer and was MBL negative. The strain was refractory to the following 13 of 20 antimicrobial agents tested: ampicillin, ampicillin-sulbactam, and carbenicillin of the penicillin class; cephalothin (first generation), cefazolin (first generation), cefotaxime (third generation), and cefoperazone (third generation) of the cephalosporin class; norfloxacin, ciprofloxacin, moxifloxacin, and gatifloxacin of the fluoroquinolone class; and streptomycin and co-trimoxazole of the aminoglycoside and trimethoprim-sulfamethoxazole classes, respectively. The strain was completely susceptible only to the macrolide antibiotic class and to five other antibiotics belonging to the aforementioned classes. Kado lysis revealed the presence of four plasmids, but only two plasmids could be identified in PCR-based replicon typing.

Phenotypic characteristics of ST131 strains NA097 and NA114. Biofilm formation is considered to be an important virulence property among pathogenic *E. coli* strains (40), especially those that cause urinary tract infections (UTIs) and sepsis. Therefore, we investigated the ability of two ST131 strains to form biofilms in M63 minimal medium. We observed significant specific biofilm formation (SBF) values for both NA097 and NA114 strains, compared to DH5α, after 48 h of incubation (Fig. 2a). Because serum resistance is reported to be a marker of virulence (41), we examined the survival of the two ST131 strains and DH5α in human serum after 3 h. The mean CFU values for growth of NA097, NA114, and DH5α were 3.9×10^8, 2.6×10^8, and 7.7×10^7 CFU, respectively, while the blank showed no growth (Fig. 2b). NA097 and NA114 were significantly resistant to serum, compared to DH5α. Moreover, the levels of serum resistance for NA097 and NA114 were comparable. Furthermore, NA097 showed mannose-resistant hemagglutination (MRHA) in human RBCs. To assess whether isolates of the ST131 lineage shared specific metabolic characteristics, we investigated the carbohydrate utilization ability of NA097 using 35 API test reagents (HiMedia, Mumbai, India). Of the 35 sugars, NA097 was able to utilize 20 sugars (57%). The metabolic capabilities of NA097 and NA114 for



S. No	Region	Size (kb)	Names of closest relative organism/ phage (if any)	Status/ coding proteins
1.	Phage1	41.8	P88	incomplete
2.	Phage2	78.6	BcepMu	intact
3.	Phage3	56.3	BP4795	intact
4.	Phage4	50.4	HK629	intact
5.	Phage5	12.7	mEp235	incomplete
6.	Phage6	49.4	ZF40	intact
7.	Phage7	11.7	P4	intact
8.	Phage8	5.9	P88	incomplete
9.	Phage9	3.1	D108	Incomplete
10.	R1	5.6	-	site specific DNA methyltransferases, phage integrase etc.
11.	R2	7	-	Sam dependent methytrasferase,mobile element protein, Autoinducers LsrG, LsrF etc.
12.	R3	7	-	XcbB, CDP glycerol etc.
13.	R4	25.2	-	mobile element protein, type II restriction enzyme methylase, ATP dependent protease etc
14.	R5	9.4	-	Type I restriction modification system
15.	R6	23	-	Flagellar synthesis apparatus
16.	R7	9	-	Autoinducers ABC transport system, mobile elements etc.
17.	R8	10	-	DNA recombinase protein, transposase, Aminoglycoside transacetylase, hypothetical proteins etc.

FIG 1 Left, BRIG figure, representing a comparison of NA097 with four reference genomes of ST131 *E. coli* downloaded from the NCBI. Blue regions (R1, R2, R3, R7, and R8) are present only in NA097, the black region (R4) is present only in NA097 and EC958, green regions (R5 and R6) are absent in SE15, and red regions are phage-associated regions. Right, details of the regions and their relevant sizes.

carbohydrate utilization were found to be comparable. However, the strains did not demonstrate enhanced utilization of a panel of metabolic substrates, compared to the data on metabolic utilization by *E. coli* strains belonging to other sequence types (15).

Adhesion and invasion assays with human epithelial cells. Because of the significance of human kidney and bladder epithelial cells in the development of acute UPEC-related complications and because we wanted to investigate whether the presence or

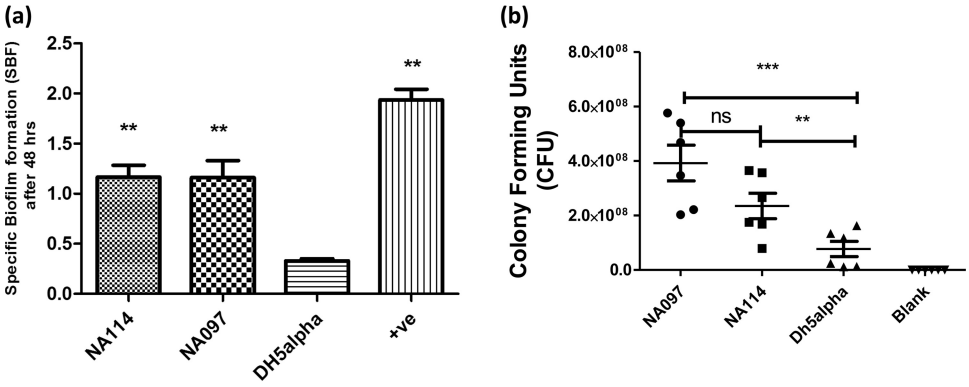


FIG 2 Biofilm formation capabilities and serum bactericidal resistance of *E. coli* strains NA114 and NA097. (a) Specific biofilm formation (SBF) after 48 h of incubation. (b) Resistance to human serum, as determined by CFU counting after 3 h of incubation. Strain DH5α was used as a negative control for both experiments, and NA434 served as a positive control (+ve) for the determination of SBF. The experiment was performed in nine replicates for analysis of biofilm formation and six replicates for the serum assay, and statistical values were obtained with the nonparametric Mann-Whitney *U* test, using GraphPad Prism 5.01. ns, nonsignificant; **, $P < 0.01$; ***, $P < 0.001$.

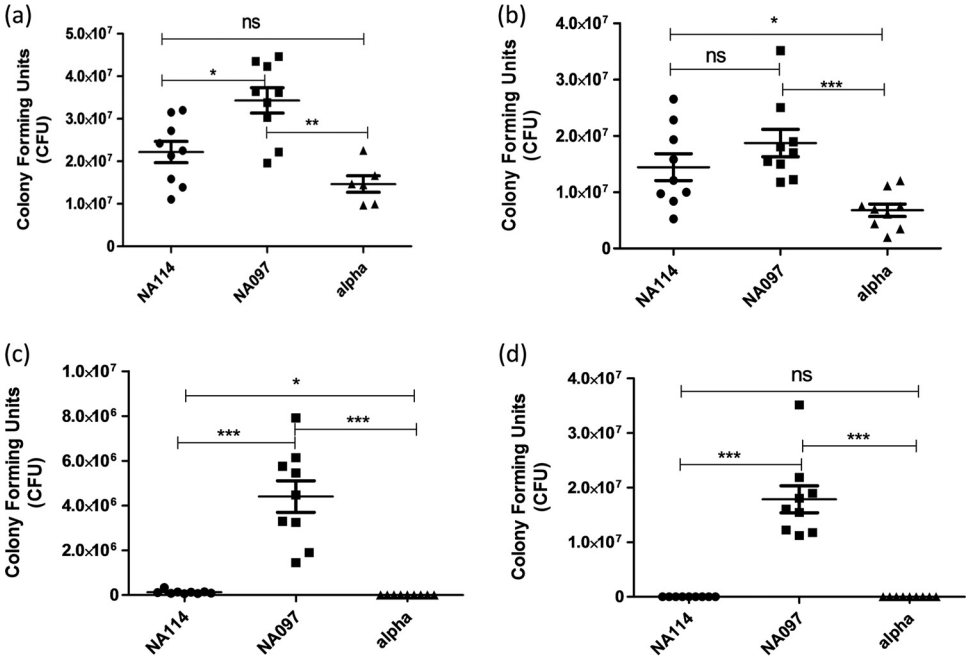


FIG 3 Adhesion and invasion capabilities of *E. coli* strains NA114 and NA097. (a and b) Adhesion on the human bladder epithelial cell line T24 (a) and the human kidney epithelial cell line ACHN (b) after 3 h of incubation. (c and d) Invasion capabilities in the T24 cell line (c) and the ACHN cell line (d) after 3 h of incubation and additional incubation for 1.5 h in medium containing gentamicin (100 μg/ml). DH5α (alpha) was used as a negative control for the adhesion and invasion assays, experiments were performed three times in triplicate, and statistical values were obtained with the nonparametric Mann-Whitney *U* test, using GraphPad Prism 5.01. ns, nonsignificant; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

absence of virulence genes affected the pathogenic nature of the bacteria (42), we conducted *in vitro* cell adhesion and invasion assays using strains NA097 and NA114, along with the negative control (DH5α), with both ACHN and T24 cells. We observed significant levels of adhesion and invasion of NA097 with both cell types, compared to DH5α (Fig. 3). In contrast, NA114 was adherent with both cell types but was less invasive (1.3×10^5 CFU) with T24 cells and was noninvasive with ACHN cells.

RT^2 gene expression profiling of antibacterial responses in THP1 cells. In order to analyze the transcription of antibacterial response genes in differentiated human THP1 macrophages infected with the ST131 *E. coli* strain NA097, the DH5α strain, and LPS, we performed quantitative PCR to determine the expression of 84 genes, corresponding to inflammatory response factors, pattern recognition receptors (PRRs), chemokines, cytokines, or downstream signaling components. The expression levels for 12 genes were unaltered in all three groups (including the control), while levels for a total of 24 transcripts were identified as being significantly altered. In all three categories, 12 genes were upregulated, while five genes were downregulated. A majority of significantly upregulated genes encoded cytokines such as interleukin 8 (IL-8), IL-1β, and tumor necrosis factor (TNF) and chemokines such as CXCL1, CXCL2, and CCL3; the remaining significantly upregulated transcripts encoded intracellular signaling factors such as BIRC3, MAP2K3, NFKB1, NFKBIA, NLRP3, and RIPK2. In contrast, transcripts corresponding to CARD6, TLR1, TLR6, PSTPIP1, and PYCARD were significantly downregulated (Fig. 4). IL-6 was upregulated more than 100-fold in all cases. Overall, the levels of transcriptionally altered genes were similar (with no significant differences) in all three groups. The rest of the genes did not show any significant fold changes, compared to the untreated control; however, a few, such as IFNB1, NOD1, SUGT1, and NAIP, were upregulated differentially. These data provide a comprehensive picture of transcriptional regulation of macrophage responses to *E. coli* infections.

DISCUSSION

Previous work by our group showed that ST131 *E. coli* strains in India have emerged as important causes of community-acquired urinary tract infections (8, 15). Molecular epidemiological analysis of these strains demonstrated that they are largely evolving clonally and are associated with several successful phenotypes that are likely involved in ST131 *E. coli* worldwide dissemination and infections (8, 15, 18, 43). This highlights the urgent need to understand this clonal pathogen, in order to develop better control strategies. In this study, we investigated the comprehensive profile of antibacterial responses of human macrophages upon ST131 *E. coli* infection. We further investigated the variations in the phenotypes of ST131 strains by analyzing the phenotypic traits versus the whole-genome sequence data for our two ST131 strains (NA097 and NA114). Whole-genome comparisons and phylogenetic comparisons were also performed with other sequenced genomes of ST131 clones.

Our data lend support to observations indicating adaptive diversification and clonal expansion of ST131 as the underlying mechanisms for the evolution of fluoroquinolone-resistant and ESBL-producing ST131 *E. coli* (44, 45). Our data also reveal that the similarity in phenotypes observed for our two ST131 strains correlates with the observation of the clonal genetic structure due to the evolution of a distinct subclade within ST131, namely, *H*30-Rx (10). The minor differences observed could be due to variations in virulence gene carriage, as observed by others (46).

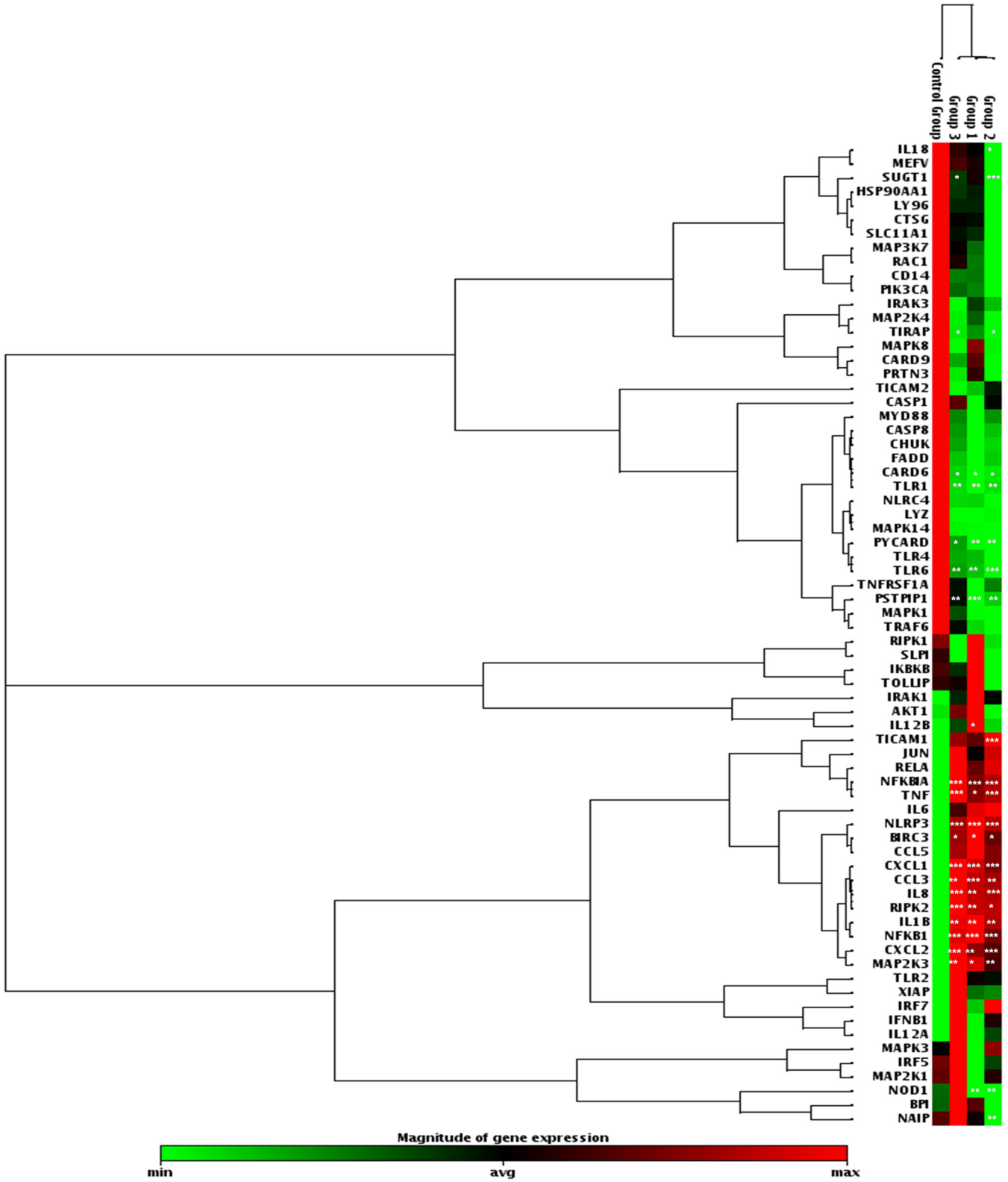


FIG 4 Cluster diagram indicating transcript levels of genes associated with bacterial infections in THP-1 cells. Group 1, LPS-treated cell sample; group 2, cells infected with *E. coli* strain DH5α; group 3, cells infected with strain NA097. Untreated THP-1 cells were used as the control. Online RT^2 Profiler array data analysis custom software (version 3.5) from Qiagen was used to obtain statistical values (Student's *t* test), according to the RT^2 Profiler instructions. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

ST131 *E. coli* is now the dominant extraintestinal pathogen throughout the world and is a major etiological agent of bladder and kidney infections (47). This organism represents a classic example of evolution through both horizontal and vertical (clonal expansion) mechanisms. Although ST131 first attracted attention because of its close association with ESBLs, especially the plasmid-mediated CTX-M-15, recent studies involving population genetic analysis of ST131 strains indicated that the global rise in extended-spectrum cephalosporin-resistant ST131 *E. coli* infections is due to the expansion of a single, highly virulent subclone, *H*30-Rx (10). Phylogenomic analysis of two of our ST131 strains (NA114 and NA097), together with several ST131 reference genomes (4), revealed that both of the Indian strains NA097 and NA114 belonged to the *H*30-Rx subclone (see Fig. S1 in the supplemental material). These results might represent only the first of many such findings and therefore warrant attention from clinicians and researchers, as our previous studies have already documented that ST131 *E. coli* strains represent a major population among clinical ExPEC isolates in India (8, 15).

Earlier studies reported the presence of mobile genetic elements (MGEs) in ST131 *E. coli* strains, which are not found in non-ST131 UPEC strains (4). Our study confirms this and documents the high level of conservation of seven prophage-associated genomic regions among all of the *H*30-Rx strains studied (Fig. 1); these seven prophage elements were clearly absent in the commensal non-*H*30-Rx ST131 SE15 genome (48). The genome of our ST131 strain NA097 revealed greater genetic affinity with the recently published EC958 ST131 genome from England (49), followed by the genome of the Indian ST131 strain NA114. This variation could possibly be due to differences in the arrangements of different genomic regions, arising from variable genomic contents encompassing mobile elements and other auxiliary genomic components. However, such a degree of clonality between two geographically distinct isolates hints at the possibility of their potential to emerge and to expand globally. Other studies have also observed phylogeographic patterns among spatially dispersed ST131 isolates (8).

Given the extensive genomic similarities observed among the strains within the *H*30-Rx clade, we sought to determine whether there were any phenotypic similarities between the two representative *H*30-Rx strains, i.e., NA114 and NA097. We compared the abilities of our two *H*30-Rx strains to form biofilms *in vitro* over 48 h. There was no significant difference in the levels of specific biofilm formation between the two strains (Fig. 2a). Furthermore, we found that the two strains were equally resistant to serum bactericidal activity (Fig. 2b). We also conducted *in vitro* cell adhesion and invasion assays with T24 bladder epithelial cells and ACHN kidney epithelial cells. In addition to the virulence-associated phenotypes, identical patterns were observed for the adhesion capabilities of the two strains on both types of epithelial cells (Fig. 3). However, the two strains differed significantly in their abilities to invade the epithelial cells. The NA097 strain was highly invasive, while NA114 showed insignificant invasion (Fig. 3). This observed difference could be partly attributable to the presence of the *dra* operon (which was found to be absent in the NA114 strain) in the NA097 strain. Goluszko et al. observed similar results, with insertional mutants of *dra* operon members being unable to invade HeLa cells (50). Finally, we compared the abilities of our two ST131 strains to metabolize 35 carbon sources. We found similar patterns of carbon utilization for the two strains. Overall, our data provide a correlation between the genetic frameworks and the virulence phenotypes of two *H*30-Rx strains. In contrast, previous studies showed that ST131 strains in general were genetically monomorphic but phenotypically diverse (46). This might be due to the fact that all such studies analyzed ST131 strains without taking into account the recent subclones of ST131, namely, *H*30-R and *H*30-Rx. When these subclones are studied separately, they might yield clearer distinct observations.

While several studies have investigated the pathogenesis of ST131 *E. coli* using *in vitro* and *in vivo* assays, the innate immune responses involved and the molecular pathways governing these responses largely remain obscure. Deciphering the innate responses against this pathogen is key to understanding the immunological basis of bacterial clearance from the host. Here we performed transcriptome profiling of innate immune signatures in the context of host responses to ST131 strains, the control strain DH5α, and LPS alone. We observed that only the key inflammatory pathways were activated upon infection with ST131, DH5α, or LPS. Therefore, we could not identify major strain-specific host responses, which implies that, after infection, capsular antigens would be primarily associated with the immune responses and thus pathogenesis could not necessarily be correlated with repertoires of virulence genes or mobile genetic elements. A previous study by Vimont et al. (16) confirmed that the inflammatory responses of kidney cells after infection with ST131 versus non-ST131 were comparable. Also, Horn et al. (51) observed similar histopathological outcomes in lung tissues after infection with pathogenic versus fecal avian *E. coli*. These observations, together with our results, suggest that the fitness and epidemiological success of ST131 strains are more relevant in the upstream steps involved at the onset of infection.

In conclusion, our study confirmed the presence of *H*30-Rx ST131 *E. coli* among Indian ST131 strains and provided more insights into the clonal diversification of genomic and phenotypic features of *H*30-Rx ST131 strains. The study also shows the immunological basis of the outcomes of infection with ST131 strains. Such expository observations could be of great significance for understanding the emergence of globally disseminated, virulent *H*30-Rx strains. Further studies are needed to investigate whether this clone undergoes evolutionary selection and whether that has any effect on its virulence and resistance phenotypes.

ACKNOWLEDGMENTS

We acknowledge funding received from the Department of Biotechnology, Government of India [grant BT/HRD/NBA/34/01/2011(ix)]. We also acknowledge support received from University of Malaya High Impact Research Grant UM.C/625/1HIR/MOHE/02 (A000002-5000 1) and support from Indo-German International Research Training Group GRK1673, an initiative of the German Research Foundation and the University of Hyderabad (India). A.R. acknowledges the Indian Council of Medical Research for the award of a junior research fellowship. N.A. is an adjunct professor at the Academy of Scientific and Innovative Research (AcSIR), New Delhi, India.

REFERENCES

1. **Tumbarello M, Sanguinetti M, Montuori E, Trecarichi EM, Posteraro B, Fiori B, Citton R, D'Inzeo T, Fadda G, Cauda R, Spanu T.** 2007. Predictors of mortality in patients with bloodstream infections caused by extended-spectrum-β-lactamase-producing *Enterobacteriaceae*: importance of inadequate initial antimicrobial treatment. Antimicrob Agents Chemother **51:**1987–1994. http://dx.doi.org/10.1128/AAC.01509-06.
2. **Lautenbach E, Patel JB, Bilker WB, Edelstein PH, Fishman NO.** 2001.

Extended-spectrum β-lactamase-producing *Escherichia coli* and *Klebsiella pneumoniae*: risk factors for infection and impact of resistance on outcomes. Clin Infect Dis **32:**1162–1171. http://dx.doi.org/10.1086/319757.

3. **Hong T, Moland ES, Abdalhamid B, Hanson ND, Wang J, Sloan C, Fabian D, Farajallah A, Levine J, Thomson KS.** 2005. *Escherichia coli*: development of carbapenem resistance during therapy. Clin Infect Dis **40:**e84–e86. http://dx.doi.org/10.1086/429822.
4. **Petty NK, Ben Zakour NL, Stanton-Cook M, Skippington E, Totsika M, Forde BM, Phan MD, Gomes Moriel D, Peters KM, Davies M, Rogers BA, Dougan G, Rodriguez-Bano J, Pascual A, Pitout JD, Upton M, Paterson DL, Walsh TR, Schembri MA, Beatson SA.** 2014. Global dissemination of a multidrug resistant *Escherichia coli* clone. Proc Natl Acad Sci U S A **111:**5694–5699. http://dx.doi.org/10.1073/pnas.1322678111.
5. **Johnson JR, Johnston B, Clabots C, Kuskowski MA, Castanheira M.** 2010. *Escherichia coli* sequence type ST131 as the major cause of serious multidrug-resistant *E. coli* infections in the United States. Clin Infect Dis **51:**286–294. http://dx.doi.org/10.1086/653932.
6. **Nicolas-Chanoine MH, Blanco J, Leflon-Guibout V, Demarty R, Alonso MP, Canica MM, Park YJ, Lavigne JP, Pitout J, Johnson JR.** 2008. Intercontinental emergence of *Escherichia coli* clone O25:H4-ST131 producing CTX-M-15. J Antimicrob Chemother **61:**273–281.
7. **Coque TM, Novais A, Carattoli A, Poirel L, Pitout J, Peixe L, Baquero F, Canton R, Nordmann P.** 2008. Dissemination of clonally related *Escherichia coli* strains expressing extended-spectrum β-lactamase CTX-M-15. Emerg Infect Dis **14:**195–200. http://dx.doi.org/10.3201/eid1402.070350.
8. **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. http://dx.doi.org/10.1128/AAC.01099-12.
9. **Banerjee R, Johnson JR.** 2014. A new clone sweeps clean: the enigmatic emergence of *Escherichia coli* sequence type 131. Antimicrob Agents Chemother **58:**4997–5004. http://dx.doi.org/10.1128/AAC.02824-14.
10. **Price LB, Johnson JR, Aziz M, Clabots C, Johnston B, Tchesnokova V, Nordstrom L, Billig M, Chattopadhyay S, Stegger M, Andersen PS, Pearson T, Riddell K, Rogers P, Scholes D, Kahl B, Keim P, Sokurenko EV.** 2013. The epidemic of extended-spectrum-β-lactamase-producing *Escherichia coli* ST131 is driven by a single highly pathogenic subclone, *H*30-Rx. mBio **4**(6):e00377-13. http://dx.doi.org/10.1128/mBio.00377-13.
11. **Poirel L, Dortet L, Bernabeu S, Nordmann P.** 2011. Genetic features of bla_{NDM-1}-positive *Enterobacteriaceae*. Antimicrob Agents Chemother **55:**5403–5407. http://dx.doi.org/10.1128/AAC.00585-11.
12. **Johnson JR, Miller S, Johnston B, Clabots C, Debroy C.** 2009. Sharing of *Escherichia coli* sequence type ST131 and other multidrug-resistant and urovirulent *E. coli* strains among dogs and cats within a household. J Clin Microbiol **47:**3721–3725. http://dx.doi.org/10.1128/JCM.01581-09.
13. **Platell JL, Cobbold RN, Johnson JR, Heisig A, Heisig P, Clabots C, Kuskowski MA, Trott DJ.** 2011. Commonality among fluoroquinolone-resistant sequence type ST131 extraintestinal *Escherichia coli* isolates from humans and companion animals in Australia. Antimicrob Agents Chemother **55:**3782–3787. http://dx.doi.org/10.1128/AAC.00306-11.
14. **Ewers C, Bethe A, Semmler T, Guenther S, Wieler LH.** 2012. Extended-spectrum β-lactamase-producing and AmpC-producing *Escherichia coli* from livestock and companion animals, and their putative impact on public health: a global perspective. Clin Microbiol Infect **18:**646–655. http://dx.doi.org/10.1111/j.1469-0691.2012.03850.x.
15. **Hussain A, Ranjan A, Nandanwar 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. http://dx.doi.org/10.1128/AAC.03320-14.
16. **Vimont S, Boyd A, Bleibtreu A, Bens M, Goujon JM, Garry L, Clermont O, Denamur E, Arlet G, Vandewalle A.** 2012. The CTX-M-15-producing *Escherichia coli* clone O25b:H4-ST131 has high intestine colonization and urinary tract infection abilities. PLoS One **7:**e46547. http://dx.doi.org/10.1371/journal.pone.0046547.
17. **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. http://dx.doi.org/10.1038/nrmicro1889.
18. **Avasthi TS, Kumar N, Baddam R, Hussain A, Nandanwar N, Jadhav S, Ahmed N.** 2011. Genome of multidrug-resistant uropathogenic *Escherichia coli* strain NA114 from India. J Bacteriol **193:**4272–4273. http://dx.doi.org/10.1128/JB.05413-11.
19. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res **18:**821–829. http://dx.doi.org/10.1101/gr.074492.107.
20. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. J Mol Biol **215:**403–410. http://dx.doi.org/10.1016/S0022-2836(05)80360-2.
21. **Aziz RK, Devoid S, Disz T, Edwards RA, Henry CS, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, Stevens RL, Vonstein V, Xia F.** 2012. SEED servers: high-performance access to the SEED genomes, annotations, and metabolic models. PLoS One **7:**e48053. http://dx.doi.org/10.1371/journal.pone.0048053.
22. **Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B.** 2000. Artemis: sequence visualization and annotation. Bioinformatics **16:**944–945. http://dx.doi.org/10.1093/bioinformatics/16.10.944.
23. **Besemer J, Lomsadze A, Borodovsky M.** 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes: implications for finding sequence motifs in regulatory regions. Nucleic Acids Res **29:**2607–2618. http://dx.doi.org/10.1093/nar/29.12.2607.
24. **Schattner P, Brooks AN, Lowe TM.** 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res **33:**W686–W689. http://dx.doi.org/10.1093/nar/gki366.
25. **Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW.** 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res **35:**3100–3108. http://dx.doi.org/10.1093/nar/gkm160.
26. **Treangen TJ, Ondov BD, Koren S, Phillippy AM.** 2014. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol **15:**524. http://dx.doi.org/10.1186/s13059-014-0524-x.
27. **Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA.** 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics **12:**402. http://dx.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.
29. **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 LJ, 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. http://dx.doi.org/10.1128/AAC.00419-13.
30. **Akhter S, Aziz RK, Edwards RA.** 2012. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. Nucleic Acids Res **40:**e126. http://dx.doi.org/10.1093/nar/gks406.
31. **Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS.** 2011. PHAST: a fast phage search tool. Nucleic Acids Res **39:**W347–W352. http://dx.doi.org/10.1093/nar/gkr485.
32. **Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M.** 2006. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res **34:**D32–D36. http://dx.doi.org/10.1093/nar/gkj014.
33. **Langille MG, Brinkman FS.** 2009. IslandViewer: an integrated interface for computational identification and visualization of genomic islands. Bioinformatics **25:**664–665. http://dx.doi.org/10.1093/bioinformatics/btp030.
34. **Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH.** 2011. CDD: a conserved domain database for the functional annotation of proteins. Nucleic Acids Res **39:**D225–D229. http://dx.doi.org/10.1093/nar/gkq1189.
35. **Clermont O, Bonacorsi S, Bingen E.** 2000. Rapid and simple determination of the *Escherichia coli* phylogenetic group. Appl Environ Microbiol **66:**4555–4558. http://dx.doi.org/10.1128/AEM.66.10.4555-4558.2000.
36. **Clermont O, Dhanji H, Upton M, Gibreel T, Fox A, Boyd D, Mulvey MR, Nordmann P, Ruppe E, Sarthou JL, Frank T, Vimont S, Arlet G, Branger C, Woodford N, Denamur E.** 2009. Rapid detection of the O25b-ST131 clone of *Escherichia coli* encompassing the CTX-M-15-

producing strains. J Antimicrob Chemother **64:**274–277. http://dx.doi.org/10.1093/jac/dkp194.
37. **Carattoli A, Bertini A, Villa L, Falbo V, Hopkins KL, Threlfall EJ.** 2005. Identification of plasmids by PCR-based replicon typing. J Microbiol Methods **63:**219–228. http://dx.doi.org/10.1016/j.mimet.2005.03.018.
38. **Nandanwar N, Janssen T, Kuhl 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. http://dx.doi.org/10.1016/j.ijmm.2014.06.009.
39. **Siegfried L, Kmetova M, Puzova H, Molokacova M, Filka J.** 1994. Virulence-associated factors in *Escherichia coli* strains isolated from children with urinary tract infections. J Med Microbiol **41:**127–132. http://dx.doi.org/10.1099/00222615-41-2-127.
40. **Beloin C, Roux A, Ghigo JM.** 2008. *Escherichia coli* biofilms. Curr Top Microbiol Immunol **322:**249–289.
41. **Nolan LK, Horne SM, Giddings CW, Foley SL, Johnson TJ, Lynne AM, Skyberg J.** 2003. Resistance to serum complement, *iss*, and virulence of avian *Escherichia coli*. Vet Res Commun **27:**101–110. http://dx.doi.org/10.1023/A:1022854902700.
42. **Bien J, Sokolova O, Bozko P.** 2012. Role of uropathogenic *Escherichia coli* virulence factors in development of urinary tract infection and kidney damage. Int J Nephrol **2012:**681473.
43. **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. http://dx.doi.org/10.1371/journal.pone.0018063.
44. **Paul S, Linardopoulou EV, Billig M, Tchesnokova V, Price LB, Johnson JR, Chattopadhyay S, Sokurenko EV.** 2013. Role of homologous recombination in adaptive diversification of extraintestinal *Escherichia coli*. J Bacteriol **195:**231–242. http://dx.doi.org/10.1128/JB.01524-12.
45. **Banerjee R, Johnson JR.** 2013. *Escherichia coli* ST131: variations on a theme of clonal expansion. Enferm Infecc Microbiol Clin **31:**355–356. http://dx.doi.org/10.1016/j.eimc.2013.01.004.
46. **Clark G, Paszkiewicz K, Hale J, Weston V, Constantinidou C, Penn C, Achtman M, McNally A.** 2012. Genomic analysis uncovers a phenotypically diverse but genetically homogeneous *Escherichia coli* ST131 clone circulating in unrelated urinary tract infections. J Antimicrob Chemother **67:**868–877. http://dx.doi.org/10.1093/jac/dkr585.
47. **Nicolas-Chanoine MH, Bertrand X, Madec JY.** 2014. *Escherichia coli* ST131, an intriguing clonal group. Clin Microbiol Rev **27:**543–574. http://dx.doi.org/10.1128/CMR.00125-13.
48. **Forde BM, Ben Zakour NL, Stanton-Cook M, Phan MD, Totsika M, Peters KM, Chan KG, Schembri MA, Upton M, Beatson SA.** 2014. The complete genome sequence of *Escherichia coli* EC958: a high quality reference sequence for the globally disseminated multidrug resistant *E. coli* O25b:H4-ST131 clone. PLoS One **9:**e104400. http://dx.doi.org/10.1371/journal.pone.0104400.
49. **Totsika M, Beatson SA, Sarkar S, Phan MD, Petty NK, Bachmann N, Szubert M, Sidjabat HE, Paterson DL, Upton M, Schembri MA.** 2011. Insights into a multidrug resistant *Escherichia coli* pathogen of the globally disseminated ST131 lineage: genome analysis and virulence mechanisms. PLoS One **6:**e26578. http://dx.doi.org/10.1371/journal.pone.0026578.
50. **Goluszko P, Popov V, Selvarangan R, Nowicki S, Pham T, Nowicki BJ.** 1997. Dr fimbriae operon of uropathogenic *Escherichia coli* mediate microtubule-dependent invasion to the HeLa epithelial cell line. J Infect Dis **176:**158–167. http://dx.doi.org/10.1086/514018.
51. **Horn F, Corrêa AMR, Barbieri NL, Glodde S, Weyrauch KD, Kaspers B, Driemeier D, Ewers C, Wieler LH.** 2012. Infections with avian pathogenic and fecal *Escherichia coli* strains display similar lung histopathology and macrophage apoptosis. PLoS One **7:**e41031. http://dx.doi.org/10.1371/journal.pone.0041031.

AMERICAN SOCIETY FOR MICROBIOLOGY | Applied and Environmental Microbiology®

EVOLUTIONARY AND GENOMIC MICROBIOLOGY

Genome Dynamics and Molecular Infection Epidemiology of Multidrug-Resistant *Helicobacter pullorum* Isolates Obtained from Broiler and Free-Range Chickens in India

Shamsul Qumar,[a] Mohammad Majid,[a] Narender Kumar,[a] Sumeet K. Tiwari,[a] Torsten Semmler,[b] Savita Devi,[a] Ramani Baddam,[a,b] Arif Hussain,[a] Sabiha Shaik,[a] Niyaz Ahmed[a]

Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, University of Hyderabad, Gachibowli, Hyderabad, India[a]; Robert Koch Institute, Berlin, Germany[b]

ABSTRACT Some life-threatening, foodborne, and zoonotic infections are transmitted through poultry birds. Inappropriate and indiscriminate use of antimicrobials in the livestock industry has led to an increased prevalence of multidrug-resistant bacteria with epidemic potential. Here, we present a functional molecular epidemiological analysis entailing the phenotypic and whole-genome sequence-based characterization of 11 *H. pullorum* isolates from broiler and free-range chickens sampled from retail wet markets in Hyderabad City, India. Antimicrobial susceptibility tests revealed all of the isolates to be resistant to multiple antibiotic classes such as fluoroquinolones, cephalosporins, sulfonamides, and macrolides. The isolates were also found to be extended-spectrum β-lactamase producers and were even resistant to clavulanic acid. Whole-genome sequencing and comparative genomic analysis of these isolates revealed the presence of five or six well-characterized antimicrobial resistance genes, including those encoding a resistance-nodulation-division efflux pump(s). Phylogenetic analysis combined with pan-genome analysis revealed a remarkable degree of genetic diversity among the isolates from free-range chickens; in contrast, a high degree of genetic similarity was observed among broiler chicken isolates. Comparative genomic analysis of all publicly available *H. pullorum* genomes, including our isolates ($n = 16$), together with the genomes of 17 other *Helicobacter* species, revealed a high number (8,560) of *H. pullorum*-specific protein-encoding genes, with an average of 535 such genes per isolate. *In silico* virulence screening identified 182 important virulence genes and also revealed high strain-specific gene content in isolates from free-range chickens (average, 34) compared to broiler chicken isolates. A significant prevalence of prophages (ranging from 1 to 9) and a significant presence of genomic islands (0 to 4) were observed in free-range and broiler chicken isolates. Taken together, these observations provide significant baseline data for functional molecular infection epidemiology of nonpyloric *Helicobacter* species such as *H. pullorum* by unraveling their evolution in chickens and their possible zoonotic transmission to humans.

IMPORTANCE Globally, the poultry industry is expanding with an ever-growing consumer base for chicken meat. Given this, food-associated transmission of multidrug-resistant bacteria represents an important health care issue. Our study involves a critical baseline approach directed at genome sequence-based epidemiology and transmission dynamics of *H. pullorum*, a poultry pathogen having established zoonotic potential. We believe our studies would facilitate the development of surveillance systems that ensure the safety of food for humans and guide public health

Received 4 August 2016 **Accepted** 18 October 2016

Accepted manuscript posted online 4 November 2016

Citation Qumar S, Majid M, Kumar N, Tiwari SK, Semmler T, Devi S, Baddam R, Hussain A, Shaik S, Ahmed N. 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. https://doi.org/10.1128/AEM.02305-16.

Editor Shuang-Jiang Liu, Chinese Academy of Sciences

Copyright © 2016 American Society for Microbiology. All Rights Reserved.

Address correspondence to Niyaz Ahmed, ahmed.nizi@gmail.com.

S.Q. and M.M. contributed equally to this work.

Downloaded from http://aem.asm.org/ on December 19, 2016 by guest

policies related to the use of antibiotics in animal feed in countries such as India. We sequenced 11 new genomes of *H. pullorum* as a part of this study. These genomes would provide much value in addition to the ongoing comparative genomic studies of helicobacters.

KEYWORDS Helicobacters, genomics, molecular epidemiology, poultry

Helicobacter pullorum is a urease-negative, bile-resistant bacterium that was first isolated from the ceca of asymptomatic poultry and was also recovered from the intestinal contents and livers of laying hens with enteritis and vibrionic hepatitis (1). This nonpyloric *Helicobacter* species has also been isolated from people suffering from gastroenteritis (1–4) and is likely to be associated with inflammatory bowel disease (5, 6) and chronic liver disease (7–12). Laharie et al. (13) also detected DNA of *H. pullorum* and *H. canadensis* in Crohn's disease patients, as well as in controls. Moreover, several reports suggest that some enterohepatic *Helicobacter* species, such as *Helicobacter hepaticus*, *Helicobacter bilis*, *Helicobacter canis*, *Helicobacter cinaedi*, and *H. pullorum*, produce a common, well-characterized bacterial virulence factor, the cytolethal distending toxin (CDT) (14–18). This bacterial protein triggers G_2/M cell cycle arrest in a wide range of mammalian cell lines, leading to enlarged or distended cells dying with apoptosis (19). A recent finding also highlighted that CDT in *H. pullorum* is responsible for major cytopathogenic effects *in vitro*, reinforcing its role as a major virulence factor in pathogenesis (20).

The infectious disease burden is very high globally, and inappropriate and irrational use of antimicrobial agents against these infectious diseases has already resulted in the emergence of multidrug-resistant (MDR) bacteria. The prevalence of MDR organisms in livestock and their products has traditionally been linked to the disproportionate use of antimicrobials in veterinary practice and animal husbandry settings in the form of prophylactic agents and growth promoters (21). These drug-resistant bacteria greatly reduce treatment efficacy and cause increased morbidity and mortality in livestock, as well as in humans (22). The problem of antimicrobial resistance (AMR) is even more complex in developing countries because of high population density, poor sanitation, and less stringent antibiotic policies. In addition to this, there are only a few novel antimicrobial agents that are expected to be available for use in the next few years (23).

The problem of AMR is one of the greatest threats to public health. The resistant bacteria and their associated genes can move within and between populations of humans and animals, which makes AMR a contentious issue. As this problem continues to grow, it is imperative to characterize organisms that are resistant to multiple antimicrobial agents so that epidemiological surveillance data can be reliably obtained.

Chickens are used as a major source of food proteins worldwide. In India, consumption of poultry meat is growing at about 12% a year, making it one of the fastest-growing markets in the world (National Sample Survey Office of India report, 2014). In order to fulfill the high demand for chicken meat, poultry farmers or feed manufacturers may resort to the use of growth promoters, including various antibiotics, sometimes indiscriminately, for faster weight gain by chickens. Moreover, epidemiological investigations suggest that the majority of the foodborne bacterial infections caused by different enteric pathogens majorly spread through foods of animal origin (24, 25). The mainstream medical literature lacks any report from India or other developing countries on systematic epidemiological tracking of bacterial pathogens from poultry, in particular, *H. pullorum*. In poultry, *H. pullorum* was found to colonize predominantly the cecum (26). Recently, Borges et al. isolated *H. pullorum* directly from chicken meat (27). The presence of *H. pullorum* on poultry carcasses has also been reported, which might be due to contamination during the poultry rearing, handling, and slaughtering processes (28). Therefore, with the risk of zoonosis together with its ability to spread AMR, *H. pullorum* could be an emerging foodborne human pathogen that needs to be deciphered at the genomic and molecular levels, particularly by elucidating its zoonotic potential.

Downloaded from http://aem.asm.org/ on December 19, 2016 by guest

TABLE 1 Genome statistics of the whole-genome sequences of the 11 *H. pullorum* isolates in this study

Strain	Avg fold genome coverage	No. of contigs	Genome size (bp)	No. of CDSs	Avg CDS length (bp)	% coding capacity	% G+C	No. of genes	
								rRNA	tRNA
NAP1W4	202	162	1,889,077	1,947	885	88.2	34.16	3	37
NAP2W5	212	140	1,804,031	1,814	886	89.1	34.38	3	36
NAP3W17	286	74	1,733,930	1,746	906	91.2	34.44	3	37
NAP5W19	166	127	1,731,995	1,719	893	88.6	34.45	3	36
NAP6W24	171	108	1,691,767	1,688	904	90.2	34.56	5	38
NAP8W25	110	134	1,783,256	1,819	883	90	34.25	3	37
NAP10B8	119	105	1,717,485	1,705	907	90.1	34.46	4	36
NAP11B31	80	139	1,862,853	1,886	876	88.7	34.08	3	37
NAP12B32	97	130	1,856,360	1,883	877	88.9	34.11	2	36
NAP13B35	83	123	1,856,511	1,881	885	89.6	34.11	2	36
NAP14B36	76	101	1,855,374	1,888	879	89.4	34.11	3	36

In this study, we attempted to rigorously characterize a modest set of *H. pullorum* isolates with respect to their multidrug resistance potential and their genome sequence-based phylogeny and epidemiology by analyzing their core and pan-genomes, mobilome, and virulence gene repertoire.

RESULTS

Isolation, characterization, and genome sequencing of *H. pullorum* isolates. We isolated 11 *H. pullorum* isolates from 100 chicken samples that were collected from seven different retail wet markets in Hyderabad City, India. These 11 isolates originated from four different markets that were spread out in different places in Hyderabad City. Of these 11 strains, 6 were isolated from free-range chickens and were designated NAP1W4, NAP2W5, NAP3W17, NAP5W19, NAP6W24, and NAP8W25 and 5 were from broiler chickens and named NAP10B8, NAP11B31, NAP12B32, NAP13B35, and NAP14B36. The broiler chickens we sampled were reared on farms that employ formula feed supplemented with antibiotics, while the free-range chickens originated in rural areas. Gram staining and urease testing revealed small, curved, Gram-negative, and urease-negative bacilli that were confirmed as *H. pullorum* by species-specific PCR and 16S rRNA gene sequencing. Whole-genome sequencing (WGS) of 11 *H. pullorum* isolates produced high-quality genomic assemblies. The detailed genome characteristics of the 11 sequenced *H. pullorum* isolates are shown in Table 1, and an alignment of their whole genome sequences with the *H. pullorum* MIT985489 genome as a reference is shown in Fig. 1. The result of BLAST ring image generator (BRIG) analysis of this whole-genome alignment demonstrated that the 11 *H. pullorum* isolates of free-range and broiler chickens together with 1 *H. pullorum* isolate from a human were homogeneous in their genetic architecture, with no significant genomic divergence despite their geographic and host differences.

Antimicrobial susceptibility profiles and identification of AMR genes. All of the isolates from both broiler and free-range chickens were found to be MDR, as all 11 isolates were resistant to at least three different antibiotic classes. The isolates from broiler and free-range chickens did not show any significant difference in their antibiotic resistance profiles, as shown in Table 2. The AMR rates of broiler and free-range chicken isolates were found to be 100% for nalidixic acid, enrofloxacin, co-trimoxazole, and cefotaxime, while 80% of the broiler and 83% of the free-range chicken isolates were resistant to ciprofloxacin. Also, 80% of the broiler and 67% of the free-range chicken isolates were resistant to clarithromycin. However, all of the isolates were susceptible to tetracycline, neomycin, chloramphenicol, and colistin. Phenotypically, the rate of extended-spectrum β-lactamase (ESBL) production in broiler and free-range chicken isolates was found to be 100%. In addition, all of the ESBL producers were resistant to the beta-lactamase inhibitor clavulanic acid. Correlating with the phenotypic observations, the genomic data of all of the isolates showed the presence of five or six well-characterized AMR genes linked to resistance to various antibiotic classes (see Table S2 in the supplemental material). All of the strains, irrespective of their

Downloaded from http://aem.asm.org/ on December 19, 2016 by guest

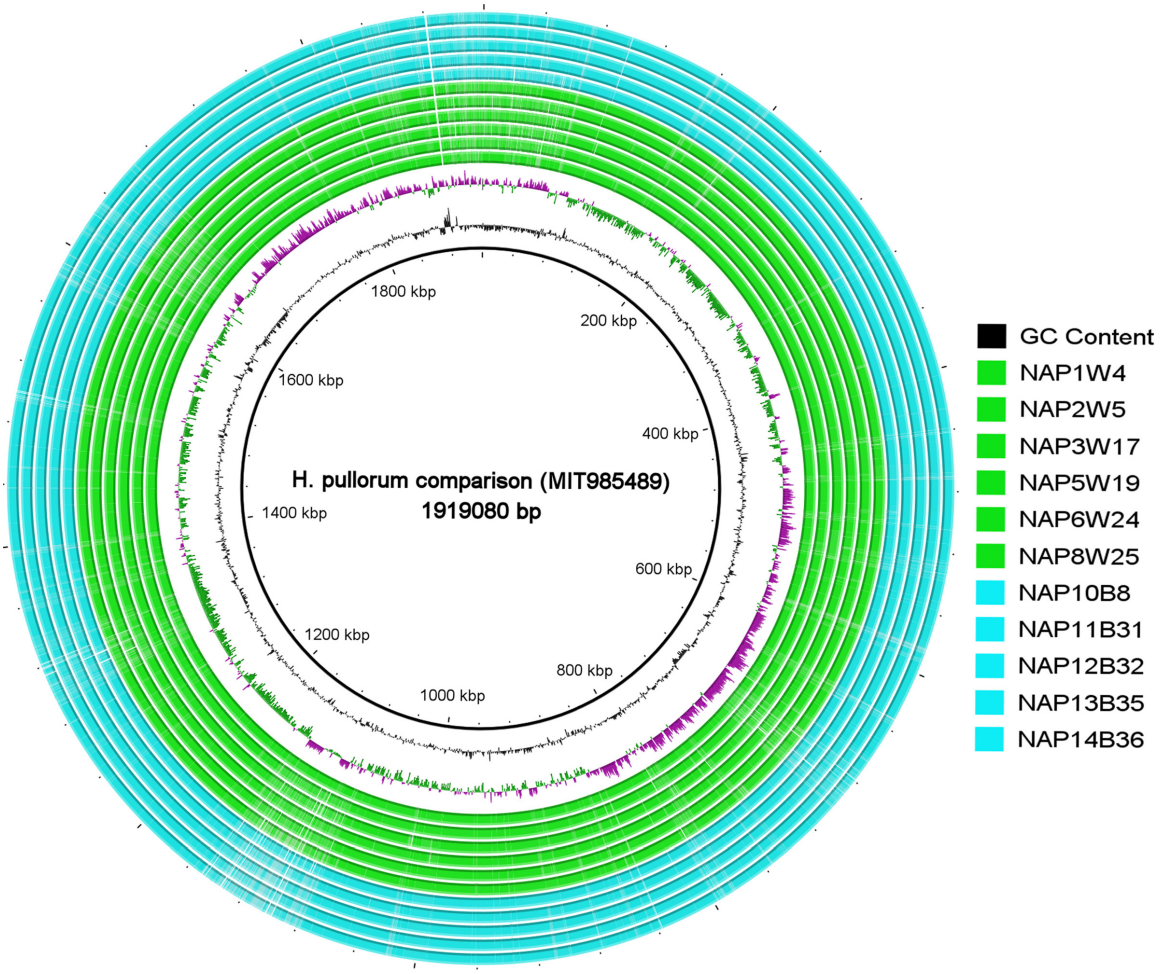


FIG 1 Whole-genome alignment of *H. pullorum* comprising six free-range (green) and five broiler (blue) chicken isolates generated by BRIG with the *H. pullorum* MIT985489 genome positioned as a reference. The key to the right lists the isolates from top to bottom as they appear from the inside of the circle toward the outside.

isolation source, were found to harbor common AMR-associated genes such as the elfamycin resistance gene *tufA*, a quinolone/fluoroquinolone resistance-related gene encoding DNA gyrase subunit A, and a multidrug resistance-nodulation-division (RND) family transporter efflux pump gene (*cmeB*). All of the isolates carried aminoglycoside resistance gene *aph*(*3″*)-*Ib*, except for one free-range chicken isolate, NAP8W25. The broiler chicken isolates carried one or two additional genes encoding aminoglycoside resistance (see Table S2 in the supplemental material).

Phylogenetic analysis. To determine the genetic relatedness of heterogeneous *Helicobacter* species and *H. pullorum* strains, we performed a consensus bootstrapped core genome maximum-likelihood-based phylogenetic analysis. The results showed that all 33 *Helicobacter*, *Wolinella*, and *Campylobacter* genomes could be broadly separated into two distinct clusters (Fig. 2). *H. canadensis* and *H. rodentium* clustered closely with *H. pullorum* isolates, revealing high similarity to *H. pullorum* among all of the other *Helicobacter* species analyzed. This is in agreement with the previous reports of 16S rRNA gene-based phylogenetic analyses (29, 30). Interestingly, the genomes of *Wolinella succinogenes* and *Campylobacter jejuni*, although they belong to different genera, coclustered with *H. pullorum*, exhibiting their genetic relatedness to the latter.

The phylogenetic diversity within *H. pullorum* isolates is depicted by a tree in Fig. 3. From the phylogenetic tree, we could glean that all of our *H. pullorum* isolates were intermixed with each other, forming mixed clusters, irrespective of their geographic origins and isolation sources. Hence, our results suggest that our *H. pullorum* strains did

Downloaded from http://aem.asm.org/ on December 19, 2016 by guest

TABLE 2 Prevalence of antimicrobial resistance in *H. pullorum* isolates obtained from free-range and broiler chickens

Phenotype	Overall prevalence of resistance phenotype in population (%)	Estimated fraction due to isolates from:	
		Free-range chickens	Broiler chickens
Quinolone/fluoroquinolone resistance			
Ciprofloxacin	82	0.55	0.44
Nalidixic acid	100	0.54	0.45
Enrofloxacin	100	0.54	0.45
Sulfonamide/trimethoprim (co-trimoxazole) resistance	100	0.54	0.45
Cephalosporin (cefotaxime) resistance	100	0.54	0.45
Macrolide (clarithromycin) resistance	73	0.50	0.50
Aminoglycoside (neomycin) resistance	0	0	0
Phenicol (chloramphenicol) resistance	0	0	0
Tetracycline resistance			
Oxytetracycline	0	0	0
Chlortetracycline	0	0	0
Doxycycline hydrochloride	0	0	0
Miscellaneous (colistin) resistance	0	0	0
ESBL production	100	0.54	0.45
Clavulanic acid resistance	100	0.54	0.45
Multidrug resistance	100	0.54	0.45

not show geographically restricted lineages, unlike *H. pylori*. Additionally, *H. pullorum* strains from broiler and free-range chickens and a single human isolate showed mixed clusters, indicating that *H. pullorum* strains from different hosts and geographical locations share significant genetic features with each other and they particularly did not exhibit a host-specific genomic architecture. The results also demonstrated that the *H. pullorum* isolates from broiler chickens were genetically more similar than *H. pullorum* isolates from free-range chickens (Fig. 3).

Core and pan-genome analysis. The core genome contents of 11 sequenced *H. pullorum* strains, *H. pullorum* strain 229313, and human-derived strain MIT985489 were analyzed by OrthoMCL. A total of 25,438 proteins were predicted from the 13 strains. These predicted proteins formed 2,188 orthologous gene clusters. Of these, around 1,359 orthologous gene clusters were found to constitute the core genome. Of the 1,359 core gene clusters, 1,041 could be characterized into various functional categories according to the Clusters of Orthologous Groups of proteins (COG) database, as shown in Fig. 4. The COG functional classification revealed a large proportion of genes belonging to the J, E, and M functional classes. Class J contains genes having functions related to translation, ribosomal structure, and biogenesis, while classes E and M consist of genes involved in amino acid transport/metabolism, and cell wall/envelope biogenesis, respectively. The core genome comprised genes encoding multidrug efflux pumps, which contribute to intrinsic drug resistance in bacteria (31). The genes encoding metallo-β-lactamase enzymes that confer resistance to β-lactam antibiotics, including carbapenems, were also identified in the core genome. *H. pullorum*-specific gene content analysis revealed a total of 8,560 *H. pullorum*-specific proteins among all 16 *H. pullorum* isolates, with an average of 535 specific proteins per isolate. Of the total of 8,560 *H. pullorum*-specific genes, 83% (7,136) belong to a hypothetical category whereas the remaining 17% (1,424) were classified into different COG categories, with the majority of them being assigned to cell wall/membrane/envelope biogenesis, transcription, defense mechanisms, and other multiple classes.

Further analysis of our *H. pullorum* isolates for the presence of strain-specific content revealed a total of 327 strain-specific genes. *H. pullorum* strains from free-range chickens were found to contain 204 strain-specific genes, with an average of 34 genes

Downloaded from http://aem.asm.org/ on December 19, 2016 by guest

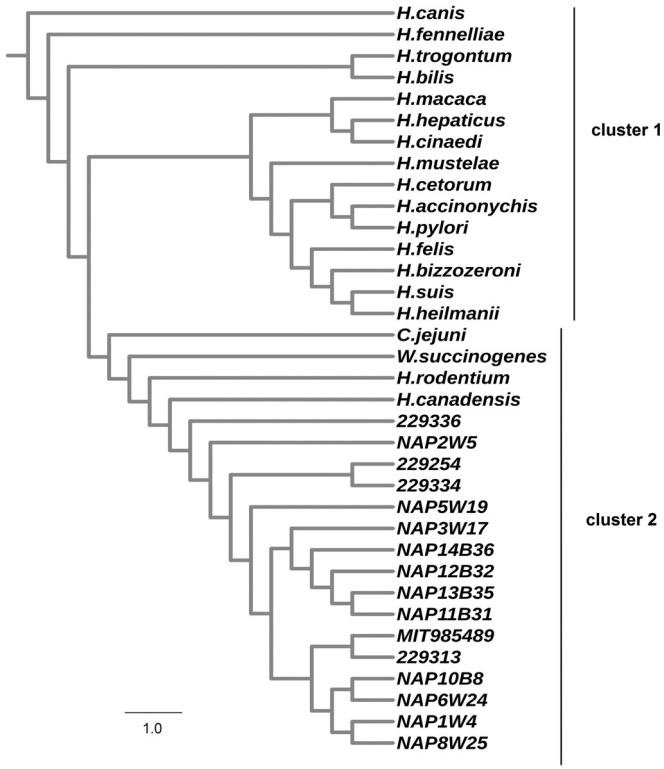


FIG 2 Core genome-based consensus maximum-likelihood phylogenetic tree of 11 in-house and 5 publicly available *H. pullorum* genomes with 1 *C. jejuni*, 1 *W. succinogenes*, and 17 other *Helicobacter* species genomes. The tree was generated after 1,000 replicates/bootstraps with RAxML. The output of RAxML was visualized with the iTOL tool.

per genome, while *H. pullorum* from broiler chicken strains contained only 14 strain-specific genes. The genomes of *H. pullorum* strains 229313 and MIT985489 carried 59 and 50 strain-specific genes, respectively (Fig. 5). Analysis of these strain-specific genes revealed that a majority of them were predicted to code for hypothetical proteins, prophages, transposons, and functions related to cell wall/envelope biogenesis and

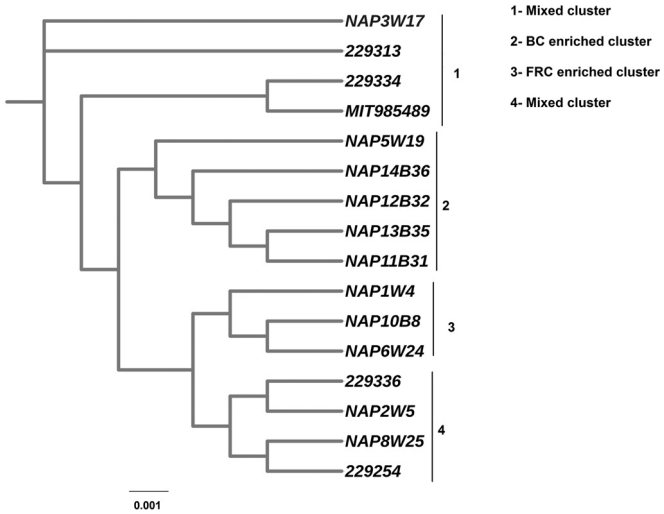


FIG 3 Core genome-based consensus maximum-likelihood phylogenetic tree of only *H. pullorum* isolates comprising 11 in-house and 5 publicly available *H. pullorum* genomes. The tree was generated after 1,000 replicates/bootstraps with RAxML and visualized by the iTOL tool. FRC, free-range chicken isolates; BC, broiler chicken isolates.

Downloaded from http://aem.asm.org/ on December 19, 2016 by guest

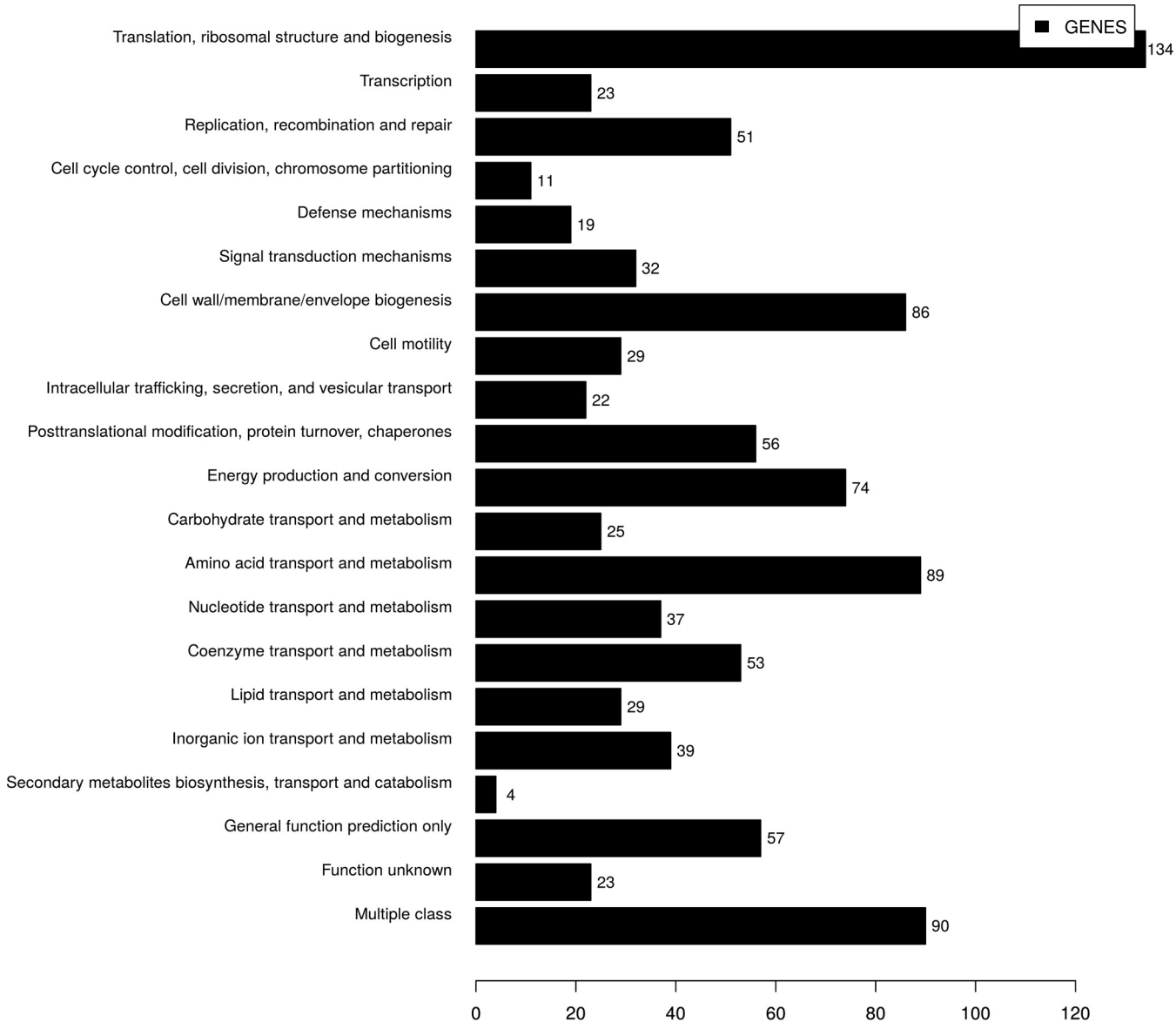


FIG 4 COG functional classification of 1,041 core genes obtained by comparing 13 *H. pullorum* isolates (11 in-house and 2 publicly available genomes). On the *x* axis is the number of genes in each functional category on the *y* axis.

transcription. In addition, some of these genes also code for type IV DNA secretion systems and type I restriction-modification systems. Moreover, a few strain-specific genes were also found to be associated with bacterial defense mechanisms, replication, recombination, and signal transduction.

Identification of virulence factors and mobile genetic elements. To understand the pathogenicity of *H. pullorum* and the most prevalent virulence profiles, we investigated the virulence gene contents of all of our isolates. A total of 182 different virulence factors were predicted from among our *H. pullorum* isolates (see Table S3 in the supplemental material). The results revealed that around 116 virulence genes were conserved in all of the *H. pullorum* isolates; the other 66 genes were present in some but absent from other strains of *H. pullorum*. It was found that about 40 such virulence genes encode functions related to bacterial motility that could be essential for the colonization of chicken ceca by bacteria. The presence of a histidine kinase (FlgS) and a response regulator protein (FlgR) also indicated the existence of bacterial two-component regulatory (TCR) systems. A total of 20 genes were found to encode

Downloaded from http://aem.asm.org/ on December 19, 2016 by guest

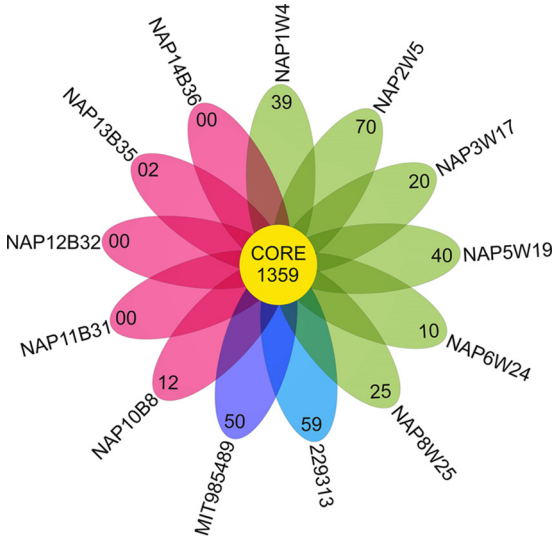


FIG 5 Diagrammatic representation of core and specific gene contents of free-range and broiler chicken isolates and two other *H. pullorum* strains (MIT985489 and 229313).

adhesion function proteins with the unique characteristic of being heat stable. Other virulence factors identified in *H. pullorum* strains included 18 genes that are part of a glycosylation system and 7 genes coding for hypothetical lipopolysaccharide endotoxins. Most importantly, all of the isolates encoded the CdtB subunit of a common cell cycle-arresting cytotoxic protein named CDT.

The draft genome sequences of *H. pullorum* strains from both broiler and free-range chickens were also screened for the presence of putative phage sequences, genomic islands, and insertion elements. The genomes of broiler chicken isolates NAP11B31, NAP12B32, and NAP13B35 carried various phage remnants along with one intact phage region. While NAP14B36 carried only one intact phage sequence, NAP10B8 harbored four incomplete phage regions. The sizes of the intact phage regions present in NAP11B31, NAP12B32, NAP13B35, and NAP14B36 were found to be 8.7, 6.4, 7.7, and 19.8 kb, respectively, containing 18, 12, 15, and 24 coding sequences (CDS), respectively. All intact phage regions contained *Escherichia* phage vB_EcoM-ep3, except for the intact region of NAP11B31, which contained *Stenotrophomonas* phage Smp131 of *Stenotrophomonas maltophilia*.

All of the *H. pullorum* isolates from free-range chickens contained incomplete phage-associated regions of various sizes, ranging from 5.6 to 15.8 kb. *Cronobacter sakazakii* phage GAP32 and *Cafeteria roenbergensis* phage BV-PW1 were the common phage-associated regions found in these isolates. These phage-associated regions consisted of genes encoding putative functions relevant to bacterial metabolism, membrane-associated protein functions, restriction-modification systems, and metalloproteases together with phage-related tail and base plate assembly proteins and some hypothetical proteins. For detailed descriptions of the phages identified in the genomes of *H. pullorum* and their gene contents, see Table S4 in the supplemental material.

Several genomic islands ranging in size from 4 to 67 kb were identified in 7 of the 11 *H. pullorum* genomes sequenced. No genomic island was detected in the NAP2W5, NAP3W17, NAP6W24, or NAP10B8 isolate. These islands mainly contain genes associated with bacterial outer membrane proteins, type II/IV and VI secretory systems, restriction-modification system(s), and putative hypothetical proteins. Table S5 in the supplemental material presents a detailed account of all of the genomic islands identified in seven isolates along with their major predicted functions.

DISCUSSION

From the above observations, it is evident that our study exhaustively unraveled, as a pioneering effort, the genotypic, phenotypic, chromosomal, and molecular charac-

Downloaded from http://aem.asm.org/ on December 19, 2016 by guest

teristics of *H. pullorum* isolates from the intestines of broiler and free-range chickens in India. Other studies (26, 32) previously reported on the prevalence of *H. pullorum* in the raw meat, cecum, and liver of poultry from different countries, which could possibly pose a high health risk to humans via its transmission through food similar to that of *Campylobacter* species. However, the prevalence and molecular characteristics of *H. pullorum* from chicken carcasses have not been studied in greater detail, including the genomic and epidemiological attributes of field strains. Our study also gains significance in the Indian context, as there is currently no effective implementation of antibiotic use guidelines in the poultry industry in this vast country that has a predominantly agriculture-based and rural-area-focused economy. On the other hand, the rate of consumption of chicken meat is increasing exponentially. The reason for the paucity of work in this area could be that *H. pullorum* is a fastidious organism and its isolation is troublesome; this has, in many ways, hindered the establishment of this organism as a foodborne pathogen. In this study, we validated and reassessed the usefulness of a convenient method for isolation of *H. pullorum* from the poultry cecum, a membrane filter technique described earlier (26).

Further, we have comprehensively characterized the AMR profiles of *H. pullorum* isolates from both broiler and free-range chickens by testing them against several different classes of antibiotics and also compared their genomic features. Previous studies involving antimicrobial susceptibilities of *H. pullorum* by Zanoni et al. (26) and Ceelen et al. (33) did not detect resistance to the macrolide and fluoroquinolone classes. Although Borges et al. (27) reported *H. pullorum* resistance against fluoroquinolones, they only included four strains isolated from a single poultry source and did not examine the ESBL production of *H. pullorum* strains. In our study, we found that all of the isolates were MDR, irrespective of their isolation source. Our findings showed that all of our Indian isolates were prominent in harboring resistance to various classes of antibiotics (quinolones/fluoroquinolones, cephalosporins, sulfonamides, and macrolides), unlike previous reports (26, 33). Moreover, all of our isolates were also resistant to clavulanic acid. We were expecting a lower prevalence of AMR in free-range chicken isolates because there is no intensive use of antibiotics in their rearing process, unlike that of broiler chickens. Surprisingly, our broiler and free-range chicken isolates showed similar resistance patterns, which might be due to the ability of these bacteria to acquire genes horizontally (34) from resistant microorganisms in the surrounding environment as the exposure of free-range chickens to environmental contamination is much greater than that of farm-reared chickens (35, 36). This could also be due to the indiscriminate use of antibiotics in humans and animals, which results in high selection pressure. Free-range chickens have high exposure to human and animal excreta and their wastes. Because they inhabit human communities, these birds pick up antibiotic residues/resistance determinants, thus making their gut microbiota resistant to different antibiotics. We believe that this is how the free-range and broiler chicken isolates in this study came to have similar AMR profiles.

Resistance to clavulanate might be due to hyperproduction of the enzyme beta-lactamase, which overwhelms the action of this enzyme inhibitor (37). However, the presence of an RND family efflux pump(s) and its involvement in clavulanate resistance cannot be ruled out because of its wide range of efflux substrates (38, 39). In addition to performing phenotypic antimicrobial susceptibility tests, we also genotypically confirmed the presence of molecular determinants of AMR toward elfamycin, quinolone/fluoroquinolone resistance genes, and a multidrug RND family transporter efflux pump in all of our isolates, with the aminoglycoside resistance genes being predominant (92% of our isolates). In particular, more aminoglycoside resistance genes were detected in *H. pullorum* from broiler chickens than in *H. pullorum* from free-range chickens.

The core genome is the part of a genome that is present in all of the isolates of a species and is crucial for survival in a particular habitat. Core genome analysis plays a key role in determining population structure, which in turn sheds light on the evolutionary trajectories of strains (40). Core genome-based phylogenetic analysis clearly

Downloaded from http://aem.asm.org/ on December 19, 2016 by guest

demonstrated an interspecies genetic relatedness of *Helicobacter* isolates in cluster 1 of Fig. 2. High intraspecies genomic diversity of *H. pullorum* from free-range chickens was observed, as shown in Fig. 3. Analysis of phylogenetic trees revealed that the genomes of *W. succinogenes*, *C. jejuni*, *H. canadensis*, *H. rodentium*, and *H. pullorum* clustered together, showing high conservation of their core genome content compared to that of other *Helicobacter* species, which indicates their common ancestry and suggests that *H. pullorum* could have descended from or shared lineages with any of the above species, particularly *H. canadensis*. The habitat of an organism also plays an important role in the evolution of bacterial species, as some of them, such as *H. canadensis* (29), *C. jejuni*, and *H. pullorum*, have the same host, and this might be the reason for shared core genome contents compared to others. The presence of a remarkable number of genomic islands and a large inventory of prophages in *H. pullorum* isolates points toward plausibility of its ability to acquire genes via horizontal gene transfer, which has a potential impact on the evolutionary adaptation of bacterial species (41).

Our analyses also made it clear that the genomes of *H. pullorum* isolates exhibited some degree of genomic diversity, which is consistent with a previous study based on two fingerprinting techniques, amplified fragment length polymorphism and pulsed-field gel electrophoresis (42). We found that strains from free-range chickens were genetically more diverse than broiler chicken isolates, which could possibly be due to the unrestricted movement of free-range chickens in their surroundings, where continuous exposure to a large number of microorganisms from a variety of foods and contaminated water and air shapes their gut microbiota composition. In contrast, broiler chickens were restricted at a specific place and fed formula feed with a clean water supply and compulsory exposure to antibiotics used as anabolics and administered also for prophylaxis and treatment. Because of this restricted environment and the continuous selection pressure of antibiotics, it was obvious that the guts of broiler chickens on poultry farms would contain a largely clonal and identical repertoire of microbial species that would be very different from that of their free-ranging counterparts. Hence, we saw less diversity in *H. pullorum* isolates obtained from broiler chickens.

Corroborating the results of whole-genome phylogeny, the results of strain-specific gene content analysis confirmed the increased genomic diversity of *H. pullorum* from free-range chickens. Consequently, we identified a higher number of specific genes in all of our free-range chicken isolates than in our broiler chicken isolates, three of which were completely lacking specific genes, while only 12 and 2 specific genes were observed in two such isolates. According to our restricted/unrestricted-movement assumption, free-range chickens have an opportunity to make their gut microbiota more dynamic and more prone to acquire genetically diverse bacteria from their surroundings than broiler chickens.

Although the pathogenic mechanisms of *H. pullorum* are not well understood, it is known that adherence, cell invasion, and intracellular resistance, along with toxin-producing factors, are prerequisites for establishing a bacterial infection and for its outcome. As motility is an essential process involved in direct colonization (43), the presence of a notable number of virulence genes associated with bacterial motility and signal transduction indicates *H. pullorum*'s potential to colonize chickens, as well as humans. Also, the presence of genes for a TCR system in *H. pullorum* may suggest its importance in supporting bacterial growth at various temperatures in different hosts (44).

A noteworthy number of genes were found to be associated with bacterial adherence, immune evasion, and glycosylation systems. These genes mostly correspond to bacterial membranous lipopolysaccharides and capsule biosynthesis and are considered heat-stable antigens and therefore may be required for the pathogen's stability and proper function at the high body temperature of chickens. In addition, the presence of about 20 glycosylation system-related genes makes it highly probable that they are required to change the carbohydrate structure and thus orient carbohydrates

Downloaded from http://aem.asm.org/ on December 19, 2016 by guest

to mimic surface-exposed host carbohydrate molecules, as in Guillain-Barre syndrome caused by *C. jejuni* (45).

In our virulence factor analysis, we found genes for synthesizing a common subunit of CDT, CdtB. We identified it as a *C. jejuni*-like toxin subunit; CdtB in *H. pullorum* was found to possess characteristics of type I DNases. Brief exposure of cultured cells to this subunit results in remarkable chromatin disruption, induction of cytoplasmic distention, and eventually cell cycle arrest (46). So, the existence of this protein in *H. pullorum* might play an important role in several complications associated with human, as well as animal, health.

Besides these virulence factors, the core genomes of all of our *H. pullorum* isolates were found to have an RND family efflux pump, which is not restricted only to drugs but could also play an important role in bacterial pathogenicity. In addition to their wide range of efflux substrates, these multiple drug resistance pumps also facilitate bacterial adaptation to different niches (47). The presence of AcrAB-TolC and CmeABC RND efflux pumps suggests *H. pullorum*'s extraordinary defense against a wide range of host-derived substrates, and these pumps could also play an important role in bacterial adherence to and invasion of host epithelial cells, as observed in *Salmonella* Typhimurium (48) and *C. jejuni* (49) infections, respectively. Because of the wide range of antibiotic resistances associated with them, multiple-drug efflux pumps may also serve as potential antibacterial targets, as their inhibition may help to achieve the maximum effectiveness of current and future antibiotics (50).

In conclusion, our study suggests that chickens could be a major source for transmission of the emerging MDR pathogen *H. pullorum* from poultry to humans. Additionally, our study provides a strong argument that *H. pullorum* is an emerging foodborne pathogen because of its high genomic relatedness to some seasoned pathogens such as *C. jejuni*. Given that India has the world's fifth largest poultry industry and a large consumer base of approximately 500 million customers for chicken meat, food-associated acquisition of AMR and MDR genes and virulence genotypes and phenotypes represents an alarming situation, and our study constitutes a baseline pioneering effort at dissecting this situation from a pathogen biology and epidemiology point of view. This would likely underpin increased and augmented surveillance systems for the safety of food for humans and also for monitoring animal health. Further, the genomes newly sequenced in this study would be an important resource for future studies involving the comparative genomics and molecular epidemiology of pyloric and nonpyloric *Helicobacter* isolates in different hosts.

MATERIALS AND METHODS

Sampling, isolation, and identification of strains. Samples from the gastrointestinal tracts of 55 broiler chickens and 45 free-range chickens from commercial retail poultry outlets were collected in separate sterilized plastic tubes and processed within 3 to 4 h. The ceca were aseptically severed, and their surfaces were washed with 1× phosphate-buffered saline to minimize contamination. The cecal contents were removed, and tissues were cut into small pieces and put into a sterile tissue homogenizer containing 5 ml of brain heart infusion (BHI) broth to get a homogeneous suspension. Tissue homogenates (100 μl) were diluted in 300 μl of BHI broth. Diluted samples were inoculated onto BHI agar supplemented with 10% horse serum and antibiotics (vancomycin [10 mg/liter], amphotericin B [3 mg/liter], and cefoperazone [32 mg/liter]) by the modified filter technique of Steele and McDermott (51). In brief, sterile cellulose acetate membrane filters (0.65-μm pore size) were applied with sterile tweezers onto the surface of a BHI agar plate and 100 μl of diluted homogenate was placed in the middle of each membrane filter. The plate was then incubated for 1 to 2 h in a microaerobic atmosphere (90% N_2, 5% O_2, 5% CO_2) at 37°C. The membrane filters were then removed with sterile tweezers, and the filtrate was streaked onto the agar with a sterile loop. The plates were then incubated under the same conditions for 3 to 4 days. Approximately 5 to 10 grayish white bacterial colonies were observed on the plates in each case.

Colony PCR assays for the *H. pullorum* 16S rRNA gene with specific primers (forward, 5′ ATG AAT GCTAGT TGT TGT CAG 3′; reverse, 5′ GAT TGG CTC CACTTC ACA 3′) were performed for confirmation to the species level. Amplified products were run on 1.2% agarose gel and visualized with a gel documentation system (Major Science, USA). Amplified products were also sequenced for further confirmation.

Antimicrobial susceptibility testing and identification of AMR genes. The standard Kirby-Bauer disk diffusion method was used to test the antimicrobial susceptibility of all isolates. The antimicrobial drugs (HiMedia, Mumbai, India) used for this experiment included oxytetracycline (30 μg), doxycycline

Downloaded from http://aem.asm.org/ on December 19, 2016 by guest

hydrochloride (30 μg), chlortetracycline (30 μg), ciprofloxacin (5 μg), nalidixic acid (30 μg), enrofloxacin (5 μg), chloramphenicol (30 μg), neomycin (10 μg), co-trimoxazole (25 μg), cefotaxime (30 μg), clarithromycin (15 μg), and colistin (10 μg). To standardize the bacterial suspension (in BHI broth medium), the bacterial density was adjusted to a 0.5 McFarland standard and spread over the entire surface of each BHI agar plate with a sterile cotton swab. Antimicrobial discs were placed on the agar surface with sterile tweezers, and then the plates were incubated at 37°C for 3 to 4 days under microaerobic conditions as described before. A double-disc synergy test was used to check all isolates for ESBL production (52). Thereafter, both results were interpreted according to Clinical and Laboratory Standards Institute recommendations (73). In addition, the amino acid sequences encoded by the predicted genes of 11 *H. pullorum* strains were compared with those in the comprehensive antibiotic resistance database (http://arpcard.mcmaster.ca) by using BLASTp with identity and query coverage of >50% and >80%, respectively.

Isolation of genomic DNA and WGS. Genomic DNA was extracted with the Qiagen DNeasy blood and tissue kit (Qiagen, Germany). The yield and quality of DNA were assessed with a NanoDrop spectrophotometer (Thermo Scientific). WGS of all isolates was carried out with the Illumina MiSeq sequencer, which generated approximately 0.5 to 1.2 million paired-end reads of 300 bp, with an average insert size of 400 to 500 bp. Filtering and trimming of the paired-end reads were done with the help of NGS QC Toolkit (53) and FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), respectively. To determine the optimum hash length and the best assembly, VelevetOptimiser.pl script and Velvet (v1.2.10), respectively, were employed (54). Gene prediction and functional annotation of the draft genomes were carried out with RAST (Rapid Annotation Using Subsystem Technology) (55–57), followed by validation with GeneMarkS (58) and Glimmer (59). Genome statistics of these strains were gleaned with ARTEMIS (60). The numbers of tRNA and rRNA genes were identified with tRNAscan-SE (61) and RNAmmer (62), respectively.

Phylogenetic analysis. The predicted amino acid sequences of 11 *H. pullorum* isolates along with 5 other strains of *H. pullorum*, 17 other *Helicobacter* species, 1 *C. jejuni* strain, and 1 *Wolinella succinogenes* strain from the National Center for Biotechnology Information (see Table S1 in the supplemental material) were compared all versus all with BLASTp, followed by the Markov clustering algorithm to identify orthologs (63). Here, the percent identity threshold and E value were adjusted to 70% and 0.00001, respectively. Only genes comprising a minimum of 50 amino acids were included in ortholog analysis. The orthologous gene clusters that did not have any paralogs were considered for further core genome-based phylogeny. Core orthologous genes were aligned with Mafft by using default parameters, followed by removal of gaps with TrimAL (64, 65). The processed alignments were concatenated by using an in-house perl script and used as the input for RAxML (66). The program was run by using the general time-reversible nucleotide substitution model with gamma correction. A consensus maximum-likelihood-based phylogenetic tree was generated after 1,000 replicates/bootstraps. Similarly, we constructed a separate phylogenetic tree exclusively for *H. pullorum* isolates of free-range and broiler chickens to get a better understanding of their genotypic resemblance. The output of RAxML was visualized with the interactive tree of life (iTOL) tool (67). Further, the whole genomes of free-range and broiler chicken isolates of *H. pullorum* were also compared by using BRIG (68) to visualize the approximate similarity within their genomes with BLASTn.

Core and pan-genome analysis. The sequenced *H. pullorum* isolates along with two others (human *H. pullorum* isolate MIT985489 and poultry isolate 229313) were analyzed to calculate their core and accessory genome contents. The orthologs were determined on the basis of the threshold values as mentioned in the above section. The orthologous gene clusters that carried orthologs in all of the genomes constituted the core content, whereas others were considered accessory gene content.

The specific gene content of *H. pullorum* was analyzed by pooling the predicted amino acid sequences of 17 different species of *Helicobacter* (see Table S1 in the supplemental material) together with all of our *H. pullorum* isolates ($n = 16$). Core and accessory gene contents were identified with the OrthoMCL program. *H. pullorum* accessory genes were extracted from the pooled accessory gene content of all *Helicobacter* species. BLASTp analysis was performed by using *H. pullorum* accessory proteins as the query with *Helicobacter* accessory proteins as the database. The best hits of BLASTp output for each protein were obtained, and proteins <70% identical were considered *H. pullorum*-specific proteins. Further, *H. pullorum* strain-specific genes were also identified by using an in-house perl script. The functional categories of the strain-specific genes were predicted by using the COG database (69). The genes were classified into multiple classes that comprise more than one domain of a COG functional category, while those genes that did not show any significant hit against the database were identified as hypothetical genes.

Identification of virulence genes and detection of mobile genetic elements. The whole genomes of all of our isolates were screened with the BLASTp program for the presence of virulence genes listed in the Virulence Factor Database (VFDB, http://www.mgc.ac.cn/VFs/). The amino acid sequences of the predicted genes from all of the 11 strains were compared with the database of virulence-related genes of *C. jejuni* obtained from the VFDB. The identity cutoff and query coverage values were kept at >50% and >80%, respectively.

In addition, all of these strains were screened for the presence of mobile genetic elements. Phages or phage-like elements were identified by the PHAST tool (70). Genomic islands in these strains were determined with IslandViewer (71), which makes use of three algorithms (IslandPick, IslandPath-DIMOB, and SIGI-HMM) for prediction and facilitates the visualization and analysis of these islands. Insertion elements were predicted with the ISfinder web tool (72), which compares the query sequences against a database of insertion sequence elements.

Downloaded from http://aem.asm.org/ on December 19, 2016 by guest

Accession number(s). This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under accession numbers MAOZ00000000 (*H. pullorum* NAP10B8), MAPA00000000 (*H. pullorum* NAP8W25), MAPB00000000 (*H. pullorum* NAP6W24), MAPC00000000 (*H. pullorum* NAP5W19), MAPD00000000 (*H. pullorum* NAP3W17), MAPE00000000 (*H. pullorum* NAP2W5), MANJ00000000 (*H. pullorum* NAP13B35), MANK00000000 (*H. pullorum* NAP14B36), MAJF00000000 (*H. pullorum* NAP11B31), MAJG00000000 (*H. pullorum* NAP12B32), and LXWI00000000 (*H. pullorum* NAP1W4), including all 11 of the isolates sequenced in this study.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/AEM.02305-16.

TEXT S1, PDF file, 1.2 MB.

ACKNOWLEDGMENTS

We acknowledge funding received from the Department of Biotechnology, Government of India (BT/PR6921/MED/29/699/2013), and support from the Indo-German International Research Training Group, Internationales Graduiertenkolleg (GRK1673)—functional molecular infection epidemiology, an initiative of the German Research Foundation (DFG) and the University of Hyderabad (India), for which N.A. served as a speaker from the Indian side. N.A. is an adjunct professor at the Academy of Scientific and Innovative Research (ACSIR), India. S.Q. acknowledges the award of a junior research fellowship from the University Grants Commission of India.

We thank Lothar H. Wieler for his advice and discussions.

REFERENCES

1. Stanley J, Linton D, Burnens AP, Dewhirst FE, On SL, Porter A, Owen RJ, Costas M. 1994. Helicobacter pullorum sp. nov.-genotype and phenotype of a new species isolated from poultry and from human patients with gastroenteritis. Microbiology 140(Pt 12):3441–3449. https://doi.org/10.1099/13500872-140-12-3441.
2. Steinbrueckner B, Haerter G, Pelz K, Weiner S, Rump JA, Deissler W, Bereswill S, Kist M. 1997. Isolation of Helicobacter pullorum from patients with enteritis. Scand J Infect Dis 29:315–318. https://doi.org/10.3109/00365549709019053.
3. Burnens AP, Stanley J, Morgenstern R, Nicolet J. 1994. Gastroenteritis associated with Helicobacter pullorum. Lancet 344:1569–1570.
4. Bascuñana P, Padrones I, Picazo JJ, Velasco AC. 2011. Diarrhea associated with Helicobacter pullorum in a child. Rev Esp Quimioter 24:50–51. (In Spanish.)
5. Bohr UR, Glasbrenner B, Primus A, Zagoura A, Wex T, Malfertheiner P. 2004. Identification of enterohepatic *Helicobacter* species in patients suffering from inflammatory bowel disease. J Clin Microbiol 42:2766–2768. https://doi.org/10.1128/JCM.42.6.2766-2768.2004.
6. Veijola L, Nilsson I, Halme L, Al-Soud WA, Makinen J, Ljungh A, Rautelin H. 2007. Detection of Helicobacter species in chronic liver disease and chronic inflammatory bowel disease. Ann Med 39:554–560. https://doi.org/10.1080/07853890701545714.
7. Karagin PH, Stenram U, Wadström T, Ljungh A. 2010. Helicobacter species and common gut bacterial DNA in gallbladder with cholecystitis. World J Gastroenterol 16:4817–4822. https://doi.org/10.3748/wjg.v16.i38.4817.
8. Ponzetto A, Pellicano R, Leone N, Cutufia MA, Turrini F, Grigioni WF, D'Errico A, Mortimer P, Rizzetto M, Silengo L. 2000. Helicobacter infection and cirrhosis in hepatitis C virus carriage: is it an innocent bystander or a troublemaker? Med Hypotheses 54:275–277. https://doi.org/10.1054/mehy.1999.0987.
9. Rocha M, Avenaud P, Menard A, Le Bail B, Balabaud C, Bioulac-Sage P, de Magalhaes Queiroz DM, Megraud F. 2005. Association of Helicobacter species with hepatitis C cirrhosis with or without hepatocellular carcinoma. Gut 54:396–401. https://doi.org/10.1136/gut.2004.042168.
10. Pellicano R, Menard A, Rizzetto M, Megraud F. 2008. Helicobacter species and liver diseases: association or causation? Lancet Infect Dis 8:254–260. https://doi.org/10.1016/S1473-3099(08)70066-5.
11. Casswall TH, Nemeth A, Nilsson I, Wadstrom T, Nilsson HO. 2010. Helicobacter species DNA in liver and gastric tissues in children and adolescents with chronic liver disease. Scand J Gastroenterol 45:160–167. https://doi.org/10.3109/00365520903426915.
12. Fox JG, Dewhirst FE, Shen Z, Feng Y, Taylor NS, Paster BJ, Ericson RL, Lau CN, Correa P, Araya JC, Roa I. 1998. Hepatic Helicobacter species identified in bile and gallbladder tissue from Chileans with chronic cholecystitis. Gastroenterology 114:755–763. https://doi.org/10.1016/S0016-5085(98)70589-X.
13. Laharie D, Asencio C, Asselineau J, Bulois P, Bourreille A, Moreau J, Bonjean P, Lamarque D, Pariente A, Soule JC, Charachon A, Coffin B, Perez P, Megraud F, Zerbib F. 2009. Association between entero-hepatic Helicobacter species and Crohn's disease: a prospective cross-sectional study. Aliment Pharmacol Ther 30:283–293. https://doi.org/10.1111/j.1365-2036.2009.04034.x.
14. Chien CC, Taylor NS, Ge Z, Schauer DB, Young VB, Fox JG. 2000. Identification of cdtB homologues and cytolethal distending toxin activity in enterohepatic Helicobacter spp. J Med Microbiol 49:525–534. https://doi.org/10.1099/0022-1317-49-6-525.
15. Young VB, Knox KA, Schauer DB. 2000. Cytolethal distending toxin sequence and activity in the enterohepatic pathogen *Helicobacter hepaticus*. Infect Immun 68:184–191. https://doi.org/10.1128/IAI.68.1.184-191.2000.
16. Taylor NS, Ge Z, Shen Z, Dewhirst FE, Fox JG. 2003. Cytolethal distending toxin: a potential virulence factor for Helicobacter cinaedi. J Infect Dis 188:1892–1897. https://doi.org/10.1086/379837.
17. Young VB, Knox KA, Pratt JS, Cortez JS, Mansfield LS, Rogers AB, Fox JG, Schauer DB. 2004. In vitro and in vivo characterization of *Helicobacter hepaticus* cytolethal distending toxin mutants. Infect Immun 72:2521–2527. https://doi.org/10.1128/IAI.72.5.2521-2527.2004.
18. Young VB, Chien CC, Knox KA, Taylor NS, Schauer DB, Fox JG. 2000. Cytolethal distending toxin in avian and human isolates of Helicobacter pullorum. J Infect Dis 182:620–623. https://doi.org/10.1086/315705.
19. Jinadasa RN, Bloom SE, Weiss RS, Duhamel GE. 2011. Cytolethal distending toxin: a conserved bacterial genotoxin that blocks cell cycle progression, leading to apoptosis of a broad range of mammalian cell lineages. Microbiology 157:1851–1875. https://doi.org/10.1099/mic.0.049536-0.
20. Varon C, Mocan I, Mihi B, Pere-Vedrenne C, Aboubacar A, Morate C, Oleastro M, Doignon F, Laharie D, Megraud F, Menard A. 2014. Helicobacter pullorum cytolethal distending toxin targets vinculin and cortactin and triggers formation of lamellipodia in intestinal epithelial cells. J Infect Dis 209:588–599. https://doi.org/10.1093/infdis/jit539.
21. Shah PM, Schafer V, Knothe H. 1993. Medical and veterinary use of antimicrobial agents: implications for public health. A clinician's view on antimicrobial resistance. Vet Microbiol 35:269–274.

Downloaded from http://aem.asm.org/ on December 19, 2016 by guest

22. Klevens RM, Morrison MA, Nadle J, Petit S, Gershman K, Ray S, Harrison LH, Lynfield R, Dumyati G, Townes JM, Craig AS, Zell ER, Fosheim GE, McDougal LK, Carey RB, Fridkin SK, Active Bacterial Core Surveillance (ABCs) MRSA Investigators. 2007. Invasive methicillin-resistant Staphylococcus aureus infections in the United States. JAMA 298:1763–1771. https://doi.org/10.1001/jama.298.15.1763.
23. Boucher HW, Talbot GH, Bradley JS, Edwards JE, Gilbert D, Rice LB, Scheld M, Spellberg B, Bartlett J. 2009. Bad bugs, no drugs: no ESKAPE! An update from the Infectious Diseases Society of America. Clin Infect Dis 48:1–12. https://doi.org/10.1086/595011.
24. Schroeder CM, Zhao C, DebRoy C, Torcolini J, Zhao S, White DG, Wagner DD, McDermott PF, Walker RD, Meng J. 2002. Antimicrobial resistance of *Escherichia coli* O157 isolated from humans, cattle, swine, and food. Appl Environ Microbiol 68:576–581. https://doi.org/10.1128/AEM.68.2.576-581.2002.
25. Angulo FJ, Johnson KR, Tauxe RV, Cohen ML. 2000. Origins and consequences of antimicrobial-resistant nontyphoidal Salmonella: implications for the use of fluoroquinolones in food animals. Microb Drug Resist 6:77–83. https://doi.org/10.1089/mdr.2000.6.77.
26. Zanoni RG, Rossi M, Giacomucci D, Sanguinetti V, Manfreda G. 2007. Occurrence and antibiotic susceptibility of Helicobacter pullorum from broiler chickens and commercial laying hens in Italy. Int J Food Microbiol 116:168–173. https://doi.org/10.1016/j.ijfoodmicro.2006.12.007.
27. Borges V, Santos A, Correia CB, Saraiva M, Menard A, Vieira L, Sampaio DA, Pinheiro M, Gomes JP, Oleastro M. 2015. Helicobacter pullorum isolated from fresh chicken meat: antibiotic resistance and genomic traits of an emerging foodborne pathogen. Appl Environ Microbiol 81:8155–8163. https://doi.org/10.1128/AEM.02394-15.
28. González A, Piqueres P, Moreno Y, Canigral I, Owen RJ, Hernandez J, Ferrus MA. 2008. A novel real-time PCR assay for the detection of Helicobacter pullorum-like organisms in chicken products. Int Microbiol 11:203–208.
29. Fox JG, Chien CC, Dewhirst FE, Paster BJ, Shen Z, Melito PL, Woodward DL, Rodgers FG. 2000. *Helicobacter canadensis* sp. nov. isolated from humans with diarrhea as an example of an emerging pathogen. J Clin Microbiol 38:2546–2549.
30. Waldenström J, On SL, Ottvall R, Hasselquist D, Harrington CS, Olsen B. 2003. Avian reservoirs and zoonotic potential of the emerging human pathogen *Helicobacter canadensis*. Appl Environ Microbiol 69:7523–7526. https://doi.org/10.1128/AEM.69.12.7523-7526.2003.
31. Olivares J, Bernardini A, Garcia-Leon G, Corona F, Sanchez MB, Martinez JL. 2013. The intrinsic resistome of bacterial pathogens. Front Microbiol 4:103. https://doi.org/10.3389/fmicb.2013.00103.
32. Manfreda G, Parisi A, Lucchi A, Zanoni RG, De Cesare A. 2011. Prevalence of *Helicobacter pullorum* in conventional, organic, and free-range broilers and typing of isolates. Appl Environ Microbiol 77:479–484. https://doi.org/10.1128/AEM.01712-10.
33. Ceelen L, Decostere A, Devriese LA, Ducatelle R, Haesebrouck F. 2005. In vitro susceptibility of Helicobacter pullorum strains to different antimicrobial agents. Microb Drug Resist 11:122–126. https://doi.org/10.1089/mdr.2005.11.122.
34. 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.
35. Schwartz T, Kohnen W, Jansen B, Obst U. 2003. Detection of antibiotic-resistant bacteria and their resistance genes in wastewater, surface water, and drinking water biofilms. FEMS Microbiol Ecol 43:325–335. https://doi.org/10.1111/j.1574-6941.2003.tb01073.x.
36. Oluyege JO, Dada AC, Odeyemi AT. 2009. Incidence of multiple antibiotic resistant Gram-negative bacteria isolated from surface and underground water sources in south western region of Nigeria. Water Sci Technol 59:1929–1936. https://doi.org/10.2166/wst.2009.219.
37. Gulay Z, Bicmen M, Amyes SG, Yulug N. 2000. Beta-lactamase patterns and betalactam/clavulanic acid resistance in Escherichia coli isolated from fecal samples from healthy volunteers. J Chemother 12:208–215. https://doi.org/10.1179/joc.2000.12.3.208.
38. Pumbwe L, Randall LP, Woodward MJ, Piddock LJ. 2004. Expression of the efflux pump genes cmeB, cmeF and the porin gene porA in multiple-antibiotic-resistant Campylobacter jejuni. J Antimicrob Chemother 54:341–347. https://doi.org/10.1093/jac/dkh331.
39. Elkins CA, Mullis LB. 2006. Mammalian steroid hormones are substrates for the major RND- and MFS-type tripartite multidrug efflux pumps of *Escherichia coli*. J Bacteriol 188:1191–1195. https://doi.org/10.1128/JB.188.3.1191-1195.2006.
40. Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. 2003. The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet 4:981–994. https://doi.org/10.1038/nrg1226.
41. Juhas M, van der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW. 2009. Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol Rev 33:376–393. https://doi.org/10.1111/j.1574-6976.2008.00136.x.
42. Gibson JR, Ferrus MA, Woodward D, Xerry J, Owen RJ. 1999. Genetic diversity in Helicobacter pullorum from human and poultry sources identified by an amplified fragment length polymorphism technique and pulsed-field gel electrophoresis. J Appl Microbiol 87:602–610.
43. Ottemann KM, Lowenthal AC. 2002. *Helicobacter pylori* uses motility for initial colonization and to attain robust infection. Infect Immun 70:1984–1990. https://doi.org/10.1128/IAI.70.4.1984-1990.2002.
44. Brás AM, Chatterjee S, Wren BW, Newell DG, Ketley JM. 1999. A novel *Campylobacter jejuni* two-component regulatory system important for temperature-dependent growth and colonization. J Bacteriol 181:3298–3302.
45. Allos BM. 1998. Campylobacter jejuni infection as a cause of the Guillain-Barre syndrome. Infect Dis Clin North Am 12:173–184.
46. Lara-Tejero M, Galan JE. 2000. A bacterial toxin that controls cell cycle progression as a deoxyribonuclease I-like protein. Science 290:354–357. https://doi.org/10.1126/science.290.5490.354.
47. Blanco P, Hernando-Amado S, Reales-Calderon J, Corona F, Lira F, Alcalde-Rico M, Bernardini A, Sanchez M, Martinez J. 2016. Bacterial multidrug efflux pumps: much more than antibiotic resistance determinants. Microorganisms 4:E14. https://doi.org/10.3390/microorganisms4010014.
48. Buckley AM, Webber MA, Cooles S, Randall LP, La Ragione RM, Woodward MJ, Piddock LJ. 2006. The AcrAB-TolC efflux system of Salmonella enterica serovar Typhimurium plays a role in pathogenesis. Cell Microbiol 8:847–856. https://doi.org/10.1111/j.1462-5822.2005.00671.x.
49. Lin J, Sahin O, Michel LO, Zhang Q. 2003. Critical role of multidrug efflux pump CmeABC in bile resistance and in vivo colonization of *Campylobacter jejuni*. Infect Immun 71:4250–4259. https://doi.org/10.1128/IAI.71.8.4250-4259.2003.
50. Vila J, Martinez JL. 2008. Clinical impact of the over-expression of efflux pump in nonfermentative Gram-negative bacilli, development of efflux pump inhibitors. Curr Drug Targets 9:797–807. https://doi.org/10.2174/138945008785747806.
51. Steele TW, McDermott SN. 1984. The use of membrane filters applied directly to the surface of agar plates for the isolation of Campylobacter jejuni from feces. Pathology 16:263–265.
52. Hussain A, Ranjan A, Nandanwar 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.
53. 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.
54. Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829. https://doi.org/10.1101/gr.074492.107.
55. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9:75. https://doi.org/10.1186/1471-2164-9-75.
56. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R. 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res 42:D206–D214. https://doi.org/10.1093/nar/gkt1226.
57. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, Shukla M, Thomason JA, III, Stevens R, Vonstein V, Wattam AR, Xia F. 2015. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Sci Rep 5:8365. https://doi.org/10.1038/srep08365.
58. Besemer J, Lomsadze A, Borodovsky M. 2001. GeneMarkS: a self-training

Downloaded from http://aem.asm.org/ on December 19, 2016 by guest

method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res 29:2607–2618.
59. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27:4636–4641.
60. Mural RJ. 2000. ARTEMIS: a tool for displaying and annotating DNA sequence. Brief Bioinform 1:199–200. https://doi.org/10.1093/bib/1.2.199.
61. Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964.
62. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35:3100–3108. https://doi.org/10.1093/nar/gkm160.
63. Li L, Stoeckert CJ, Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189. https://doi.org/10.1101/gr.1224503.
64. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010.
65. Capella-Gutiérrez S, Silla-Martinez JM, Gabaldon 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.
66. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033.
67. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44:W242–W245. https://doi.org/10.1093/nar/gkw290.
68. 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.
69. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH. 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 39:D225–D229. https://doi.org/10.1093/nar/gkq1189.
70. 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.
71. Langille MG, Brinkman FS. 2009. IslandViewer: an integrated interface for computational identification and visualization of genomic islands. Bioinformatics 25:664–665. https://doi.org/10.1093/bioinformatics/btp030.
72. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. 2006. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res 34:D32–D36. https://doi.org/10.1093/nar/gkj014.
73. Clinical and Laboratory Standards Institute. 2009. Performance standards for antimicrobial disk susceptibility tests; approved standard—10th edition. CLSI document M02-A10. CLSI, Wayne, PA.

Downloaded from http://aem.asm.org/ on December 19, 2016 by guest

SCIENTIFIC REPORTS



OPEN

SUBJECT AREAS:
PATHOGENS
COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

Received
4 September 2014

Accepted
24 November 2014

Published
12 December 2014

Correspondence and requests for materials should be addressed to N.A. (niyaz.ahmed@uohyd.ac.in)

Genome dynamics and evolution of *Salmonella* Typhi strains from the typhoid-endemic zones

Ramani Baddam, Narender Kumar, Sabiha Shaik, Aditya Kumar Lankapalli & Niyaz Ahmed

Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad, Gachibowli, Hyderabad 500046, India.

Typhoid fever poses significant burden on healthcare systems in Southeast Asia and other endemic countries. Several epidemiological and genomic studies have attributed pseudogenisation to be the major driving force for the evolution of *Salmonella* Typhi although its real potential remains elusive. In the present study, we analyzed genomes of *S.* Typhi from different parts of Southeast Asia and Oceania, comprising of isolates from outbreak, sporadic and carrier cases. The genomes showed high genetic relatedness with limited opportunity for gene acquisition as evident from pan-genome structure. Given that pseudogenisation is an active process in *S.* Typhi, we further investigated core and pan-genome profiles of functional and pseudogenes separately. We observed a decline in core functional gene content and a significant increase in accessory pseudogene content. Upon functional classification, genes encoding metabolic functions formed a major constituent of pseudogenes as well as core functional gene clusters with SNPs. Further, an in-depth analysis of accessory pseudogene content revealed the existence of heterogeneous complements of functional and pseudogenes among the strains. In addition, these polymorphic genes were also enriched in metabolism related functions. Thus, the study highlights the existence of heterogeneous strains in a population with varying metabolic potential and that *S.* Typhi possibly resorts to metabolic fine tuning for its adaptation.

Salmonella are enteric bacteria that can infect a broad range of host species causing various infectious diseases. Presently, there are two well recognized species of *Salmonella* - *S. bongori* and *S. enterica*. Further, based on Kauffman-White classification scheme, *S. enterica* is divided into six distinct subspecies and more than 2500 serovars[1]. However, serovar *S.* Typhi of *Salmonella enterica* subspecies *enterica* infects only humans resulting in a systemic infection, typhoid fever[2]. This infection is of immense concern to public health worldwide as it is responsible for about 21.6 million cases of which 1% become fatal, on an average, per year[3]. About 90% of this morbidity and mortality stems from Asia due to high endemicity of typhoid fever in developing countries where drinking water quality and sewage treatment facilities are poor[4]. The control and prevention strategies are also severely hampered due to the emergence of antibiotic resistant strains in these regions, which are responsible for periodic outbreaks and sporadic cases causing severe complications and mortality[5]. Further, the presence of these antimicrobial resistance genes carried on mobile elements such as integrons and self-transmissible plasmids like that of IncH1, which was reported to be associated with many strains from endemic zones such as Vietnam, pose a constant threat to public health world-wide[6,7].

For a host adapted strain like *S.* Typhi, survival in the host and dissemination are vital for establishing persistent infections. Some infected individuals even serve as asymptomatic reservoirs who continue to shed the bacterium in stools for a long period of time[8,9]. The studies on transmission dynamics of *salmonellae* have emphasized on monitoring of the carrier isolates for effective epidemiological tracking and surveillance[10]. The carrier isolates have also been shown to exhibit similar pulsed field gel electrophoresis (PFGE) profiles with other *S.* Typhi isolates from various regions of Southeast Asia. Therefore, it appears that the spread of *S.* Typhi occurs mostly through carrier individuals[11].

In the past, various genomic studies have attributed signatures like pseudogenisation (loss of gene function) or gene deletion for host restriction in pathogenic bacteria[12]. It was also reported that in human-restricted serovar *S.* Typhi, pseudogenisation is an active process compared to other generalist serovars like *S.* Typhimurium[13]. Further, the extent of this pseudogenisation also varies considerably even among host restricted serovars[14–16]. Pseudogenes constitute up to 4.5% of *S.* Typhi gene pool, making them an important driver of genome

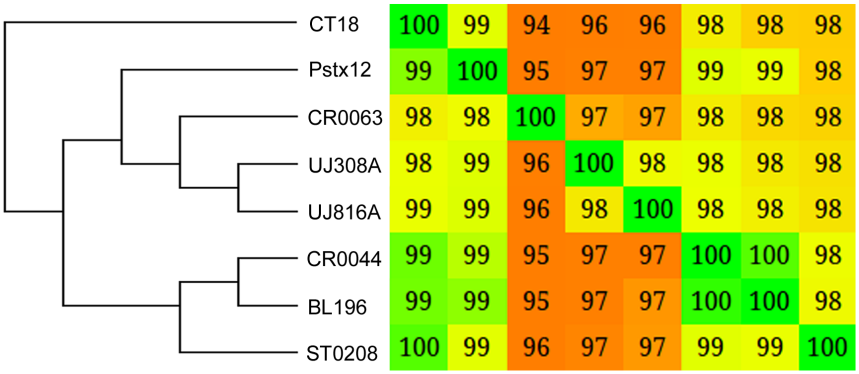


Figure 1 | Phylogenomic Tree. The whole genome information was used to build the distance matrix using Gegenees. The phylogenetic tree was developed using SplitsTree by NJ method. This revealed close similarity among genomes and also co-clustering of strains isolated from the same regions.

re-assortment over time[17]. However the potential role of pseudogenisation in persistence and adaptation of *S.* Typhi still remains elusive.

Given this, it is important to characterize *S.* Typhi's pan-genome, more importantly with respect to functional and pseudogene complements and investigate their gene-frequency distributions among various strains. The pan-genome of a species is the complete inventory of genes in the population and is always significantly greater than the gene content of an individual[18]. The pan genome is composed of both 'core genome' and 'accessory genome' where accessory part is comprised of genes shared by some but not all strains. This accessory or dispensable part confers various selective advantages such as antibiotic resistance, niche adaptation, pathogenicity and host specificity[19,20]. However, the residual core part of genome that keeps a very high sequence similarity of about 95% ANI (Average Nucleotide Identity), encodes all the fundamental biological processes essential for survival[18]. Thus, the pan-genome analysis helps us to better understand the population genetic structure and provides cues about the mechanisms underlying adaptation and evolution of bacteria. Studies based on whole genome comparative analyses carried out at the population level involving other *S. enterica* serovars such as Paratyphi A and Agona have recently provided significant insights into the evolution of these serovars[21,22].

The whole genome sequences corresponding to eight strains previously isolated from different endemic regions of Southeast Asia and Oceania were extensively analyzed in this study. These strains were associated with different clinical manifestations - outbreaks, sporadic cases, carrier strains and fatal episodes. The strain BL196 was associated with a large outbreak in Kelantan, Malaysia in 2005[23]. Strains CR0044 and CR0063 were isolated from carrier individuals in 2007 after an outbreak and were reported to share PFGE profiles with the strain BL196[24,25]. Strain ST0208 was associated with a sporadic case in Kuala Lumpur, Malaysia[26]. The previous findings have also recorded shared PFGE patterns among the isolates from Southeast Asia[11]. Strains UJ308A and UJ816A were isolated in Papua New Guinea from fatal and non-fatal cases, respectively, in 1998[27]. Genomes of multi-drug resistant strains, CT18 from Vietnam, and P-stx-12 strain isolated from a carrier individual in India[28,29], were also analyzed. Some of the earlier studies based on PFGE observed minimal to moderate diversity among the isolates from Papua New Guinea and elsewhere in Asia[30,31], thus verifying the limited observed diversity if not a clonal nature of this organism. Herein, we analyzed genomes of the strains described in some of the above pioneering studies. These genomes, although limited in number, were chosen owing to their being most authentic available representatives of geographically distinct populations from different endemic countries such as India, Vietnam, Papua New Guinea and Malaysia and thus were used for extensive genomic analyses hitherto unreported for these unique strains. Our comprehensive genomic analyses reported herein highlight the possible evolutionary mechanisms and in particular, the impact of pseudogenes on the evolution of *S.* Typhi in different patient types from the typhoid-endemic countries of the east.

Results

Phylogeny. The whole genome based phylogenetic tree allowed us to understand the close genetic relationship among various strains as shown in Figure 1. The strain BL196 isolated during the outbreak, and the carrier strain CR0044 isolated a year later, co-clustered revealing close similarity. This suggests that the strain CR0044 could have emerged due to clonal expansion of BL196, whereas another carrier strain CR0063 might have accumulated enough variations allowing it to cluster separately. The two strains isolated from Papua New Guinea, UJ816A and UJ308A, also clustered together with respect to all other strains. This observation by whole genome based phylogeny corroborates with the PFGE based analysis of Thong *et al*, where *S.* Typhi strains from Papua New Guinea showed highly similar PFGE patterns exhibiting limited genetic diversity among the strains[30]. As typhoid cases were rarely detected in Papua New Guinea before 1985, the limited observed diversity might be due to clonal expansion of a single ancestral strain[30]. The strains CT18, P-stx-12, ST0208 have shown up independently in the tree. The close similarity of these genomes is also reflected in whole genome alignment as depicted (Figure 2). This analysis once again reinforces the genetically monomorphic nature of this pathogen and our observations are in concurrence with the previous findings based on Multilocus sequence typing and other techniques[17,32]. A similar co-clustering pattern was also observed with Maximum Likelihood based phylogenetic tree constructed using core gene clusters without paralogs (Supplementary Figure S1).

Mobile elements. The phages and insertion sequence (IS) elements of the two complete genomes, CT18 and P-stx-12 have been already reported[28,29]. The IS elements belonging to the family IS200/605, IS3, IS256 were commonly observed in all the other draft genomes we analyzed herein. However, the strain CR0063 also contained copies that belonged to IS1 family. The determination of exact copy number of these IS elements was difficult because of the draft status of the genomes. Further search for putative phage elements revealed presence of 4 intact phage sequences together with various phage remnants in each of the genomes (Supplementary table S1). Gifsy-2 and fels-2 phage sequences were common in most of the genomes. The atypical regions which encode genes mostly associated with virulence are designated as *Salmonella* pathogenicity islands (SPI). BRIG[33] was used to represent the status of major pathogenicity islands in these genomes as shown in Supplementary Figure S2. A list of genomic islands detected in these genomes as well as of the genes encoded by them is provided (Supplementary table S2). The plasmid related genes were not found in any strains other than CT18 and P-stx-12. The characteristics of the plasmids present in these



Figure 2 | **Genome alignment.** The whole genome alignment of all eight genomes was generated using progressiveMauve[50]. Each colored block represents similar sequences in the respective genomes.

strains along with the orthologous genes shared by them have already been discussed previously[28,29].

Pan-genome analysis. The pan-genome content measured up to a total of 5426 genes, 1.07 times higher than the average number of genes per individual strain. The pan-genome extrapolation was carried out in accordance with Heap's law[34]. The Heap's law can be represented by the following equation:

$n = k*N^{-\alpha}$, where n is pan-genome size, N is the number of genomes and k, γ are curve specific constants where $\alpha = 1 - \gamma$.

The exponential term α determines whether pan-genome of a bacterial species is closed or open. For $\alpha > 1$ ($\gamma < 0$) the pan-genome is considered closed i.e. sampling more genomes will not affect the pan-genome size, whereas for $\alpha < 1$ ($0 < \gamma > 1$) the pan-genome remains open and addition of more genomes would increase its size. In this study, the k and γ values were determined as 4486 and 0.087 respectively. The pan-genome analysis of *S.* Typhi strains revealed an α value of 0.913 implying a highly conservative nature of these endemic isolates (Figure 3a).

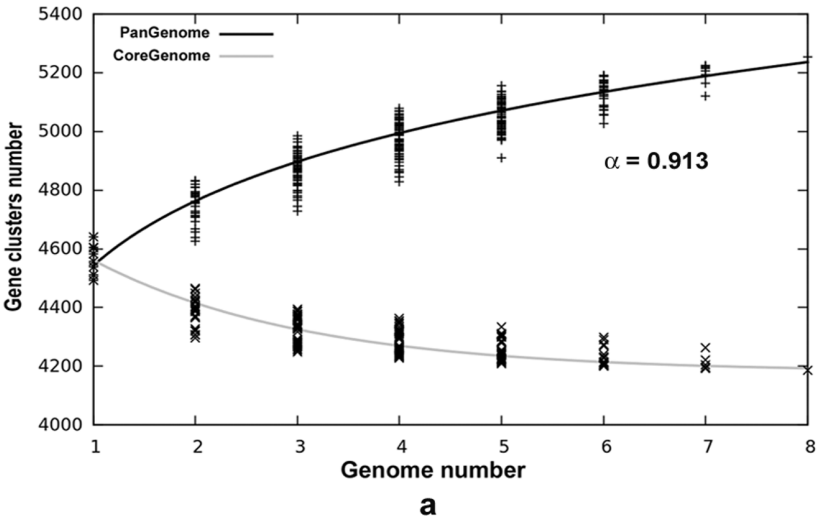
Further, to investigate the effect of pseudogenisation on gene frequency distributions of functional and pseudogenes, their pan and core genome components were determined separately. The pan-genome of functional genes contained a total of 4632 genes which was 1.03 times the average functional gene content per strain, whereas the pan-genome of pseudogenes contained a total of 857 genes which was 2.49 times the average pseudogene content per strain. This increased proportion of pseudogene content compared to functional pan-genome suggests that pseudogenisation is an active process in *S.* Typhi and this increase is also reflected in the pan-genome curve of pseudogenes (Figure 3c).

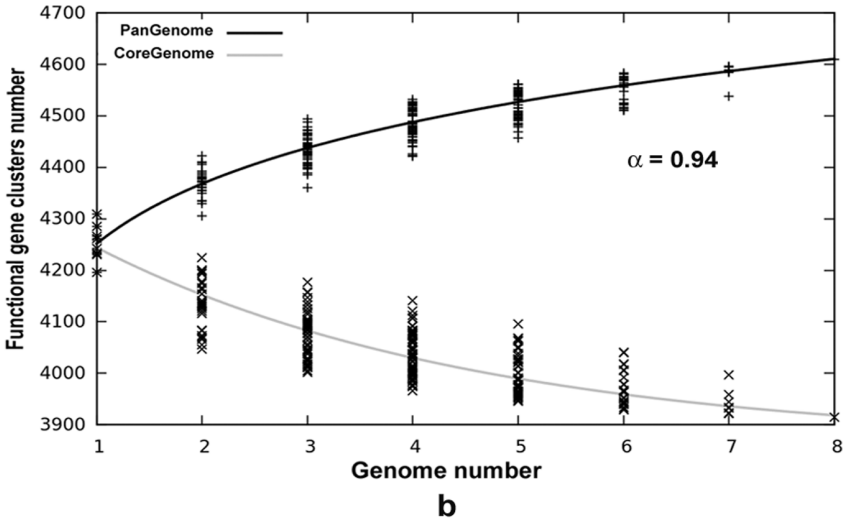
The values k and γ for functional gene clusters after curve fitting were determined as 4054 and 0.06 respectively, with a α value of 0.94. In contrast with what we observed in functional genes scenario, after curve fitting, α value for pseudogenes was 0.556 (Figure 3b, 3c), with k and γ values determined as 342.5 and 0.444 respectively. The α value of 0.94 indicates that the pan-genome of the functional genes is highly restricted in nature to allow any significant intake of foreign DNA and thus corroborates with high collinearity observed among these endemic strains. However the pan-genome of pseudogenes with an α value of 0.556 showed a very non-conservative nature as shown in Figure 3c and thus reemphasizes that pseudogenisation of functional genes is an ongoing process in *S.* Typhi.

The core genome of *S.* Typhi. The core genome of a species includes a subset of genes that are shared by all strains. The core genome of our endemic strains contained 4131 genes. This core genome size tended to decrease upon increasing the number of genomes; therefore, the curve fitting and extrapolation was done by least square fit of the exponential regression decay. This equation is written as: $n = k * \exp[-\frac{N}{\tau}] + tg(\theta)$, where n is the expected core genome size, N denotes number of genomes and *k*, $tg(\theta)$ are constants that fit the curve. In this equation, the first term $k * \exp[-\frac{N}{\tau}]$ will tend towards zero and the second term $tg(\theta)$ tends to converge towards a specific value. The analysis revealed a convergence value of 4124 genes which indicates a minimal genome content retained by the bacteria to perform basic biological processes (Figure 3a).

The core genome of functional and pseudogenes was also determined separately and these distributions provided some significant pointers. The core genome of functional genes was determined to be around 3558 genes and was still decreasing as shown in core genome curve of functional genes (Figure 3b) with the convergence value $tg(\theta)$ as 3495 obtained upon solving the equation. However, in the case of pseudogenes, the core genome has already reached its convergence with a $tg(\theta)$ value of 166 genes (Figure 3c). Further, the core genome profiles of functional and pseudogenes (Figure 3b, 3c) imply that core genome of pseudogenes constitutes a minor component of the total pseudogene content unlike that of core genome of functional genes. Moreover, these findings also stress on the need to analyze the role of these high number of accessory pseudogenes which are causing a steep increase in its pan-genome.

The pseudogenes identified in this analysis (Supplementary table S3) included various fimbrial proteins, methyl accepting chemotaxis proteins and certain secreted effector proteins. Some of these





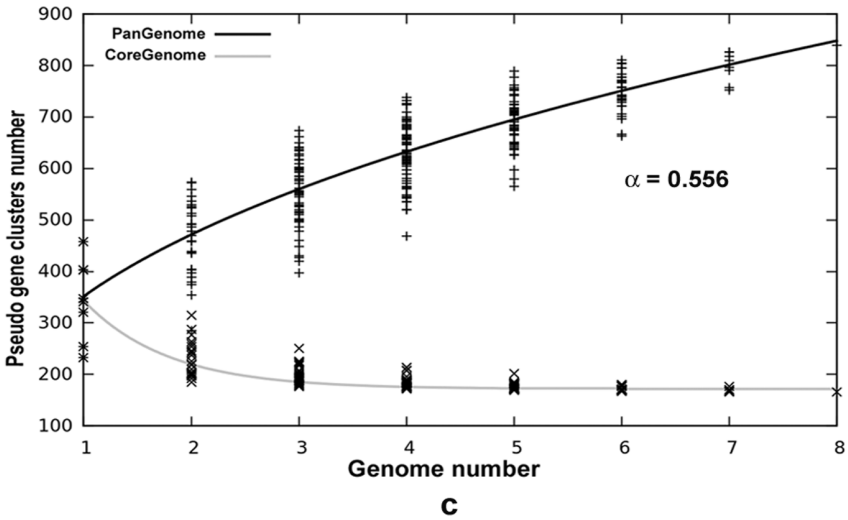


Figure 3 | Pan and Core Genome Distribution. (a). Pan and core genome developments using median values of the combinations of all eight genomes. (b). Pan and core genome developments of functional genes of these eight isolates. Here it can be observed that core genome is decreasing sharply. (c). Pan and core genome developments of pseudogenes of these eight isolates. It can be seen that pan genome of pseudogenes is highly non conservative with a steep increase in accessory content while the core genome reached convergence.

pseudogenes were potentially homologous to the genes found to be associated with important cellular functions such as anaerobic metabolism – ethanolamine utilization, being precursors of vitB_{12} synthesis, or acting as electron donors (formate dehydrogenase, galactarate dehydrogenase, succinyl glutamic semialdehyde dehydrogenase) and acceptors (tetrathionate reductase, trimethylamine-N-oxide reductase, nitric oxide reductase). The affordability to dispense such genes in intracellular bacteria like *S.* Typhi has already been previously reported and discussed[35,36].

Further to evaluate pseudogene distribution among various functional classes, they were classified into COG functional categories based on RPS BLAST. This analysis showed that, of those functionally classified, majority of the pseudogenes were related to metabolic processes: carbohydrate metabolism, amino acid transport and metabolism, inorganic ion transport etc. (Figure 4a). Further, when core functional gene clusters with SNPs (604clusters out of 3333 core clusters) were assigned COG classification, a higher proportion of functionally classified genes was observed in the same functional categories related to metabolic functions as observed in case of pseudogenes (Figure 4b). This enrichment of pseudogenes observed in the metabolism related genes was also statistically significant according to the proportionality z-test. Thus, from our observations, as depicted (Figure 4a, 4b), it can be inferred that metabolism related gene repertoire is under constant fine tuning and might relate to a rapid adaptation to the immediate local niche.

To gain further insights into the differential pseudogene content among various strains, we focused on the accessory pseudogene clusters marked by absence of a corresponding ortholog in at least one or more strains. For this analysis, only those accessory pseudogene clusters which do not have paralogs were considered. The absence of an ortholog in these clusters indicated only two possibilities: either the ortholog is not present in the strain or there exists a functional complement in the strain. Therefore, status of these accessory pseudogenes in each cluster was marked as P for pseudogene, F for functional complement and N for absence of a gene. Finally, we considered only those clusters where the orthologs were present in P or F states and removed those which had N in any of the strains. The plot of these pseudogene clusters along with their respective status in the genomes revealed a mixed profile (Figure 5). This analysis provides evidence for the existence of a heterogeneous mixture of functional and pseudogene complements in the population. Further, COG classification of those pseudogene clusters with P or F status in at least two query strains has shown that these were also enriched in metabolism related functions and this proportion was statistically significant (Figure 6). Thus the comparison points at the existence of heterogeneous strains with varying metabolic potential and might confer an adaptive advantage for the persistence of the pathogen.

We also performed pairwise comparisons of the strains of different clinical spectrum in order to determine if there are any state specific genes that entail different clinical level phenotypes of the strains. For this, we considered a total of four different pair-wise comparisons: outbreak versus carrier strains in 2 sets (BL196 & CR0044 and, BL196&CR0063) from Kelantan, Malaysia; a pair of strains (BL196 & ST0208) associated with an outbreak (BL196) and sporadic case (ST0208) from Malaysia; and a pair of strains (UJ308A & UJ816A) associated with fatal (UJ308A) and non-fatal (UJ816A) cases from Papua New Guinea. The core and specific functional gene content as well as pseudogene content were determined for all these strains. Further, after identifying the specific functional and pseudogenes of each strain in comparison, we also checked if a given strain in comparison carried an ortholog in a different functional state (functional or pseudo) than that of its corresponding strain, or vice versa. In this way, gene contents of all the strains were analyzed. Although this analysis helped us develop pairwise inventories of complimentary functional genes and pseudogenes, it did not identify any specific pattern of potential associations that could be attributed to a strain of a certain clinical spectrum conveying an acute or a carrier stage, for example.

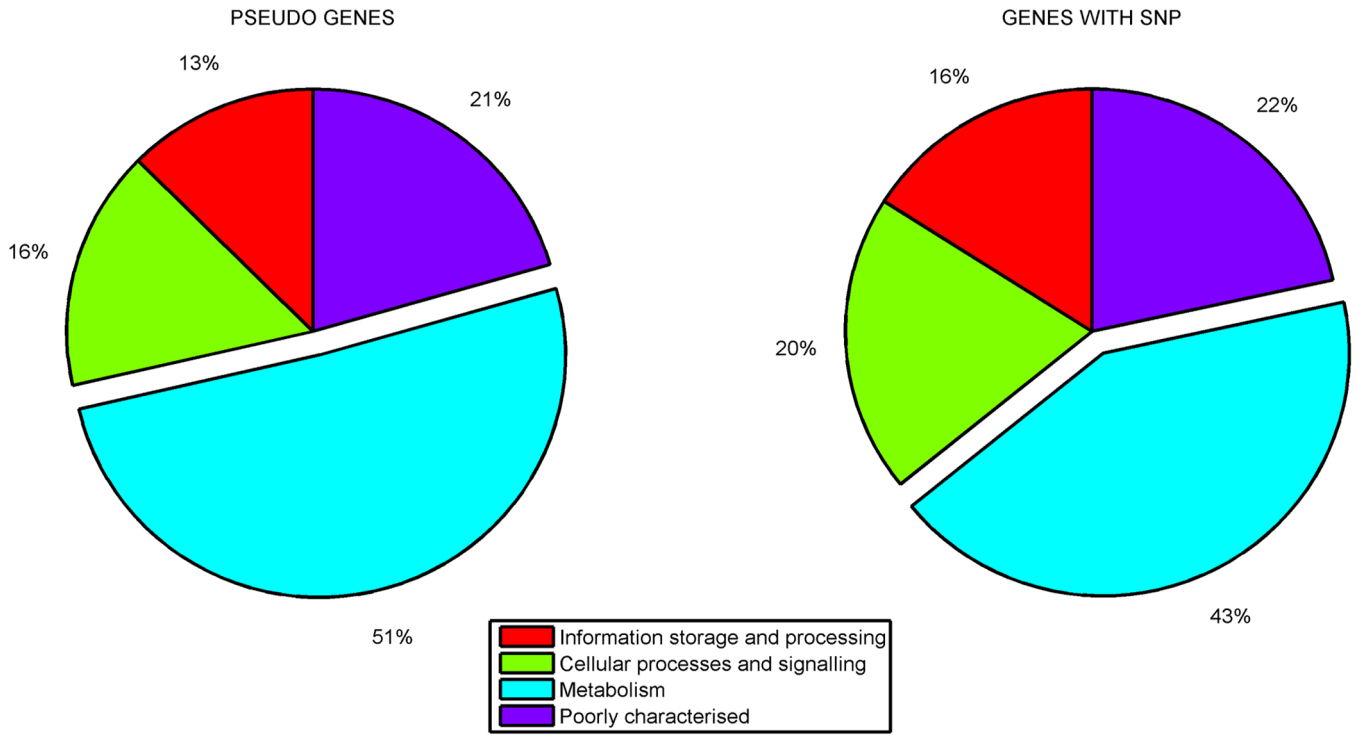


Figure 4 | The proportion of functionally classified pseudogenes and the functional genes with SNPs according to COG classification. The pie chart represents the proportion of various functional classes among the pseudogenes and the functional genes with SNPs. The figure clearly shows the enrichment of metabolism related genes in pseudogenes and the functional genes with SNPs.

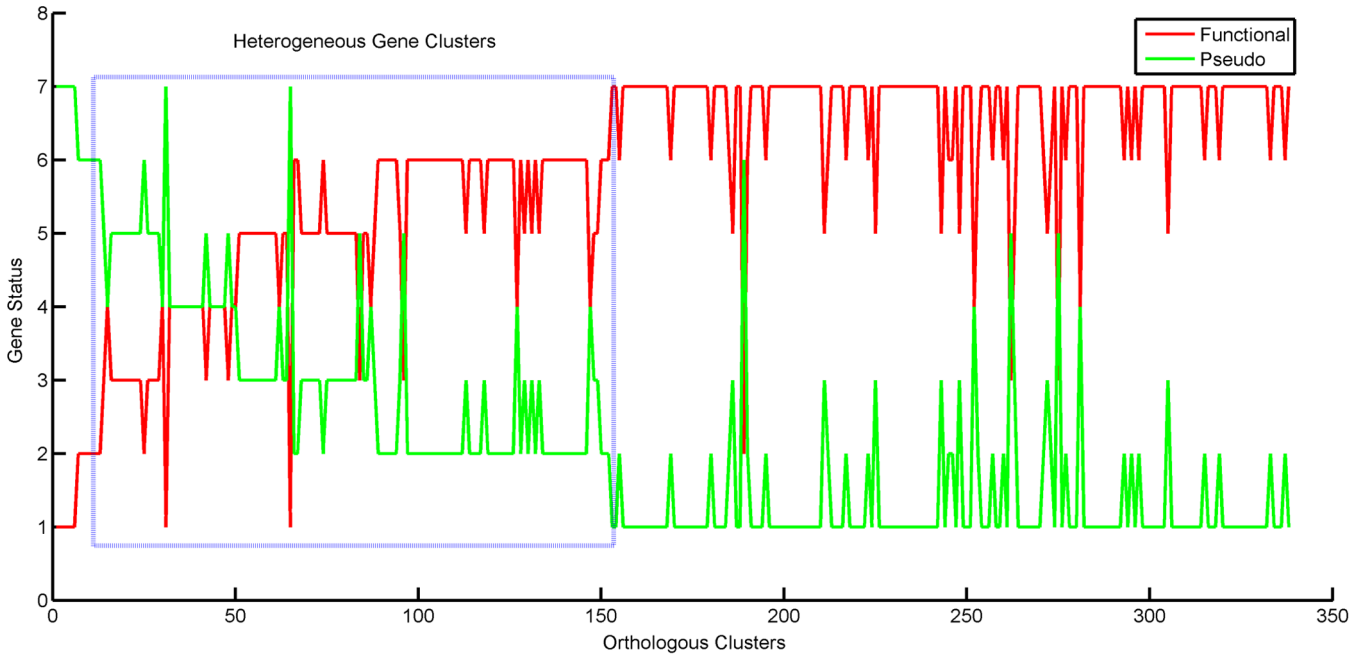


Figure 5 | Accessory pseudogene clusters analysis. The status of genes in each accessory pseudogene cluster was marked as P for pseudogene, F for functional complement and N for absence of gene. The clusters where the orthologs were present in P or F states were considered in plot. This shows the heterogeneous existence of functional and pseudogene complements in the population.

Discussion

The previous whole genome based study on *S.* Typhi by Holt *et al*[17] revealed that these genomes are highly clonal with minimal genomic variation due to SNPs, recombination and horizontal gene acquisition. In the present study also we observed a close genetic relatedness among the strains from different endemic zones of Southeast Asia/Oceania sharing similarity of 94–98% as shown (Figures 1, 2). In the past, comparative genomic studies have proposed that pseudogenisation is the main driving force in evolution of this organism when compared to others like acquiring foreign genetic material through HGT (Horizontal Gene Transfer) or gain of function[17,28,35]. Therefore, we attempted to understand the pseudogene pool of various isolates in greater detail and cues they could provide about various adaptation and survival mechanisms harnessed by this pathogen.

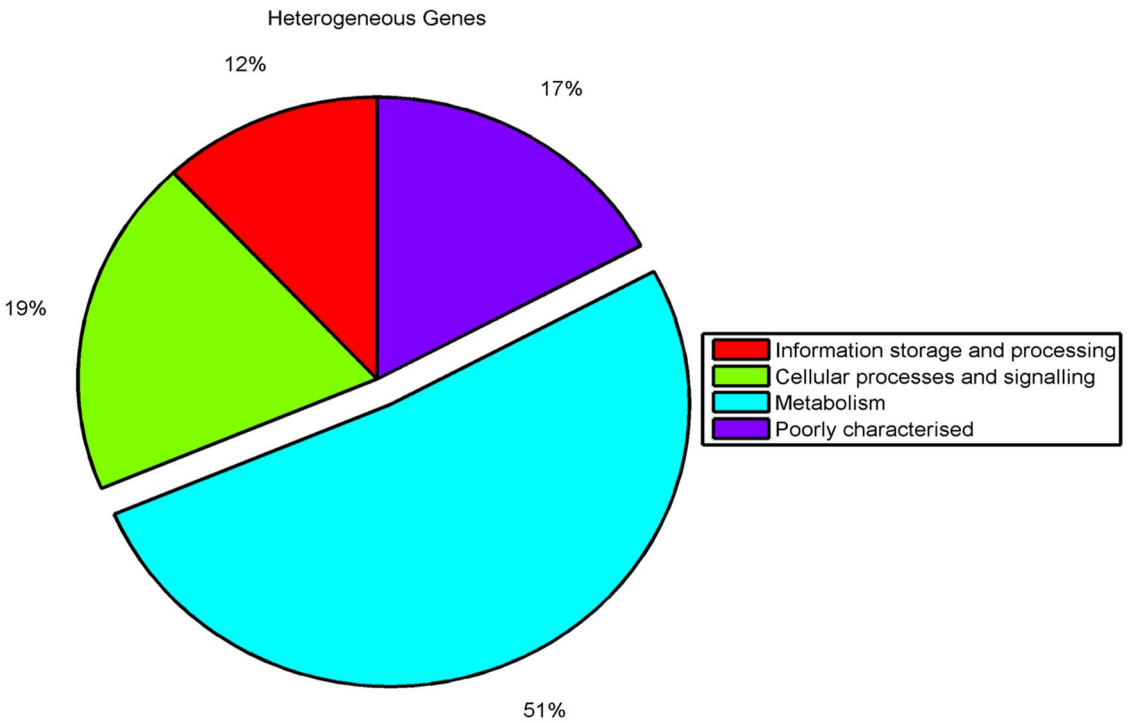


Figure 6 | Proportion of heterogeneous genes classified according to COG functional categories. The figure represents the distribution of accessory pseudogenes (those having variable functional and pseudogene status in atleast two strains) among various COG functional categories. The genes related to metabolism were clearly enriched in the accessory pseudogenes.

We identified most of the pseudogenes including those caused due to frameshift mutations, as these were not detected by Holt *et al* because of the low quality of sequence data[17].

The pan-genome analysis of these isolates has revealed limited potential for horizontal gene acquisition. This characteristic of the gene pool is a commonly observed phenomenon in case of intracellular organisms as they have limited contact with the potential gene donors[37]. Moreover, when the same analysis was carried out for functional and pseudogenes separately, it was observed that the core content of functional genes is still declining whereas the pseudogene content recorded steady increase in pan-genome (Figure 3). This decreasing trend of functional core genome and an increase in the pan-pseudogene content indicates that potential loss of functional genes might be a consequence of active pseudogenisation. Though an active pseudogenisation was observed, we could not detect any significant reduction in genome size or gene content indicating that pseudogenisation perhaps does not entail concurrent or consequent gene deletion in case of *S.* Typhi in contrast to other important human pathogens such as *Mycobacterium leprae* wherein pseudogenisation is followed by deletion, thus downsizing the genome[38]. Further, this finding could just be a reflection of different inactivating mutation rates or varying negative or positive selection pressures experienced by different isolates and/or lineages. Another possible explanation could be due to reversion of pseudogenes[39] although any gain of function may be rare, or only occurring in a small number of genes through point mutations[17]. Given this situation, a definitive mechanism can be confirmed only through genetic and functional studies involving serial isolates. Collectively, from these findings, we believe that over the course of evolution, *S.* Typhi has resorted to maintain its genome size through a fine balance between functional and pseudogenes.

Further, when the core functional gene clusters with SNPs were functionally classified, it was observed that these genes majorly belonged to metabolism related functions. A significant number of pseudogenes also belonged to same functional category as core functional genes with SNPs (Figure 4). This convergence of the core functional genes with SNPs and the pseudogenes indeed emphasizes the stress on the metabolic machinery. In addition, the analysis of accessory pseudogene clusters identified 338 clusters with mixed profile of pseudo and functional gene complements in various strains (Figure 5). Upon functional classification, even these polymorphic genes were found to be enriched in metabolism related functions (Figure 6). This could be an advantageous mechanism for the bacterium to modulate its metabolic repertoire through pseudogenisation depending on its specific local niche[40]. Similar survival strategy is reported in other pathogenic bacteria where virulence optimization is achieved at the cost of certain metabolic genes[41,42]. Given this, it can be surmised that this modulation could be one of the major mechanisms underlying the carrier state. Hence, it is important to focus on characterizing the metabolic potential and its implications on virulence of *S.* Typhi.

The heterogeneity displayed by functional and pseudogene content of these isolates, especially even in those collected from the same region (Kelantan, Malaysia) over a period of time, provides explanation for the interplay between them and supports previous hypothesis that the restoration of function might be occurring through mutation[17]. However this can be only further proved with functional studies on serial isolates from a single individual.

S. Typhi encounters drastically different environments from its initial point of entry into the small intestine up to its final colonization of internal organs like gall bladder for chronic carriage[43]. To succeed in these varying environments, it might be very important to optimize its metabolism through loss of function, conferring an advantage within its immediate local niche.

The above observations regarding pan and core genome distributions of functional and pseudogenes lend support to the idea that *S.* Typhi maintains an efficient balance through various mechanisms, such that its genome is not degraded beyond a certain level. At the same time, a heterogeneous profile of functional and pseudo gene complements could possibly culminate in a more hospitable metabolic environment. Collectively, these orchestrated genome dynamics most likely appear to aid in persistence and host adaptation. Genome analysis of this limited but important collection of strains could provide us some significant pointers regarding adaptation of this organism which appears to be possibly influenced by a conserved nature of its genome. Further, inclusion of more number of genomes in the analysis would possibly enhance the understanding of these observations.

Nevertheless, given these findings, it will be possible to advance the current knowledge of the carrier state in *Salmonella* pathogens underlying continuous emergence and reemergence of typhoid in endemic regions.

Methods

Sequence information. The *S.* Typhi strains chosen for the analysis have been isolated from various countries of Southeast Asia/Oceania and were isolated at different time points by different researchers. The genome collection included two complete genomes - CT18, P-stx-12 strains and six draft genomes - UJ308A, UJ816A, BL196, CR0063, CR0044 and ST0208 which were available in public domain through NCBI. Few other strains from Southeast Asia which are available in NCBI could not be included in the analysis because of the low sequence coverage and poor quality of data.

Refinement of assembly and annotation. The contigs from WGS master records of *S.* Typhi genomes (UJ308A, UJ816A, BL196, CR0044, CR0063, ST0208) were downloaded from NCBI. These contigs were ordered according to a reference using standalone BLAST. The high quality filtered reads of respective strains were mapped to the contigs using BWA alignment tool[44]. The alignment file was visualized using Tablet alignment viewer[45] to sort the scaffolds in correct order based on paired-end read information. The regions with low coverage were manually inspected before including them into final genome.

After finalizing the order of the contigs, they were linked using a linker sequence (NNN NNC ATT CCA TTC ATT AAT TAA TTA ATG AAT GAA TGN NNN N) that encodes start and stop codons in all six frames. The contigs thus obtained was submitted to ISGA pipeline for annotation[46]. The two complete genomes CT-18 and P-stx-12 were also re-annotated to homogenize the data with a single annotation platform. This annotation pipeline uses Glimmer 3 for prediction of ORFs and BLASTx for searches based on sequence similarity[47]. The predicted ORFs were scanned to identify protein domains using HMMPfam (http://hmmer.janelia.org/). The tRNAscan-SE and RNAmmer were used respectively for the detection of tRNA and rRNA[48,49]. The whole genome alignment of all these genomes was generated using progressive Mauve[50].

For the identification of pseudogenes, BLASTn was performed for all the nucleotide sequences of query ORFs against the functional proteins of *Salmonella* strains, which were submitted to NCBI as complete genomes. The corresponding protein sequences of the best five hits of nucleotide BLAST were considered for performing BLASTx against individual query ORFs. Then, to mark the latter as a pseudogene, based on above results, threshold of more than 60% coverage of query length and 98% identity were applied. The above method could detect all the pseudogenes that originated due to a nonsense mutation resulting in early termination of translation. To identify pseudogenes that were formed due to potential frame shifts causing protein fragmentation were identified using an inbuilt module of PanOCT[51]. The BLASTp result of query ORFs against the functional genes of *Salmonella* strains was provided as input to PanOCT. The number of BLAST matches needed to confirm a protein fragment/frame-shift was set to 1 and the frame-shift overlap parameter as 1.33. In the case of proteins which are split due to frame-shifts, the major fragment was considered in the final pseudogene list, so that the number did not over-represent.

Phylogenomic analysis. The whole genome based phylogeny was performed for all eight strains using Gegenees (version 1.1.4)[52] which employs a fragmented all-against-all comparison of the genomes and builds a distance matrix file suitable to construct a phylogenetic tree and heat map. The phylogenetic tree was built by NJ (Neighbor-joining) method using SplitsTree software (version 1.1.4)[53]. The detailed methodology of core genome based phylogeny using maximum likelihood is explained in Supplementary information.

Detection of mobile elements. All of these strains included various mobile phages or phage like elements. To identify these elements, *PhiSpy*[54], an algorithm that combines both similarity and composition based strategies, was used. These predictions were compared with results obtained from PHAST (A Fast Phage search tool)[55]. IS elements were identified using IS finder[56]. The genomic islands in these strains were identified using IslandViewer[57].

Pan-genome analysis. Pan-genome analysis represents the variation in gene content of different strains. The determination of pan and core genome requires correct identification of orthologous clusters of all selected strains. This was done using OrthoMCL[58] which is mainly developed for clustering of orthologous protein sequences based on user defined percent match cutoff and minimum protein length.

Further, the pan-genome and core genome of the two strains A and B (AB) were calculated as follows: pan-genome AB is composed of the sum of gene sets of A and B (strain A and non-orthologous genes of strain B) and the core genome AB is composed of orthologous genes that are present in both A and B. Upon addition of more genomes, pan-genome was estimated in an additive manner whereas the core genome was determined in a reductive manner. The median values of all possible combinations of genomes were considered to further examine the patterns of pan- and core genomes. The curve fitting of pan-genome was done using Heap's law whereas that of core genome using least square fit of the exponential regression decay as described previously by Tettelin *et al*[34].

Initially, orthologous clusters of all strains were generated with the percent match threshold of 85% and minimum protein length of 50 amino acids. Later, the functional genes and pseudogenes were analyzed separately as mentioned by Liang *et al*[59], but their respective orthologous clusters were generated using OrthoMCL with same percent match cutoff. However, minimum protein length considered for generating functional gene clusters was 50 amino acids whereas for pseudogenes it was set to 10 amino acids. The extrapolations of pan-genome and core genome of functional and pseudogenes was done as mentioned above, but individually for each of them. The detailed explanation of COG classification using RPS BLAST (NCBI) is given in supplementary information. The statistical two sample z-proportionality test was applied to calculate the significance for the enrichment of various functional classes.

Detection of SNP in core genome. The core functional gene clusters were identified as those which contain only one representative protein from each of the query strain. SNP detection in these core functional gene clusters was done by aligning the corresponding sequences of each cluster using ClustalW[60].

1. Coburn, B., Grassl, G. A. & Finlay, B. B. *Salmonella*, the host and disease: a brief review. *Immunol Cell Biol* **85**, 112–118 (2007).
2. Boyle, E. C., Bishop, J. L., Grassl, G. A. & Finlay, B. B. *Salmonella*: from pathogenesis to therapeutics. *J Bacteriol* **189**, 1489–1495 (2007).
3. Crump, J. A., Luby, S. P. & Mintz, E. D. The global burden of typhoid fever. *Bull World Health Organ* **82**, 346–353 (2004).
4. Gopinath, S., Carden, S. & Monack, D. Shedding light on *Salmonella* carriers. *Trends Microbiol* **20**, 320–327 (2012).
5. Chandel, D. S., Chaudhry, R., Dhawan, B., Pandey, A. & Dey, A. B. Drug-resistant Salmonella enterica serotype Paratyphi A in India. *Emerg Infect Dis* **6**, 420–421 (2000).
6. Holt, K. E. *et al*. Temporal fluctuation of multidrug resistant salmonella typhi haplotypes in the Mekong river delta region of Vietnam. *PLoS Negl Trop Dis* **5**, e929 (2011).
7. Ploy, M. C. *et al*. Integron-associated antibiotic resistance in Salmonella enterica serovar typhi from Asia. *Antimicrob Agents Chemother* **47**, 1427–1429 (2003).
8. Ruby, T., McLaughlin, L., Gopinath, S. & Monack, D. *Salmonella*'s long-term relationship with its host. *FEMS Microbiol Rev* **36**, 600–615 (2012).
9. Kalai Chelvam, K., Chai, L. C. & Thong, K. L. Variations in motility and biofilm formation of Salmonella enterica serovar Typhi. *Gut Pathog* **6**, 2 (2014).
10. Lanzas, C. *et al*. The effect of heterogeneous infectious period and contagiousness on the dynamics of *Salmonella* transmission in dairy cattle. *Epidemiol Infect* **136**, 1496–1510 (2008).
11. Thong, K. L. *et al*. Analysis of Salmonella typhi isolates from southeast Asia by pulsed-field gel electrophoresis. *J Clin Microbiol* **33**, 1938–1941 (1995).
12. Baumler, A. & Fang, F. C. Host specificity of bacterial pathogens. *Cold Spring Harb Perspect Med* **3**, a010041 (2013).
13. McClelland, M. *et al*. Complete genome sequence of Salmonella enterica serovar Typhimurium LT2. *Nature* **413**, 852–856 (2001).
14. Holt, K. E. *et al*. Pseudogene accumulation in the evolutionary histories of Salmonella enterica serovars Paratyphi A and Typhi. *BMC Genomics* **10**, 36 (2009).
15. Chiu, C. H. *et al*. The genome sequence of Salmonella enterica serovar Choleraesuis, a highly invasive and resistant zoonotic pathogen. *Nucleic Acids Res* **33**, 1690–98 (2005).
16. Thomson, N. R. *et al*. Comparative genome analysis of Salmonella Enteritidis PT4 and Salmonella Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. *Genome Res.* **18**, 1624–1637 (2008).
17. Holt, K. E. *et al*. High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. *Nat Genet* **40**, 987–993 (2008).
18. Rodriguez-Valera, F. U. D. Is the pan-genome also a pan-selectome? *F1000Res* **1**, 16 (2012).
19. Tettelin, H. *et al*. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial 'pan-genome'. *Proc Natl Acad Sci U S A* **102**, 16530–16530 (2005).
20. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Curr Opin Genet Dev* **15**, 589–594 (2005).
21. Zhou, Z. *et al*. Transient Darwinian selection in Salmonella enterica serovar Paratyphi A during 450 years of global spread of enteric fever. *Proc Natl Acad Sci U S A* **111**, 12199–12204 (2014).
22. Zhou, Z. *et al*. Neutral genomic microevolution of a recently emerged pathogen, Salmonella enterica serovar Agona. *PLoS Genet* **9**, e1003471 (2013).
23. Baddam, R. *et al*. Genetic Fine Structure of a Salmonella enterica Serovar Typhi Strain Associated with the 2005 Outbreak of Typhoid Fever in Kelantan, Malaysia. *J Bacteriol* **194**, 3565–3566 (2012).
24. Yap, K. P. *et al*. Genome Sequence and Comparative Pathogenomics Analysis of a Salmonella enterica Serovar Typhi Strain Associated with a Typhoid Carrier in Malaysia. *J Bacteriol* **194**, 5970–5971 (2012).
25. Baddam, R. *et al*. Genome sequencing and analysis of Salmonella enterica serovar Typhi strain CR0063 representing a carrier individual during an outbreak of typhoid fever in Kelantan, Malaysia. *Gut Pathog* **4**, 20 (2012).
26. Yap, K. P. *et al*. Insights from the Genome Sequence of a Salmonella enterica Serovar Typhi Strain Associated with a Sporadic Case of Typhoid Fever in Malaysia. *J Bacteriol* **194**, 5124–5125 (2012).
27. Baddam, R. *et al*. Whole-Genome Sequences and Comparative Genomics of Salmonella enterica Serovar Typhi Isolates from Patients with Fatal and Nonfatal Typhoid Fever in Papua New Guinea. *J Bacteriol* **194**, 5122–5123 (2012).
28. Parkhill, J. *et al*. Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. *Nature* **413**, 848–852 (2001).
29. Ong, S. Y. *et al*. Complete Genome Sequence of Salmonella enterica subsp enterica Serovar Typhi P-stx-12. *J Bacteriol* **194**, 2115–2116 (2012).
30. Thong, K. L. *et al*. Molecular analysis of isolates of Salmonella typhi obtained from patients with fatal and nonfatal typhoid fever. *J Clin Microbiol* **34**, 1029–1033 (1996).
31. Mirza, S., Kariuki, S., Mamun, K. Z., Beeching, N. J. & Hart, C. A. Analysis of plasmid and chromosomal DNA of multidrug-resistant Salmonella enterica serovar typhi from Asia. *J Clin Microbiol* **38**, 1449–1452 (2000).
32. Kidgell, C. *et al*. Salmonella typhi, the causative agent of typhoid fever, is approximately 50,000 years old. *Infect Genet Evol* **2**, 39–45 (2002).
33. Alikhan, N. F., Petty, N. K., Ben Zakour, N. L. & Beatson, S. A. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC genomics* **12**, 402 (2011).
34. Tettelin, H., Riley, D., Cattuto, C. & Medini, D. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol* **11**, 472–477 (2008).
35. McClelland, M. *et al*. Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of Salmonella enterica that cause typhoid. *Nat Genet* **36**, 1268–1274 (2004).
36. Nuccio, S. P. & Baumler, A. J. Comparative analysis of Salmonella genomes identifies a metabolic network for escalating growth in the inflamed gut. *MBio* **5**, e00929–14 (2014).
37. Kuenne, C. *et al*. Reassessment of the Listeria monocytogenes pan-genome reveals dynamic integration hotspots and mobile genetic elements as major components of the accessory genome. *BMC genomics* **14**, 47 (2013).
38. Cole, S. T. *et al*. Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007–1011 (2001).
39. Olson, M. V. When less is more: gene loss as an engine of evolutionary change. *Am J Hum Genet* **64**, 18–23 (1999).
40. Rohmer, L., Hocquet, D. & Miller, S. I. Are pathogenic bacteria just looking for food? Metabolism and microbial pathogenesis. *Trends Microbiol* **19**, 341–348 (2011).
41. Maurelli, A. T., Fernandez, R. E., Bloch, C. A., Rode, C. K. & Fasano, A. "Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of Shigella spp. and enteroinvasive Escherichia coli. *Proc Natl Acad Sci U S A* **95**, 3943–3948 (1998).
42. Touchon, M. *et al*. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS genet* **5**, e1000344 (2009).
43. Reis, R. S. & Horn, F. Enteropathogenic Escherichia coli, Samonella, Shigella and Yersinia: cellular aspects of host-bacteria interactions in enteric diseases. *Gut Pathog* **2**, 8 (2010).
44. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
45. Milne, I. *et al*. Tablet-next generation sequence assembly visualization. *Bioinformatics* **26**, 401–402 (2010).
46. Hemmerich, C., Buechlein, A., Podicheti, R., Revanna, K. V. & Dong, Q. F. An Ergatis-based prokaryotic genome annotation web server. *Bioinformatics* **26**, 1122–1124 (2010).
47. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**, 4636–4641 (1999).
48. Schattner, P., Brooks, A. N. & Lowe, T. M. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* **33**, W686–W689 (2005).
49. Lagesen, K. *et al*. RNammer: consistent annotation of rRNA genes in genomic sequences. *Nucleic Acids Res* **35**, 3100–3108 (2000).
50. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PloS one* **5**, e11147 (2010).
51. Fouts, D. E., Brinkac, L., Beck, E., Inman, J. & Sutton, G. PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic

analysis of bacterial strains and closely related species. *Nucleic Acids Res* **40**, e172 (2012).
52. Agren, J., Sundstrom, A., Hafstrom, T. & Segerman, B. Gegenees: Fragmented Alignment of Multiple Genomes for Determining Phylogenomic Distances and Genetic Signatures Unique for Specified Target Groups. *PloS one* **7**, e39107 (2012).
53. Huson, D. H. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68–73 (1998).
54. Akhter, S., Aziz, R. K. & Edwards, R. A. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res* **40**, e126 (2012).
55. Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. & Wishart, D. S. "PHAST: A Fast Phage Search Tool" *Nucleic Acids Res* **39**, W347–352 (2011).
56. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* **34**, D32–D36 (2006).
57. Langille, M. G. & Brinkman, F. S. IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* **25**, 664–665 (2009).
58. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178–2189 (2003).
59. Liang, W. *et al*. Pan-genomic analysis provides insights into the genomic variation and evolution of Salmonella Paratyphi A. *PLoS One* **7**, e45346 (2012).
60. Larkin, M. A. *et al*. Clustal W and clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).

Acknowledgments

We would like to thankfully acknowledge Prof. Kwai-Lin Thong for information about some of the strains studied herein and to all those authors who contributed to data about these strains as available in the public domain. We would also like to thank Donepudi RaviTeja for his help with gnuplot and R. We thankfully acknowledge funding received from Department of Biotechnology, Government of India (Ref. No. BT/HRD/NBA/34/01/2011(ix) and BT/PR6921/MED/29/699/2013). RB would like to acknowledge the UGC-RFSMS fellowship.

Author contributions

R.B. and N.A. designed and conducted the study. N.K. helped in analysis of data and preparation of the manuscript. A.K.L. and S.S. provided help in writing scripts.

Additional information

Accession codes CT18 (NC_003198), P-stx-12 (NC_016832), UJ308A (AJTD00000000), UJ816A (AJTE00000000), BL196 (AJGK00000000), CR0063 (AKIC00000000), CR0044 (AKZO00000000) and ST0208 (AJXA00000000).

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Baddam, R., Kumar, N., Shaik, S., Lankapalli, A.K. & Ahmed, N. Genome dynamics and evolution of *Salmonella* Typhi strains from the typhoid-endemic zones. *Sci. Rep.* **4**, 7457; DOI:10.1038/srep07457 (2014).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder in order to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/

324–335 *Nucleic Acids Research, 2015, Vol. 43, No. 1* *Published online 1 December 2014*
doi: 10.1093/nar/gku1271

Comparative genomic analysis of *Helicobacter pylori* from Malaysia identifies three distinct lineages suggestive of differential evolution

Narender Kumar[1], Vanitha Mariappan[2], Ramani Baddam[1], Aditya K. Lankapalli[1], Sabiha Shaik[1], Khean-Lee Goh[3], Mun Fai Loke[2], Tim Perkins[4], Mohammed Benghezal[4], Seyed E. Hasnain[5], Jamuna Vadivelu[2], Barry J. Marshall[4] and Niyaz Ahmed[1,6,*]

[1]Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, University of Hyderabad, Gachibowli, Hyderabad, 500046, India, [2]Department of Medical Microbiology, Faculty of Medicine, University of Malaya, 50603, Kuala Lumpur, Malaysia, [3]Department of Medicine, Faculty of Medicine, University of Malaya, 50603, Kuala Lumpur, Malaysia, [4]School of Pathology and Laboratory Medicine, University of Western Australia, Nedlands 6009, Western Australia, Australia, [5]Kusuma School of Biological Sciences, Indian Institute of Technology, Hauz Khas, New Delhi, 110016, India and [6]Institute of Biological Sciences, University of Malaya, 50603, Kuala Lumpur, Malaysia

Received September 30, 2014; Revised November 12, 2014; Accepted November 19, 2014

ABSTRACT

The discordant prevalence of *Helicobacter pylori* and its related diseases, for a long time, fostered certain enigmatic situations observed in the countries of the southern world. Variation in *H. pylori* infection rates and disease outcomes among different populations in multi-ethnic Malaysia provides a unique opportunity to understand dynamics of host–pathogen interaction and genome evolution. In this study, we extensively analyzed and compared genomes of 27 Malaysian *H. pylori* isolates and identified three major phylogeographic lineages: hspEastAsia, hpEurope and hpSouthIndia. The analysis of the virulence genes within the core genome, however, revealed a comparable pathogenic potential of the strains. In addition, we identified four genes limited to strains of East-Asian lineage. Our analyses identified a few strain-specific genes encoding restriction modification systems and outlined 311 core genes possibly under differential evolutionary constraints, among the strains representing different ethnic groups. The *cagA* and *vacA* genes also showed variations in accordance with the host genetic background of the strains. Moreover, restriction modification genes were found to be significantly enriched in East-Asian strains. An understanding of these variations in the genome content would provide significant insights into various adaptive and host modulation strategies harnessed by *H. pylori* to effectively persist in a host-specific manner.

INTRODUCTION

Helicobacter pylori, the human gastric pathogen, colonizes almost 50% of the world population (∼70% of the population of developing countries and ∼40% of developed countries) (1,2). It is a major etiological agent for a wide range of gastric diseases such as gastritis, peptic ulcers, gastric carcinoma and mucosa-associated lymphoid tissue lymphoma (3,4). Generally acquired during the childhood (intra-familial transfer), *H. pylori* establishes a lifelong persistent infection unless cleared by antibiotics (5).

The analysis of *H. pylori* strains has revealed existence of populations that are geographically localized (6,7). These populations have been classified based on multi-locus sequence typing (MLST) (7–11) into seven major lineages or genotypes depending on their regional prevalence: hpEurope, hpSahul, hpEastAsia, hpAfrica1, hpAfrica2, hpWestAfrica and hpAsia2. The ability to undergo frequent mutation and recombination serves as one of the major contributors for the observed genetic heterogeneity among various *H. pylori* isolates (12–14). It also allows the bacterium to quickly adapt to the changing gastric niches and establish a persistent infection. Certain countries with people of different ethnicities, cultures, lifestyles and religions present a pertinent model to examine the effects of migration and co-evolution on bacteria–host interaction. Studies entailing such settings would provide a better understanding of the evolutionary and adaptive strategies employed by *H. pylori* which might aid in design of intervention strategies (15).

*To whom correspondence should be addressed. Tel: +91 40 23134585; Fax: +91 40 23134585; Email: niyaz.ahmed@uohyd.ac.in; ahmed.nizi@gmail.com

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Malaysia is one such multicultural, developing nation with a population comprising four major ethnic groups: Malay/'Bumiputera', Chinese, Indians and others (http://www.statistics.gov.my). In general, the Malays are considered natives (Bumiputera) and are in majority. The Malaysian-Chinese comprise the second largest ethnic group and are documented to have migrated from Southern China while the Malaysian-Indian group is comprised of migrants from Southern India (16). Apart from these major ethnic groups, there are a number of indigenous groups ('Orang Asli') living together, particularly in East Malaysia, Sabah and Sarawak who do not share the same ethnic origin as the Malays (17).

Previous reports have shown high prevalence of *H. pylori* infection among the Malaysian-Indians (69–75%), followed by Malaysian-Chinese (45–60%) and Malays (8–43%), and a minuscule number of inter-racial/inter-community or inter-religion marriages result in a putatively reduced chance of cross-infection occurring between ethnic groups (18,19). A majority of the *H. pylori* isolates from Malays and Malaysian-Indians were suggested to be of a recent common origin, while those from Malaysian-Chinese exhibited East-Asian ancestry (6,19). Generally, the *H. pylori* isolate collections representative of Asian populations are composed of strains from hpEastAsia, hpEurope and hpAsia2 (10,19). A significant proportion of the Malay isolates were found similar to their Indian counterparts, suggesting a possible acquisition of *H. pylori* from Indians (19), although there is not enough genomic evidence to support this interpretation. Further, the reason for low prevalence of *H. pylori* infection among Malay/'Bumiputera' population remains unclear and is likely to involve a number of environmental, genetic and host-related factors (20).

Recent sequencing efforts have reported multiple genome sequences of *H. pylori* isolates from patients of different ethnicities and various disease manifestations from Malaysia (21–23). To date, there are 29 Malaysian *H. pylori* genomes (27 clinical strains and two mice-adapted strains) available in NCBI database, a majority of them sequenced and deposited as a part of this work. In this study, we carried out an in-depth whole genome comparative analysis of 27 clinical isolates. The comparison of their core and accessory gene pools demonstrated close similarity among the strains according to their respective host genetic backgrounds. The status of various virulence genes and outer membrane proteins (OMPs) was also compared among the strains in order to unleash novel co-ordinates of adaptive evolution. The study aimed at understanding the genomic heterogeneity among these isolates and their possible role in observed enigmas related to disease outcomes in the region. Further, the analysis of strain-specific genes would allow us to better understand the disease biology and might open avenues for developing effective control strategies.

MATERIALS AND METHODS

Strain collection and ethics approval

Gastric biopsy samples were obtained from five non-ulcer dyspepsia patients of different ethnicities [two Malaysian-Chinese (UM007 and UM034), two Malaysian-Indians (UM018 and UM054) and one Malay (UM045)] at the University of Malaya Medical Centre (UMMC). All biopsies were obtained with written informed consents of the patients attending the Endoscopy Unit, at UMMC. This study was approved by the Human Ethics Committee of the University of Malaya, Kuala Lumpur, Malaysia (Ref. No. 943.2).

Bacterial culture and DNA isolation

The *H. pylori* isolates were cultured from gastric biopsies by inoculating them on chocolate agar fortified with 4% blood base agar No. 2 (Oxoid) containing defibrinated horse blood (Oxoid) and antimicrobials such as trimethoprim, vancomycin and polymyxin B added to it at standard concentrations. Primary cultures were kept for incubation for up to 10 days (with daily observation) at 37°C in an incubator with 10% CO_2. For isolation of pure cultures, a single colony was identified and sub-cultured on chocolate agar for 3–5 days. Morphological identification of *H. pylori* isolates was carried out based on microscopic features and based on characteristic biochemical tests for the detection of enzymes such as urease, oxidase and catalase. A plateful of *H. pylori* culture was suspended into 500 μl of Tris buffer. The suspension was centrifuged at 5000 rpm for 10 min and the resulting pellet was collected. The *H. pylori* DNA was isolated using the QIAamp DNA Mini kit (Qiagen) according to the manufacturer's instruction.

Genome sequencing, assembly and annotation

Whole genome sequencing of the collected strains was carried out on Illumina GAIIx sequencer. The 100-bp paired-end sequencing run generated ∼1-GB read data per strain with an average insert size of ∼400 bp. The raw reads were then filtered using NGS QC toolkit (threshold quality >20) and were assembled into contigs using Velvet *de novo* assembler (24,25). The contigs were aligned against the NCBI refSeq database to identify a suitable complete genome as a reference. The contigs were sorted and re-oriented according to the chosen reference using in-house written scripts utilizing BLASTn. The sort-order was subjected to manual curation based on the paired-end information which helped us to order most of the unaligned contigs. These ordered contigs were joined together to form a draft genome by inserting a linker sequence (NNNNNCACACACTTAATT AATTAAGTGTGTGNNNNN) encoding start and stop codons in all six frames at the ends. The draft genomes were submitted for gene prediction and annotation to RAST annotation server, and the results were validated using GeneMarkS, Easygene and Glimmer (26–31). Genome statistics of respective strains were obtained through Artemis (32). tRNAs were identified using tRNAscan-SE program while rRNA genes were identified by using RNAmmer program (33,34).

Functional annotation, calculation of core and specific content

The predicted genes were scored using BLASTp against a protein database consisting of genes from 36 complete genomes from the NCBI refseq database. The output was

filtered with an identity and query coverage of 90 and 70%, respectively. The proteins were then assigned functional categories based on their best hit obtained in the Basic Local Alignment Search Tool (BLAST) alignment. The other 22 draft genomes reported from Malaysia were downloaded and their gene prediction was performed as mentioned in the previous section. The core genome was calculated by identifying orthologs in every genome by applying Markov cluster (MCL) algorithm included in the OrthoMCL program (35). The parameters for deciding orthologs such as identity and e-value cutoff were set to 80% and 0.00001, respectively. The genes with less than 50 amino acids were excluded from the analysis. The clusters that contained orthologs in all the strains constituted the core, while those that did not have corresponding ortholog in any of the other genomes were considered as strain specific. The identified gene clusters were assigned functional categories after comparison with the COG database using rpsBLAST program followed by manual curation of the results (36,37).

Phylogenetic analysis

A whole genome alignment of 27 genomes of the Malaysian isolates was carried out with 43 other *H. pylori* genomes (draft or complete) from NCBI database using Gegenees tool (38,39). The tool utilizes a fragmented alignment algorithm to calculate average similarity among the compared genomes using BLASTn. The fragment size can be optimized according to the user. The tool was run with the fragment size set to 200 and a step size of 100 using BLASTn. The average similarity was calculated with a BLAST score threshold of 40% generating a heat plot matrix that was further used to deduce phylogenetic relationships exported in the form of a .nexus file. This nexus tree file was supplied as an input to SplitsTree (40) program for building an unrooted phylogenetic tree employing Neighbor-Joining algorithm.

Virulence genes and phage detection

The available whole genomes were screened for the presence of virulence genes enlisted in the Virulence Factor Database (VFDB) using BLAST program (41). The cutoff identity and query coverage was set to 70 and 60%, respectively. Further, comparison of the amino acid sequences of CagA and VacA was carried out using sequence-similarity-based alignments. Phage-related sequences in the genome were identified using PHAST server that integrates the analysis against various phage databases and compares key phage attributes to detect similar phage sequences in the query genome sequence (42).

RESULTS AND DISCUSSION

Genome statistics and phylogenetic analysis

The whole genome sequencing of five Malaysian isolates (Figure 1) revealed their chromosome sizes ranging from 1.56 to 1.62 Mb. The genomes also revealed a low G+C content of 39% which is characteristic of *H. pylori*. The draft genomes were predicted to encode ∼1600 genes with an average coding DNA sequence (CDS) measuring up to 930 bp. All the sequenced genomes harbored three rRNA operons as well as 36 tRNA genes. Two out of five sequenced strains (UM045 and UM054) also harbored phage sequences encoding 136 and 12 putative phage genes, respectively. A detailed genome statistics of these five isolates has been mentioned in Table 1 and a comparison with the remaining 22 genomes under the study is given in Supplementary Table S1. The genomes were also compared using BLASTn against a reference strain G27 as shown in Figure 1.

The genomes of sequenced isolates were pooled together with others from NCBI database to construct a whole genome based phylogenetic tree. The phylogenetic tree demonstrated a similar clustering pattern of various isolates as reported by other MLST-based phylogenetic trees (5,7,9). The strains co-clustered according to their genetic relatedness, exhibited by the formation of distinct clusters, and could be grouped according to their geographical affinities as shown in Figure 2. The strains affiliated to European countries formed hpEurope cluster, whereas those from African continent clustered into hpAfrica1 and hpAfrica2. The East-Asian genotype (hpEastAsia) has been further subdivided into three subpopulations: hspEastAsia (found in Japan and China), hspAmerind (found among Native Americans) and hspMaori (found among Taiwanese aboriginals, Melanesians and Polynesians) (43). Although studies based on MLST have indicated existence of three lineages (6,10,19) in South Asia: hpEurope, hpAsia2 and hpEastAsia, a comprehensive understanding of their phylogeny, evolution and adaptation could not be achieved perhaps because of scarcity of available genome sequences. A rapid increase in the number of genome sequences being available provided a better opportunity to achieve greater resolution in classifying *H. pylori* strains by whole genome comparative studies.

The isolates from South India and those from Malaysian-Indians clustered tightly forming a group that we named as hpSouthIndia. This close phylogenetic association among Indian and Malaysian-Indian strains is in accordance with the findings of Tay *et al.* based on MLST typing (19). Moreover, UM045 and UM037 isolated from Malay and Malaysian-Indian patients, respectively, clustered with hpEurope genotype suggesting their European ancestry (11). Further analysis of only Malaysian *H. pylori* genomes also revealed a bipartite clustering that was supplemented by a similarity score matrix, as shown in Figure 3. All the strains of Malay and Malaysian-Indian (European) origin exhibited more similarity to each other allowing them to cluster away from the Malaysian-Chinese (East-Asian) strains. These findings also appear to support the hypothesis of ancient human migration entailing ancestral Indians (11) and their subsequent migration to South Asia including Malaysia (6,19). Similarly, the affinity of Malaysian-Chinese strains with hspEastAsia reiterated their common ancestry. On the whole, the phylogenetic analysis explained a mixed population genetic structure of *H. pylori* existing in Malaysian population. These differential genotypes might explain the observed discrepancy in the colonization rates and disease outcome among various ethnic groups (44–46).

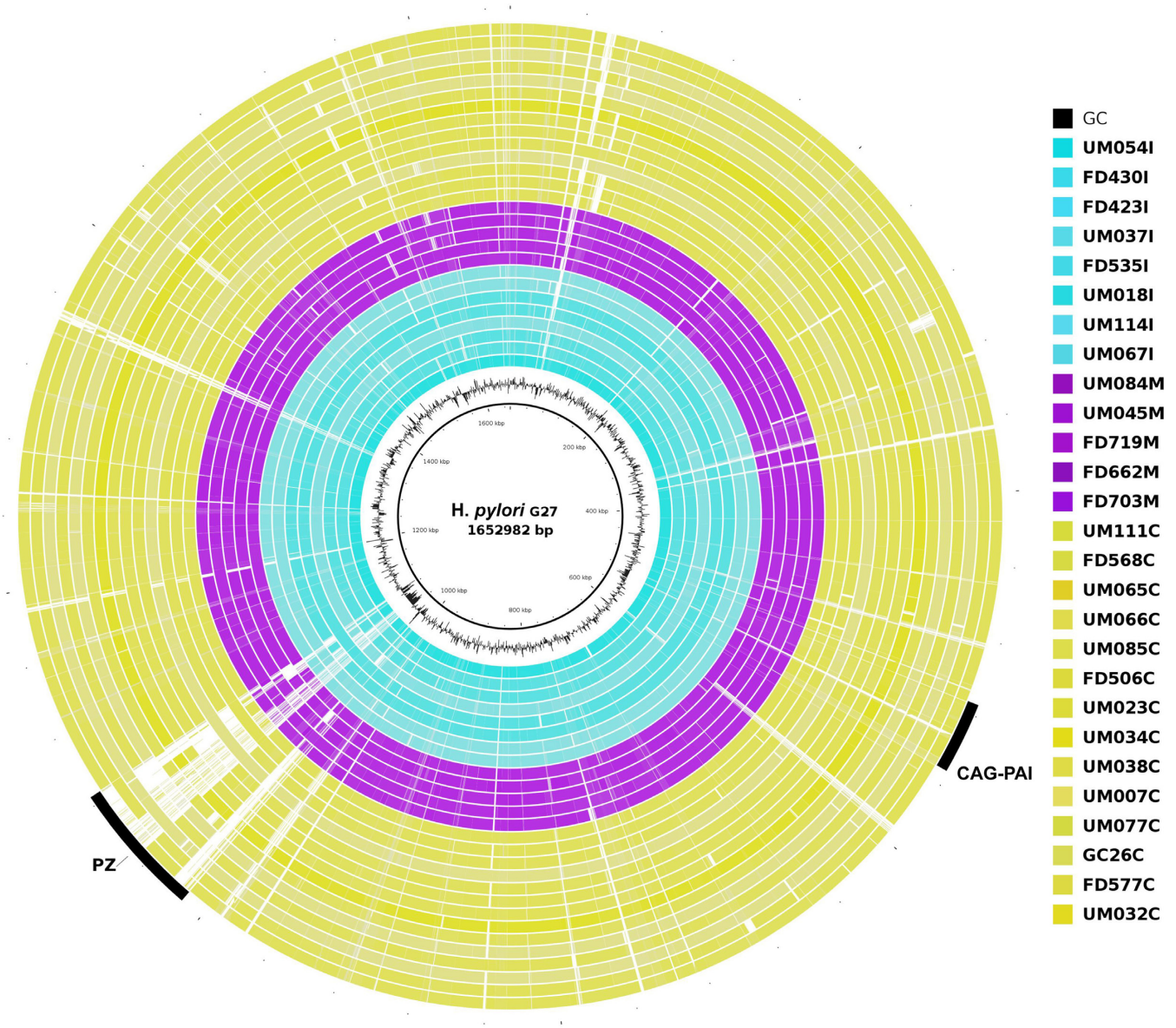


Figure 1. A circular representation of the genomes of Malaysian isolates: the draft genomes of 27 strains were aligned against the genome of reference strain *H. pylori* G27. Each genome is represented by a ring. The yellow rings represent *H. pylori* genomes from Malaysian-Chinese, purple represents those from Malays and light blue represents genomes of Malaysian-Indian strains. The G+C content (%) of the reference genome (strain G27) is represented by a ragged inner circle in black (GC). The variable regions such as plasticity zones (PZ) and *cag*PAI are compared across all the genomes using BRIG image generator (http://brig.sourceforge.net).

Virulence potential

The observed phylogenetic distinction among the Malaysian isolates was further investigated for the presence of differential virulence gene content. Various comparative studies have reported high polymorphism among different *H. pylori* lineages. Therefore, we sought to analyze the status of OMPs among 27 Malaysian *H. pylori* isolates. All the Malaysian genomes revealed a conserved nature for most of the 62 OMPs with minor exceptions. The BLASTn similarity percentage for these genes varied from 84 to 100 indicating their polymorphic nature. Few of the genes such as *hopZ*, *hopMN*, *hopQ* (*sabB*) varied among the strains, but we could not succeed in identifying a lineage/group-specific pattern among the East-Asian and other strains. The genes such as *homA* and *homB* were also found to variably exist among the genomes. The status of genes encoding these OMPs has been shown in Supplementary Table S2. Some of these genes correspond to critical virulence determinants induced upon host cell contact. These OMPs play an important role in adhesion and are reported to be associated with increased pro-inflammatory responses (47). OMPs in *H. pylori* have been categorized into five different families based on their structural composition and are known to carry out various functions ranging from host–surface interactions to non-selective porins for import of ions (48,49).

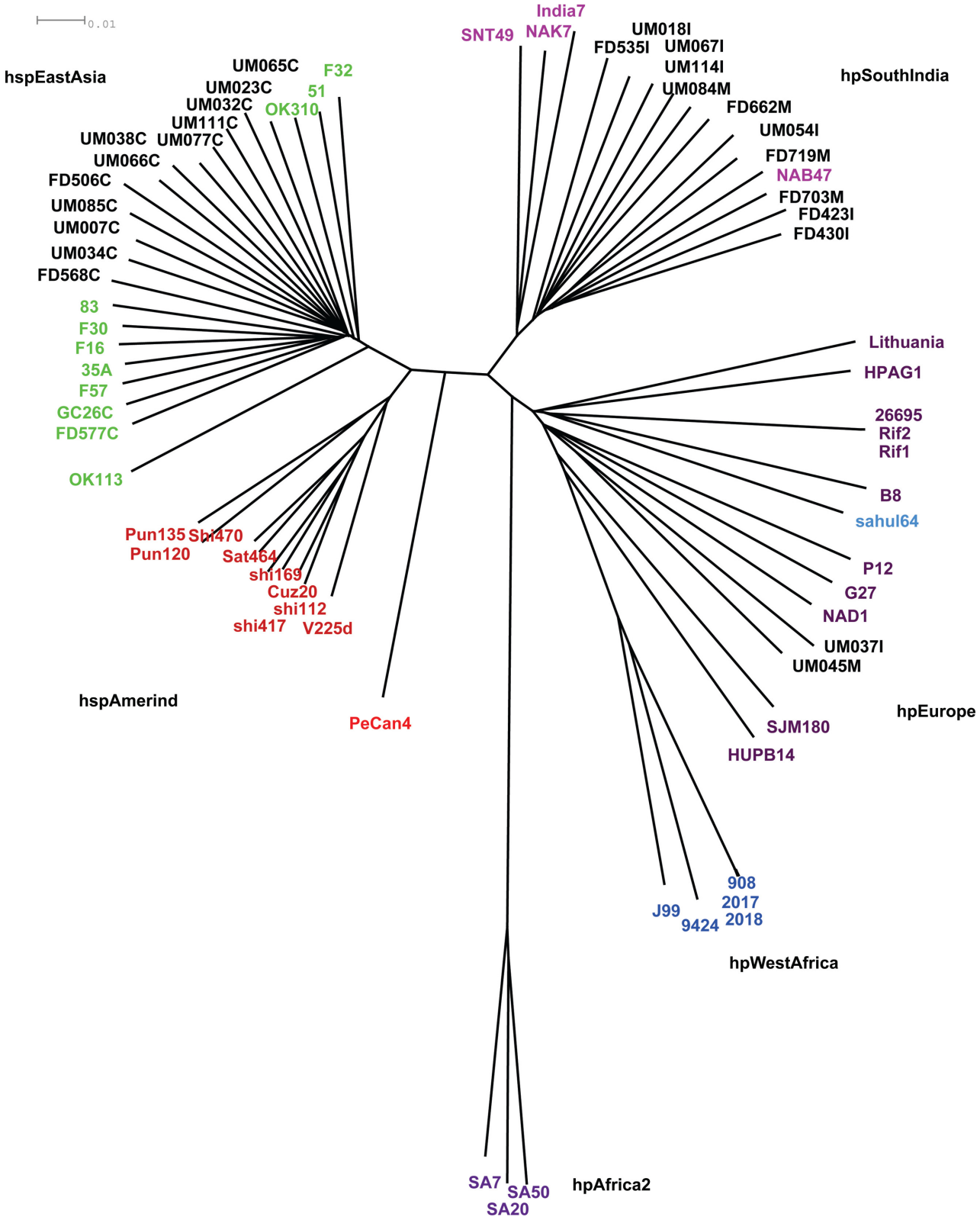


Figure 2. The whole genome phylogenetic analysis: the figure represents a whole genome comparison-based phylogenetic tree of various complete and draft *H. pylori* genomes from different geographical regions. The tree was constructed based on neighbor joining algorithm using SplitsTree. The Malaysian strains used in the analysis are labeled in black whereas the other genomes are colored to represent their genotypes.

Table 1. Genome statistics of the sequenced Malaysian *H. pylori* isolates

	UM018	UM054	UM007	UM034	UM045
Origin	Malaysian-Indian	Malaysian-Indian	Chinese	Chinese	Malay
Avg. genome coverage	170X	170X	150X	180X	200X
No. of contigs	72	89	80	27	28
Genome size	1 617 433	1 603 218	1 568 678	1 714 278	1 623 876
G+C	39.05	39.12	38.84	38.59	38.96
CDS	1579	1585	1557	1669	1595
Avg. length	937	926	924	940	933
Coding%	91.5	91.6	91.7	91.5	91.7
rRNA	3	4	3	4	3
tRNA	36	36	36	36	36
cagA (EPIYA-motif)	AB-C	AB-C	AB-D	AB-D	AB-C
vacA	s1m1	s2m2	s1m2	s1m2	s1m2
Prophage	Absent	Present (incomplete)	Absent	Absent	Present (intact)

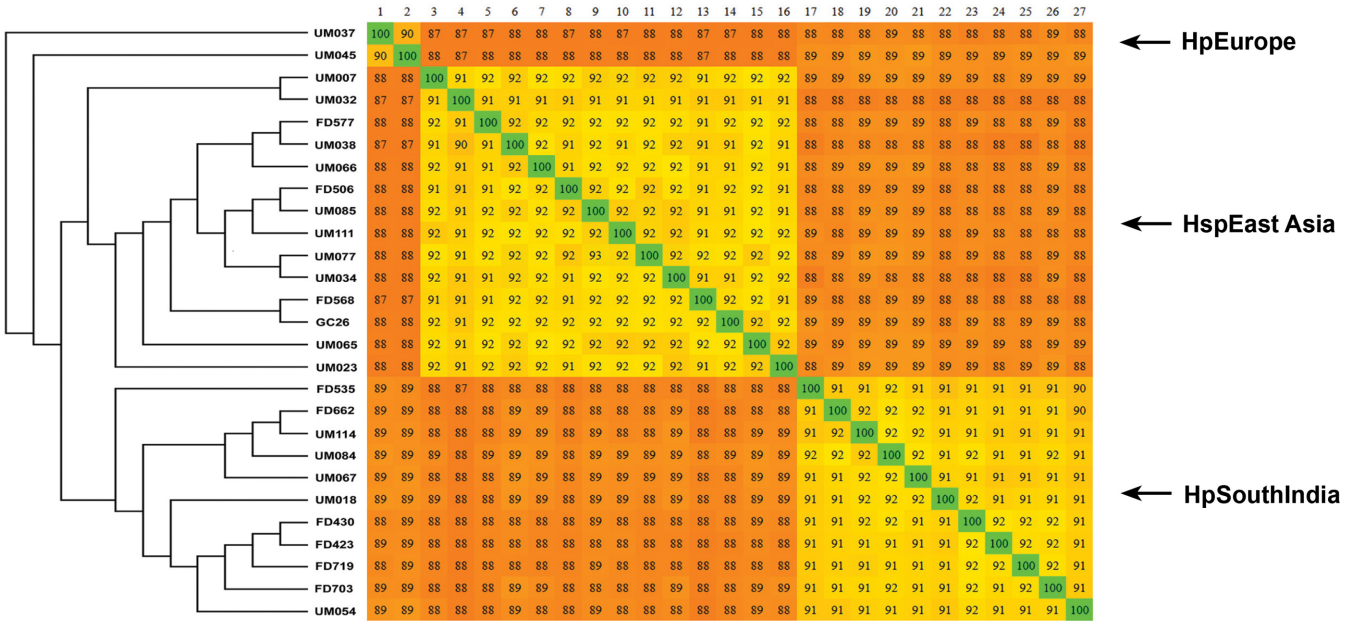


Figure 3. The analysis of 27 Malaysian genomes: the Neighbor-Joining phylogenetic tree constructed after the alignment of 27 Malaysian *H. pylori* isolates representing various ethnic groups. The heat plot shows average similarity values among the strains.

The high recombination capability of *H. pylori* (50) and its natural competence (51) makes it difficult to draw conclusive inferences about its virulence apparatus. Therefore, various computational and functional efforts have revealed a number of genes implicated in pathogenesis of *H. pylori*. The virulence factor database (VFDB) (41) lists all the reported and predicted virulence markers for pathogenic organisms including *H. pylori*. We determined the status of all 57 virulence markers in the Malaysian *H. pylori* genomes using BLASTp (Supplementary Table S3). All the strains harbored intact *cag*PAI including a conserved *cagA* gene. Other virulence markers such as *oipA*, *vacA* and *flgG* which have been associated with severe disease phenotypes were also consistently present. Further, all the genomes possessed components of the urease cluster which allows *H. pylori* to survive under low pH conditions. The analysis thus revealed a high virulence potential encoded by the genomes irrespective of the ethnic groups that they represented. However, analysis of gene polymorphisms in *cag*A revealed lineage-specific patterns. CagA protein encoded by the *cag*PAI is highly correlated with severe gastric outcomes (45,52). The extraordinary virulence potential of CagA has earned its name as a bacterial 'oncoprotein' (53). Phylogenetic analysis of CagA could clearly differentiate East-Asian (Malaysian-Chinese) strains from their non-East-Asian (Malay and Malaysian Indian) counterparts (Supplementary Figure S1). The analysis of alignment revealed a lineage-specific variation not only at C-terminal EPIYA motifs but also at N-terminal region. The Malaysian-Indian and Malay strains possessed AB-C-type EPIYA motifs, whereas all the Malaysian-Chinese strains had AB-D-type motifs. These findings are in line with others and suggest a differential evolution of this protein among isolates of different lineages and its probable role in the observed disparity of the disease outcomes (13).

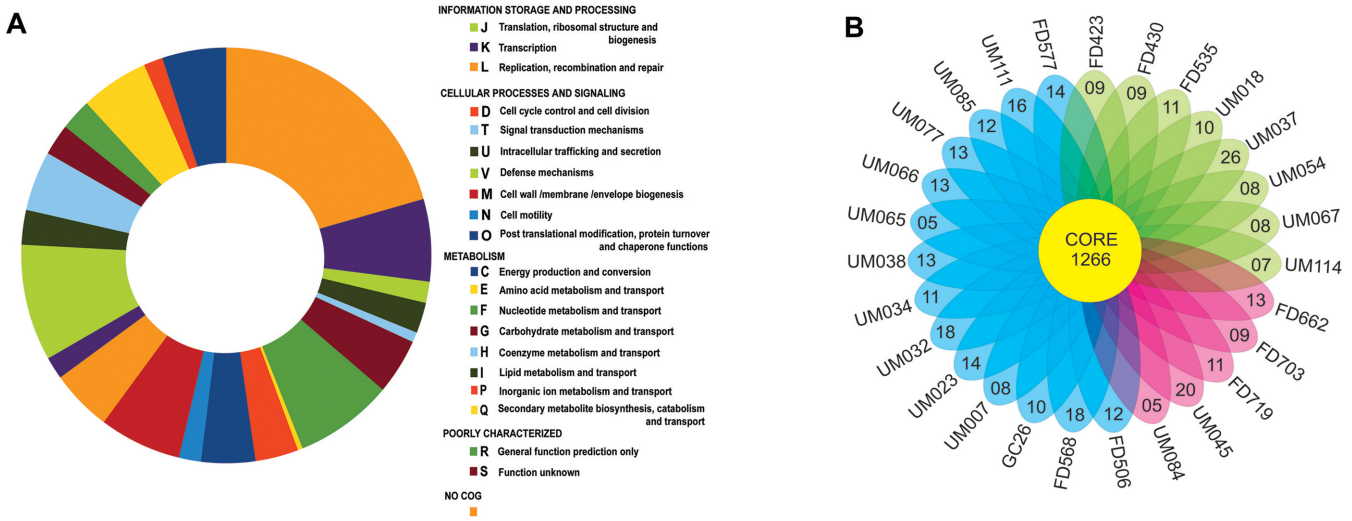


Figure 4. The functional COG classification of genes: (**A**) the COG functional classification representing core genome of the Malaysian *H. pylori* isolates. (**B**) Core and specific gene content observed among various strains compared in the study.

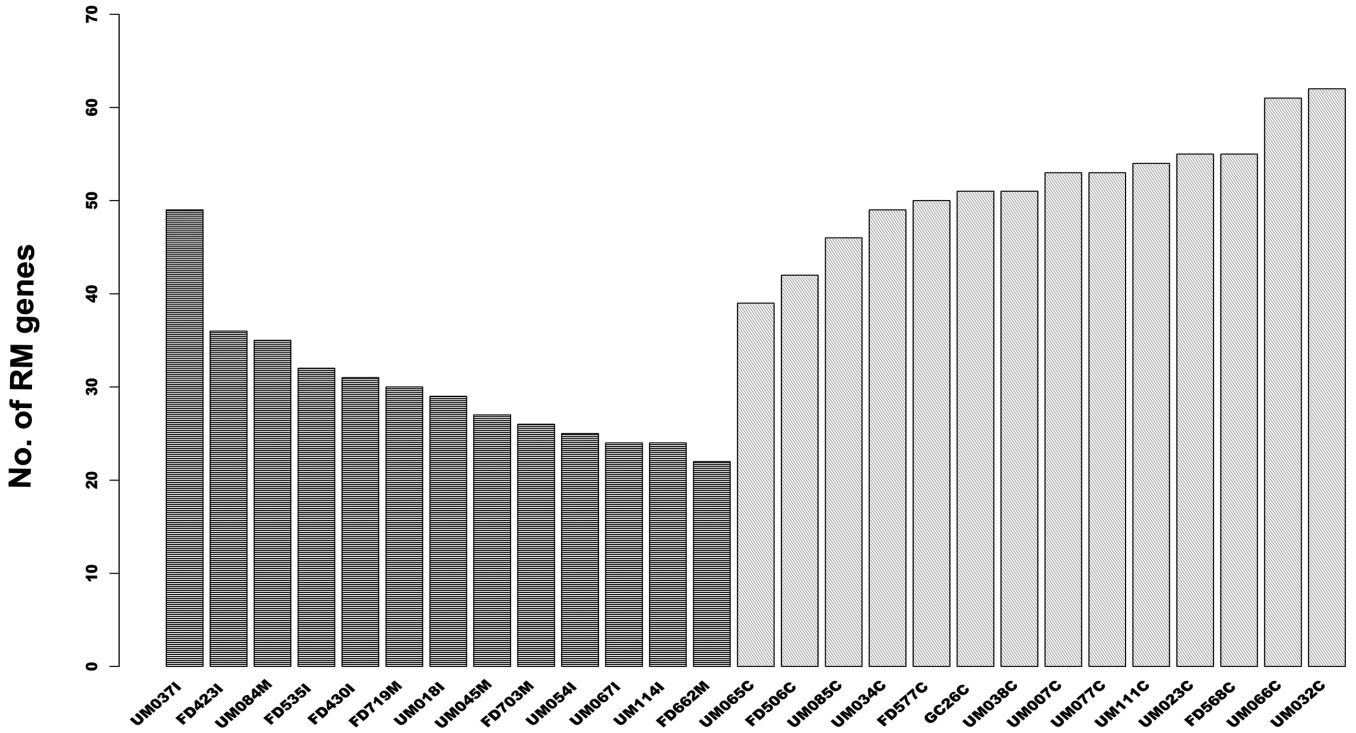


Figure 5. The distribution of RM genes in various strains: the graph shows the number of genes annotated to encode putative RM functions in each strain. The Y-axis represents the number of genes and the X-axis denotes strain names.

The core genome of Malaysian *H. pylori*

The gene content analysis of Malaysian isolates was carried out by calculating the core and accessory genome content. The genes that shared orthologs in all the genomes constituted the core while the accessory gene pool was constituted by those gene clusters which did not have orthologs in all the genomes. All the genes from the 27 strains formed a total of 1993 orthologous gene clusters. Among them, 1266 clusters comprised orthologs in all the genomes representing the core gene pool, whereas the remaining 727 gene clusters formed the variable or accessory gene pool which is also in accordance with the previous reports (54). Out of 1266 core gene clusters, 1005 clusters did find a significant match with the COG database and were assigned functional categories as shown in Figure 4A and the rest 261 gene clusters remained uncategorized. Among these 261 gene clusters, a majority were found to be encoding putative hypothetical proteins based on their comparison with other *H. pylori* genes. Out of 1005 functionally categorized gene clus-

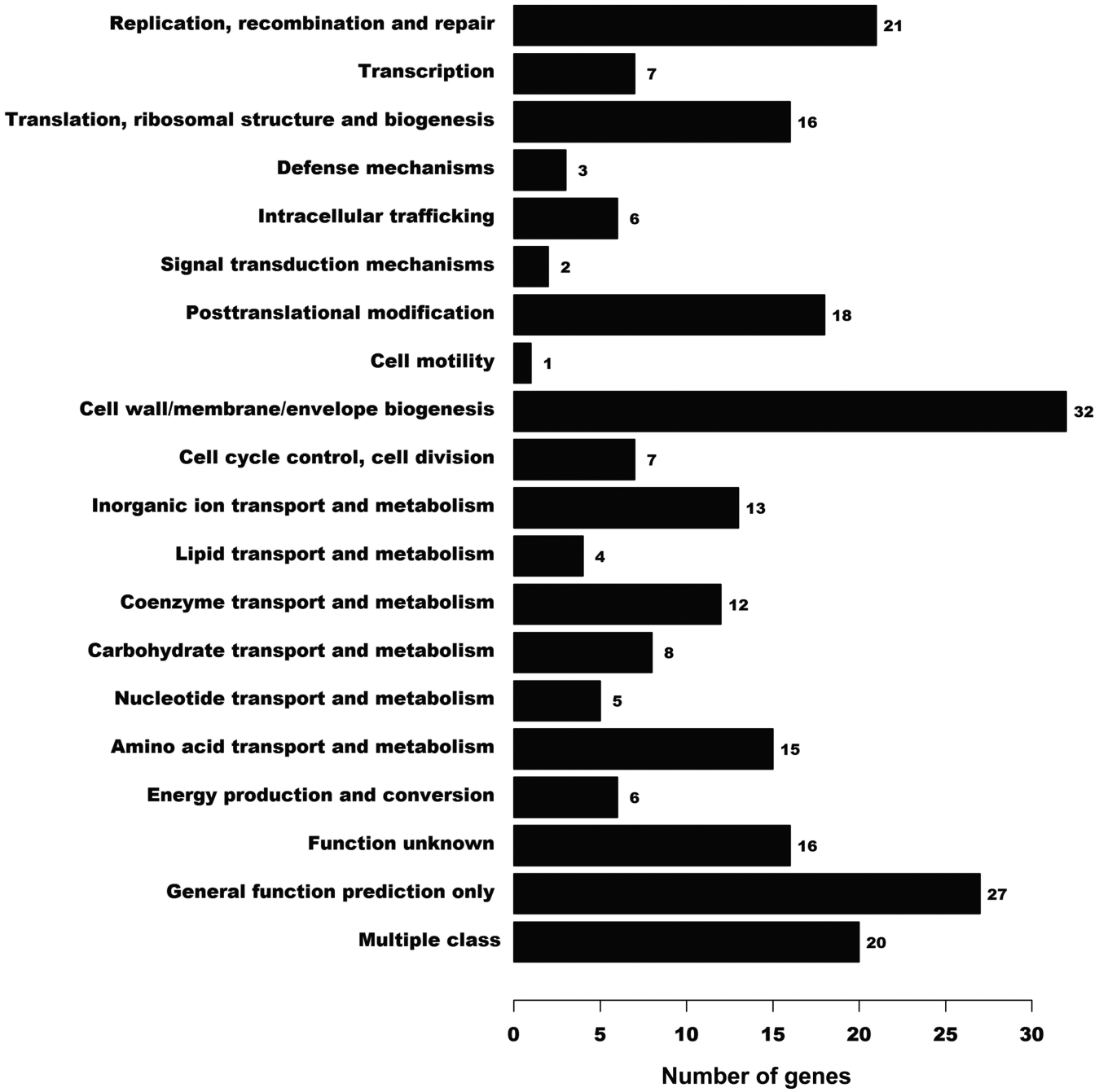


Figure 6. The functional classification of differentially evolving genes: the graph shows the COG functional classification of various core genes with signs of differential evolution among East-Asian and non-East-Asian strains. The Y-axis denotes the functional category and the X-axis represents the number of genes in a particular functional category.

ters, 115 clusters encoded proteins involved in translation, ribosomal structure and biogenesis. Other than performing housekeeping functions, studies have shown that some proteins like Pol I also aid in generating genome plasticity (55). We found the core genome to be enriched with the genes related to cell wall biosynthesis and amino acid/ion transport. The existence of a significant proportion of these transport related genes may be suggestive of an increased dependence of *H. pylori* on the host metabolites which could possibly be a result of its long association with the host. Interestingly, 81 core gene clusters were identified as belonging to multiple functional classes. These genes could represent the proteins involved in multiple pathways (56). Multifunctional proteins could also be advantageous to *H. pylori* with a small genome and limited coding potential, but this requires further functional validation. Moreover, a high proportion of hypothetical proteins in the core genome also warrants their functional characterization to ascertain their role in the biology and pathogenesis of this gastric pathogen.

H. pylori has been reported to possess a high strain-specific gene content majorly localized in hypervariable regions known as plasticity zones (57). Few genes from plasticity regions have been reported to be associated with increased pro-inflammatory secretion in cell culture studies. Recent studies on *jhp0940* and *hp0986* have provided strong evidence for their role in induction of pathogenic phenotypes (58–60). Of the 27 Malaysian strains, 11 harbored *hp0986* and belonged to non-East-Asian genotype. Further, *jhp0940* was found to be present in seven of the East-Asian strains and one non-East-Asian strain. The presence of these genes also reflects their importance in the pathogenesis of this pathogen. The strain-specific content of Malaysian genomes varied from 10 to 12 genes per genome (Figure 4B). A high proportion of these strain specific genes was predicted to encode hypothetical proteins while few of them encoded putative restriction-modification (RM) systems in some strains. A total of 749 strain-specific genes were identified among the Malaysian strains of which 15 genes were found to encode putative type II restriction or modification related functions. Interestingly, these were mostly prevalent among the Malaysian-Chinese strains that appear highly virulent and therefore warrant further functional characterization of their possible role in the pathogenesis of *H. pylori.*

Lineage-specific genes

Our phylogenetic analysis classified Malaysian isolates into three distinct genotypes. The strains belonging to hpEurope shared a close similarity to hpSouthIndia compared to hspEastAsia strains. We divided the strains into two groups in accordance with their phylogenetic clustering and genomic identity. In total, 14 Malaysian-Chinese strains represented the East-Asian group, while 13 Malaysian-Indian and Malay strains constituted the non-East-Asian group. The core genome content was calculated for each group separately from the same orthologous cluster file. The East-Asian core genome possessed 1299 orthologous gene clusters, whereas 1301 gene clusters formed the core content among non-East-Asian genomes. The comparison of these two core genomes revealed 33 clusters conserved among East-Asian but varied among non-East-Asian genomes. Out of these 33 gene clusters, four gene clusters did not have orthologs in any of the non-East-Asian strains; one of them was predicted to encode a putative lysozyme-like protein (Table 2). The lysozyme-like proteins have been observed to be upregulated during DNA-damage-induced stress in *H. pylori* (50). The other three encoded hypothetical proteins await further functional characterization. A clear understanding of the proteins encoded by these genes could provide significant insights into the underlying distinction between East-Asian and non-East-Asian genotypes and associated disease outcomes (61).

RM genes

Previous studies on *H. pylori* genomes revealed a proportion of genes encoding RM systems (62). In our collection of 27 Malaysian genomes, a total of 1077 genes were predicted to encode RM-related genes. Further, clustering by UCLUST (63) with an identity of 80% arranged these genes into 149 clusters. We then analyzed the RM gene content of East-Asian (Malaysian-Chinese) and non-East-Asian (Malay and Malaysian-Indian) strains separately. It was observed that East-Asian strains together contained 698 genes, whereas non-East-Asian strains had only 379 genes. Thus, East-Asian strains harbored, on average, 52 RM genes per strain, much higher compared to 29 RM genes per strain for non-East-Asian strains (Figure 5). In addition, the distribution of RM genes among compared genomes revealed a higher proportion of genes in East-Asian strains as shown in Figure 5. This analysis, therefore, clearly outlines the extent of diversity both in terms of numbers and allelic diversity of RM genes present in *H. pylori*. This might explain the observed strain to strain diversity in *H. pylori*. Higher proportion of RM genes in case of East-Asian strains is striking and warrants further functional validation. The role of RM genes in regulating gene expression and virulence of *H. pylori* is being earnestly pursued. Recent studies have proved that inactivation of the RM genes leads to changes in the expression of several genes in *H. pylori* (64). Moreover, these RM genes have also been shown to exhibit phase variation (65). Therefore, a clear understanding of the roles played by these RM genes in a host/lineage-specific manner would be necessary to better understand the mechanisms of differential host adaptation in *H. pylori*.

Differentially evolving genes

It has been proposed that pathogenic bacteria that resort to long-term adaptation to a particular niche modulate their core gene repertoire in synchrony with their virulence complement to gain fitness advantage (66). Therefore, we attempted to identify core gene clusters that show some evidence of differential evolution between East-Asian and non-East-Asian strains. The core gene clusters were analyzed by constructing a gene-based phylogeny to search for those gene clusters which distinguished East-Asian strains from non-East-Asian strains. The analysis identified 311 out of 1266 core gene clusters with possible signs of differential evolution. Out of 311, only 239 genes could be assigned to functional categories while the rest did not find a significant hit with the COG database. Their functional categorization revealed an enrichment of genes with functions related to cell-wall/membrane biogenesis, recombination and repair, plus some others with poorly characterized functions (Figure 6). The latter included various OMPs that have been proven to be differentially evolving among East-Asian and non-East-Asian genomes (67). Even *cagA* and *vacA* that are known to be differentially evolving among East-Asian-type and non-East-Asian-type strains showed up in our analysis. The core genome thus possesses a significant number of differentially evolving genes. This also mirrors the differential adaptive and evolutionary pressures experienced by these isolates.

CONCLUSION AND FUTURE PERSPECTIVES

This study was aimed at understanding the genetic structure of *H. pylori* in Malaysia and cues obtained therefrom to gain insights into observed variation in disease outcomes. The

Table 2. Genes differentially present among the core of East-Asian (EA) and non-East-Asian (Non-EA) *H. pylori*

	Status in *H. pylori* strains			
Cluster ID	EA ($n = 14$)	Non-EA ($n = 13$)	Predicted functions	Orthologs in 26695
2426	14	0	Lysozyme family protein	HP0339
2425	14	0	Hypothetical protein	HP0344
2424	14	0	Hypothetical protein	HP0346
2423	14	0	Hypothetical protein	Absent
2403	14	2	Lipopolysaccharide biosynthesis protein	Absent
2376	14	4	Type II restriction endonuclease	HP1537
2375	14	4	Type II methylase	Absent
2370	14	5	Glycosyltransferase	Absent
2364	14	4	DNA methyltransferase	HP0051
2402	1	13	Hypothetical protein	Absent

whole genome phylogenetic analysis resolved the strains into three lineages representing patients/individuals from various ethnic groups in a multicultural setting such as Malaysia. The conservation of most of the virulence related genes in the core genome revealed a high pathogenic potential of the strains. Few genes were found to be more prevalent in East-Asian strains as compared to others but await further confirmation considering the draft status of the genomes we analyzed. Further investigation of the core gene pool revealed a significant proportion of genes differentially represented/evolving among East-Asian and non-East-Asian strains. These differentially evolving genes included RM genes and OMPs. Given these findings, it is tempting to believe that *H. pylori* could possibly harness various mechanisms like surface antigen variation and virulence gene regulation to effectively evade the inhospitable microenvironment of the host. A careful analysis of these molecular interactions would also open avenues for the development of specific control strategies and drug intervention for *H. pylori*. A functional level understanding of the preponderances and interplay of the virulence and core gene complements among different strains/lineages would allow us to gain better understanding of the pathogen biology and host–pathogen interactions in different endemic settings.

ACCESSION NUMBERS

The whole genome sequences of five Malaysian strains sequenced in this study have been submitted to NCBI genome database with the following accession numbers: UM018 (AONK00000000), UM054 (AONL00000000), UM007 (AONM00000000), UM034 (AONN00000000) and UM045 (AONO00000000).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

We would like to thank members of the Ahmed lab, Vadivelu lab and Marshall labs for constructive suggestions and discussions. N.A. is a Visiting Professor at the University of Malaya, Malaysia, and is an Adjunct Professor of the Academy of Scientific and Innovative Research (AcSIR), India. We are thankful to several researchers for the use of genome sequence data from NCBI.

FUNDING

Department of Biotechnology, Government of India [BT/PR6921/MED/29/699/2013 to N.A.]; Senior Research Fellowship, Council for Scientific and Industrial Research, India [to N.K.]; HIR Project of the University of Malaya [UM.C/625/1/HIR/MOHE/CHAN-02-Molecular Genetics to J.V., K.L.G., N.A., M.B.G., B.J.M.]. Funding for openaccess charge: University of Malaya [UM.C/625/1/HIR/MOHE/CHAN-02-Molecular Genetics to J.V., K.L.G., N.A., M.B.G., B.J.M.].
Conflict of interest statement. None declared.

REFERENCES

1. Perez-Perez,G.I., Rothenbacher,D. and Brenner,H. (2004) Epidemiology of *Helicobacter pylori* infection. *Helicobacter*, **9**(Suppl. 1), 1–6.
2. Khalifa,M.M., Sharaf,R.R. and Aziz,R.K. (2010) *Helicobacter pylori*: a poor man's gut pathogen? *Gut Pathog.*, **2**, 2.
3. Amieva,M.R. and El-Omar,E.M. (2008) Host-bacterial interactions in *Helicobacter pylori* infection. *Gastroenterology*, **134**, 306–323.
4. Cover,T.L. and Blaser,M.J. (2009) *Helicobacter pylori* in health and disease. *Gastroenterology*, **136**, 1863–1873.
5. Kuipers,E.J., Israel,D.A., Kusters,J.G., Gerrits,M.M., Weel,J., van Der Ende,A., van Der Hulst,R.W., Wirth,H.P., Hook-Nikanne,J., Thompson,S.A. *et al.* (2000) Quasispecies development of *Helicobacter pylori* observed in paired isolates obtained years apart from the same host. *J. Infect. Dis.*, **181**, 273–282.
6. Breurec,S., Guillard,B., Hem,S., Brisse,S., Dieye,F.B., Huerre,M., Oung,C., Raymond,J., Tan,T.S., Thiberge,J.M. *et al.* (2011) Evolutionary history of *Helicobacter pylori* sequences reflect past human migrations in Southeast Asia. *PloS one*, **6**, e22058.
7. Linz,B., Balloux,F., Moodley,Y., Manica,A., Liu,H., Roumagnac,P., Falush,D., Stamer,C., Prugnolle,F., van der Merwe,S.W. *et al.* (2007) An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature*, **445**, 915–918.
8. Falush,D., Wirth,T., Linz,B., Pritchard,J.K., Stephens,M., Kidd,M., Blaser,M.J., Graham,D.Y., Vacher,S., Perez-Perez,G.I. *et al.* (2003) Traces of human migrations in *Helicobacter pylori* populations. *Science*, **299**, 1582–1585.
9. Achtman,M., Azuma,T., Berg,D.E., Ito,Y., Morelli,G., Pan,Z.J., Suerbaum,S., Thompson,S.A., van der Ende,A. and van Doorn,L.J. (1999) Recombination and clonal groupings within *Helicobacter pylori* from different geographical regions. *Mol. Microbiol.*, **32**, 459–470.
10. Moodley,Y., Linz,B., Yamaoka,Y., Windsor,H.M., Breurec,S., Wu,J.Y., Maady,A., Bernhoft,S., Thiberge,J.M., Phuanukoonnon,S. *et al.* (2009) The peopling of the Pacific from a bacterial perspective. *Science*, **323**, 527–530.
11. Devi,S.M., Ahmed,I., Francalacci,P., Hussain,M.A., Akhter,Y., Alvi,A., Sechi,L.A., Megraud,F. and Ahmed,N. (2007) Ancestral

European roots of *Helicobacter pylori* in India. *BMC Genomics*, **8**, 184.
12. Blaser,M.J. and Berg,D.E. (2001) *Helicobacter pylori* genetic diversity and risk of human disease. *J. Clin. Invest.*, **107**, 767–773.
13. Suerbaum,S. and Josenhans,C. (2007) *Helicobacter pylori* evolution and phenotypic diversification in a changing host. *Nat. Rev. Microbiol.*, **5**, 441–452.
14. Suzuki,R., Shiota,S. and Yamaoka,Y. (2012) Molecular epidemiology, population genetics, and pathogenic role of *Helicobacter pylori*. *Infect. Genet. Evol.*, **12**, 203–213.
15. Stein,M., Ruggiero,P., Rappuoli,R. and Bagnoli,F. (2013) *Helicobacter pylori* CagA: from pathogenic mechanisms to its use as an anti-cancer vaccine. *Front. immunol.*, **4**, 328.
16. Andaya,L.Y. (2001) The search for the 'origins' of Melayu. *J. Southeast Asian Stud.*, **32**, 315–330.
17. Hill,C., Soares,P., Mormina,M., Macaulay,V., Meehan,W., Blackburn,J., Clarke,D., Raja,J.M., Ismail,P., Bulbeck,D. *et al.* (2006) Phylogeography and ethnogenesis of aboriginal Southeast Asians. *Mol. Biol. Evol.*, **23**, 2480–2491.
18. Tan,H.J., Rizal,A.M., Rosmadi,M.Y. and Goh,K.L. (2005) Distribution of *Helicobacter pylori* cagA, cagE and vacA in different ethnic groups in Kuala Lumpur, Malaysia. *J. Gastroenterol. Hepatol.*, **20**, 589–594.
19. Tay,C.Y., Mitchell,H., Dong,Q., Goh,K.L., Dawes,I.W. and Lan,R. (2009) Population structure of *Helicobacter pylori* among ethnic groups in Malaysia: recent acquisition of the bacterium by the Malay population. *BMC Microbiol.*, **9**, 126.
20. Lee,Y.Y., Mahendra Raj,S. and Graham,D.Y. (2013) *Helicobacter pylori* infection–a boon or a bane: lessons from studies in a low-prevalence population. *Helicobacter*, **18**, 338–346.
21. Rehvathy,V., Tan,M.H., Gunaletchumy,S.P., Teh,X., Wang,S., Baybayan,P., Singh,S., Ashby,M., Kaakoush,N.O., Mitchell,H.M. *et al.* (2013) Multiple genome sequences of *Helicobacter pylori* strains of diverse disease and antibiotic resistance backgrounds from Malaysia. *Genome Announcements*, **1**, doi:10.1128/genomeA.00687-13.
22. Khosravi,Y., Rehvathy,V., Wee,W.Y., Wang,S., Baybayan,P., Singh,S., Ashby,M., Ong,J., Amoyo,A.A., Seow,S.W. *et al.* (2013) Comparing the genomes of *Helicobacter pylori* clinical strain UM032 and Mice-adapted derivatives. *Gut Pathog.*, **5**, 25.
23. Gunaletchumy,S.P., Teh,X., Khosravi,Y., Ramli,N.S., Chua,E.G., Kavitha,T., Mason,J.N., Lee,H.T., Alias,H., Zaidan,N.Z. *et al.* (2012) Draft genome sequences of *Helicobacter pylori* isolates from Malaysia, cultured from patients with functional dyspepsia and gastric cancer. *J. Bacteriol.*, **194**, 5695–5696.
24. Patel,R.K. and Jain,M. (2012) NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PloS one*, **7**, e30619.
25. Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.
26. Overbeek,R., Olson,R., Pusch,G.D., Olsen,G.J., Davis,J.J., Disz,T., Edwards,R.A., Gerdes,S., Parrello,B., Shukla,M. *et al.* (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.*, **42**, D206–D214.
27. Besemer,J., Lomsadze,A. and Borodovsky,M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.*, **29**, 2607–2618.
28. Larsen,T.S. and Krogh,A. (2003) EasyGene—a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics*, **4**, 21.
29. Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.
30. Salzberg,S.L., Delcher,A.L., Kasif,S. and White,O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.*, **26**, 544–548.
31. Aziz,R.K., Devoid,S., Disz,T., Edwards,R.A., Henry,C.S., Olsen,G.J., Olson,R., Overbeek,R., Parrello,B., Pusch,G.D. *et al.* (2012) SEED servers: high-performance access to the SEED genomes, annotations, and metabolic models. *PloS one*, **7**, e48053.
32. Rutherford,K., Parkhill,J., Crook,J., Horsnell,T., Rice,P., Rajandream,M.A. and Barrell,B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.
33. Lagesen,K., Hallin,P., Rodland,E.A., Staerfeldt,H.H., Rognes,T. and Ussery,D.W. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.*, **35**, 3100–3108.
34. Schattner,P., Brooks,A.N. and Lowe,T.M. (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.*, **33**, W686–W689.
35. Li,L., Stoeckert,C.J. Jr and Roos,D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.
36. Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.
37. Natale,D.A., Galperin,M.Y., Tatusov,R.L. and Koonin,E.V. (2000) Using the COG database to improve gene recognition in complete genomes. *Genetica*, **108**, 9–17.
38. Agren,J., Sundstrom,A., Hafstrom,T. and Segerman,B. (2012) Gegenees: fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups. *PloS one*, **7**, e39107.
39. Kumar,N., Mukhopadhyay,A.K., Patra,R., De,R., Baddam,R., Shaik,S., Alam,J., Tiruvayipati,S. and Ahmed,N. (2012) Next-generation sequencing and de novo assembly, genome organization, and comparative genomic analyses of the genomes of two *Helicobacter pylori* isolates from duodenal ulcer patients in India. *J. Bacteriol.*, **194**, 5963–5964.
40. Huson,D.H. (1998) SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, **14**, 68–73.
41. Chen,L., Yang,J., Yu,J., Yao,Z., Sun,L., Shen,Y. and Jin,Q. (2005) VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.*, **33**, D325–D328.
42. Zhou,Y., Liang,Y., Lynch,K.H., Dennis,J.J. and Wishart,D.S. (2011) PHAST: a fast phage search tool. *Nucleic Acids Res.*, **39**, W347–W352.
43. Moodley,Y., Linz,B., Bond,R.P., Nieuwoudt,M., Soodyall,H., Schlebusch,C.M., Bernhoft,S., Hale,J., Suerbaum,S., Mugisha,L. *et al.* (2012) Age of the association between *Helicobacter pylori* and man. *PLoS Pathog.*, **8**, e1002693.
44. Kaur,G. and Naing,N.N. (2003) Prevalence and ethnic distribution of *Helicobacter pylori* infection among endoscoped patients in north eastern peninsular malaysia. *Malays. J. Med. Sci.*, **10**, 66–70.
45. Peek,R.M. Jr and Crabtree,J.E. (2006) Helicobacter infection and gastric neoplasia. *J. Pathol.*, **208**, 233–248.
46. Ahmed,N., Loke,M.F., Kumar,N. and Vadivelu,J. (2013) *Helicobacter pylori* in 2013: multiplying genomes, emerging insights. *Helicobacter*, **18**(Suppl. 1), 1–4.
47. Gerhard,M., Rad,R., Prinz,C. and Naumann,M. (2002) Pathogenesis of *Helicobacter pylori* infection. *Helicobacter*, **7**(Suppl. 1), 17–23.
48. Alm,R.A., Ling,L.S., Moir,D.T., King,B.L., Brown,E.D., Doig,P.C., Smith,D.R., Noonan,B., Guild,B.C., deJonge,B.L. *et al.* (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature*, **397**, 176–180.
49. Alm,R.A., Bina,J., Andrews,B.M., Doig,P., Hancock,R.E. and Trust,T.J. (2000) Comparative genomics of *Helicobacter pylori*: analysis of the outer membrane protein families. *Infect. Immun.*, **68**, 4155–4168.
50. Dorer,M.S., Fero,J. and Salama,N.R. (2010) DNA damage triggers genetic exchange in *Helicobacter pylori*. *PLoS Pathog.*, **6**, e1001026.
51. Dorer,M.S., Cohen,I.E., Sessler,T.H., Fero,J. and Salama,N.R. (2013) Natural competence promotes *Helicobacter pylori* chronic infection. *Infect. Immun.*, **81**, 209–215.
52. Wiedemann,T., Loell,E., Mueller,S., Stoeckelhuber,M., Stolte,M., Haas,R. and Rieder,G. (2009) *Helicobacter pylori* cag-Pathogenicity island-dependent early immunological response triggers later precancerous gastric changes in Mongolian gerbils. *PloS one*, **4**, e4754.
53. Ohnishi,N., Yuasa,H., Tanaka,S., Sawa,H., Miura,M., Matsui,A., Higashi,H., Musashi,M., Iwabuchi,K., Suzuki,M. *et al.* (2008) Transgenic expression of *Helicobacter pylori* CagA induces gastrointestinal and hematopoietic neoplasms in mouse. *Proc. Natl Acad. Sci. U.S.A.*, **105**, 1003–1008.
54. Lu,W., Wise,M.J., Tay,C.Y., Windsor,H.M., Marshall,B.J., Peacock,C. and Perkins,T. (2014) Comparative analysis of the full genome of *Helicobacter pylori* isolate Sahul64 identifies genes of high divergence. *J. Bacteriol.*, **196**, 1073–1083.

55. Garcia-Ortiz,M.V., Marsin,S., Arana,M.E., Gasparutto,D., Guerois,R., Kunkel,T.A. and Radicella,J.P. (2011) Unexpected role for *Helicobacter pylori* DNA polymerase I as a source of genetic variability. *PLoS Genet.*, **7**, e1002152.
56. Boneca,I.G., de Reuse,H., Epinat,J.C., Pupin,M., Labigne,A. and Moszer,I. (2003) A revised annotation and comparative analysis of *Helicobacter pylori* genomes. *Nucleic Acids Res.*, **31**, 1704–1714.
57. Kersulyte,D., Lee,W., Subramaniam,D., Anant,S., Herrera,P., Cabrera,L., Balqui,J., Barabas,O., Kalia,A., Gilman,R.H. *et al.* (2009) *Helicobacter pylori*'s plasticity zones are novel transposable elements. *PloS one*, **4**, e6859.
58. Devi,S., Ansari,S.A., Vadivelu,J., Megraud,F., Tenguria,S. and Ahmed,N. (2014) *Helicobacter pylori* antigen HP0986 (TieA) interacts with cultured gastric epithelial cells and induces IL8 secretion via NF-kappaB mediated pathway. *Helicobacter*, **19**, 26–36.
59. Alvi,A., Ansari,S.A., Ehtesham,N.Z., Rizwan,M., Devi,S., Sechi,L.A., Qureshi,I.A., Hasnain,S.E. and Ahmed,N. (2011) Concurrent proinflammatory and apoptotic activity of a *Helicobacter pylori* protein (HP986) points to its role in chronic persistence. *PloS one*, **6**, e22530.
60. Rizwan,M., Alvi,A. and Ahmed,N. (2008) Novel protein antigen (JHP940) from the genomic plasticity region of *Helicobacter pylori* induces tumor necrosis factor alpha and interleukin-8 secretion by human macrophages. *J. Bacteriol.*, **190**, 1146–1151.
61. 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.
62. Lin,L.F., Posfai,J., Roberts,R.J. and Kong,H. (2001) Comparative genomics of the restriction-modification systems in *Helicobacter pylori*. *Proc. Natl Acad. Sci. U.S.A.*, **98**, 2740–2745.
63. Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.
64. Furuta,Y., Namba-Fukuyo,H., Shibata,T.F., Nishiyama,T., Shigenobu,S., Suzuki,Y., Sugano,S., Hasebe,M. and Kobayashi,I. (2014) Methylome diversification through changes in DNA methyltransferase sequence specificity. *PLoS Genet.*, **10**, e1004272.
65. Krebes,J., Morgan,R.D., Bunk,B., Sproer,C., Luong,K., Parusel,R., Anton,B.P., Konig,C., Josenhans,C., Overmann,J. *et al.* (2014) The complex methylome of the human gastric pathogen *Helicobacter pylori*. *Nucleic Acids Res.*, **42**, 2415–2432.
66. Rohmer,L., Hocquet,D. and Miller,S.I. (2011) Are pathogenic bacteria just looking for food? Metabolism and microbial pathogenesis. *Trends Microbiol.*, **19**, 341–348.
67. Duncan,S.S., Valk,P.L., McClain,M.S., Shaffer,C.L., Metcalf,J.A., Bordenstein,S.R. and Cover,T.L. (2013) Comparative genomic analysis of East Asian and non-Asian *Helicobacter pylori* strains identifies rapidly evolving genes. *PloS one*, **8**, e55120.

genomeA
Journals.ASM.org

Genomes of Two Clinical Isolates of *Mycobacterium tuberculosis* from Odisha, India

Mohammad Majid,[a,d] Narender Kumar,[a] Asifa Qureshi,[d] Priyadarshini Yerra,[a] Ashutosh Kumar,[a] Mandala Kiran Kumar,[a] Suma Tiruvayipati,[a,c] Ramani Baddam,[a] Sabiha Shaik,[a] Aparna Srikantam,[b] Niyaz Ahmed[a,c]

Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, University of Hyderabad, Hyderabad, India[a]; Blue Peter Research Centre (Lepra Society), Cherlapally, Hyderabad, India[b]; Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia[c]; Environmental Genomics Division, CSIR–National Environmental Engineering Research Institute (NEERI), Nagpur, India[d]

We report whole-genome sequences of two clinical isolates of *Mycobacterium tuberculosis* isolated from patients in Odisha, India. The sequence analysis revealed that these isolates are of an ancestral type and might represent some of the "pristine" isolates in India that have not admixed with other lineages.

Received 20 February 2014 **Accepted** 27 February 2014 **Published** 20 March 2014

Citation Majid M, Kumar N, Qureshi A, Yerra P, Kumar A, Kumar MK, Tiruvayipati S, Baddam R, Shaik S, Srikantam A, Ahmed N. 2014. Genomes of two clinical isolates of *Mycobacterium tuberculosis* from Odisha, India. Genome Announc. 2(2):e00199-14. doi:10.1128/genomeA.00199-14.

Copyright © 2014 Majid et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported license.

Address correspondence to Narender Kumar, kumarnaren13@gmail.com, or Niyaz Ahmed, niyaz.ahmed@uohyd.ac.in.

Tuberculosis caused by *Mycobacterium tuberculosis* is a chronic infectious disease that is often fatal if not effectively treated. Every year, 8 to 9 million new infections and a death toll of 1.5 million are recorded worldwide. It is estimated that about one-third of the human population is infected with *M. tuberculosis* (1). Comparative genomic studies have provided deeper insights into the genetic diversity and clonal architecture of *M. tuberculosis* (2). Recent studies conducted on isolates from India have shown that highly concentrated reservoirs of the ancestral *M. tuberculosis* lineages prevail in South and Central India (3–6). Only a limited number of *M. tuberculosis* genomes from India are sequenced. The whole-genome analysis of ancestral and modern lineages would facilitate deciphering of the genetic variability and evolutionary mechanisms of this obligate parasite. We describe the whole-genome sequences of two *M. tuberculosis* strains, NA-A0008 and NA-A0009, isolated in 2008 from patients in rural Odisha, India.

Genomic DNAs of both strains were isolated using the Qiagen kit method. Whole-genome sequencing was carried out on an Ion Torrent sequencing platform (Life Technologies). The process generated 3 million and 2.9 million reads amounting to 89× and 93× genome coverage for NA-A0008 and NA-A0009, respectively, with a mean read length of 250 bp. The reads after filtration were assembled into 280 and 310 contigs for NA-A0008 and NA-A0009, respectively, using the MIRA v.2 *de novo* assembler. These contigs were ordered and reoriented according to the *M. tuberculosis* CCDC 5180 genome using in-house written scripts. The resulting draft genomes were annotated using the RAST annotation server (7), and CDSs were validated by comparing outputs from EasyGene (8) and Glimmer (9), as done previously (10–13). The number of rRNA operons were predicted in both strains using RNAmmer (14), while tRNAscan-SE (15) was used to identify tRNA sequences. Artemis (16) was used to glean the genome statistics of both the strains. The genome sizes of NA-A0008 and NA-A0009 were 4,259,206 and 4,271,739 bp, with coding percentages of 89.4% and 89.3%, respectively. The G+C contents of both strains were high, as usually observed for the *M. tuberculosis* complex, 65.31% (NA-A0008) and 65.28% (NA-A0009). The two genomes, NA-A0008 and NA-A0009, were predicted to encode 4,400 and 4,453 CDSs with average lengths of 866 and 857 bp, respectively. Both of them contained a single rRNA operon and 45 tRNA genes.

The availability of these genome sequences would definitely complement the gene pool analysis of the Indian strains from different parts of the country. Besides this, comparative genomic analysis and phylogenetic study of these isolates with other *M. tuberculosis* strains might give us important insights into the biology and molecular epidemiology of this organism.

Nucleotide sequence accession numbers. The *M. tuberculosis* NA-A0008 and NA-A0009 whole-genome shotgun projects have been deposited in the GenBank database under the accession numbers ALYG00000000 and ALYH00000000, respectively. The BioProject designations for these projects are PRJNA168604 and PRJNA168605, respectively.

ACKNOWLEDGMENTS

We acknowledge the DFG program GRK1673 and financial support under the DBT-COE grant (BT/01/COE/07/02) from the Department of Biotechnology, Government of India. Mohammad Majid is grateful to Vijay N. Charde, M. Uma, and Mumtaz Baig for their help and support.

REFERENCES

1. **Frieden TR, Sterling TR, Munsiff SS, Watt CJ, Dye C.** 2003. Tuberculosis. Lancet **362:**887–899. http://dx.doi.org/10.1016/S0140-6736(03)14333-4.
2. **Ilina EN, Shitikov EA, Ikryannikova LN, Alekseev DG, Kamashev DE, Malakhova MV, Parfenova TV, Afanas'ev MV, Ischenko DS, Bazaleev NA, Smirnova TG, Larionova EE, Chernousova LN, Beletsky AV, Mardanov AV, Ravin NV, Skryabin KG, Govorun VM.** 2013. Comparative genomic analysis of *Mycobacterium tuberculosis* drug resistant strains from Russia. PLoS One **8:**e56577. http://dx.doi.org/10.1371/journal.pone.0056577.
3. **Narayanan S, Gagneux S, Hari L, Tsolaki AG, Rajasekhar S, Narayanan**

PR, Small PM, Holmes S, Deriemer K. 2008. Genomic interrogation of ancestral *Mycobacterium tuberculosis* from south India. Infect. Genet. Evol. **8:**474–483. http://dx.doi.org/10.1016/j.meegid.2007.09.007.

4. **Rao KR, Kauser F, Srinivas S, Zanetti S, Sechi LA, Ahmed N, Hasnain SE.** 2005. Analysis of genomic downsizing on the basis of region-of-difference polymorphism profiling of *Mycobacterium tuberculosis* patient isolates reveals geographic partitioning. J. Clin. Microbiol. **43:**5978–5982. http://dx.doi.org/10.1128/JCM.43.12.5978-5982.2005.
5. **Ahmed N, Saini V, Raghuvanshi S, Khurana JP, Tyagi AK, Tyagi AK, Hasnain SE.** 2007. Molecular analysis of a leprosy immunotherapeutic bacillus provides insights into *Mycobacterium* evolution. PLoS One **2:**e968. http://dx.doi.org/10.1371/journal.pone.0000968.
6. **Thomas SK, Iravatham CC, Moni BH, Kumar A, Archana BV, Majid M, Priyadarshini Y, Rani PS, Valluri V, Hasnain SE, Ahmed N.** 2011. Modern and ancestral genotypes of *Mycobacterium tuberculosis* from Andhra Pradesh, India. PLoS One **6:**e27584. http://dx.doi.org/10.1371/journal.pone.0027584.
7. **Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O.** 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics **9:**75. http://dx.doi.org/10.1186/1471-2164-9-75.
8. **Larsen TS, Krogh A.** 2003. EasyGene—a prokaryotic gene finder that ranks ORFs by statistical significance. BMC Bioinformatics **4:**21. http://dx.doi.org/10.1186/1471-2105-4-21.
9. **Delcher AL, Harmon D, Kasif S, White O, Salzberg SL.** 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. **27:**4636–4641. http://dx.doi.org/10.1093/nar/27.23.4636.
10. **Avasthi TS, Kumar N, Baddam R, Hussain A, Nandanwar N, Jadhav S, Ahmed N.** 2011. Genome of multidrug-resistant uropathogenic *Escherichia coli* strain NA114 from India. J. Bacteriol. **193:**4272–4273. http://dx.doi.org/10.1128/JB.05413-11.
11. **Avasthi TS, Devi SH, Taylor TD, Kumar N, Baddam R, Kondo S, Suzuki Y, Lamouliatte H, Mégraud F, Ahmed N.** 2011. Genomes of two chronological isolates (*Helicobacter pylori* 2017 and 2018) of the West African *Helicobacter pylori* strain 908 obtained from a single patient. J. Bacteriol. **193:**3385–3386. http://dx.doi.org/10.1128/JB.05006-11.
12. **Baddam R, Thong KL, Avasthi TS, Shaik S, Yap KP, Teh CS, Chai LC, Kumar N, Ahmed N.** 2012. Whole-genome sequences and comparative genomics of *Salmonella enterica* serovar. *Typhi* isolates from patients with fatal and nonfatal typhoid fever in Papua New Guinea. J. Bacteriol. **194:**5122–5123. http://dx.doi.org/10.1128/JB.01051-12.
13. **Kumar N, Mukhopadhyay AK, Patra R, De R, Baddam R, Shaik S, Alam J, Tiruvayipati S, Ahmed N.** 2012. Next-generation sequencing and de novo assembly, genome organization, and comparative genomic analyses of the genomes of two *Helicobacter pylori* isolates from duodenal ulcer patients in India. J. Bacteriol. **194:**5963–5964. http://dx.doi.org/10.1128/JB.01371-12.
14. **Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW.** 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. **35:**3100–3108. http://dx.doi.org/10.1093/nar/gkm160.
15. **Lowe TM, Eddy SR.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25:**955–964. http://dx.doi.org/10.1093/nar/25.5.0955.
16. **Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B.** 2000. Artemis: sequence visualization and annotation. Bioinformatics **16:**944–945. http://dx.doi.org/10.1093/bioinformatics/16.10.944.

Gut Pathogens

RESEARCH Open Access

Genome anatomy of the gastrointestinal pathogen, Vibrio parahaemolyticus of crustacean origin

Suma Tiruvayipati[1,2], Subha Bhassu[1,3*], Narender Kumar[2], Ramani Baddam[2], Sabiha Shaik[2], Anil Kumar Gurindapalli[2], Kwai Lin Thong[1,4] and Niyaz Ahmed[1,2*]

Abstract

Vibrio parahaemolyticus, an important human pathogen, is associated with gastroenteritis and transmitted through partially cooked seafood. It has become a major concern in the production and trade of marine food products. The prevalence of potentially virulent and pathogenic *V. parahaemolyticus* in raw seafood is of public health significance. Here we describe the genome sequence of a *V. parahaemolyticus* isolate of crustacean origin which was cultured from prawns in 2008 in Selangor, Malaysia (isolate PCV08-7). The next generation sequencing and analysis revealed that the genome of isolate PCV08-7 has closest similarity to that of *V. parahaemolyticus* RIMD2210633. However, there are certain unique features of the PCV08-7 genome such as the absence of TDH-related hemolysin (TRH), and the presence of HU-alpha insertion. The genome of isolate PCV08-7 encodes a thermostable direct hemolysin (TDH), an important virulence factor that classifies PCV08-7 isolate to be a serovariant of O3:K6 strain. Apart from these, we observed that there is certain pattern of genetic rearrangements that makes *V. parahaemolyticus* PCV08-7 a non-pandemic clone. We present detailed genome statistics and important genetic features of this bacterium and discuss how its survival, adaptation and virulence in marine and terrestrial hosts can be understood through the genomic blueprint and that the availability of genome sequence entailing this important Malaysian isolate would likely enhance our understanding of the epidemiology, evolution and transmission of foodborne Vibrios in Malaysia and elsewhere.

Keywords: *Vibrio parahaemolyticus*, Genomics, Malaysia, Seafood, Comparative genomics

Background

Vibrio parahaemolyticus inhabits the estuarine, marine and brackish water ecosystems. It is an important human pathogen associated with gastroenteritis linked to contaminated seafood consumption. Since this species is abundant in marine products, it has become a significant concern in the production and trade of seafood worldwide [1]. In Southeast Asian countries, including Malaysia, virulent *V. parahaemolyticus* in raw seafood have been reported [2,3]. Numerous cases of *V. parahaemolyticus* infection were reported in North America, South East Asia and Japan including some places in East Asia [4-10] giving the illness a pandemic status affecting thousands of people. Thus, the prevalence of pathogenic Vibrios in seafood is of public health concern and is an open ended issue.

The pathogenic *V. parahaemolyticus* strains are differentiated from non-pathogenic ones by their ability to cause beta-haemolysis on Wagatsuma agar, an activity known as 'Kanagawa phenomenon'. This effect is mediated by the activity of thermostable direct hemolysin (TDH) encoded by the *tdh* genes [8]. A pandemic clone of *V. parahaemolyticus* can broadly be defined as the one that is positive for TDH and exhibits the Kanagawa phenomenon [10].

V. parahaemolyticus strains are classified based on the types and variants of their O antigen and flagellar antigen (K). There are 13 O-serogroups and 71 K antigens and various combinations of these give rise to a wide variety of serovars which have been recognized as the causative agents of the disease. A clone of serovar O3:K6 has recently emerged and was associated with outbreaks

* Correspondence: subhabhassu@um.edu.my; niyaz.ahmed@uohyd.ac.in
[1]Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia
[2]Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, University of Hyderabad, Hyderabad, India
Full list of author information is available at the end of the article

BioMed Central

© 2013 Tiruvayipati et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

in India and Japan [7]. Frequent recombination events that promote clonal diversification suggest a scenario whereby a subset of O3:K6 strains might continue to evolve [11]. Consequently, different groups of related O3:K6 clonal strains have now been globally disseminated in Asia, North and South America, Africa and Europe [7].

The genomes of *V. parahaemolyticus* strains are said to have undergone a number of recombination events that could have been the reason for serotype conversion from O3:K6 to O4:K68 [12]. Regions of recombination likely involve a genetic element larger than the gene clusters encoding O and K-antigens. More than 20 serovariants which include O3:K6, O4:K68, O1:K25, O6:K18 and O1:KUT [13,14] emerged from an original pandemic strain, O3:K6. The pandemic group of these bacteria has evolved through a number of deletions, substitutions and acquisitions of regions primarily corresponding to TDH or a TDH-related hemolysin (TRH). It is the presence of either of these two virulence factors that confer potential to cause gastroenteritis in human populations. The pandemic clone is said to have emerged from a pre-pandemic clone which was positive for TRH and negative for TDH genes and harbored a new sequence of toxR (GS-PCR). The intermediate clone is described as being negative for both TRH and TDH, but positive for GS-PCR.

It has been observed that *V. parahaemolyticus* contains two chromosomes; *V. parahaemolyticus* RIMD2210633 has 3.2 Mb and 1.8 Mb of genome sizes for chromosome1 and 2 respectively [15]. There are several *V. parahaemolyticus* genomes which have been sequenced and deposited in Genbank as whole genomes or shotgun submissions (WGS) and sequence read archives (SRA). The only fully annotated submissions entail *V. parahaemolyticus* RIMD2210633 and *V. parahaemolyticus* BB220P. The *V. parahaemolyticus* RIMD2210633 genome harbors a Type III secretion system as a central virulence factor which is found in most diarrhea-causing bacteria [15]. As mentioned above, many studies link to the evolutionary aspects of the present pandemic clone formed from a pre-pandemic clone with a drastic change in its gene content i.e., the evolution from a TDH negative/TRH positive to a TDH positive/TRH negative strain and the occurrence of several serovariants in the *V. parahaemolyticus* species. The present isolate (*V. parahaemolyticus* PCV08-7) has been recovered from seafood (prawn) in 2008 which were purchased from a wet market in Selangor, Malaysia.

The main purpose of this study was to analyze the PCV08-7 genome that originates from Malaysia, a large peninsular as well as archipelagic country having a thriving seafood business and that it experiences several food borne outbreaks each season. Unfortunately, there are no markers based on native genome(s) to guide detection of *V. parahaemolyticus* in wet market, in the aquaculture farms and from human excreta and blood. We hope that this genome sequence will be helpful in identifying markers relevant in diagnostic development and molecular epidemiology/transmission dynamics of this significant bacterium in Malaysia and elsewhere.

Methods

Source, isolation and culture of V. parahaemolyticus PCV08-7

The *V. parahaemolyticus* PCV08-7 (VPPCV08-7) isolate was identified and characterized by obtaining pure cultures on selective media followed by analysis through biochemical tests, Analytical Profile Index (API) tests and genetic confirmation by PCR. The bacterial culture was maintained by streak plate on a Thiosulfate-Citrate-Bile-Sucrose (Difco, France) agar plates. After incubation at 37°C for 21 – 24 hr, characteristic bacterial colonies appeared with blue-green colored boundaries. An isolated bacterial colony was cultured in Luria-Bertani (LB) broth with 2% Sodium Chloride (NaCl) and incubated overnight at 37°C for 16 – 18 hr. This bacterial culture was further maintained as glycerol stocks at −80°C in 20% glycerol. The genomic DNA was isolated from a pure, single colony. The bacterial identity was confirmed by sequence analysis of the 16S rRNA.

Genomic DNA isolation and Next-Generation Sequencing

The genomic DNA was isolated using Qiagen DNeasy Blood & Tissue kit (Qiagen, Germany) and the genome sequence was determined by Illumina genome analyzer at the Genotypic Technology Pvt. Ltd. Bengaluru, India (GA2x, pipeline version 1.6). The sequencing data comprised of 100 bp paired-end reads with an insert size corresponding to approximately 240 bp. The genome coverage obtained was approximately about 80X with per base quality of reads in a range of 25 – 40. A total of 3.8 million reads were generated. Bioinformatics analysis was carried out with the help of protocols, algorithms and scripts developed, customized and tested in Ahmed Labs.

Assembly and alignment

Various strategies were applied to resolve the difficulties in dealing with the two chromosomes to be assembled from the sequence reads. The following main approaches were adopted:

1. *Velvet* [16]: Contigs were generated using the sequence reads which consisted of information from both the chromosomes of the isolate PCV08-7. This was checked by manually comparing contigs against the NCBI database by BLAST to check the highest similarity hit. *V. parahaemolyticus* RIMD2210633 was found to be the closest match in each search. The contigs showed unique hits to chromosome 1

(CHR1) and chromosome 2 (CHR2) as well as few common hits at both the chromosomes. The strategy of using the contigs together representing a whole genome (i.e., CHR1 and CHR2 together) or using the contigs separately as CHR1 and CHR2 was found to be challenging for further analysis to assemble them separately into two chromosomal sequences.

2. *OSLAY* [17]: All the contigs were compared against both the chromosomes of the genome of RIMD2210633 individually and were then used to form supercontigs for both the chromosomes separately. This procedure was found to be problematic as the supercontig files generated from CHR1 and CHR2 (separately) revealed that the preliminary contigs mapped to sequences in both the supercontig files. This was perhaps due to the input file comprising assembled whole genome contigs used against CHR1 and CHR2. The second strategy under OSLAY was to attach CHR1 and CHR2 of the reference genome RIMD2210633 as follows: CHR1 and CHR2 were concatenated (as a 'whole genome stretch') and then further used as one full length single sequence. Using this whole genome stretch for BLAST analysis, supercontigs were generated using Velvet contigs and the BLAST results. This also eventually proved inefficient since the supercontigs contained some sequences with several 'N' representing a gap in this case and such supercontigs had to be sorted to their own positions on the genome.
3. *SSPACE* [18]: Scaffolding was performed on velvet assembled contigs. As explained above, scaffolds were obtained separately from both CHR1 and CHR2 as well as with the whole genome stretch. All the scaffolds were then BLAST analyzed against both CHR1 and CHR2 of the reference genome individually, as well as at the level of the whole genome stretch. The difficulty faced with scaffolding was similar to that of OSLAY. Hence, the option of separately identifying the scaffolds with respect to CHR1 and CHR2 and dealing with them separately remained a problem.
4. *Mauve* [19]: Velvet assembled contigs were used at this step and exported as sorted contigs by performing an alignment against the whole genome stretch. The results obtained as aligned sorted contigs were taken through a stand-alone BLAST protocol against the whole genome of RIMD2210633. Then the BLAST results were carefully checked for their positions corresponding to both CHR1 and CHR2. The contigs were carefully divided as belonging to CHR1 and CHR2 sequences of PCV08-7 draft genome. The issues faced here were limited to identifying and dealing with the sequences other than those present in the contigs, but which were common to both RIMD2210633 and PCV08-7 genomes. While working on the above strategies, BWA alignment [20] was performed using sequence reads against the whole genome stretch of VPRIMD2210633. Using SAMTOOLS [21] a *.sam* file was generated with which the whole genome of RIMD2210633/FASTA sequence was loaded on Tablet viewer [22] to manually inspect the presence of common genes and to position the draft genome of PCV08-7.

The sequencing reads obtained by us were primarily passed through a quality control step using FASTX toolkit [23] to obtain high quality reads free from adaptor and primer contamination which was further standardized to an optimal parameter p value of 70. High quality reads thus obtained were assembled de-novo [22,23] using the Velvet assembly tool which produced 83 contigs with a hash length optimized to 71. These contigs were used to run OSLAY to form supercontigs with the reference genome RIMD2210633. Alignment of the reads against the reference genome was performed using BWA. The pre-assembled reads were also formed into scaffolds using SSPACE. Perl scripts written in house and modified after Baddam *et al.* [24] were used to re-order the contigs, supercontigs and scaffolds into their individual files. These approaches were put together to finalize the draft genome of *V. parahaemolyticus* PCV08-7 (Figure 1).

Results and discussion

Genome assembly

The 100 bp paired end reads were assembled using Velvet assembly tool that effectively utilized approximately 3.7 million reads. The N50 value observed was 261989 bp. The contig with the maximum length was 704232 bp and the total number of bases in the genome were 5184164 bp. The genome was artificially closed.

The genomes with multiple chromosomes pose technical difficulties during assembly. It is a known fact that Vibrios – *V. cholerae, V. parahaemolyticus* and *V. vulnificus* contain two circular chromosomes [26]. The reference genome used in this study, *V. parahaemolyticus* RIMD2210633 also consists of two chromosomes [13]. As studied previously [13], the origin of replication in chromosome 1 with the presence of *dnaA* gene shows its similarity to many genomes of prokaryotic origin and the origin of replication of chromosome 2 shows homology with that present on *V. cholerae* chromosome 2. The identification of distinct replication sites is of utmost importance for assembling bacterial genomes with two chromosomes which in the case of *V. cholerae* have been studied earlier [27]. Previous studies explain need for a more accurate procedure to handle data to correctly assemble two chromosomes and assign gene

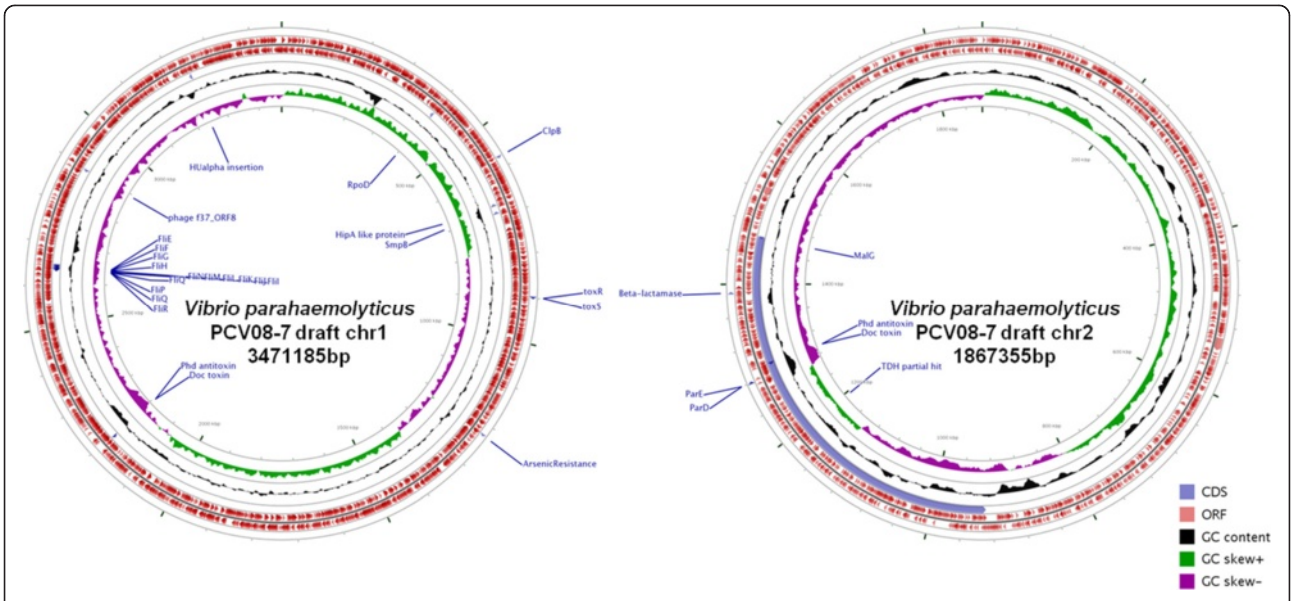


Figure 1 Circular view of *Vibrio parahaemolyticus* PCV08-7 draft genome. Diagrammatic representation of major genes carried by the two chromosomes of *Vibrio parahaemolyticus* PCV08-7 genome using CGview [25].

locations. The reads were assembled into a total of 83 contigs which were separated based on the assemble strategy as explained in the materials and methods section. Dealing with the present data, we observed that many of the genes of significant virulence or fitness importance were located on the chromosomes rather than showing any significant homology to the *Vibrio* plasmids. The presence of the Phd-Doc toxin antitoxin gene in our genome makes it interesting as the antitoxin gene has been previously reported related to plasmids [28] while a recent study [29] described

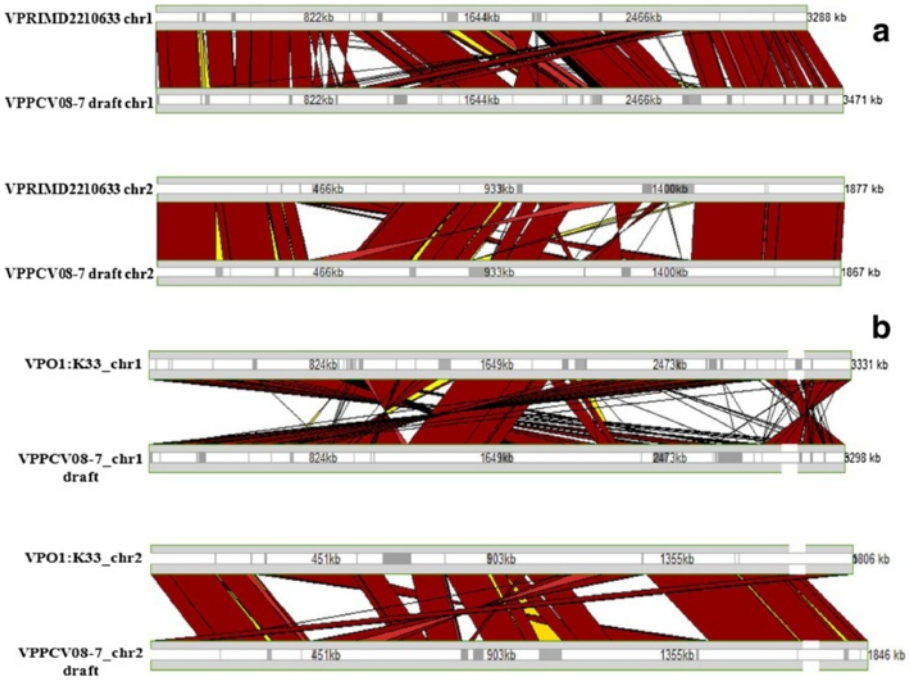


Figure 2 Alignment of the genome of strain RIMD2210633 against that of isolate PCV08-7 and strain O1:K33. (a) Comparison of chromosomes of strain RIMD2210633 (VPRIMD2210633 chr1, VPRIMD2210633 chr2) with the draft chromosomes of PCV08-7 (VPPCV08-7 draft chr1, VPPCV08-7 draft chr2) using M-GCAT. **(b)** Comparison of chromosomes of strain O1:K33 (VPO1:K33 chr1, VPO1:K33 chr2) with the draft chromosomes of PCV08-7 (VPPCV08-7 draft chr1, VPPCV08-7 draft chr2) using M-GCAT.

its occurrence on the chromosome of *Vibrio* species. However, we agree that the exact source of these genes can be mapped only when the plasmids will be sequenced and or analyzed separately.

Genome statistics and annotation

The draft assembled genome was annotated using the RAST server [30]. Statistics of the *V. parahaemolyticus* PCV08-7 draft genome were derived using Artemis [31], RNAmmer [32] and tRNAscanSE [33]: the sizes of chromosome 1 and chromosome 2 of the isolate were 3471185 bp and 1867355 bp respectively with G + C content of 45.35%. The tRNA and rRNA genes were 102 and 31 for chromosome 1, and 13 and 3 for chromosome 2, respectively. The chromosome 1 revealed a coding percentage of 85 with an average gene length of 943 bp while the chromosome 2 had a coding percentage of 86.2 with an average gene length of 950 bp.

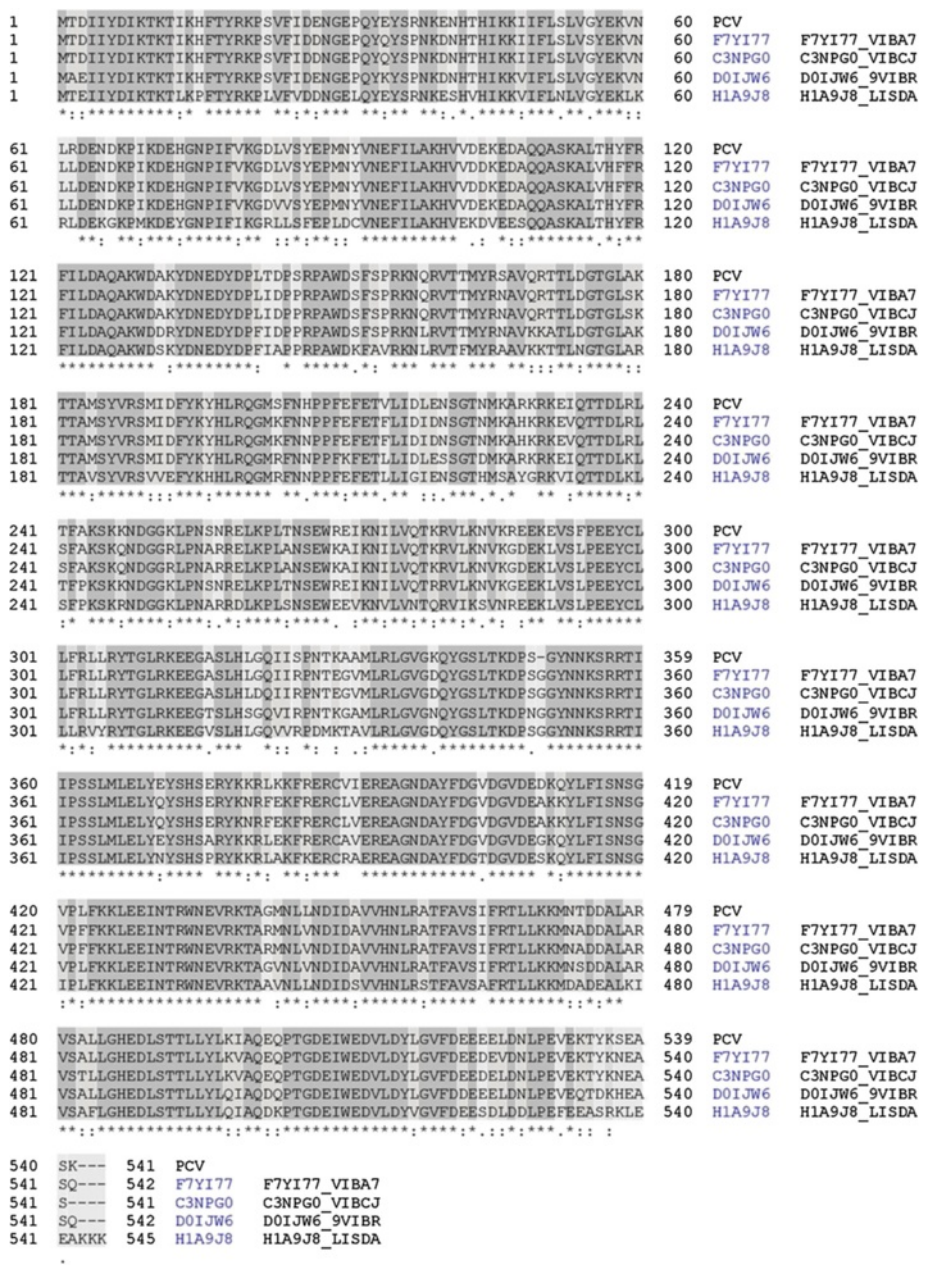


Figure 3 Alignment of a unique PCV08-7 protein sequence similar to *Photobacterium damselae subsps. damselae*. A unique sequence from PCV08-7 genome showed similarity with putative uncharacterized proteins of *V. anguillarum* 775 (F7YI77), *V. cholera* MJ-1236 (C3NPG0) and *Vibrio sp.* RC586 (D0IJW6) and similarity to a phage integrase of *Photobacterium damselae subsps. damselae* (H1A9J8).

The alignment of *V. parahaemolyticus* PCV08-7 genome with that of the *V. parahaemolyticus* RIMD2210633 genome using M-GCAT [34] showed visible rearrangements in the sequences of the two chromosomes of PCV08-7 isolate (Figure 2). The chromosome 1 of the draft genome carried phage shock proteins A, B and C, and bacteriophage f237 ORF8. It contained an integrated *tmRNA* gene with the closest element encoding the ribonuclease H. A site-specific recombinase *IntI4* and a gene encoding beta-lactamase were present. The draft genome also revealed genes responsible for fatty acid and amino acid metabolism. An important outer membrane protein *OmpU* was also identified. Genes coding for gyrase B (*gyrB*), HU-alpha insertion and putative sigma factors such as *rpoD, rpoE, rpoS, rpoN and rpoH* were also found in our analysis. The chromosome 2 carried a TDH pathogenicity island with many deletions and substitutions and displayed a *malG* gene on one of the flanking regions of the pathogenicity island. This region also contained genes coding for nutrient uptake and metabolism. We documented the presence of vibrio ferrin receptor *pvuA* and ferrichrome ABC transport *pvuB*, *pvuC*, *pvuD* and *pvuE* encoding genes, and the related *pvsA*, *pvsB*, *pvsC*, *pvsD* and *pvsE* genes. The analysis of the genome further revealed presence of a cobalt-zinc-cadmium resistance protein and a Rhodanese related sulfur transferase (as also present in RIMD2210633 genome) and a lead-cadmium-zinc-mercury transporting ATPase enzyme (as seen in the *V. parahaemolyticus* BB220P genome). Phd antitoxin and Doc toxin [28] which fall under the programmed cell death systems were also uniquely identified. Studies in *E. coli* have shown the presence of a stress related protein *clpB* along with *rpoS* and a few other genes [35] which help cope with stress conditions and help in survival. Our analysis detected the presence of *clpB, rpoS* and *hipA* genes in the present genome as was also seen in the reference genome of RIMD2210633. There were two types of Type III secretion systems observed in *V. parahaemolyticus* RIMD2210633 [36]; T3SS1 and T3SS2. Our genome analysis remains open ended with respect to the presence of such type III secretion systems.

Identification of novel gene content and comparative analysis

Our genome analysis revealed some unique sequences which have good similarity to hypothetical proteins of other Vibrio species such as *Vibrio anguillarum* and *Vibrio cholerae*. A 6315 bp nucleotide sequence showed identity to a *V. anguillarum* hypothetical protein and a *V. cholera* hypothetical protein on NCBI-BLASTN. One of the coding proteins in this stretch revealed similarity to the annotated phage integrase encoding gene of *Photobacterium damselae subsp. damselae* plasmid pAQU1 DNA (Figure 3). A parD gene (antitoxin to parE) was also found which showed closer identity to other *Vibrio* species such as *Vibrio vulnificus*, *Vibrio mimicus* and *Vibrio orientalis*. parD when aligned against *V. vulnificus* and *V. mimicus* revealed an identity of 76 bp out of 80 bp (95%) (e-value 2e-48) and with *V. orientalis* an identity of 72 bp out of 80 bp (90%) (e-value 2e-45) on NCBI-BLASTN. A few newer hypothetical proteins with no reported annotation were identified. The genome also contained a gene relevant to arsenic resistance, possibly important in the adaptation of the bacterium to a high arsenic environment. Our analysis of the genome revealed presence of a partially similar sequence of TDH Pathogenicity Island, as compared to *V. parahaemolyticus* RIMD2210633. This island revealed genetic instability due to various insertion/deletion and substitution events we documented. The presence of toxS and toxR genes was also observed.

The old pandemic O3:K6 strain of *V. parahaemolyticus* is said to have gained gene clusters VPaI1-VPaI7 [37] to develop into a new pandemic clone of which VPaI4-VPaI6 are said to be putative virulence factors and may be potential pathogenicity islands. These regions are said to carry along with them a type VI secretion system (VP1386-1420). Our PCV08-7 genome analysis revealed that only one cluster, VPaI2, was detected completely, whereas VPaI3 and VPaI7 were partially present (Table 1). This perhaps shows that our strain could be possibly a new serovariant of a non-pandemic O3:K6 strain like the *V. parahaemolyticus* AQ3810 [8]. While

Table 1 Table representing pathogenicity related clusters and other VP clusters in *V. parahaemolyticus* PCV08-7: (1) pathogenicity related clusters (VPaI1-VPaI7) in the genome of strain RIMD2210633 that signify it to be a pandemic O3:K6 strain and their presence or absence in the genome of PCV08-7 isolate, (2) various other VP clusters and their occurrence in the genome of PCV08-7

(1) *Vibrio parahaemolyticus* RIMD2210633	***V. parahaemolyticus* PCV08-7**
VPaI1 (VP0380-0403)	Absent
VpaI2 (VP0635-0643)	Present
VpaI3 (VP1071-1094)	Partially present
VpaI4 (VP2131-2144)	Absent
VpaI5 (VP2900-2910)	Absent
VpaI6 (VPA1254-1270)	Absent
VpaI7 (VP1312-1398)	Present
Type VI secretion system (VP1386-1420)	Absent
(2) Other VP clusters	***V. parahaemolyticus* PCV08-7**
VP1355-1368	Partially present
VPA0074-0089	Present
VPA0713-0732	Present
VPA1194-1210	Present

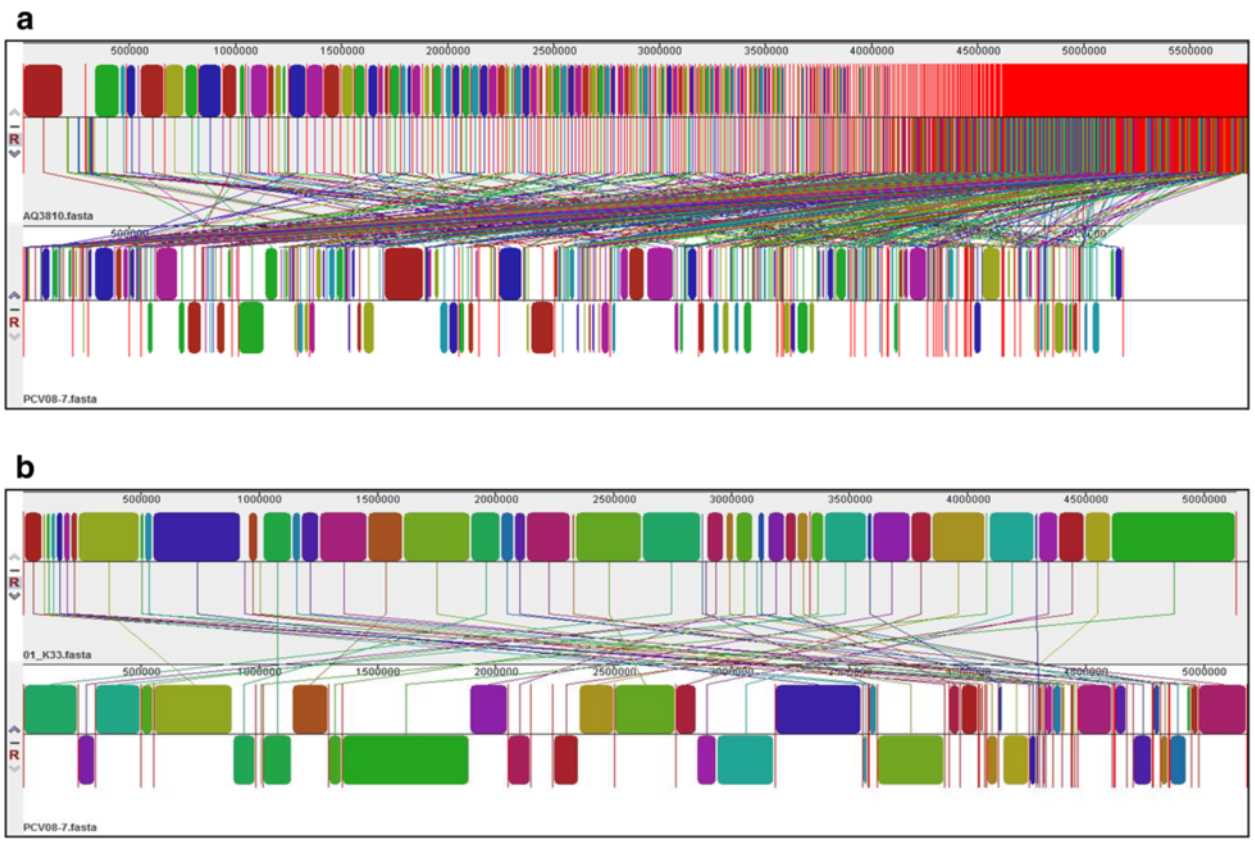


Figure 4 (a) ***V. parahaemolyticus*** **AQ3810 alignment against PCV08-7 genome: Concatenated chromosome 1 and chromosome 2 of** ***V. parahaemolyticus*** **AQ3810 (AQ3810.fasta) against** ***V. parahaemolyticus*** **PCV08-7 (PCV08-7.fasta) (b)** ***V. parahaemolyticus*** **O1:K33 alignment against PCV08-7.** Concatenated chromosome 1 and chromosome 2 of *V. parahaemolyticus* O1:K33 (O1_K33.fasta) against *V. parahaemolyticus* PCV08-7 (PCV08-7.fasta).

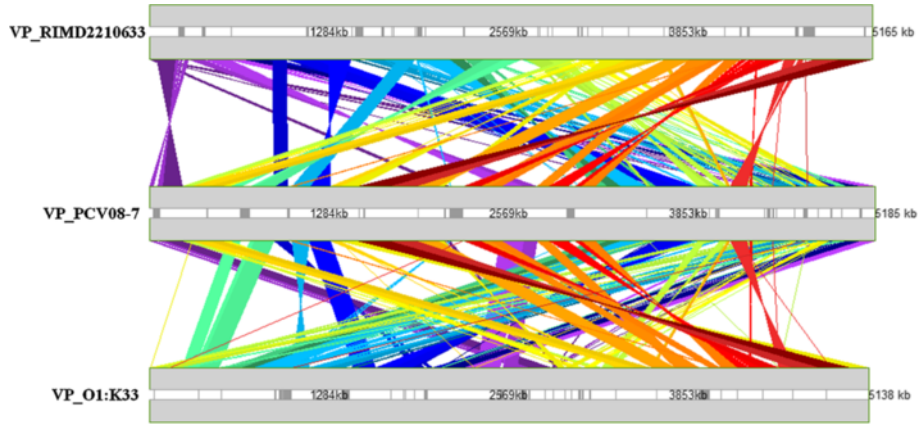


Figure 5 Comparison of whole genome sequences of strains RIMD2210633, PCV08-7 and O1:K33. Alignment of complete genomes of *V. parahaemolyticus* RIMD2210633, *V. parahaemolyticus* PCV08-7 and *V. parahaemolyticus* O1:K33, showing PCV08-7 being more similar to RIMD2210633.

variability of different gene clusters (Table 1) portrays a probably novel serovariant of *V. parahaemolyticus* with the presence of ribonuclease H encoding element (previously thought to be present only in *V. parahaemolyticus* RIMD2210633 and absent in *V. parahaemolyticus* AQ3810 [12]). A further comparative study between the *V. parahaemolyticus* PCV08-7 and the non-pandemic *V. parahaemolyticus* AQ3810 (O3:K6 strain) and the newest *V. parahaemolyticus* O1:K33 (trh+/ tdh + genotype) strain showed that *V. parahaemolyticus* PCV08-7 has more genetic relatedness towards a trh+/ tdh + strain (Figure 4). But, alignments of the *V. parahaemolyticus* PCV08-7 contig data against the *V. parahaemolyticus* O1:K33 and *V. parahaemolyticus* RIMD2210633 (Figure 2) strains show that it is closer to O3:K6 serotype (Figure 5).

From the above thesis, it becomes probably apparent that the genome of *V. parahaemolyticus* PCV08-7 meaningfully adds to the battery of important genomic sequences representing enteropathogenic bacteria. The genome of an arthropod derived, foodborne *Vibrio* should be important to understand adaptation to a crustacean host and a human host.

Epilogue and future directions

A first account of the genome of *V. parahaemolyticus* PCV08-7 has been presented. The draft genome and its annotation as described would be able to explain the lifestyle of pathogenic *Vibrio* species. The experience of assembling this genome and the difficulties associated with separating the data with respect to two chromosomes would certainly be helpful to the community in the follow-up studies. Further, a host of new molecular markers as gleaned by our analysis would be relevant in the diagnostic development and molecular epidemiology. The present genome and the ensuing comparative genomics would be able to rekindle our thoughts on the survival and virulence as well as transmission potentials of *V. parahaemolyticus* and also on their adaptation to different hosts and the niches thereof. Our results clearly reveal a significantly novel gene content which could presumably have been acquired through a horizontal gene transfer mechanism. Our analysis revealed the presence of not only the conserved genomic regions among different *V. parahaemolyticus* bacteria, but also dissects some of the unique sets of genes that hold relevance to virulence. We propose to finish and polish the genome in the near future also with the help of further coverage using alternative sequencing platforms and by employing a hybrid assembly approach. Also, it will be possible to determine the true extent of the diversity of *V. parahaemolyticus* strains obtained from seafood as compared to those isolated from human cases. Such a diversity analysis would focus on 1) genomic coordinates relevant to colonization of and adaptation to different hosts in different ecosystems; 2) genome dynamics relative to bacterial fitness shaping over time and with transmission across different hosts; and 3) profile of genomic rearrangements including additive and reductive genome evolution and their significance in the evolution of pathogenic *Vibrio* species. Presently, the epidemiology of *V. parahaemolyticus* infection in resource-poor countries largely entails a classical serology concocted with guess work as to the type of strain involved and its source. Our genomic data would hopefully contribute to this situation also.

Availability of supporting data

The *Vibrio parahaemolyticus* PCV08-7 whole genome shotgun project was deposited in Genbank under the accession AOCL00000000. The version described in this paper is the first version, AOCL01000000. This consists of sequences from AOCL01000000 – AOCL01000083 (http://www.ncbi.nlm.nih.gov/nuccore/AOCL00000000).

Competing interests

NA and TKL are the editors of Gut Pathogens.

Authors' contributions

NA and SB: Designed and supervised the study and written and edited the manuscript, TS: performed genomic DNA preparation, sequencing analysis, annotation and comparative genomics, AKG: performed initial bioinformatics analysis, RB and SS: provided tools and IT support for the study, NK: contributed to quality control of the NGS data and assembly. TKL: isolated and maintained the strain and provided inputs on lifestyle and evolution of the organism. All authors read and approved the final manuscript.

Acknowledgements

TS was supported by a doctoral fellowship from University of Malaya under the Bright Sparks program (BSP 226(3)-12). SB would like to thank University of Malaya for the support from the PPP grant PG088-2012B. SB and KLT would like to acknowledge research support received from University of Malaya under different funding instruments. NA would like to acknowledge partial support from the UM-HIR project of the University of Malaya. NA is an Academy Professor (Adjunct) of the Academy of Scientific and Innovative Research, India and visiting Professor at the Institute of Biological Sciences, University of Malaya, Kuala Lumpur.

Author details

[1]Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia. [2]Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad, Hyderabad, India. [3]Center of Biotechnology for Agriculture (CEBAR), University of Malaya, Kuala Lumpur, Malaysia. [4]Laboratory of Biomedical Science and Molecular Microbiology, UMBIO Research Cluster, University of Malaya, Kuala Lumpur, Malaysia.

Received: 3 October 2013 Accepted: 29 November 2013
Published: 11 December 2013

References

1. DePaola A, Nordstrom JL, Dalsgaard A, Forslund A, Oliver J, Bates T, Bourdage KL, Gulig PA: **Analysis of Vibrio vulnificus from market oysters and septicemia cases for virulence markers.** *Appl Environ Microbiol* 2003, **69**(7):4006–4011.
2. Sujeewa AKW, Norrakiah AS, Laina M: **Prevalence of toxic genes of Vibrio parahaemolyticus in shrimps (Penaeus monodon) and culture environment.** *International Food Research Journal* 2009, **16**:89–95.

3. Paydar M, Teh CS, Thong KL: **Prevalence and characterisation of potentially virulent Vibrio parahaemolyticus in seafood in Malaysia using conventional methods, PCR and REP-PCR.** *Food Control* 2013, **32:**13–18.
4. **Guidelines for national human immunodeficiency virus case surveillance, including monitoring for human immunodeficiency virus infection and acquired immunodeficiency syndrome. Centers for Disease Control and Prevention.** *MMWR Recomm Rep: Morbidity and mortality weekly report Recommendations and reports/Centers for Disease Control* 1999, **48**(RR-13):1–27. 29–31.
5. Bag PK, Nandi S, Bhadra RK, Ramamurthy T, Bhattacharya SK, Nishibuchi M, Hamabata T, Yamasaki S, Takeda Y, Nair GB: **Clonal diversity among recently emerged strains of Vibrio parahaemolyticus O3:K6 associated with pandemic spread.** *J Clin Microbiol* 1999, **37**(7):2354–2357.
6. Nair GB, Hormazabal JC: **The Vibrio parahaemolyticus pandemic.** *Revista chilena de infectologia: organo oficial de la Sociedad Chilena de Infectologia* 2005, **22**(2):125–130.
7. Nair GB, Ramamurthy T, Bhattacharya SK, Dutta B, Takeda Y, Sack DA: **Global dissemination of Vibrio parahaemolyticus serotype O3:K6 and its serovariants.** *Clin Microbiol Rev* 2007, **20**(1):39–48.
8. Nishibuchi M, Kaper JB: **Thermostable direct hemolysin gene of Vibrio parahaemolyticus: a virulence gene acquired by a marine bacterium.** *Infect Immun* 1995, **63**(6):2093–2099.
9. Okuda J, Ishibashi M, Hayakawa E, Nishino T, Takeda Y, Mukhopadhyay AK, Garg S, Bhattacharya SK, Nair GB, Nishibuchi M: **Emergence of a unique O3: K6 clone of Vibrio parahaemolyticus in Calcutta, India, and isolation of strains from the same clonal group from Southeast Asian travelers arriving in Japan.** *J Clin Microbiol* 1997, **35**(12):3150–3155.
10. Han H, Wong HC, Kan B, Guo Z, Zeng X, Yin S, Liu X, Yang R, Zhou D: **Genome plasticity of Vibrio parahaemolyticus: microevolution of the 'pandemic group'.** *BMC genomics* 2008, **9:**570.
11. Gonzalez-Escalona N, Martinez-Urtaza J, Romero J, Espejo RT, Jaykus LA, DePaola A: **Determination of molecular phylogenetics of Vibrio parahaemolyticus strains by multilocus sequence typing.** *J Bacteriol* 2008, **190**(8):2831–2840.
12. Chen Y, Stine OC, Badger JH, Gil AI, Nair GB, Nishibuchi M, Fouts DE: **Comparative genomic analysis of Vibrio parahaemolyticus: serotype conversion and virulence.** *BMC Genomics* 2011, **12:**294.
13. Chowdhury NR, Chakraborty S, Ramamurthy T, Nishibuchi M, Yamasaki S, Takeda Y, Nair GB: **Molecular evidence of clonal Vibrio parahaemolyticus pandemic strains.** *Emerg Infect Dis* 2000, **6**(6):631–636.
14. Chowdhury NR, Stine OC, Morris JG, Nair GB: **Assessment of evolution of pandemic Vibrio parahaemolyticus by multilocus sequence typing.** *J Clin Microbiol* 2004, **42**(3):1280–1282.
15. Makino K, Oshima K, Kurokawa K, Yokoyama K, Uda T, Tagomori K, Iijima Y, Najima M, Nakano M, Yamashita A, *et al*: **Genome sequence of Vibrio parahaemolyticus: a pathogenic mechanism distinct from that of V cholerae.** *Lancet* 2003, **361**(9359):743–749.
16. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18**(5):821–829.
17. Richter DC, Schuster SC, Huson DH: **OSLay: optimal syntenic layout of unfinished assemblies.** *Bioinformatics* 2007, **23**(13):1573–1579.
18. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W: **Scaffolding pre-assembled contigs using SSPACE.** *Bioinformatics* 2011, **27**(4):578–579.
19. Darling AE, Mau B, Perna NT: **ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement.** *PLoS One* 2010, **5**(6):e11147.
20. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2010, **26**(5):589–595.
21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **Genome project data processing S: the sequence alignment/ Map format and SAMtools.** *Bioinformatics* 2009, **25**(16):2078–2079.
22. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D: **Tablet–next generation sequence assembly visualization.** *Bioinformatics* 2010, **26**(3):401–402.
23. Taylor J, Schenck I, Blankenberg D, Nekrutenko A: **Using galaxy to perform large-scale interactive data analyses.** *Curr Protoc Bioinformatics* 2007, **10:**10.5.
24. Baddam R, Kumar N, Shaik S, Suma T, Ngoi ST, Thong KL, Ahmed N: **Genome sequencing and analysis of Salmonella enterica serovar Typhi strain CR0063 representing a carrier individual during an outbreak of typhoid fever in Kelantan, Malaysia.** *Gut Pathogens* 2012, **4**(1):20.
25. Stothard P, Wishart DS: **Circular genome visualization and exploration using CGView.** *Bioinformatics* 2005, **21:**537–539.
26. Yamaichi Y, Iida T, Park KS, Yamamoto K, Honda T: **Physical and genetic map of the genome of Vibrio parahaemolyticus: presence of two chromosomes in Vibrio species.** *Mol Microbiol* 1999, **31**(5):1513–1521.
27. Egan ES, Waldor MK: **Distinct replication requirements for the two Vibrio cholerae chromosomes.** *Cell* 2003, **114**(4):521–530.
28. McKinley JE, Magnuson RD: **Characterization of the Phd repressor-antitoxin boundary.** *J Bacteriol* 2005, **187**(2):765–770.
29. Guerout AM, Iqbal N, Mine N, Ducos-Galand M, Van Melderen L, Mazel D: **Characterization of the phd-doc and ccd Toxin-Antitoxin Cassettes from Vibrio Superintegrons.** *J Bacteriol* 2013, **195**(10):2270–2283.
30. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, *et al*: **The RAST Server: rapid annotations using subsystems technology.** *BMC Genomics* 2008, **9:**75.
31. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16**(10):944–945.
32. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW: **RNAmmer: consistent and rapid annotation of ribosomal RNA genes.** *Nucleic Acids Res* 2007, **35**(9):3100–3108.
33. Schattner P, Brooks AN, Lowe TM: **The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs.** *Nucleic Acids Res* 2005, **33**(Web Server issue):W686–W689.
34. Treangen TJ, Messeguer X: **M-GCAT: interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species.** *BMC Bioinformatics* 2006, **7:**433.
35. Wang X, Wood TK: **Toxin-antitoxin systems influence biofilm and persister cell formation and the general stress response.** *Appl Environ Microbiol* 2011, **77**(16):5577–5583.
36. Park KS, Ono T, Rokuda M, Jang MH, Okada K, Iida T, Honda T: **Functional characterization of two type III secretion systems of Vibrio parahaemolyticus.** *Infect Immun* 2004, **72**(11):6659–6665.
37. Hurley CC, Quirke A, Reen FJ, Boyd EF: **Four genomic islands that mark post-1995 pandemic Vibrio parahaemolyticus isolates.** *BMC Genomics* 2006, **7:**104.

doi:10.1186/1757-4749-5-37
Cite this article as: Tiruvayipati *et al.*: Genome anatomy of the gastrointestinal pathogen, Vibrio parahaemolyticus of crustacean origin. *Gut Pathogens* 2013 **5**:37.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

Gut Pathogens

RESEARCH Open Access

Genome sequencing and analysis of *Salmonella enterica* serovar Typhi strain CR0063 representing a carrier individual during an outbreak of typhoid fever in Kelantan, Malaysia

Ramani Baddam[1], Narender Kumar[1], Sabiha Shaik[1], Tiruvayipati Suma[1,2], Soo Tein Ngoi[2,3], Kwai-Lin Thong[2,3] and Niyaz Ahmed[1,2*]

Abstract

Salmonella Typhi is a human restricted pathogen with a significant number of individuals as asymptomatic carriers of the bacterium. *Salmonella* infection can be effectively controlled if a reliable method for identification of these carriers is developed. In this context, the availability of whole genomes of carrier strains through high- throughput sequencing and further downstream analysis by comparative genomics approaches is very promising. Herein we describe the genome sequence of a *Salmonella* Typhi isolate representing an asymptomatic carrier individual during a prolonged outbreak of typhoid fever in Kelantan, Malaysia. Putative genomic coordinates relevant in pathogenesis and persistence of this carrier strain are identified and discussed.

Background

Salmonella enterica serovar Typhi, the aetiologic agent of typhoid fever is still posing a major health problem for the developing world, as about 16 million new cases are reported each year [1]. *S.* Typhi causes systemic infections (typhoid fever) as well as chronic infections (asymptomatic carriers) in humans, the latter serve as the source of infection [2]. The transmission of *S.* Typhi is primarily through faecal-oral route and a significant number of infected individuals become chronic asymptomatic carriers and keep shedding *S.* Typhi in faeces for decades [3]. This results in endemicity of *S.* Typhi in regions of the world with underdeveloped sanitation and community hygiene [4].

Carrier identification becomes extremely important as some of the ancestral haplotypes were observed in recent isolates suggesting their persistence in these asymptomatic carriers [5]. Traditional methods such as culturing of bacteria from faecal samples are not fool proof as the carriers shed bacteria intermittently. Serological tests to detect specific antibodies such as anti-H and anti-O are unable to differentiate between carriers and individuals who have recovered from the infection [6]. Especially, in areas endemic for *S.* Typhi, due to high background levels of these antibodies, serological tests cannot be adopted for the identification of a carrier [7]. Thus, there is an urgent need for inexpensive and efficient detection methods for the establishment of carrier state, perhaps based on genomic markers.

The genetic typing tools such as PFGE, AFLP, ribotyping etc. can resolve limited genetic variation occurring within specific sites, and therefore are incapable of differentiating highly clonal strains such as outbreak related strains from the ones not associated with the outbreak (carrier isolates) [8-10]. High-throughput sequencing technologies have already been employed as a high resolution molecular epidemiologic tool to discern microevolution of highly related strains [11].

In this study, we attempted to determine if whole genome sequencing of *S.* Typhi isolated from a carrier individual can provide insights related to persistence and or adaptation mechanisms. We describe the genome

* Correspondence: ahmed.nizi@gmail.com
[1]Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad, Hyderabad, India
[2]Institute of Biological sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia
Full list of author information is available at the end of the article

BioMed Central

© 2012 Baddam et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

sequence of a *Salmonella enterica* serovar Typhi strain (ST CR0063) isolated from a carrier individual during a prolonged outbreak of typhoid fever in Kelantan, Malaysia.

Results and discussion

Genome statistics

The size of the draft genome of *Salmonella* Typhi (ST CR0063) is 4,585,851 bp with a coding percentage of 86.1%. The G + C content of this strain is about 51.71%. The total number of CDS determined are 4946 with an average length of gene about 798 nucleotides. The genome of ST CR0063 revealed 77 tRNA and 22 rRNA genes. The subsystems distribution of basic metabolic machinery of this strain is represented in Figure 1. The assembled draft genome shows high degree of similarity and shared core genome regions with *Salmonella* Typhi ST BL196 [12], the one identified as associated with a typhoid outbreak in Kelantan during the same period (Figure 2).

Virulence factors

The gene *shdA*, a key factor predicted to be involved in persistence of the bacterium in the intestines [14] by binding to its extracellular matrix, was identified and annotated. This gene, by mimicking the host heparin, is able to bind to the extracellular matrix proteins, fibronectin and collagen, and probably plays an important role in carriers by contributing to prolonged faecal shedding [15]. The *fim* gene cluster [16] of chaperone –usher family involved in adhesion to non-phagocytic cells was detected along with its negative regulator fimW. Type IV pili and *agf* operon [17,18] encoding curli fimbriae which aid in attachment of the bacterium to intestinal villi and also with each other, were found in the genome. These adherence factors determine the sites of bacterial colonisation and thereby adaptation and pathogenicity of a particular strain [19,20].

The *S.* Typhi strain ST CR0063 genome also revealed *viaA* and *viaB* loci, the prime regulators of Vi antigen expression. The *viaB* locus contains all genes for the biosynthesis (*tviA-E*) and export (*vexA-E*) of the Vi antigen, a well-known virulence factor [21,22]. The *mgtC* gene involved in Magnesium uptake and ferric uptake regulators (fur) [23] were also identified in ST CR0063. The PhoPQ regulon [24], which induces cytokine secretion and cationic antimicrobial peptide resistance, was also found to be conserved in our carrier strain. The RpoS sigma factor needed to cope up with external stress and nutrient depletion conditions [25] was also identified and annotated. The co-ordinates of these virulence factors in the genome of ST CR0063 are depicted in Figure 3.

Phages and pathogenicity islands (PAIs)

The phages gifsy-1 and fels-2 [27] together with many phage proteins and a few hypothetical proteins were identified in the genome of ST CR0063 by various algorithms (See Methods for details). It is expected that these phages are acquired by horizontal gene transfer (HGT) events as they were embedded in some of the genomic islands recognized. The phage encoding SopE effector protein of SPI-1 (Salmonella Pathogenicity Island) was present in ST CR0063 as recognized in other Typhi genomes [28,29].

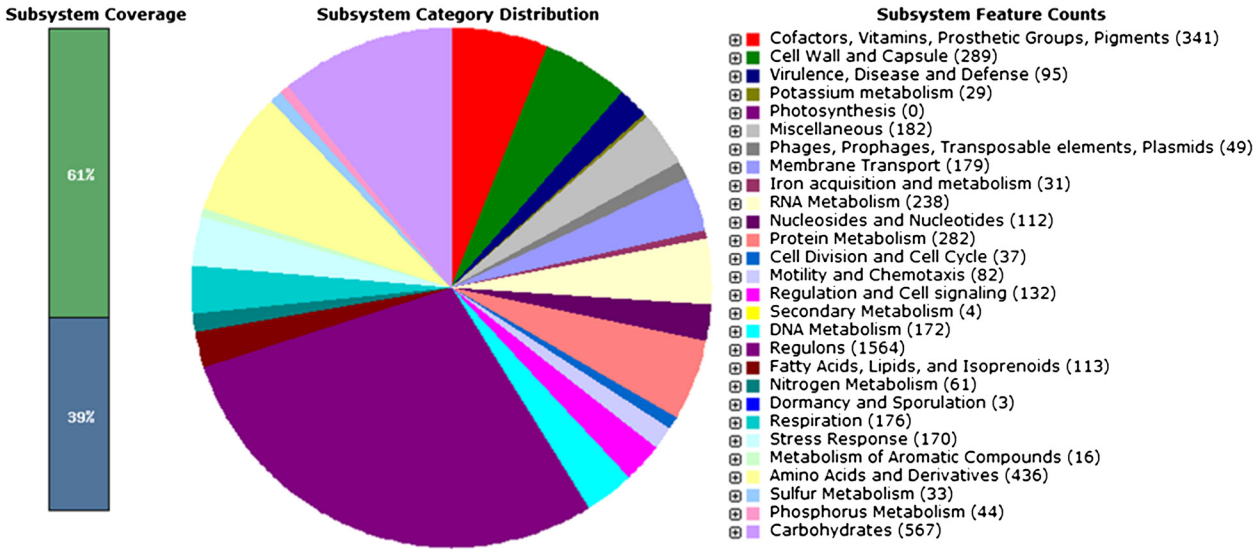


Figure 1 Subsystem distribution of ST CR0063. The subsystem statistics of ST CR0063 based on genome annotations performed according to RAST conventions.

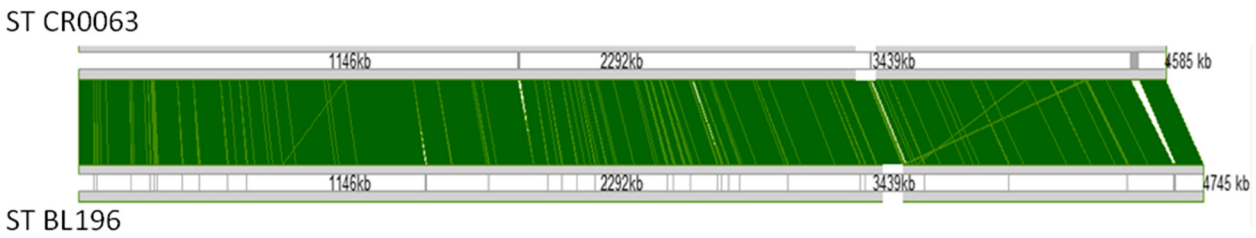


Figure 2 Comparison of *Salmonella* Typhi strains ST CR0063 and ST BL196. Comparison of whole genome sequences of *S. Typhi* strains using MG-CAT – one strain was isolated from a carrier individual (ST CR0063) and another from an infected individual (ST BL196) during a prolonged outbreak of Typhoid fever in Kelantan [13].

More than 15 PAIs that encode clusters of virulence associated genes have been identified across various serovars of *Salmonella enterica*. Ten pathogenicity islands have been identified by us in ST CR0063 and as expected [30], they were characterised by different G + C content and bounded by t-RNA genes. The SPI-1 type III secretion system (TTSS) structural genes spaM-NOPQRS and invABCEFGH and their regulatory proteins HilA, HilC, HilD [31] were also identified and annotated. The SPI-1 secreted effector proteins SopE,

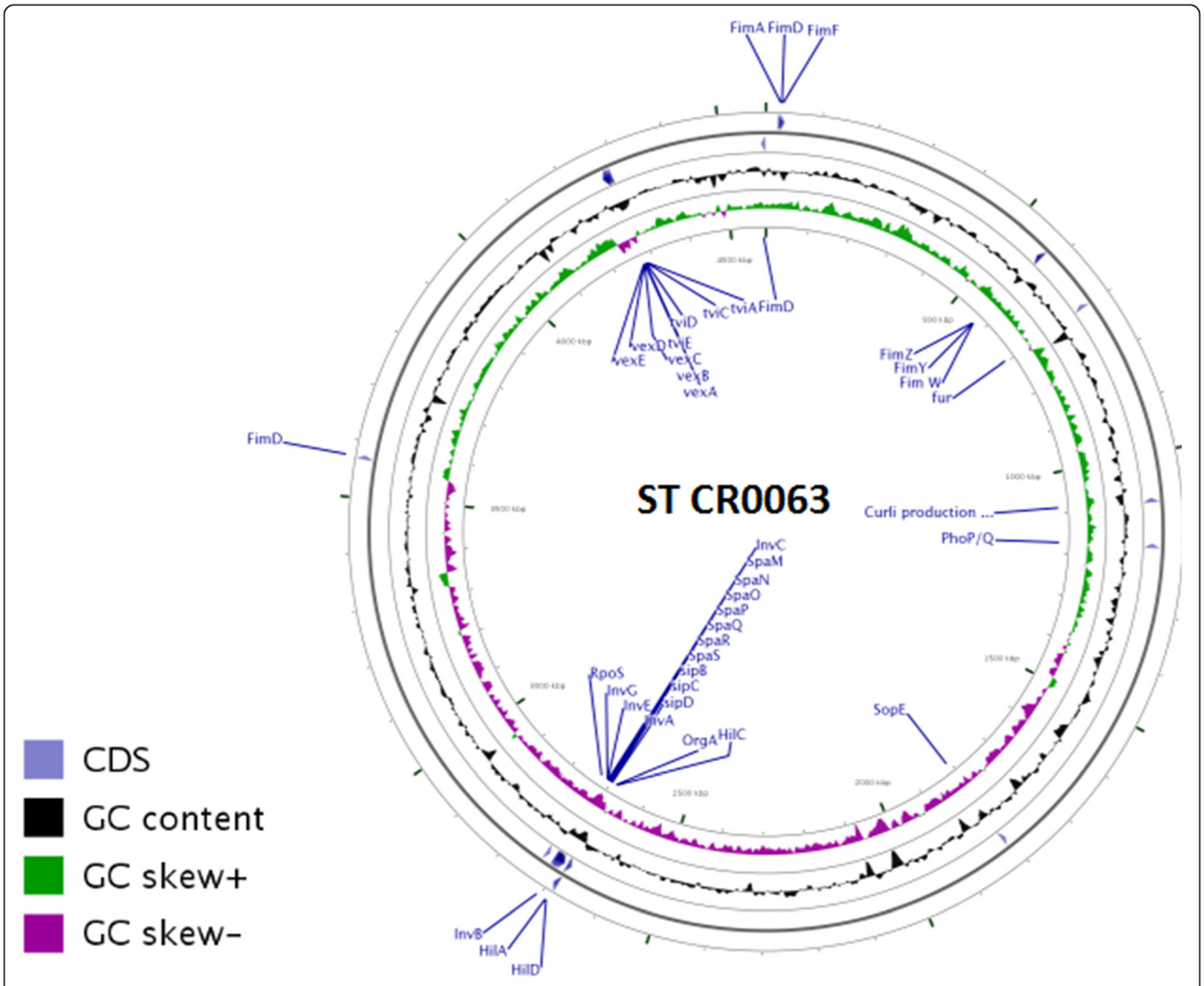


Figure 3 Circular Genome view of ST CR063. Positions of some of the major virulence factors and their regulators identified in ST CR0063 marked in the circular genome generated using CGview [26].

SopE2, SipA, SipB, SipC and SptP required for endothelial uptake and invasion [32] are also present. The genes SpiC, SseF, SseG, SifA, SifB secreted by SPI-2 TTSS and that are needed for survival in macrophages and colonisation of host organs [33] were also recognised in the present genome. The known regulators of SPI-2, OmpR-EnvZ and PhoP-PhoQ [34] were present. SPI-3, identified by us, contained magnesium transport genes *mgtC* and *marT* which are required for survival in macrophages [35]. Type I secretion system and its associated proteins encoded by SPI-4, and that are involved in the invasion of the intestinal epithelium [36], were also located in the present genome. The SPI-1 effector proteins SopB and PipB associated with enteritis and coded by SPI-5 [37] were also detected and annotated. The chaperone-usher fimbrial operons carried by SPI-6, SPI-10 and bacteriocin immunity proteins carried by SPI-8 [38] were identified. The SPI-7 and SPI-9 were identified in the ST CR0063 genome and were found to encode *viaB* locus, type IV pili formation proteins and TISS [38,39].

Conclusions and prospective

The genomic blueprint of *Salmonella* Typhi isolate ST CR0063 was elucidated in this study. The genome sequence information presented herein may be harnessed to guide comparative genomics and identification of novel and specific diagnostic markers. However, further studies involving large scale genome sequencing of the strains from several of the endemic countries and especially those from carrier individuals of different socio-economical settings is needed to develop a reliable approach to decipher the characteristics of a carrier state. Also, it will be required to determine the true extent of the diversity of carrier strains as juxtaposed to their acutely pathogenic forms in terms of 1) gene gain/loss during colonization and adaptation; 2) dynamics of virulence acquisition/attenuation; 3) possible genomic rearrangements; and 4) the relative preponderance of carrier and virulent strains circulating in different endemic regions of the world. Finally, an in-depth analysis of the host-pathogen interactions and their influence on gut microbiota can only explain the adaptation and persistence mechanisms of the (asymptomatic) carrier strains.

Methods

Genome sequencing

DNA was isolated from the stool sample of an asymptomatic carrier individual from Kelantan, Malaysia in 2007 during a prolonged outbreak. The draft genome sequence of this strain (STCR0063) was determined by Illumina Genome Analyzer (GAIIx, pipe- line ver l.6). The 100 bp paired-end sequencing was done with an insert size of 300 bp. About 67X genome coverage was achieved and 1.9 gigabytes of data were obtained.

Assembly and annotation

The sequence data were assembled *de-novo* in the same way as described previously [40-45] into 538 contigs using Velvet [46] at optimal hash length 39. SSPACE [47] was used for scaffolding the pre-assembled contigs using paired-end data. The gaps within these scaffolds were filled using Gapfiller by aligning the reads against already generated Scaffolds by SSPACE [48].

A reference guided assembly was generated by aligning reads to *Salmonella* Typhi str. CT18 [GenBank: AL513382.1] using bwa tools [49]. This reference guided assembly was used to re-order the scaffolds generated in de-novo way. In-house written Perl scripts were used for this re-ordering process and to finalize the gaps. The de novo and reference guided approaches were used to finalize the consensus draft genome. The reference guided assembly and reordered scaffolds were loaded on to Tablet – NGS data visualisation tool, to visualise the repeats, insertions and deletions [50].

The final draft nucleotide sequence after manual curation was annotated in our laboratory using RAST [51] and ISGA pipeline [52]. The genome statistics were gleaned using Artemis [53]. The data were further validated using gene prediction tools such as Glimmer [54] and EasyGene [55]. The RNAmmer [56] and tRNAscan-SE [57] were used to identify rRNA and tRNA respectively.

Phages and PAIs

Prophages and putative phage like elements in the genome were identified using PhiSpy [58] and Prophage Finder [59]. The putative HGT events were determined using Alien Hunter tool [60]. An integrated interface Island Viewer was used to predict putative genomic islands within the genome [61].

Sequence data access

The Salmonella enterica subsp. enterica serovar Typhi str. CR0063 whole genome shotgun (WGS) project has been submitted to the GenBank and has the project accession AKIC00000000. The project version entailing draft assembly described herein has the accession number AKIC01000000, and consists of sequences AKIC01000001-AKIC01000538.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

NA designed the study, interpreted the results and edited the manuscript. RB and NK managed Illumina sequencing, made the assemblies, analyzed the genome, and performed annotations. SS and TS provided computational tools and contributed to automation of the analysis process. KT provided

inputs related to the outbreak and the strain features, characterized the strain and maintained it in pure cultures. STN contributed to microbiology of the strain and prepared high molecular weight DNA for genome sequencing. All the authors read and approved the manuscript prior to submission.

Acknowledgements
We thankfully acknowledge support received from the University of Malaya High Impact Research Grant (Ref. UM.C/625/1HIR/MOHE/02 [A000002-5000 1]) - MOLECULAR GENETICS.

Author details
[1]Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad, Hyderabad, India. [2]Institute of Biological sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia. [3]Laboratory of Biomedical Science and Molecular Microbiology, UMBIO Research Cluster, University of Malaya, Kuala Lumpur, Malaysia.

Received: 14 November 2012 Accepted: 29 November 2012
Published: 13 December 2012

References

1. Parry CM, Hien TT, Dougan G, White NJ, Farrar JJ: **Typhoid fever.** *N Engl J Med* 2002, **347:**1770–1782.
2. Boyle EC, Bishop JL, Grassl GA, Finlay BB: **Salmonella: from pathogenesis to therapeutics.** *J Bacteriol* 2007, **189:**1489–1495.
3. Gonzalez-Escobedo G, Marshall JM, Gunn JS: **Chronic and acute infection of the gall bladder by Salmonella Typhi: understanding the carrier state.** *Nat Rev Microbiol* 2011, **9:**9–14.
4. Gopinath S, Carden S, Monack D: **Shedding light on Salmonella carriers.** *Trends Microbiol* 2012, **20:**320–327.
5. Roumagnac P, Weill FX, Dolecek C, Baker S, Brisse S, Chinh NT, Le TA, Acosta CJ, Farrar J, Dougan G, Achtman M: **Evolutionary history of Salmonella Typhi.** *Science* 2006, **314:**1301–1304.
6. Olopoenia LA, King AL: **Widal agglutination test – 100 years later: still plagued by controversy.** *Postgrad Med J* 2000, **76:**80–84.
7. Gupta A, My Thanh NT, Olsen SJ, Sivapalasingam S, My Trinh TT, Phuong Lan NT, Hoekstra RM, Bibb W, Minh NT, Danh TP, Cam PD, Mintz ED: **Evaluation of community-based serologic screening for identification of chronic Salmonella Typhi carriers in Vietnam.** *Int J Infect* 2006, **10:**309–314.
8. Nair S, Schreiber E, Thong KL, Pang T, Altwegg M: **Genotypic characterization of Salmonella typhi by amplified fragment length polymorphism fingerprinting provides increased discrimination as compared to pulsed-field gel electrophoresis and ribotyping.** *J Microbiol Methods* 2000, **41:**35–43.
9. Thong KL, Puthucheary S, Yassin RM, Sudarmono P, Padmidewi M, Soewandojo E, Handojo I, Sarasombath S, Pang T: **Analysis of *Salmonella* typhi isolates from southeast Asia by pulsed-field gel electrophoresis.** *J Clin Micro boil* 1995, **33:**1938–1941.
10. Baddam R, Thong KL, Avasthi TS, Shaik S, Yap KP, Teh CS, Chai LC, Kumar N, Ahmed N: **Whole-genome sequences and comparative genomics of *Salmonella enterica* serovar Typhi isolates from patients with fatal and nonfatal typhoid fever in Papua New Guinea.** *J Bacteriol* 2012, **194:**5122–5123.
11. Aziz RK, Nizet V: **Pathogen microevolution in high resolution.** *Sci Transl Med* 2010, **2**(16):16–4.
12. Baddam R, Kumar N, Thong KL, Ngoi ST, Teh CS, Yap KP, Chai LC, Avasthi TS, Ahmed N: **Genetic fine structure of a Salmonella enterica serovar Typhi strain associated with the 2005 outbreak of typhoid fever in Kelantan, Malaysia.** *J Bacteriol* 2012, **194:**3565–3566.
13. Treangen T, Messeguer X: **M-GCAT: Interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species.** *BMC Bioinformatics* 2006, **7:**433.
14. Kingsley RA, Santos RL, Keestra AM, Adams LG, Bäumler AJ: **Salmonella enterica serotype Typhimurium ShdA is an outer membrane fibronectin-binding protein that is expressed in the intestine.** *Mol Microbiol* 2002, **43:**895–905.
15. Kingsley RA, van Amsterdam K, Kramer N, Bäumler AJ: **The *shdA* gene is restricted to serotypes of Salmonella enterica subspecies I and contributes to efficient and prolonged fecal shedding.** *Infect Immun* 2000, **68:**2720–2727.
16. Muscas P, Rossolini GM, Chiesurin A, Santucci A, Satta G: **Purification and characterization of type 1 fimbriae of Salmonella typhi.** *Microbiol Immunol* 1994, **38:**353–358.
17. Craig L, Pique ME, Tainer JA: **Type IV pilus structure and bacterial pathogenicity.** *Nat Rev Microbiol* 2004, **2:**363–378.
18. Collinson SK, Clouthier SC, Doran JL, Banser PA, Kay WW: **Salmonella enteritidis agfBAC operon encoding thin, aggregative fimbriae.** *J Bacteriol* 1996, **178:**662–667.
19. Duncan MJ, Mann EL, Cohen MS, Ofek I, Sharon N, Abraham SN: **The distinct binding specificities exhibited by enterobacterial Type 1 fimbriae are determined by their fimbrial shafts.** *J Biol Chem* 2005, **280:**37707–37716.
20. Guo A, Cao S, Tu L, Chen P, Zhang C, Jia A, Yang W, Liu Z, Chen H, Schifferli DM: **FimH alleles direct preferential binding of Salmonella to distinct mammalian cells or to avian cells.** *Microbiology* 2009, **155:**1623–1633.
21. Virlogeux I, Waxin H, Ecobichon C, Popoff MY: **Role of the viaB locus in synthesis, transport and expression of Salmonella typhi Vi antigen.** *Microbiology* 1995, **141:**3039–3047.
22. Robbins JD, Robbins JB: **Re examination of the protective role of the capsular polysaccharide (Vi antigen) of Salmonella typhi.** *J Infect Dis* 1984, **150:**436–449.
23. Moncrief MB, Maguire ME: **Magnesium transport in prokaryotes.** *J Biol Inorg Chem* 1999, **4:**523–527.
24. Guo L, Lim KB, Gunn JS, Bainbridge B, Darveau RP, Hackett M, Miller SI: **Regulation of lipid A modifications by Salmonella typhimurium virulence genes phoP-phoQ.** *Science* 1997, **276:**250–253.
25. Chen CY, Eckmann L, Libby SJ, Fang FC, Okamoto S, Kagnoff MF, Fierer J, Guiney DG: **Expression of Salmonella typhimurium rpoS and rpoS-dependent genes in the intracellular environment of eukaryotic cells.** *Infect Immun* 1996, **64:**4739–4743.
26. Stothard P, Wishart DS: **Circular genome visualization and exploration using CGView.** *Bioinformatics* 2005, **21:**537–539.
27. Stanley TL, Ellermeier CD, Slauch JM: **Tissue-specific gene expression identifies a gene in the lysogenic phage Gifsy-1 that affects Salmonella enterica serovar typhimurium survival in Peyer's patches.** *J Bacteriol* 2000, **182:**4406–4413.
28. Mirold S, Rabsch W, Rohde M, Stender S, Tschape H, Russmann H, Igwe E, Hardt WD: **Isolation of a temperate bacteriophage encoding the type III effector protein SopE from an epidemic Salmonella typhimurium strain.** *Proc Natl Acad Sci U S A* 1999, **96:**9845–9850.
29. Hardt WD, Urlaub H, Galan JE: **A substrate of the centisome 63 type III protein secretion system of Salmonella typhimurium is encoded by acryptic bacteriophage.** *Proc Natl Acad Sci U S A* 1998, **95:**2574–2579.
30. Marcus SL, Brumell JH, Pfeifer CG, Finlay BB: **Salmonella pathogenicity islands: big virulence in small packages.** *Microbes Infect* 2000, **2:**145–156.
31. Lee VT, Schneewind O: **Type III secretion machines and the pathogenesis of enteric infections caused by Yersinia and Salmonella spp.** *Immunol Rev* 1999, **168:**241–255.
32. McGhie EJ, Brawn LC, Hume PJ, Humphreys D, Koronakis V: **Salmonella takes control: effector driven manipulation of the host.** *Curr Opin Microbiol* 2009, **12:**117–124.
33. Waterman SR, Holden DW: **Functions and effectors of the Salmonella pathogenicity island 2 type III secretion system.** *Cell Microbiol* 2003, **5:**501–511.
34. Garmendia J, Beuzón CR, Ruiz-Albert J, Holden DW: **The roles of SsrA-SsrB and OmpR-EnvZ in the regulation of genes encoding the Salmonella typhimurium SPI-2 type III secretion system.** *Microbiology* 2003, **149:**2385–2396.
35. Blanc-Potard AB, Solomon F, Kayser J, Groisman EA: **The SPI-3 Pathogenicity Island of Salmonella enterica.** *J Bacteriol* 1999, **181:**998–1004.
36. Gerlach RG, Claudio N, Rohde M, Jackel D, Wagner C, Hensel M: **Cooperation of Salmonella pathogenicity islands 1 and 4 is required to breach epithelial barriers.** *Cell Microbiol* 2008, **10:**2364–2376.
37. Wood MW, Jones MA, Watson PR, Hedges S, Wallis TS, Galyov EE: **Identification of a pathogenicity island required for Salmonella enteropathogenicity.** *Mol Microbiol* 1998, **29:**883–891.
38. Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, Churcher C, Mungall KL, Bentley SD, Holden MT, Sebaihia M, Baker S, Basham D, Brooks K, Chillingworth T, Connerton P, Cronin A, Davis P, Davies RM, Dowd L, White N, Farrar J, Feltwell T, Hamlin N, Haque A, Hien TT, Holroyd S, Jagels K, Krogh A, Larsen TS, Leather S, Moule S, O'Gaora P, Parry C, Quail M, Rutherford K, Simmonds M, Skelton J, Stevens K, Whitehead S, Barrell BG: **Complete genome sequence of a multiple drug resistant Salmonella enterica serovar TyphiCT18.** *Nature* 2001, **413:**848–852.
39. Pickard D, Wain J, Baker S, Line A, Chohan S, Fookes M, Barron A, Gaora PO, Chabalgoity JA, Thanky N, Scholes C, Thomson N, Quail M, Parkhill J,

Dougan G: **Composition, acquisition, and distribution of the Vi exopolysaccharide-encoding Salmonella enterica pathogenicity island SPI-7.** *J Bacteriol* 2003, **185:**5055–5065.
40. Avasthi TS, Devi SH, Taylor TD, Kumar N, Baddam R, Kondo S, Suzuki Y, Lamouliatte H, Mégraud F, Ahmed N: **Genomes of two chronological isolates (Helicobacter pylori 2017 and 2018) of the West African Helicobacter pylori strain 908 obtained from a single patient.** *J Bacteriol* 2011, **193:**3385–3386.
41. Avasthi TS, Kumar N, Baddam R, Hussain A, Nandanwar N, Jadhav S, Ahmed N: **Genome of multidrug-resistant uropathogenic Escherichia coli strain NA114 from India.** *J Bacteriol* 2011, **193:**4272–4273.
42. Devi SH, Taylor TD, Avasthi TS, Kondo S, Suzuki Y, Megraud F, Ahmed N: **Genome of Helicobacter pylori strain 908.** *J Bacteriol* 2010, **192:**6488–6489.
43. Siddavattam D, Karegoudar TB, Mudde SK, Kumar N, Baddam R, Avasthi TS, Ahmed N: **Genome of a novel isolate of Paracoccus denitrificans capable of degrading N, N-dimethylformamide.** *J Bacteriol* 2011, **193:**5598–5599.
44. Yap KP, Teh CS, Baddam R, Chai LC, Kumar N, Avasthi TS, Ahmed N, Thong KL: **Insights from the genome sequence of a Salmonella enterica serovar Typhi strain associated with a sporadic case of typhoid fever in Malaysia.** *J Bacteriol* 2012, **194:**5124–5125.
45. Yap KP, Gan HM, Teh CS, Baddam R, Chai LC, Kumar N, Tiruvayipati SA, Ahmed N, Thong KL: **Genome sequence and comparative pathogenomics analysis of a Salmonella enterica serovar Typhi strain associated with a typhoid carrier in Malaysia.** *J Bacteriol* 2012, **194:**5970–5971.
46. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18:**821–829.
47. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W: **Scaffolding pre-assembled contigs using SSPACE.** *Bioinformatics* 2011, **27:**578–579.
48. Nadalin F, Vezzi F, Policriti A: **GapFiller: a de novo assembly approach to fill the gap within paired reads.** *BMC Bioinformatics* 2012, **13**(14):8. doi:10.1186/1471-2105-13-S14-S8.
49. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler Transform.** *Bioinformatics* 2009, **25:**1754–1760.
50. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D: **Tablet–next generation sequence assembly visualization.** *Bioinformatics* 2010, **26:**401–402.
51. Aziz RK, Devoid S, Disz T, Edwards RA, Henry CS, *et al*: **SEED Servers: high-performance access to the SEED genomes, annotations, and metabolic models.** *PLoS ONE* 2012, **7**(10):48053. doi:10.1371/journal.pone.0048053.
52. Hemmerich C, Buechlein A, Podicheti R, Revanna KV, Dong Q: **An Ergatis-based prokaryotic genome annotation web server.** *Bioinformatics* 2010, **26:**1122–1124.
53. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16:**944–945.
54. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL: **Improved microbial gene identification with GLIMMER.** *Nucleic Acids Res* 1999, **27:**4636–4641.
55. Larsen TS, Krogh A: **EasyGene—a prokaryotic gene finder that ranks ORFs by statistical significance.** *BMC Bioinformatics* 2003, **4:**21.
56. Lagesen K, Hallin PF, Rødland E, Stærfeldt HH, Ussery DW RT: **RNammer: consistent annotation of rRNA genes in genomic sequences.** *Nucleic Acids Res* 2007, **35:**3100–3108.
57. Schattner P, Brooks AN, Lowe TM: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res* 2005, **33:**W686–W689.
58. Akhter S, Aziz RK, Edwards RA: **PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies.** *Nucleic Acids Res* 2012, **40:**e126.
59. Bose M, Barber RD: **Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences.** *In Silico Biol* 2006, **6:**223–227.
60. Vernikos GS, Parkhill J: **Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands.** *Bioinformatics* 2006, **22:**2196–2203.
61. Langille MG, Brinkman FS: **IslandViewer: an integrated interface for computational identification and visualization of genomic islands.** *Bioinformatics* 2009, **25:**664–665.

doi:10.1186/1757-4749-4-20
Cite this article as: Baddam *et al.*: Genome sequencing and analysis of *Salmonella enterica* serovar Typhi strain CR0063 representing a carrier individual during an outbreak of typhoid fever in Kelantan, Malaysia. *Gut Pathogens* 2012 **4**:20.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

JB Journals.ASM.org

GENOME ANNOUNCEMENT

Next-Generation Sequencing and *De Novo* Assembly, Genome Organization, and Comparative Genomic Analyses of the Genomes of Two *Helicobacter pylori* Isolates from Duodenal Ulcer Patients in India

Narender Kumar,[a] Asish K. Mukhopadhyay,[b] Rajashree Patra,[b] Ronita De,[b] Ramani Baddam,[a] Sabiha Shaik,[a] Jawed Alam,[b] Suma Tiruvayipati,[a,c] and Niyaz Ahmed[a,c,d]

Pathogen Biology Laboratory, Department of Biotechnology, School of Life Sciences, University of Hyderabad, Gachibowli, Hyderabad, India[a]; Division of Bacteriology, National Institute of Cholera and Enteric Diseases (Indian Council of Medical Research), Kolkata, India[b]; Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia[c]; and Institute of Life Sciences, University of Hyderabad Campus, Gachibowli, Hyderabad, India[d]

The prevalence of different *H. pylori* genotypes in various geographical regions indicates region-specific adaptations during the course of evolution. Complete genomes of *H. pylori* from countries with high infection burdens, such as India, have not yet been described. Herein we present genome sequences of two *H. pylori* strains, NAB47 and NAD1, from India. In this report, we briefly mention the sequencing and finishing approaches, genome assembly with downstream statistics, and important features of the two draft genomes, including their phylogenetic status. We believe that these genome sequences and the comparative genomics emanating thereupon will help us to clearly understand the ancestry and biology of the Indian *H. pylori* genotypes, and this will be helpful in solving the so-called Indian enigma, by which high infection rates do not corroborate the minuscule number of serious outcomes observed, including gastric cancer.

Helicobacter pylori's coevolution with its host (10, 11, 16) and its tight compartmentalization (13, 16, 18, 19) into several different populations and subpopulations have delivered an excellent premise to pursue the idea of geographic evolution/spread of humans and their pathogens from Africa and to gain insights into pathogen adaptation mechanisms (1, 3). Based partly on these conventions, Indian *H. pylori* isolates have shown to have European origins (9) and are widely held as mostly innocuous or only mildly pathogenic. The severity of *H. pylori*-induced gastro-duodenal diseases and their outcomes vary in different geographic regions and populations, which may be significantly attributable to different genetic compositions of the underlying bacterial strains. More data based on genome sequences from many of strains from different countries are needed to clearly establish the genetic makeup, colonization potential, and virulence characteristics of a particular strain or genotype. In view of this, genome sequence-based characterizations of strains prevalent in different locales is necessary (2).

We describe genomes of *H. pylori* strains NAB47 (Bangalore) and NAD1 (Delhi) from duodenal ulcer patients. Illumina sequencing was performed as described previously (4, 8); briefly, about 3 gigabytes and 1.8 gigabytes of data comprising 72-bp paired-end reads (insert size, 300 bp) provided genome coverages of approximately 300× and 200×, respectively. The raw reads were filtered using the FASTX tool kit (17) and assembled using Velvet (20); the reads yielded 107 (NAB47) and 103 (NAD1) contigs with a hash length set to 37. These contigs were joined into 34 (NAB47) and 48 (NAD1) scaffolds by using SSPACE (6). The scaffolds were aligned and ordered according to their closest reference genome and confirmed using BLAST (12) and Mummer (14). The draft genomes were submitted to RAST (5) for annotation, and the output was validated by using Glimmer (7) and EasyGene (15).

The draft genomes of *H. pylori* NAB47 and NAD1 had sizes of about 1,590,862 bp and 1,588,938 bp, respectively, with G+C contents of 39.17 and 39.03%, respectively. The genomes revealed coding percentages of 91.5% (NAB47) and 91.3% (NAD1) and encoded 1,572 and 1,567 proteins, respectively; each of the genomes contained 36 tRNA genes and 6 rRNA genes. The average lengths for protein-coding genes were found to be 929 bp and 922 bp, respectively. Major virulence markers, such as *cagA*, *vacA*, the whole *cag* pathogenicity island, and several outer membrane proteins of the Hop family, were annotated. In addition, NAD1 harbored two plasmids of 16 kb and 10 kb each that carried genes for transposase, IS*606*, and mobilization proteins, together with replication protein A. CagA protein in both of the strains contained EPIYA D-type motifs, which are typical of Indo-European strains. Important plasticity region genes, such as jhp0940, jhp0947, and *dupA*, were absent, and hp0986 was detected only in NAB47. Finally, whole-genome phylogeny incorporating all the available genomes reconfirmed an Indo-European ancestry (HpEurope).

We believe that the genomes described herein are likely to rekindle our knowledge of the genetic makeup and evolutionary relationships of *H. pylori* in India. Comparative genomic analyses extending out to other unexplored strains from the tribal and mainstream populations will facilitate understanding of the true pathogenic potential (amid adaptive evolution) of the Indian *H. pylori*. Furthermore, they will be immensely helpful in global epidemiological studies and also for the development of diagnostic tools tailored to a particular host population.

Received 31 July 2012 **Accepted** 17 August 2012

Address correspondence to Niyaz Ahmed, niyazSL@uohyd.ernet.in.

N.K. and A.K.M. contributed equally to this work.

Copyright © 2012, American Society for Microbiology. All Rights Reserved.

doi:10.1128/JB.01371-12

Nucleotide sequence accession numbers. The genome sequences of *H. pylori* NAB47 and NAD1 have been deposited with GenBank and assigned accession numbers AJFA00000000 and AJGJ00000000, respectively. The updated sequences/contigs are also available for download from the International Society for Genomic and Evolutionary Microbiology (ISOGEM) server (http://isogem.org/HPNAB47.txt and http://isogem.org/HPNAD1.txt).

ACKNOWLEDGMENTS

We acknowledge support from the University of Malaya High Impact Research Grant (UM.C/625/1HIR/MOHE/CHAN-02)-Molecular Genetics. A.K.M. acknowledges support from the Department of Biotechnology (BT/PR10407/BRB/10/604/2008) and Indian Council of Medical Research. These genomes were completed under the wider umbrella of the Indo-German International Research Training Group, Internationales Graduiertenkolleg (GRK1673), Functional Molecular Infection Epidemiology, an initiative of the German Research Foundation (DFG) and the University of Hyderabad (India). N.K. would like to acknowledge a Junior Research Fellowship received from the Council of Scientific and Industrial Research (CSIR), India, and J.A. acknowledges ICMR for a Senior Research Fellowship.

We are also grateful to M/s Genotypic Technology Pvt. Ltd., Bengaluru, India, for their efforts with the Illumina sequencing. We acknowledge the Bioinformatics Facility (BIF) at the Department of Biotechnology, University of Hyderabad, for use of their computational infrastructure. Further, we thank Akash Ranjan for helpful discussions and for enabling access to the SUN Microsystems CDFD Centre of Excellence for some of our data analyses.

REFERENCES

1. **Ahmed N.** 2011. Coevolution and adaptation of *Helicobacter pylori* and the case for 'functional molecular infection epidemiology.' Med. Princ. Pract. **20**:497–503.
2. **Ahmed N.** 2009. A flood of microbial genomes: do we need more? PLoS One **4**:e5831. doi:10.1371/journal.pone.0005831.
3. **Atherton JC, Blaser MJ.** 2009. Coadaptation of *Helicobacter pylori* and humans: ancient history, modern implications. J. Clin. Invest. **119**:2475–2487.
4. **Avasthi TS, et al.** 2011. Genomes of two chronological isolates (*Helicobacter pylori* 2017 and 2018) of the West African *Helicobacter pylori* strain 908 obtained from a single patient. J. Bacteriol. **193**:3385–3386.
5. **Aziz RK, et al.** 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics **9**:75. doi:10.1186/1471-2164-9-75.
6. **Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W.** 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics **5**:578–579.
7. **Delcher AL, et al.** 1999. Improved microbial gene identification with Glimmer. Nucleic Acids Res. **27**:4636–4641.
8. **Devi SH, et al.** 2010. Genome of *Helicobacter pylori* strain 908. J. Bacteriol. **192**:6488–6489.
9. **Devi SM, et al.** 2007. Ancestral European roots of *Helicobacter pylori* in India. BMC Genomics **8**:184. doi:10.1186/1471-2164-8-184.
10. **Devi SM, et al.** 2006. Genomes of *Helicobacter pylori* from native Peruvians suggest admixture of ancestral and modern lineages and reveal a Western type *cag*-pathogenicity island. BMC Genomics **7**:191. doi:10.1186/1471-2164-7-191.
11. **Falush D, et al.** 2003. Traces of human migrations in *Helicobacter* pylori populations. Science **299**:1582–1585.
12. **Kent WJ.** 2002. BLAT: the BLAST-like alignment tool. Genome Res. **12**:656–664.
13. **Kersulyte D, et al.** 2010. Helicobacter pylori from Peruvian Amerindians: traces of human migrations in strains from remote Amazon, and genome sequence of an Amerind strain. PLoS One **5**:e15076. doi:10.1371/journal.pone.0015076.
14. **Kurtz S, et al.** 2004. Versatile and open software for comparing large genomes. Genome Biol. **5**:R12. doi:10.1186/gb-2004-5-2-r12.
15. **Larsen TS, Krogh A.** 2003. EasyGene: a prokaryotic gene finder that ranks ORFs by statistical significance. BMC Bioinformatics **4**:21. doi:10.1186/1471-2105-4-21.
16. **Linz B, et al.** 2007. An African origin for the intimate association between humans and *Helicobacter pylori.* Nature **445**:915–918.
17. **Taylor J, Schenck I, Blankenberg D, Nekrutenko A.** 2007. Using Galaxy to perform large-scale interactive data analyses. Curr. Protoc. Bioinformatics **Chapter 10:**Unit 10.5.
18. **Wirth T, et al.** 2004. Distinguishing human ethnic groups by means of sequences from *Helicobacter pylori*: lessons from Ladakh. Proc. Natl. Acad. Sci. U. S. A. **101**:4746–4751.
19. **Yamaoka Y.** 2009. Helicobacter pylori typing as a tool for tracking human migration. Clin. Microbiol. Infect. **15**:829–834.
20. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18**:821–829.

JB Journals.ASM.org

GENOME ANNOUNCEMENT

Whole-Genome Sequences and Comparative Genomics of *Salmonella enterica* Serovar Typhi Isolates from Patients with Fatal and Nonfatal Typhoid Fever in Papua New Guinea

Ramani Baddam,[a] Kwai-Lin Thong,[b,c] Tiruvayipati Suma Avasthi,[a,b] Sabiha Shaik,[a] Kien-Pong Yap,[b,c] Cindy Shuan Ju Teh,[c] Lay-Ching Chai,[b,c] Narender Kumar,[a] and Niyaz Ahmed[a,b,d]

Pathogen Biology Laboratory, Department of Biotechnology, School of Life Sciences (Centre for Advanced Studies—UGC-SAP-CAS-I), University of Hyderabad, Gachibowli, Hyderabad, India[a]; Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia[b]; Laboratory of Biomedical Science and Molecular Microbiology, UMBIO Research Cluster, University of Malaya, Kuala Lumpur, Malaysia[c]; and Institute of Life Sciences, University of Hyderabad Campus, Gachibowli, Hyderabad, India[d]

Many of the developing countries of the Southeast Asian region are significantly affected by endemic typhoid fever, possibly as a result of marginal living standards. It is an important public health problem in countries such as Papua New Guinea, which is geographically close to some of the foci of endemicity in Asia. The severity of the disease varies in different regions, and this may be attributable to genetic diversity among the native strains. Genome sequence data on strains from different countries are needed to clearly understand their genetic makeup and virulence potential. We describe the genomes of two *Salmonella* Typhi isolates from patients with fatal and nonfatal cases of typhoid fever in Papua New Guinea. We discuss in brief the underlying sequencing methodology, assembly, genome statistics, and important features of the two draft genomes, which form an essential step in our functional molecular infection epidemiology program centering on typhoid fever. The comparative genomics of these and other isolates would enable us to identify genetic rearrangements and mechanisms responsible for endemicity and the differential severity of pathogenic salmonellae in Papua New Guinea and elsewhere.

Typhoid fever is a major pestilence in the developing world (20), and its prevalence is significant (1,000 cases per 100,000 individuals per year) in Papua New Guinea, which is geographically close to Southeast Asia (15). DNA profiling, which was used previously (6, 14, 20, 21, 22), could not fully explore genome diversity, as *Salmonella enterica* serovar Typhi isolates from Papua New Guinea showed limited heterogeneity, perhaps because of recent clonal expansion from a single endemic/ancestral strain (the disease was rarely seen before 1985) (17, 20). Minimal selection pressure and confinement to a specific geographical region might explain this limited genetic diversity (21, 22) despite horizontal gene transfer (13).

We hypothesized that the genome sequences of *Salmonella* Typhi isolates from patients with typhoid fever due a fatal strain (UJ308A) or a nonfatal strain (UJ816A) would provide significant insights into the association among disease phenotypes and strain characteristics. Two such strains were isolated from blood samples from patients and were found to be sensitive to common antibiotics. Strain UJ308A (phage type VS1) was obtained from a patient who died of typhoid, while UJ816A (phage type DI) was from a patient who recovered.

The 73-bp paired-end sequence data (insert size, 300 bp) were determined with an Illumina Genome Analyzer (GA2x, pipeline version 1.6). About 95× and 105× coverage was achieved for strains UJ308A and UJ816A, respectively, comprising 1.9 and 2.0 Gb of data, respectively. *De novo* assembly was done as described previously (1, 2, 4, 8, 19); initial assembly generated 416 and 335 contigs for UJ308A and UJ816A, respectively, using Velvet (23) with a hash length of 39. The scaffolds were generated from contigs by using SSPACE (5) and further assembled and curated to give a consensus draft. The following statistics were gleaned upon analysis at RAST (3). The sizes of the chromosomes for UJ308A and UJ816A were approximately 4,724,875 and 4,736,723 bp, respectively, with a G+C contents of 51.89 and 51.94%, respectively. The coding percentage for both strains was ~86.8%; UJ308A and UJ816A contained approximately 4,720 and 4,710 protein coding sequences with average lengths of 869 and 873 bp, respectively. The data were further validated by Glimmer (7) and EasyGene (12). RNAmmer (11) revealed that the genome of UJ308A has 78 tRNA and 21 rRNA genes and the genome of UJ816A contains 77 tRNA and 22 rRNA genes. All of the major virulence markers encoded by pathogenicity islands and the genes relevant to the assembly of a type III secretion system (16) were identified in both the strains. The Vi antigen (10, 18), which plays major role in immune evasion, was present in both strains, as in *Salmonella* Typhi CT18 (16). The homologues of *Campylobacter* toxin *cdtB* and *Bordetella* pertussis toxin (9) were also present.

In view of this, further efforts are needed to determine the true extent of strain diversity in terms of (i) gene gains and losses over an evolutionary time scale, (ii) geographic gene flow, (iii) core versus accessory genome dynamics, (iv) virulence acquisition and attenuation, and (v) the preponderance of highly virulent versus "docile" strains across the regions of Asia where typhoid fever is endemic.

Nucleotide sequence accession numbers. The GenBank accession numbers for the genomes reported here are AJTD00000000 (UJ308A) and AJTE00000000 (UJ816A).

Received 12 June 2012 **Accepted** 29 June 2012

Address correspondence to Niyaz Ahmed, niyazSL@uohyd.ernet.in.

Copyright © 2012, American Society for Microbiology. All Rights Reserved.

doi:10.1128/JB.01051-12

ACKNOWLEDGMENTS

We thankfully acknowledge support received from the University of Malaya High Impact Research Grant (UM.C/625/1HIR/MOHE/02 [A000002-5000 1]) Molecular Genetics. This genome program was completed under the wider umbrella of the Indo-German International Research Training Group, Internationales Graduiertenkolleg (GRK1673), Functional Molecular Infection Epidemiology, an initiative of the German Research Foundation (DFG) and the University of Hyderabad (India), of which N.A. is a speaker. N.A. is an Adjunct Professor of Molecular Biosciences at the University of Malaya, Kuala Lumpur, Malaysia, and an Adjunct Professor of Chemical Biology at the Institute of Life Sciences, Hyderabad, India. We are also grateful to M/s Genotypic Technology Pvt. Ltd., Bengaluru, India, for their untiring efforts with Illumina sequencing. We acknowledge the Bioinformatics Facility (BIF) at the Department of Biotechnology, University of Hyderabad, for the use of its computational infrastructure. We are thankful to Akash Ranjan for enabling access to the high-speed computing infrastructure at CDFD, Hyderabad.

All of the members of the Ahmed and Thong labs are gratefully acknowledged for their help and support.

REFERENCES

1. **Avasthi TS, et al.** 2011. Genomes of two chronological isolates (*Helicobacter pylori* 2017 and 2018) of the West African *Helicobacter pylori* strain 908 obtained from a single patient. J. Bacteriol. **193**:3385–3386.
2. **Avasthi TS, et al.** 2011. Genome of multidrug-resistant uropathogenic *Escherichia coli* strain NA114 from India. J. Bacteriol. **193**:4272–4273.
3. **Aziz RK, et al.** 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics **9**:75. doi:10.1186/1471-2164-9-75.
4. **Baddam R, et al.** 2012. Genetic fine structure of a *Salmonella enterica* serovar Typhi strain associated with the 2005 outbreak of typhoid fever in Kelantan, Malaysia. J. Bacteriol. **194**:3565–3566.
5. **Boetzer M, et al.** 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics **27**:578–579.
6. **Combs BG, et al.** 2005. Ribotyping of *Salmonella enterica* serovar Typhi isolates from Papua New Guinea over the period 1977 to 1996. P. N. G. Med. J. **48**:158–167.
7. **Delcher AL, et al.** 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. **27**:4636–4641.
8. **Devi SH, et al.** 2010. Genome of *Helicobacter pylori* strain 908. J. Bacteriol. **192**:6488–6489.
9. **Haghjoo E, Galan JE.** 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. **101**:4614–4619.
10. **Hirose K, et al.** 1997. Survival of Vi-capsulated and Vi-deleted *Salmonella typhi* strains in cultured macrophage expressing different levels of CD14 antigen. FEMS Microbiol. Lett. **147**:259–265.
11. **Lagesen K, et al.** 2007. RNammer: consistent annotation of rRNA genes in genomic sequences. Nucleic Acids Res. **35**:3100–3108.
12. **Larsen TS, Krogh A.** 2003. EasyGene—a prokaryotic gene finder that ranks ORFs by statistical significance. BMC Bioinformatics **4**:21. doi: 10.1186/1471-2105-4-21.
13. **Liu SL, Sanderson KE.** 1995. Rearrangements in the genome of the bacterium Salmonella typhi. Proc. Natl. Acad. Sci. U. S. A. **92**:1018–1022.
14. **Nair S, Schreiber E, Thong KL, Pang T, Altwegg M.** 2000. Genotypic characterization of *Salmonella typhi* by amplified fragment length polymorphism fingerprinting provides increased discrimination as compared to pulsed-field gel electrophoresis and ribotyping. J. Microbiol. Methods **41**:35–43.
15. **Pang T, Bhutta ZA, Finlay BB, Altwegg M.** 1995. Typhoid fever and other salmonelloses: a continuing challenge. Trends Microbiol. **3**:253–255.
16. **Parkhill J, et al.** 2001. Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. Nature **413**:848–852.
17. **Passey M, et al.** 1995. Highly endemic typhoid fever in Papua New Guinea. Southeast Asian J. Trop. Med. Public Health **26**:83–84.
18. **Raffatellu M, et al.** 2006. Capsule-mediated immune evasion: a new hypothesis explaining aspects of typhoid fever pathogenesis. Infect. Immun. **74**:19–27.
19. **Siddavattam D, et al.** 2011. Genome of a novel isolate of *Paracoccus* denitrificans capable of degrading *N*,*N*-dimethylformamide. J. Bacteriol. **193**:5598–5599.
20. **Thong KL, et al.** 1995. Analysis of *Salmonella typhi* isolates from Southeast Asia by pulsed-field gel electrophoresis. J. Clin. Microbiol. **33**:1938–1941.
21. **Thong KL, et al.** 1996. Molecular analysis of isolates of *Salmonella typhi* obtained from patients with fatal and nonfatal typhoid fever. J. Clin. Microbiol. **34**:1029–1033.
22. **Thong KL, et al.** 2002. Increasing genetic diversity of *Salmonella enterica* serovar Typhi isolates from Papua New Guinea over the period from 1992 to 1999. J. Clin. Microbiol. **40**:4156–4160.
23. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18**:821–829.

High throughput computational genomics of different lineages of Extra Intestinal Pathogenic Escherichia coli

ORIGINALITY REPORT

%8	%2	%8	%1
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Sabiha Shaik, Narender Kumar, Aditya K. Lankapalli, Sumeet K. Tiwari, Ramani Baddam, Niyaz Ahmed. "Contig-Layout-Authenticator (CLA): A Combinatorial Approach to Ordering and Scaffolding of Bacterial Contigs for Comparative Genomics and Molecular Epidemiology", PLOS ONE, 2016 Publication	%1
2	Ranjan, Amit, Sabiha Shaik, Arif Hussain, Nishant Nandanwar, Torsten Semmler, Savita Jadhav, Lothar H. Wieler, and Niyaz Ahmed. "Genomic and functional portrait of a highly virulent, CTX-M-15-producing H30- Rx subclone of Escherichia coli sequence type (ST) 131", Antimicrobial Agents and Chemotherapy, 2015. Publication	%1
3	Dale, Adam P., and Neil Woodford. "Extra-intestinal pathogenic Escherichia coli (ExPEC): Disease, carriage and clones", Journal of Infection, 2015.	%1

Publication

4	Mitchell, Natalie M., James R. Johnson, Brian Johnston, Roy Curtiss, and Melha Mellata. "Zoonotic Potential of Escherichia coli Isolates from Retail Chicken Meat Products and Eggs", Applied and Environmental Microbiology, 2015. Publication	%1
5	Kumar, N., V. Mariappan, R. Baddam, A. K. Lankapalli, S. Shaik, K.-L. Goh, M. F. Loke, T. Perkins, M. Benghezal, S. E. Hasnain, J. Vadivelu, B. J. Marshall, and N. Ahmed. "Comparative genomic analysis of Helicobacter pylori from Malaysia identifies three distinct lineages suggestive of differential evolution", Nucleic Acids Research, 2014. Publication	<%1
6	journal.frontiersin.org Internet Source	<%1
7	Nicolas-Chanoine, M.-H., X. Bertrand, and J.-Y. Madec. "Escherichia coli ST131, an Intriguing Clonal Group", Clinical Microbiology Reviews, 2014. Publication	<%1
8	James R. Johnson. "Contamination of Retail Foods, Particularly Turkey, from Community Markets (Minnesota, 1999–2000) with Antimicrobial-Resistant and Extraintestinal	<%1

Pathogenic *Escherichia coli*", Foodborne Pathogens and Disease, 03/2005
Publication

9	Forde, B. M., M.-D. Phan, J. A. Gawthorne, M. M. Ashcroft, M. Stanton-Cook, S. Sarkar, K. M. Peters, K.-G. Chan, T. M. Chong, W.-F. Yin, M. Upton, M. A. Schembri, and S. A. Beatson. "Lineage-Specific Methyltransferases Define the Methylome of the Globally Disseminated Escherichia coli ST131 Clone", mBio, 2015. Publication	<%1
10	www.ncbi.nlm.nih.gov Internet Source	<%1
11	Poolman, Jan T., and Michael Wacker. "Extra-intestinal Pathogenic Escherichia coli (ExPEC), a common human pathogen: challenges for vaccine development and progress in the field", The Journal of Infectious Diseases, 2015. Publication	<%1
12	Nandanwar, Nishant, Traute Janssen, Michael Kühl, Niyaz Ahmed, Christa Ewers, and Lothar H. Wieler. "Extraintestinal pathogenic Escherichia coli (ExPEC) of human and avian origin belonging to sequence type complex 95 (STC95) portray indistinguishable virulence features", International Journal of Medical Microbiology,	<%1

2014.
Publication

13	de Been, Mark, Val F. Lanza, María de Toro, Jelle Scharringa, Wietske Dohmen, Yu Du, Juan Hu, Ying Lei, Ning Li, Ave Tooming-Klunderud, Dick J. J. Heederik, Ad C. Fluit, Marc J. M. Bonten, Rob J. L. Willems, Fernando de la Cruz, and Willem van Schaik. "Dissemination of Cephalosporin Resistance Genes between Escherichia coli Strains from Farm Animals and Humans by Specific Plasmid Lineages", PLoS Genetics, 2014. Publication	<%1
14	Peirano, G., G. Mulvey, G. Armstrong, and J. Pitout. "Virulence potential and adherence properties of Escherichia coli that produce CTX-M and NDM- -lactamases", Journal of Medical Microbiology, 2013. Publication	<%1
15	journals.plos.org Internet Source	<%1
16	Ghodousi, Arash, Celestino Bonura, Anna Maria Di Noto, and Caterina Mammina. "Extended-Spectrum ß-Lactamase, AmpC-Producing, and Fluoroquinolone-Resistant Escherichia coli in Retail Broiler Chicken Meat, Italy", Foodborne Pathogens and Disease, 2015. Publication	<%1

17	Submitted to Napier University Student Paper	<%1
18	Submitted to Higher Education Commission Pakistan Student Paper	<%1
19	dyuthi.cusat.ac.in Internet Source	<%1
20	Maciuca, Iuliana E., Nicola J. Williams, Cristina Tuchilus, Olivia Dorneanu, Eleonora Guguianu, Catalin Carp-Carare, Cristina Rimbu, and Dorina Timofte. "High Prevalence of Escherichia coli-Producing CTX-M-15 Extended-Spectrum Beta-Lactamases in Poultry and Human Clinical Isolates in Romania", Microbial Drug Resistance, 2015. Publication	<%1
21	D.D. Pitout, Johann. "Enterobacteriaceae that Produce Extended-spectrum β-lactamases and AmpC β-lactamases in the Community: The Tip of the Iceberg?", Current Pharmaceutical Design, 2013. Publication	<%1
22	gradworks.umi.com Internet Source	<%1
23	Abdulgayeid, Mahmoud, Hazem Shahin, Seham Foad, and Madiha Ibrahim. "Molecular Characterization of Escherichia Coli Isolated from Buffalo Calves in El-Behera	<%1

Governorate", Alexandria Journal of Veterinary Sciences, 2015.
Publication

---

24 Agabou, A., N. Lezzar, Z. Ouchenane, S. Khemissi, D. Satta, A. Sotto, J.-P. Lavigne, and A. Pantel. "Clonal relationship between human and avian ciprofloxacin-resistant Escherichia coli isolates in North-Eastern Algeria", European Journal of Clinical Microbiology & Infectious Diseases, 2015.
Publication

<%1

---

25 www.cnic.ac.cn
Internet Source

<%1

---

26 "Symposia and Oral Presentations", Clinical Microbiology and Infection, 04/2012
Publication

<%1

---

27 Zhao, J.-y., Y.-q. Zhu, Y.-n. Li, X.-d. Mu, L.-p. You, C. Xu, P. Qin, and J.-l. Ma. "Coexistence of SFO-1 and NDM-1 -lactamase genes and fosfomycin resistance gene fosA3 in an Escherichia coli clinical isolate", FEMS Microbiology Letters, 2015.
Publication

<%1

---

28 www.escmid.org
Internet Source

<%1

EXCLUDE QUOTES ON EXCLUDE MATCHES < 5 WORDS

EXCLUDE ON
BIBLIOGRAPHY