Genome architecture and evolutionary dynamics of *Salmonella enterica* serovar Typhi from Typhoid endemic zones

Thesis Submitted to the University of Hyderabad
For the Degree of

DOCTOR OF PHILOSOPHY

By

Ramani Baddam

(Reg. No. 11LTPH02)

Department of Biotechnology and Bioinformatics
School of Life Sciences
University of Hyderabad
Hyderabad-500046
INDIA

February, 2015

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 this thesis entitled **"Genome architecture and evolutionary dynamics of *Salmonella enterica* serovar Typhi from Typhoid endemic zones"**, 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 Dr. Niyaz Ahmed, Associate Professor. 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: Ramani Baddam

Reg. No.: 11LTPH02

University of Hyderabad

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

CERTIFICATE

This is to certify that **Ms. Ramani Baddam** has carried out the research work embodied in the present thesis under the supervision and guidance of Dr. Niyaz Ahmed, Associate Professor, for a full period prescribed under the Ph.D. ordinance of this University. We recommend this thesis entitled **"Genome architecture and evolutionary dynamics of *Salmonella enterica* serovar Typhi from Typhoid endemic zones"** for submission for the degree of Doctor of Philosophy of this University.

Date:

Dr. Niyaz Ahmed
Research Supervisor & HoD
Department of Biotechnology and Bioinformatics

Date:

Prof. P. Reddanna
Dean,
School of Life Sciences

Dedication

To my loving mother,

For believing in my education,

For tireless work in support of it!!

ACKNOWLEDGEMENTS

I am most indebted to my supervisor, Dr. Niyaz Ahmed for his faith in giving me this opportunity. His measureless efforts in training and providing exposure of the state of art in NGS field are highly admired. It would not have been possible to achieve my objectives without his extraordinary support and efforts in bringing various collaborations. I am very much thankful to him for the liberty and freedom conferred throughout my doctoral work.

I would like to thank my Doctoral committee members Dr. J.S.S. Prakash and Dr. Insaf Ahmed Qureshi for their valuable suggestions. I am thankful to present and former HODs and Deans of the school of Life sciences and all the faculty members of SLS. I am also thankful to all non-teaching staff of the Department and school of Life Sciences

I would like to acknowledge UGC- RFSMS, DBT for support through fellowship and University of Malaya- High Impact Research grant, IRTG GRK 1673, NBA 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 thankful to Dr. Akash Ranjan, CDFD for enabling access to the high-speed computing infrastructure at CDFD, Hyderabad. I would like to thank the team of Genotypic Technology Pvt. Ltd., for their help regarding Illumina sequencing. My heartfelt thanks Bionivid Technology team for their support during NGS training.

Prof. Kwai LinThong and her research group members are gratefully acknowledged for their help and support in providing samples. I would like to thank our collaborators Dr. Jamuna Vadivelu of University of Malaya and Dr. Carmen Molina Paris of University of Leeds for the research exchange visits.

My heartfelt felt thanks to all present and former PBL members for making this period so wonderful. I would like to thank Suma and Haritha Devi for all their guidance during the initial days of my work. My sincere thanks to Narender for all his patience in answering my naive queries and also for being a critical commenter throughout the project. I am grateful to Aditya for all the help extended by him, whenever I confronted difficulties due to my limited programming

skills. My special thanks to Arif, Amit, Narender, Sabiha, Priya, Nishant, Kishore, Aditya for their invaluable friendship. My special thanks to Kishore for helping me out with endless technical stuff. The wonderful memories built together with Savita as a roommate and loving friend during my PhD are highly admired.

I owe to my parents and in-laws for their exceptional support and confidence in me throughout this period. I would like to thank my grandfather for all his help in support of my education. I am grateful to Naresh for his continuous encouragement and for always being with me in making this possible. I am thankful to my brother for all his help in supporting me with everything needed.

I wish to thank my friends Nitheen, Prashanti, Manasa, Dhoni, Rahul, Arpita, Nagamani, Nelson, Abishek for making this HCU journey so wonderful and memorable. My special thanks go to Bhim, Sawmya and Bhupathi for all their care and concern.

- Ramani

INDEX

List of Abbreviations

ESBL	Extended-spectrum of β lactamases
Inc HI1	Incompatibility group HI 1
MDR	Multi drug resistance
conc.	Concentration
SPI	Salmonella Pathogenicity Island
HTS	High Throughput Sequencing
NGS	Next Generation Sequencing
MLST	Multilocus Sequence Typing
MLVA	Multilocus variable number tandem repeat analysis
PFGE	Pulsed field gel electrophoresis
AFLP	Amplified fragment length polymorphism
SNP	Single Nucleotide Polymorphism
HGT	Horizontal Gene Transfer
S. Typhi	*Salmonella enterica* subsp. *enterica* serovar Typhi
S. bongori	*Salmonella bongori*
S. Typhimurium	*Salmonella enterica* subsp. *enterica* serovar Typhimurium
S. Paratyphi A	*Salmonella enterica* subsp. *enterica* serovar Paratyphi A
S. Paratyphi A	Salmonella enterica subsp. enterica serovar Paratyphi B

S. Paratyphi A	*Salmonella enterica* subsp. *enterica* serovar Paratyphi A
E.coli	*Escherichia coli*
PCR	Polymerase Chain Recation
BLAST	Basic Local Alignment Search Tool
DNA	Deoxyribonucleic acid
NCBI	National Centre for Biotechnology Information
RAM	Random Access Memory
RAST	Rapid Annotation using Subsystem Technology
VFDB	Virulence Factors Database
ANI	Average nucleotide Identity
BWA	Burrows-wheeler aligner
ISGA	Integrative Services for Genomic Analysis
PanOCT	Pan-genome ortholog clustering tool
aa	amino acids
CDD	Conserved Domains Database
RPS BLAST	Reversed Position Specific BLAST
COG	Clusters of Orthologous groups
NJ	Neighbor Joining
IS	Insertion-sequence

SAM	Sequence Alignment/Map
bcf	Binary variant call format
VCF	Variant Call Format

List of Figures

This revealed close similarity among genomes and also co-clustering of strains isolated from the same regions

Figure 11 Core genome based phylogeny: A core genome based consensus Maximum Likelihood phylogenetic tree constructed after 1000 replicates/bootstraps

Figure 12 Genome alignment: The whole genome alignment of all eight genomes was generated using progressiveMauve. Each colored block represents similar sequences in the respective genomes

Figure 13 Comparative analysis of Salmonella Pathogenicity islands: The status of major pathogenicity islands of Salmonella Typhi was identified among all the strains. Each inner ring represents an individual genome in a particular color along with their percentage identity level shown in gradation of color. The pathogenicity islands are marked in the outer ring.

Figure 14 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.

Figure 15 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 16 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.

Figure 17 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

Figure 18 SNP analysis: Methodology followed for detection of SNPs in pairwise comparisons of outbreak and carrier strains

Figure 19 Determination of variable gene content: Schematic representation followed for the identification of variable pseudogene content in pairwise comparison of strains BL196 and CR0063

List of Tables

Chapter 1

General Introduction

1.1 The *Salmonellae*

The genus *Salmonella* consists of rod shaped, Gram negative bacteria with only two identified species – *Salmonella bongori* and *Salmonella enterica*. The species *Salmonella enterica* is further subdivided into six subspecies – *enterica* (I), *salamae* (II), *arizonae* (IIIa), *diarizonae* (IIIb), *houtenae* (IV), *indica* (VI) (Brenner et al., 2000). Among these only subspecies *enterica* (I) is associated with avian and mammalian hosts, whereas others and species *S.bongori* are commonly found in cold blooded vertebrates (Bäumler et al., 1998). Further, based on antigen presented, more than 2600 serovars are identified in *Salmonella enterica* subsp. *enterica* (I). These serovars vary widely in their host range and clinical syndromes observed (Gal-Mor et al., 2014). The self-limiting gastroenteritis is noted during infection with certain host generalist serovars like *S*. Typhimurium, but host restricted serovars like *S*.Typhi and *S*. Paratyphi A lead to an invasive systemic infection in humans and have no other identified animal or environmental reservoirs (Bäumler and Fang, 2013). Typhoid fever caused by *Salmonella enterica* subsp. *enterica* serovar Typhi (*Salmonella* Typhi) is common in various developing countries with minimal living standards (Dougan and Baker, 2014).

1.2 Typhoid fever and *Salmonella* Typhi

The transmission of *S*.Typhi is mainly through fecal-oral route and some infected individuals shed bacteria in feces for very long periods, thereby contaminating food or water in regions with poor sanitation infrastructure (Gopinath et al., 2012). The global incidence of Typhoid cases is very difficult to estimate due to lack of an accurate, suitable diagnostic that can be applied in resource poor endemic settings, where the infection burden is usually very high (Parry et al., 2002). Further, due to highly diverse non-specific clinical symptoms – fever(>39°C), chills, vomiting, head ache, cough, weight loss etc, infection is often wrongly diagnosed as malaria or any other illness leading to inaccurate estimates of global burden (Dougan and Baker, 2014). Since the complete eradication of Typhoid fever demands access to clean water and improved sanitation systems in all developing countries of the world, the solution of implementing vaccination is proposed for a long time

(Waddington et al., 2014). The development of an effective vaccine needs clear understanding of *S*.Typhi's pathogenicity and host immune response (Sztein et al., 2014).

Salmonella Typhi is classified as O-antigen O9-12, flagellar antigen H:d and Vi antigen positive according to Kaufmann-White scheme (Bale, 2007). There are some exceptions noted where *S*.Typhi isolates alternatively contain flagellar antigen H:j and do not possess Vi capsule (Baker et al., 2005)(Moshitch et al., 1992). The major virulence co-ordinates of *S*.Typhi are encoded in horizontally acquired DNA segments designated as *Salmonella Pathogenicity Islands* (SPIs) (Marcus et al., 2000). These SPIs have conferred important features like, for example, SPI7 which encodes genes for the synthesis of Vi capsule that has been implicated in immune evasion of *S*.Typhi (Wilson et al., 2008) and more than 10 SPIs which encode various functions have been identified in *S*.Typhi genomes (Parkhill et al., 2001) (Deng et al., 2003). The differences in phages and pseudogene content were observed when genomes of host restricted *S*.Typhi where compared to genomes of host generalists like *S*. Typhimurium (McClelland et al., 2001). Some of the *S*.Typhi strains contained plasmids, mainly of incompatibility group HI 1(Inc HI 1) that encode genes responsible for multidrug resistance (MDR) and some others are rarely found in certain isolates (Wain et al., 2003) (Mirza et al., 2000).

1.3 Emergence of multidrug resistance

The *S*. Typhi strains resistant to all major first line antibiotics such as chloramphenicol, ampicillin and co-trimoxazole were reported in various endemic zones of Southeast Asia by 1990s (Wain et al., 2014). This led to shift from using fluoroquinolones for treatment by clinicians, but resistance to these antibiotics due to mutations in DNA gyrase and topoisomerase IV were already reported (Turner et al., 2006). Azithromycin and ceftriaxone are majorly in use once the reduced susceptibility for fluoroquinolones was observed, but certain sporadic cases resistant even to latter drugs were noted in some countries (Dutta et al., 2014). Further, detection of ESBL (Extended-spectrum of β lactamases) producing *S*. Typhi in various countries raises concern by severely limiting the available

treatment options (Kumarasamy and Krishnan, 2012). However, certain studies in recent past have again noted reduced multidrug resistance and also identified *S*. Typhi strains susceptible to some of the first line antibiotics (Kumar et al., 2011). But the re-usage of these antibiotics for treatment has to be critically evaluated.

Some of the genes encoding resistance were carried on mobile elements such as self-transmissible IncHI1 plasmid and strong association of them with certain haplotypes like H58 is reported (Holt et al., 2011). Further, the successful spread of this dominant, resistant haplotype to Asia and Africa poses a constant threat to public health, worldwide (Emary et al., 2012) (Kariuki et al., 2010).

1.4 Typhoid burden

The revised estimates of typhoid fever burden amount to approximately 26.9 million cases globally (Buckle et al., 2012). However, accuracy of these estimated values is severely affected due to lack of proper diagnostics and surveillance systems in many developing countries. In endemic zones, serological diagnosis for the presence of antibodies usually shows more number of Typhoid cases than those observed using culturing techniques, possibly due to high background levels of antibodies (Levine et al., 1978). This total burden is distributed unevenly among various countries based on factors like socio-economic status, proper sanitation system, access to safe drinking water and hygienic food. A recent report estimates that 85% of the cases are mainly reported from India, Pakistan and Bangladesh (Maurice, 2012). Another study from five developing countries of Asia – China, India, Pakistan, Indonesia and Vietnam identified differences in the disease incidence with lowest reported from China and Vietnam (Ochiai et al., 2008). The numbers also differed significantly among various age groups in certain countries and it was noticed that mean age of affected population was low in highly burdened zones. In certain countries like Malaysia, the burden of Typhoid varied considerably among different states, for example, Kelantan state which is considered as an endemic zone of Typhoid fever noted a very high value of 50.3 per 100,000 individuals when compared to total annual incidence value of 10.2 to 17.9 in Malaysia as a whole

(Yap and Puthucheary, 1998). The Typhoid burden is not only high in South Asian countries but also in certain extended regions of Oceania like Papua New Guinea, where about 1208 cases per 100,000 individuals were reported in Goroka, Eastern Highlands province in 1995 (Passey, 1995).

There is a dire need for global initiatives to improve data collection which is presently missing from many low-income countries, especially in certain countries of Africa where the disease burden is under estimated (Darton et al., 2014).

1.5 Detection of *Salmonella* Typhi

The diagnosis of *Salmonella* Typhi is done using various methods ranging from traditional culturing techniques to highly specific molecular detection methods. However, each method has its own pros and cons along with varied potential of applicability in low-income countries with minimal infrastructure (Waddington et al., 2014). Even today, culturing of bacteria from blood samples followed by microbial characterization is mainly adopted, though it has limited sensitivity of 40-60% (Akoh, 1991). Further, the probability to detect bacteria is critically linked to time duration elapsed after the onset of infection and volume of blood isolated for diagnosis (Wain et al., 1998). Although culturing of bacteria from bone marrow offers very high sensitivity when compared to culturing from blood, it is not very feasible (Bhutta, 2006). As time needed to observe the result from a culturing based diagnostic is more, the emergence of antibody based detection techniques like Widal test, Typhidot, TUBEX etc, was observed (Lim et al., 1998) (Choo et al., 1999) (Olopoenia and King, 2000). These techniques are very time-efficient, but still have lower sensitivity compared to culturing techniques. Further, they lack specificity due to cross reactivity and existence of high background antibody titers in endemic zones (Parry et al., 2011). The sensitivity of PCR based methods was also hampered due to limitations of low copy number of DNA in samples, contamination of samples etc. As the efficiency of molecular detection methods highly rely on specificity of the target, the current genomic and functional studies are mainly focused on identification of these reliable targets (Baker et al., 2010).

As the prevalence estimates and other crucial burden measures are directly affected by non-availability of efficient diagnostic methods, there is an urgent need to develop one which is easily applicable, being cost effective and promises optimal sensitivity and specificity values even in the endemic settings.

1.6 Vaccination

The increase in antimicrobial resistance has evoke the need to prevent infection using an effective vaccine that can be administered mainly to those categorized as high risk group coming from endemic zones (Wain et al., 2014). This offers an intermediate solution for disease control before the provision of safe drinking water and proper sanitation in all low-income countries (Darton et al., 2014). Various vaccines have been introduced for this purpose, but they varied widely in their efficacy levels. The first heat-killed whole cell vaccine was developed in 1890s, but showed a minimal efficacy of 51-88% (Engels and Lau, 2000). Although, later developments led to introduction of more vaccines with better efficacy like Ty21a oral live attenuated vaccine and Vi polysaccharide vaccine, their usage was restricted to individuals more than two years old (Guzman et al., 2006). The more recent Vi antigen based conjugate vaccines like Vi *O*-Acetyl Pectin-rEPA conjugate vaccine, Typbar-TCV™ and others that are still in various stages of development pipeline, though appear somewhat better than various available ones (Anwar et al., 2014), the emergence of Vi antigen-negative *S*. Typhi strains makes them redundant. Therefore, certain issues still persist for effective implementation of vaccination in developing countries - lack of an approved vaccine which can be administered to individuals of all age groups with effectiveness for longer duration without any booster doses.

1.7 Pathogenesis and carrier state

Salmonella Typhi enters humans upon consumption of contaminated food or water. After it passes through the acidic environment of stomach, it crosses the epithelial barrier primarily at M cells of

the peyer´s patches in small intestine (Ruby et al., 2012). Further systemic spread of bacteria to inner organs like lymph nodes, spleen, liver, gall bladder and bone marrow is facilitated by macrophage uptake, where bacteria multiplies intracellularly (Haraga et al., 2008). The bacteria can persist in certain organs like gall bladder for a very long time, thereby establishing a chronic carrier state in up to 1- 4% of the infected individuals (Monack et al., 2004). The colonization of *S*.Typhi in deeper internal organs is favored by minimal intestinal inflammation and lack of neutrophil recruitment (Gal-Mor et al., 2014).

The exact immune evasion strategies and persistence mechanisms employed by the bacteria are not clearly understood, but progress in this direction is more likely with the development of recent mouse models (Mathur et al., 2012). Further the role of Vi capsule and its regulatory genes in immune evasion by *S*. Typhi has been implicated in various study reports (Pickard et al., 2013).

The two famous stories of chronic carriers - Mary Malon (Typhoid Mary), a cook in USA and Mr N, a milkman from UK have clearly portrayed the importance of carriers in transmission studies of *S*.Typhi (Marineli et al., 2013) (Mortimer, 1999). Further, various studies have reported an association between chronic persistence of *S*. Typhi and occurrence of gall stones in gall bladder (Crawford et al., 2008). The gall stones favor biofilm formation by bacteria and in turn result in protection from antimicrobial drugs, thereby leaving cholecystectomy as the only treatment option (Gonzalez-Escobedo et al., 2011) (Bäumler et al., 2011).

1.8 Evolution of *Salmonella*

It was estimated that *Salmonella* and *Escherichia coli* diverged from their last common ancestor about 100 million years ago (Meysman et al., 2013). This divergence and host adaptation were known to be facilitated by acquisition of certain pathogenicity islands in *Salmonella* (Bäumler et al., 1998). The SPI I has conferred the ability to invade intestinal epithelium for both species *bongori* and *enterica* and it was not found to be present in *E.coli* (Ochman and Groisman, 1995) (Mills et al., 1995). The species

enterica further diverged from *bongori* due to acquisition of SPI II which encoded certain genes necessary for causing systemic infection (Shea et al., 1996) (Ochman et al., 1996). Further, the expansion of host range occurred only in subspecies I of *enterica*, whereas the other subspecies of *enterica* were mainly limited to coldblooded vertebrates. The differences in host range and pathogenicity among the serovars of subspecies *enterica* were attributed to various chromosomal changes – point mutation, gene gain through horizontal gene transfer (HGT), pseudogenisation etc. (McClelland et al., 2004) (Chen et al., 2009).

1.9 Genetic variation and population structure of *Salmonella* Typhi

The identification of genetic variation in a bacterial population, subdividing it based on one of the discriminatory markers and further tracing the evolutionary history of organism considering the information provided by above has been a primary component of epidemiological research. These studies are important in case of pathogenic bacteria as they have potential to provide valuable insights for the design of effective identification tests, disease control measures and prevention regimens (Niemann and Supply, 2014). They also keep public health policy makers informed about the current disease state and determine need for surveillance systems (Darton et al., 2014).

Given this, various typing tools have been developed and used extensively for analyzing the pathogenic bacteria. These typing tools are scored variedly for their performance based on certain parameters – ability to type and differentiate majority of the individuals in a population as well as the stability, reproducibility and portability of results. Many typing tools were able to provide optimal resolution for highly recombining bacteria like *Helicobacter pylori*, *E. coli* etc., but they offered only minimal to low resolution for monomorphic bacteria like *Salmonella* Typhi, *Yersinia pestis* etc. (Achtman, 2008).

Pulsed field gel electrophoresis (PFGE), Ribotyping, Amplified fragment length polymorphism (AFLP), IS200 typing, Multilocus variable number tandem repeat analysis (MLVA) and Multilocus

sequence typing (MLST) are some of the major typing tools employed previously for analysis of *Salmonella* Typhi. PFGE was successful in investigation of many outbreaks due to *S*.Typhi and is widely used even now by surveillance systems (Thong et al., 1994). With the availability of uniform guidelines from PulseNet International, a greater comparability and stability of the PFGE has been established. Even then, this technique provides limited resolution and demands a lot of technical proficiency (Sabat et al., 2013). Nair et al, have noted that AFLP has shown higher resolution over PFGE and Ribotyping in an analysis of *S*. Typhi isolates from different countries (Nair et al., 2000). IS200 typing was of limited use in the case of *S*. Typhi and certain studies have shown that PFGE and Ribotyping offered more resolution than this method (Navarro et al., 1996). Later, MLVA was reported as more promising tool (Octavia and Lan, 2009) after testing 73 global isolates, but these results were in contradiction with what was observed from SNP typing. The above typing reports reveal that much diversity exists in *S*. Typhi as they underline distinct patterns among global isolates, but it is important to note that these tools rely only on a limited, highly variable part of the genome and do not accurately represent true relationships that would emerge when complete genome is considered. The Multilocus sequence typing, which relies on more neutral housekeeping genes was carried out by Kidgell et al., 2002 using seven housekeeping genes of 26 global isolates. This study could detect variation only at three loci and observed a highly clonal nature, thereby limiting its usage for typing of *S*. Typhi. The next alternative applied was typing based on identification of SNPs in much larger portions of the genome. A remarkable study by Roumagnac et al., 2006 on 200 gene fragments of 105 global isolates also could detect only 82 SNPs. In order to achieve any increase in resolution would require the whole genome sequences of *S*.Typhi isolates from different geographical zones. The first study of this kind was done by Holt et al., 2008 where they determined whole genome sequences of 19 *S*.Typhi isolates belonging to different haplotypes. This study could accurately represent genetic variation and provided valuable insights regarding the evolution of this pathogen.

1.10 Rationale and objectives of the study

Many of the developing countries are heavily burdened due to acute and chronic infections caused by *Salmonella* Typhi. Further, the existence of strains resistant to majority of first line antibiotics along with few susceptible ones, co-circulating in the same endemic zones, are posing a severe challenge in designing effective therapeutic interventions. The situation is more complicated due to lack of an accurate diagnostic that can accurately identify the chronic carrier state. There are still many lacunae in understanding of the immune evasion and persistence mechanisms of *S.*Typhi. Therefore, a greater understanding of the genetic variation and population structure of *S.*Typhi isolated from endemic zones would provide better insights as to the adaptation and evolution of this pathogen.

The various studies that have attempted to analyze the variation using different typing tools were not successful or could not draw correct inferences due to the highly clonal nature of this organism. This inability has driven the whole genome sequencing studies of *S.*Typhi – the first one sequenced was a multidrug resistant strain CT18 from Vietnam followed by strain Ty2 which was of vaccine related importance (Parkhill et al., 2001) (Deng et al., 2003). After the reports of these two individual strains, the advent of Next generation sequencing (NGS) technologies offered feasibility of large scale sequencing for the first time and it was carried out by Holt et al., 2008. As the first whole genome sequencing study (Holt et al., 2008) was carried out during the inception of NGS technologies, the high quality and coverage of data could not be achieved, retarding certain downstream analyses. However, this study has highlighted that relative contribution of pseudogenisation could be significant in evolution of this monomorphic pathogen when compared to genome diversification through point mutation or recombination. Further, the impact of this pseudogenisation and its dynamics at population level was not analyzed. Given the above, we wanted to understand the impact of this pseudogenisation at population level using a collection of suitable strains isolated from different disease states – outbreak, carrier and sporadic ones, as this

could provide detailed understanding of the mechanisms behind the chronic persistence and successful adaptation of this pathogen. Further, we also attempted to understand the patterns that will emerge by comparison of gene inventories, especially with respect to functional and pseudogenes of these strains. Hence we framed the following objectives:

- High-Throughput whole genome sequencing and *de novo* assembly of *Salmonella* Typhi strains isolated from endemic zones
- Determination of the pan-genome structure and its boundaries in *Salmonella* Typhi with insights into adaptation mechanisms
- Comparative genomic analysis of strains associated with an outbreak and carrier individual

Chapter 2

High-Throughput whole genome sequencing and de novo genome assembly of *Salmonella* Typhi strains isolated from endemic zones

---

Part of this chapter was published as:

1) Genome sequencing and analysis of *Salmonella enterica* serovar Typhi strain CR0063 representing a carrier individual during an outbreak of typhoid fever in Kelantan, Malaysia. (2012) **Baddam R**, Kumar N, Shaik S, Suma T, Ngoi ST, Thong KL, Ahmed N. **Gut Pathog**. 13; 4(1):20.

2) Whole-genome sequences and comparative genomics of *Salmonella enterica* serovar Typhi isolates from patients with fatal and nonfatal typhoid fever in Papua New Guinea. (2012) **Baddam R**, Thong KL, Avasthi TS, Shaik S, Yap KP, Teh CS, Chai LC, Kumar N, Ahmed N. **Journal of Bacteriol**.194(18):5122-3.

3) Genetic Fine Structure of a *Salmonella enterica* serovar Typhi strain assosciated with the 2005 outbreak of typhoid fever in Kelantan, Malaysia. (2012) **Baddam R**, Kumar N, Thong KL, Ngoi TS, Teh CS, Yap KP, Chai LC, Avasthi TS, Ahmed N. **Journal of Bacteriol**. 194(13):3565-6.

4) Insights from the genome sequence of a Salmonella enterica serovar Typhi strain associated with a sporadic case of typhoid fever in Malaysia (2012) Yap KP, Teh CS, Baddam R, Chai LC, Kumar N, Avasthi TS, Ahmed N, Thong KL. **Journal of Bacteriol**. 194(18):5124-5.

5) Genome sequence and comparative pathogenomics analysis of a Salmonella enterica Serovar Typhi strain associated with a typhoid carrier in Malaysia. (2012) Yap KP, Gan HM, Teh CS, Baddam R, Chai LC, Kumar N, Tiruvayipati SA, Ahmed N, Thong KL. **Journal of Bacteriol**. 194(21):5970-1.

2.1 Introduction

The recent advances in High throughput sequencing (HTS) technologies has significantly improved understanding of the population structure and evolution of various pathogenic bacteria, by enabling whole genome sequencing of large number of strains from different genetic backgrounds (Engstrand, 2009). In addition to being economical, the ability to generate huge amount of data in significantly very less time when compared to traditional sequencing methods, has led to widespread application of NGS in epidemiological studies (Kao et al., 2014). Also the affordability of this technique due to multiplexing has given an opportunity for individual researchers to sequence strains of interest in large numbers to glean the underlying microbial evolution (Didelot et al., 2012). In a way, HTS technologies presented a shift from a reductionist approach to a holistic one in understanding various pathogenic bacterial infections.

The comparison of major NGS platforms along with recently introduced bench top sequencers has been recently reviewed by (Liu et al., 2012) .With the huge amount of data produced by these new sequencing platforms, downstream data analysis also becomes very complex (Nagarajan and Pop, 2013). For assembling a genome in *de novo* fashion, the shorter read fragments obtained after sequencing are merged to form a contiguous overlapping sequences called "contigs" using various algorithms (Pop, 2009). Based on characteristics of the read data, computational efficiency and quality of the assembly produced (assessed using pre-defined metrics like N50); a suitable method has to be chosen from a broad range of algorithms available for this purpose. The quality of the assembly produced in turn depends on coverage depth and base quality (Flicek and Birney, 2009). The "coverage depth" represents the average of number of times each nucleotide in the genome is represented, for example coverage depth of 10X indicates that on an average each base is represented by at least 10 sequence reads. Theoretically the average genome coverage (C) can be calculated using the formula:

$$C = N \times \frac{L}{G}$$

where N denotes number of reads, L denotes average read length, G denotes genome size (Ekblom and Wolf, 2014). Further the quality of each base is represented by phred score attached with it in read data and this value represents the probability of error in base calling. This means phred quality score of 10 represents that probability of incorrect base calling is 1 in 10 (Ewing and Green, 1998). Therefore high quality read data along with optimum coverage depth improves the chances of obtaining good quality genome assembly. However due to gaps between contigs, achieving a finished genome still requires some additional PCRs followed by sequencing (Mardis et al., 2002). In the case of paired end data, where a single strand DNA fragment is sequenced from both the ends, the number of gaps between contigs can be minimized using this information to obtain a nearly complete draft genome (Mardis et al., 2002).

Salmonella Typhi, the aetiologic agent of Typhoid fever, is a human restricted pathogen. Phage typing, PFGE, AFLP, Ribotyping, MLST, SNP analysis were not optimal to understand the genetic structure and diversity of this highly monomorphic pathogen (Achtman, 2008). Therefore HTS technologies have been employed to sequence large number of strains and intial studies of the genome sequences proposed pseudogenisation as one of the major underlying mechanism of evolution in this pathogen(Holt et al., 2008). Further in order to study these mechanisms in greater detail, whole genome sequencing of Salmonella Typhi strains from endemic zones in Asia were required.

These strains were sequenced using advanced Illumina sequencing platform available at the initiation of the study, so that optimal quality and coverage of genomes could be achieved. Though the strains sequenced in the study was merely driven by availability from collaborators, these strains along with others available in NCBI were sufficient to address the major objectives defined for this study. The strains sequenced in this study were associated with different clinical manifestations - outbreaks,

carrier strains and fatal episodes. The strain BL196 was associated with a large outbreak of Typhoid fever in 2005 were approximately 735 cases and two deaths were reported in Kelantan, northeast of Peninsular Malaysia. The strain CR0063 was isolated in 2007 as part of a surveillance program from a healthy adult living in Kelantan, Malaysia and was reported to share PFGE profile with the outbreak associated strain BL196. The Papua New Guinea, which is close to some of the endemic foci of Typhoid fever in Southeast Asia, is heavily burdened with these infections especially from mid-1980s (Thong et al., 1996). The strains UJ308A and UJ816A were isolated in Papua New Guinea from fatal and non-fatal cases, respectively, in 1998.

In this part, the whole genome sequencing of the above mentioned isolates with different disease manifestations was carried out. The shorter read fragments obtained after sequencing were further assembled into contigs using the best suited freely available genome assembly algorithm. These contigs were further processed using various tools to refine genome assembly and then were finally subjected to annotation. All these sequences were deposited in NCBI database.

2.2 Materials and Methods

2.2.1 DNA Quality check

The DNA of all strains was procured from our collaborator Prof. Kwai Lin Thong's lab in University of Malaya, Malaysia. The quantity and quality of DNA was assessed before proceeding for sequencing using Nanodrop Spectrophotometer and Gel electrophoresis.

S.No	Sample Name	Absorbance value 260/280	DNA conc. (ng/µl)
1	**UJ308A**	**1.9**	**107.49**
2	**UJ816A**	**1.86**	**94.07**
3	**CR0063**	**1.82**	**103.77**
4	**BL196/05**	**1.87**	**96.81**

***Table 1**: DNA purity and concentration assessment using Nanodrop spectrophotometer.*

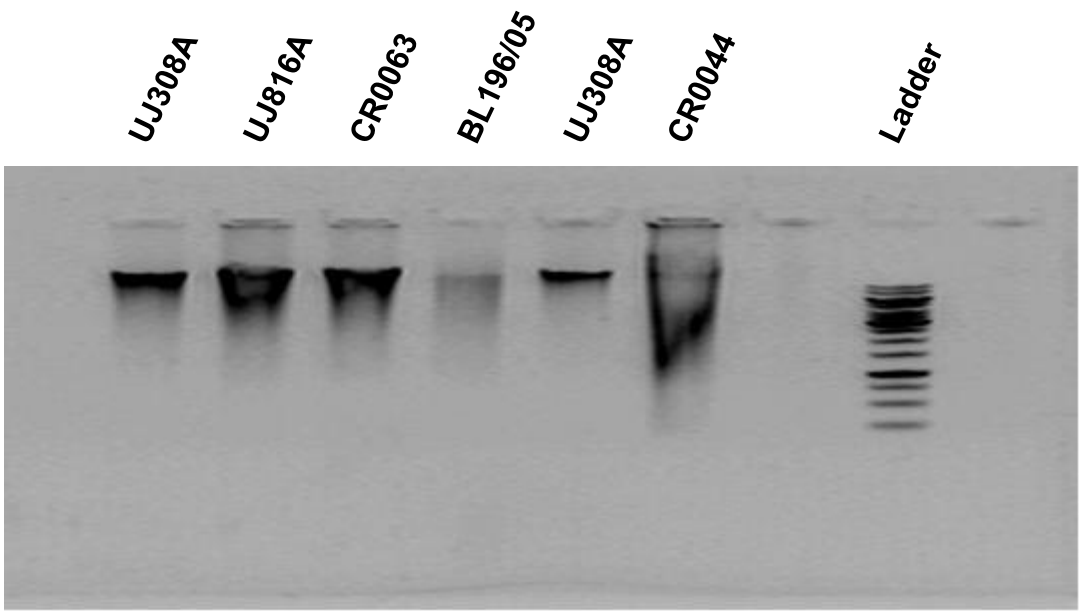


Figure 1: DNA quality analysis: *The quality of DNA was analyzed using gel electrophoresis. The DNA of sample BL196/05 was observed to be partially degraded and replacement sample were procured again.*

2.2.2 Sequencing

The whole genome sequencing of the DNA samples was conducted on Illumina Genome Analyzer IIx, which is based on principle "sequencing by synthesis". The protocol followed in this sequencing platform briefly involves - random fragmentation of DNA, ligation of customized adapters to DNA fragments and these were immobilized on a solid support called "flow cell" to create a sequencing library. For these fragments 'bridge amplification – solid phase PCR' is carried out to generate a high number of clonal DNA fragments on the flow cell. The bases in these amplified fragments are read in a cyclic manner by the sequencer utilizing fluorescent reversible nucleotide terminators and signal emitted upon incorporation of each nucleotide is recorded by a charge coupled device as depicted in Figure 2.

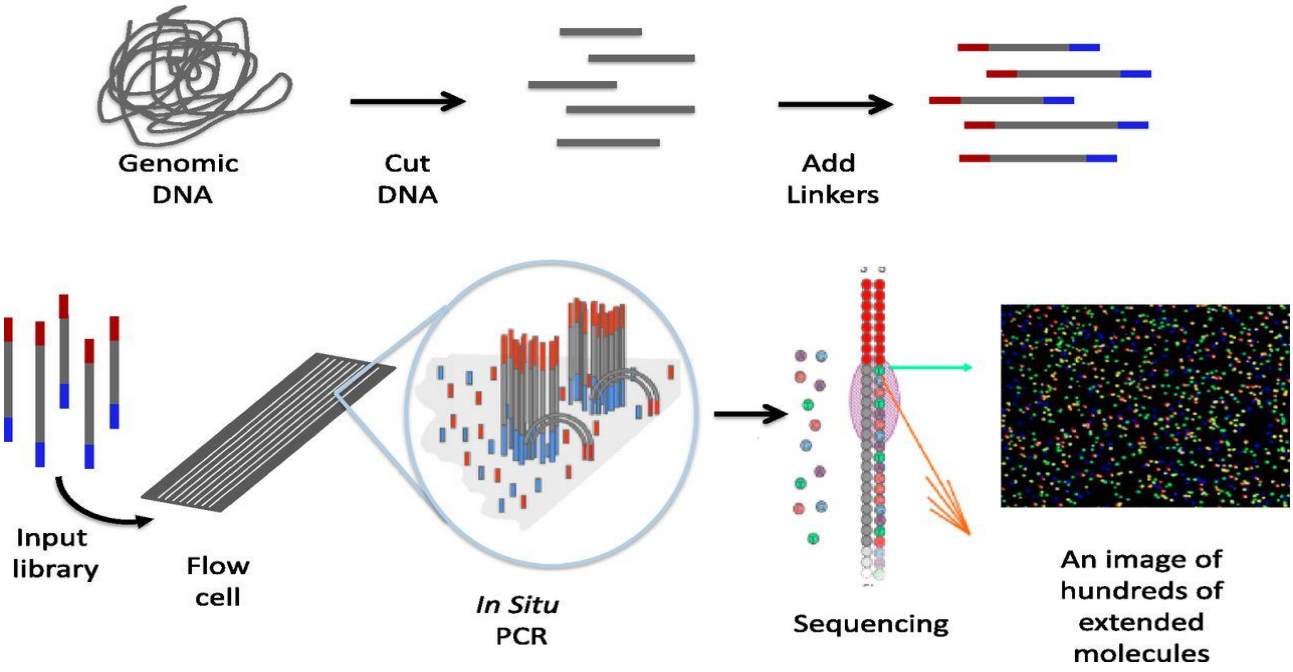


Figure 2: Schematic flow of Illumina sequencing

Source: bloodjournal.org

2.2.3 Quality Control of Data

The data from the Sequencer is generated in the form of high-resolution images which are then converted into .bcl files using proprietary software of Illumina. The read data is further converted to .qseq format and finally into FASTQ file format. The paired end read data in FASTq format was obtained for all genomes. A typical FASTQ file consists of the following information as shown in Figure 3.

The following information can be gleaned from a FASTq file –mean read length, total number of reads, total number of bases, reads with non-ATGC characters etc. The average quality score and nucleotide composition at each base position along the read was plotted in the form a graph with based on NGS QC Toolkit (Patel et al, 2012). In some cases, trimming of reads from one end was needed after visualizing the quality score at each base position along the read.

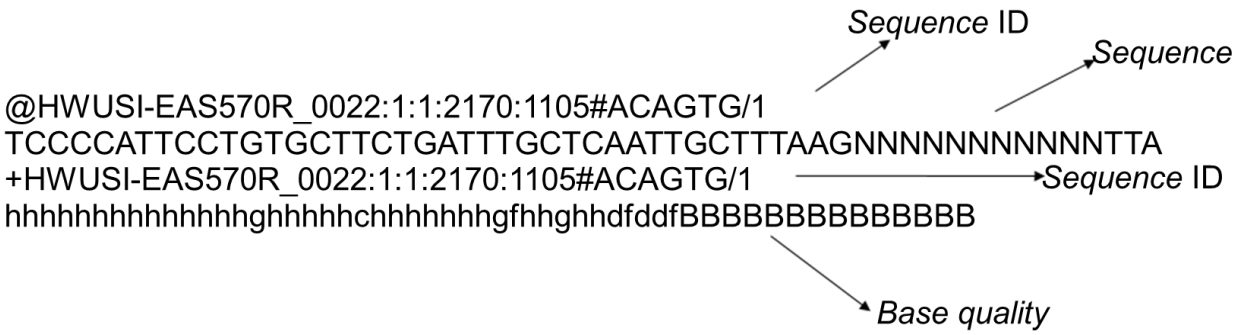


Figure 3: Schematic representation of a FASTq file

2.2.4 Genome assembly

For this study, *De Bruijn* graph based tool "velvet" was chosen, because of its efficiency in handling shorter read data with minimal RAM requirements when compared to other available algorithms (Zhang et al., 2011). This assembly algorithm mainly involves two steps (Zerbino and Birney, 2008), first one is hashing of reads which is done by a velveth command and this step produces two files- Sequences and Road maps. In the next step using velvetg command, a *De Bruijn* graph is built from hashes obtained in previous step and further error correction is applied on it to finally generate a contigs fasta file. The hash length/K-mer value denotes length of words being hashed and optimization of this parameter is necessary considering the specificity and sensitivity of the assembly. Therefore assembly is run for a range of user defined K-mer values, usually odd numbers to avoid palindromes and max K-mer value is always less than the read length. Therefore the procedure was repeated using the following command line for various hash lengths till an optimum assembly was generated for each strain.

#####command line start #####

~/Programs/velvet_1.1.05/shuffleSequences_fastq.pl Read1.fastq Read2.fastq shuffledsequences.fastq

~/Programs/velvet_1.1.05/velveth hashes K1,K2,2 -fastq -shortPaired shuffledsequences.fastq

~/Programs/velvet_1.1.05/velvetg hashes_K1/ -ins_length_sd 20 -ins_length 300 -exp_cov auto -min_contig_lgth 100 -cov_cutoff auto -read_trkg yes -scaffolding no -unused_reads yes -alignments yes

#####command line end #####

Here K1, K2 denotes the range of K-mer values considered which are incremented at each step by 2 and velvetg step is run individually for each K-mer value. Further scaffolding of contigs using paired end read information was performed by the tool SSPACE. GapFiller was used for closure of gaps within scaffolds.

2.2.5 Annotation

The annotation of each strain was carried out using RAST (Rapid Annotation using Subsystem Technology) (Aziz et al., 2008). The tRNAscan-SE and RNAmmer were used respectively for the detection of tRNA and rRNA(Schattner et al., 2005)(Lagesen K Rødland E, Stærfeldt HH, Ussery DW RT, 2000). Further all the genome sequences were submitted to NCBI in the following way:

Step 1: A new submission was registered by providing necessary details for each genome sequence and obtained Bioproject ID (PRJNAxxxx) from NCBI.

Step2: Preprocessing of Files.

- Template file(.sbt) : This file contains strain information, submitter's details and Bioproject ID
- Contig file (.fsa) : This contains contigs not less than 200bp with modified header like this

>contig001 [organism=*Salmonella* Typhi] [strain= BL196] [host=Homo sapiens] [isolation-source= stool sample] [country=Malaysia] [collection-date=2005]

Step 3: Creation of .sqn file for submission

The executable **tbl2asn** from NCBI tool box has been used to create a sqn file needed for submission using both the above files (.sbt and .fsa).

2.3 Results

2.3.1 Quality control

The following plots of average quality score and nucleotide composition at each base position of the read were generated for both the reads of all genomes using the above mentioned Script1. The high quality bases are defined as those with phred score value of greater than 20. Reads with minimum 70% of bases having phred score of greater than 20 are defined as high quality reads. With these thresholds, Read1 and Read2 of all strains are filtered separately to extract high quality reads and detailed quality control report is shown in Table 2.

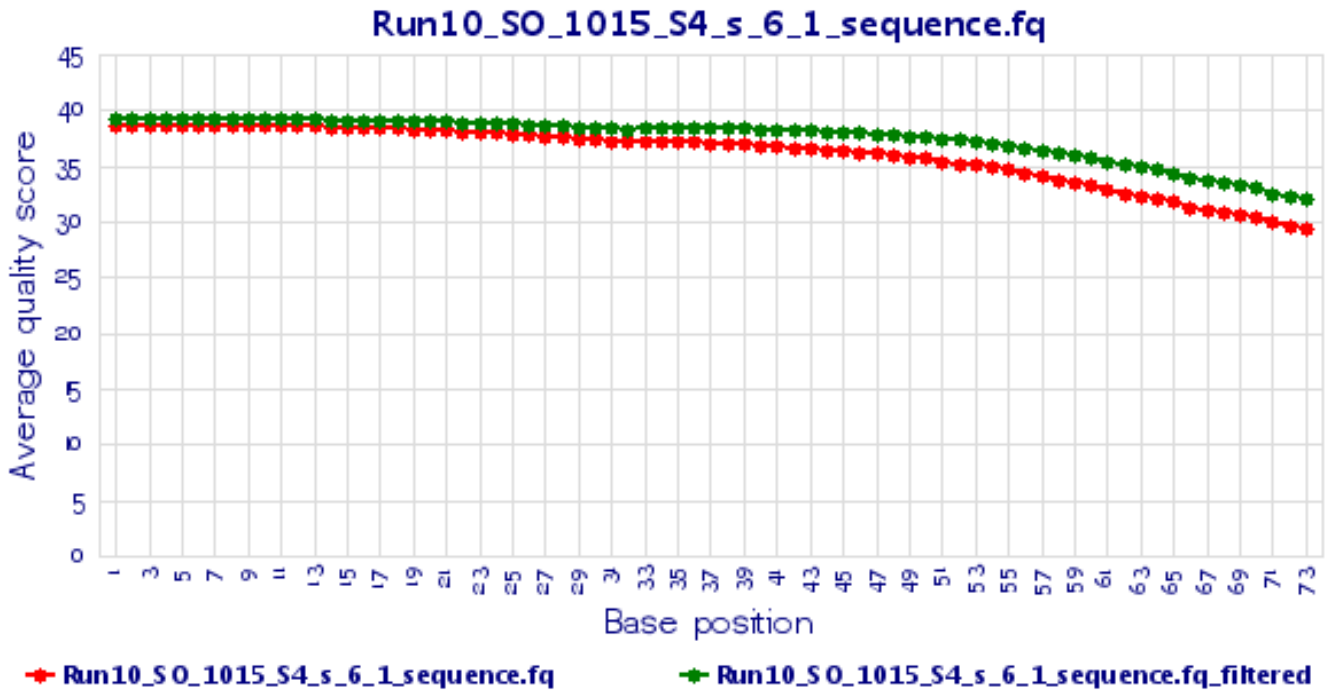


***Figure 4*: Quality score values along the read**: *The plot represents the quality scores of each base along the Read1 of strain UJ308A*

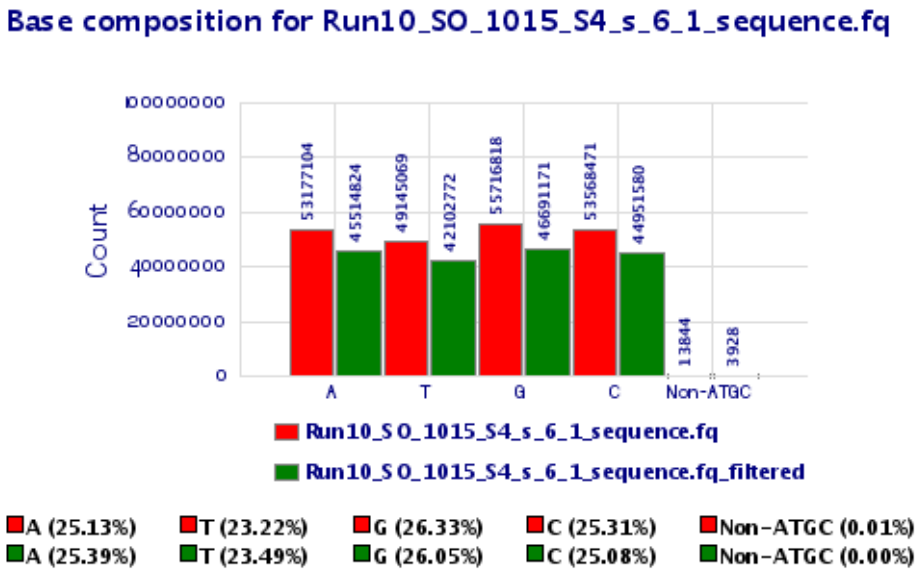


***Figure 5**: **Nucleotide distribution along the read**: The figure represents the percentage of nucleotides A, T, G, C along with N at each base position along the Read2 of strain UJ308A.*

Fastq file name	UJ308A R1	UJ308A R2	UJ816A R1	UJ816A R2	CR0063 R1	CR0063 R2	BL196 R1	BL196 R2
Max Read Length	73	73	73	73	100	100	73	73
Min Read Length	73	73	73	73	100	100	73	73
Mean Read Length	73	73	73	73	100	100	73	73
Total No. of Reads (x10^6)	3.85	3.85	4.24	4.24	1.63	1.63	2.89	2.89
No. of HQ Reads (x10^6)	3.35	3.49	3.70	3.86	1.36	1.36	2.75	2.74
% of HQ Reads	86.96	90.80	87.16	91.06	83.40	83.80	95.09	94.58
No. of Bases (x10^8)	2.81	2.81	3.09	3.09	1.63	1.63	2.11	2.11
No. of HQ Bases (x10^8)	2.50	2.58	2.76	2.85	1.42	1.41	2.02	2.01
% of HQ Bases	89.12	91.85	89.27	92.01	87.24	86.75	95.71	95.14
Sequences with Ns	183809	106246	202722	117317	51990	22977	11443	7209
% Seq with Ns	4.77	2.76	4.78	2.76	3.19	1.41	0.39	0.25
No. of 5' Adapter	118916	118388	129201	128450	72254	70863	90157	76498
No. of 3' Adapter	87555	88494	93347	93897	54466	54646	76921	65791
% 5' Adapter	3.086	3.072	3.04	3.026	4.428	4.343	3.110	2.639
% 3' Adapter	2.27	2.30	2.20	2.21	3.34	3.35	2.65	2.27
% of A	25.77	25.81	25.56	25.57	25.93	25.84	25.13	23.18
% of T	25.30	25.26	25.13	25.13	25.40	25.38	23.22	24.81
% of G	24.39	24.35	24.56	24.48	24.33	24.34	26.33	25.45
% of C	24.48	24.39	24.69	24.64	24.30	24.38	25.31	26.42
% of Non-ATGC	0.067	0.18	0.07	0.18	0.04	0.06	0.01	0.13

***Table 2**: Quality control report of sequence read data of all strains*

2.3.2 Genome Assembly

For all the strains, assembly was generated for various hash lengths in order to choose an optimum one which has a higher N50 value, lower number of contigs and utilises most of the read data. The N50 metric represents that contig length, for which all the contigs equal to or greater than that contig length collectively would represent approximately half of the genome length.

Hash length	Contigs	N50	Max contig size	Approx. genome size	Reads used	Total reads
27	**307**	**31298**	**121510**	**4661306**	**5231116**	**5279936**
29	**256**	**43519**	**106110**	**4673161**	**5235533**	**5279936**
31	**227**	**43687**	**130982**	**4676753**	**5235679**	**5279936**
33	**222**	**46029**	**124797**	**4677350**	**5239304**	**5279936**
35	**217**	**44714**	**124795**	**4680506**	**5236661**	**5279936**
37	**211**	**46589**	**120312**	**4688040**	**5230902**	**5279936**
39	**216**	**46134**	**120312**	**4687227**	**5221000**	**5279936**
41	**226**	**42475**	**120312**	**4689630**	**5209901**	**5279936**
43	**244**	**39360**	**120312**	**4689591**	**5195786**	**5279936**
45	**264**	**39360**	**119802**	**4685932**	**5178395**	**5279936**
47	**281**	**35666**	**119802**	**4682626**	**5163755**	**5279936**
49	**307**	**32293**	**119802**	**4681843**	**5148793**	**5279936**
51	**345**	**28047**	**102835**	**4679764**	**5132936**	**5279936**

Table 3: *The assembly statistics obtained for strain BL196 at various hash lengths considered.*

A clear trade off of various parameters for different hash lengths is observed. For example in the above case, hash length 37 shows optimum values for all the parameters like lower number of contigs and higher N50 value etc. Therefore contigs (>200bp) obtained at this hash length were further considered for analysis. Further, the assembly of all strains was refined using most updated versions of the tools which became available during the course of study. The scaffolding tool SSPACE was used to further merge the contigs wherever possible by utilising paired end information. In this way an improved assembly was obtained for all the strains and their statistics are shown below.

Strain Name	Hash length	Contigs	Scaffolds	N50	Max contig size	Approx. genome size	Reads used	Total reads
UJ308A	**39**	**415**	**337**	**20066**	**98328**	**4605234**	**6089035**	**6252544**
UJ816A	**39**	**334**	**259**	**27577**	**105361**	**4618857**	**6726536**	**6916574**
CR0063	**37**	**540**	**532**	**15671**	**105140**	**4534218**	**3106362**	**3263270**
BL196	**37**	**189**	**159**	**46652**	**120312**	**4688286**	**5217433**	**5279936**

Table 4: *The refined assembly statistics of all strains obtained using velvet and SSPACE.*

2.3.3 Annotation and submission to NCBI

The draft genomes of all the strains were annotated using RAST server which is based on subsystem technology.

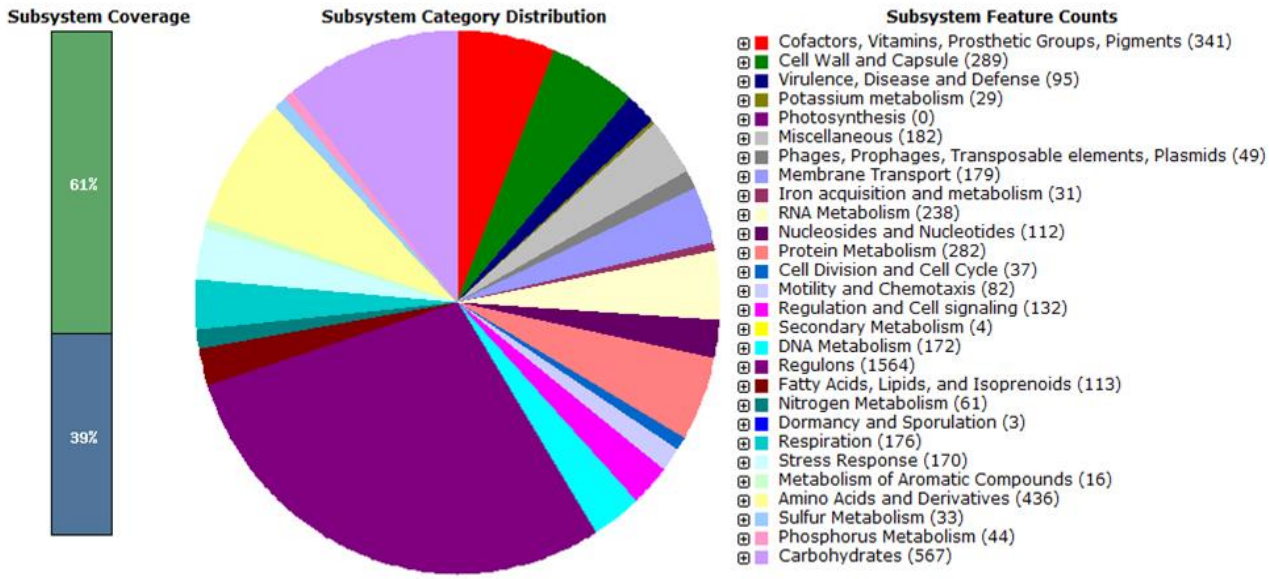


Figure 6: Subsystem distribution of ST CR0063: *The subsystem statistics of ST CR0063 based on genome annotations performed according to RAST conventions*

The prediction of tRNA and rRNA genes was done using tRNAscan-SE and RNAmmer respectively. The genbank files obtained after annotation were supplied to Artemis in order to glean the following statistics as shown in Table 5.

Strain name	% of GC content	Total CDS	Average CDS length	% of CDS	t-RNA	r-RNA
UJ308A	**51.89**	**4720**	**869**	**86.8**	**78**	**22**
UJ816A	**51.94**	**4710**	**873**	**86.8**	**77**	**22**
CR0063	**51.7**	**4946**	**798**	**86.1**	**77**	**22**
BL196/05	**53.2**	**4875**	**875**	**87.1**	**76**	**22**

Table 5: Genome Statistics

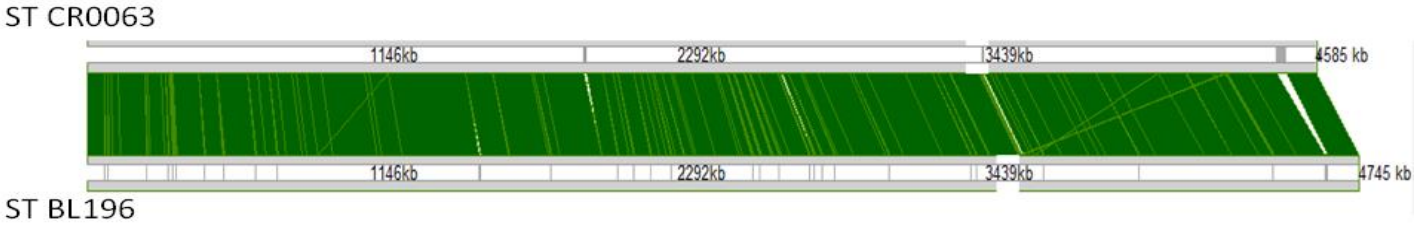
After the WGS submissions of these genomes, following accession numbers were obtained as shown in Table 6.

Strain name	BioProject ID	Accession number
UJ308A	**PRJNA157357**	**AJTD00000000**
UJ816A	**PRJNA157359**	**AJTE00000000**
CR0063	**PRJNA167146**	**AKIC00000000**
BL196/05	**PRJNA85621**	**AJGK00000000**

Table 6: NCBI accession numbers

2.4 Discussion

These strains were observed to be closely related to *Salmonella* Typhi strain CT18 upon BLAST analysis. These strains were also showed high level of similarity among themselves – a comparison of outbreak strain BL196 with the carrier strain CR0063 is shown below.



.

Figure 7: Comparison of Salmonella Typhi strains ST CR0063 and ST BL196*: Comparison of whole genome sequences of S. Typhi strains using MGCAT(Treangen and Messeguer, 2006) – 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.*

The genome sequences revealed two mar regulons, marRAB and marC, as reported also in Salmonella Typhi strains CT18 and Ty2, and these were homologous to Escherichia coli mar (multiple antibiotic resistance) regulon members (Alekshun and Levy, 1999) . The genes encoding a melittin resistance protein (PqaB), polymyxin resistance protein (PmrD) and methyl viologen resistance gene (smvA) were also located (Baker et al., 1999) (Hongo et al., 1994). The homologues of Campylobacter toxin cdtB and Bordetella pertussis toxin were identified in genomes of strains UJ308A and UJ816A (Haghjoo and Galán, 2004). They also encoded major virulence factors that are identified in *Salmonella* and enlisted in VFDB (Virulence Factors Database) (Yang et al., 2008). Gene shdA, a key factor predicted to be involved in persistence of the bacterium in the intestines by binding to its extracellular matrix, was identified and annotated (Kingsley et al., 2002). 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 fecal shedding (Kingsley et al., 2000). The fim gene cluster of chaperone –usher family involved in

adhesion to non-phagocytic cells was detected along with its negative regulator fimW (Muscas et al., 1994). Type IV pili and agf operon encoding curli fimbriae which aid in attachment of the bacterium to intestinal villi and also with each other, were found (Craig et al., 2004) (Collinson et al., 1996). These adherence factors determine the sites of bacterial colonization and thereby adaptation and pathogenicity of a particular strain (Duncan et al., 2005). The genomes 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 (Virlogeux et al., 1995). The mgtC gene involved in Magnesium uptake and ferric uptake regulators (fur) were also identified (Moncrief and Maguire, 1999). The PhoPQ regulon which induces cytokine secretion and cationic antimicrobial peptide resistance was also found (Guo et al., 1997). The RpoS sigma factor needed to cope up with external stress and nutrient depletion conditions was also identified and annotated (Chen et al., 1996). The co-ordinates of these virulence factors in the genome of ST CR0063 are depicted in Figure 8. These genomes sequence information can be further harnessed in comparative genomic studies to obtain important insights into the pathogenesis and evolution of this pathogenic bacterium.

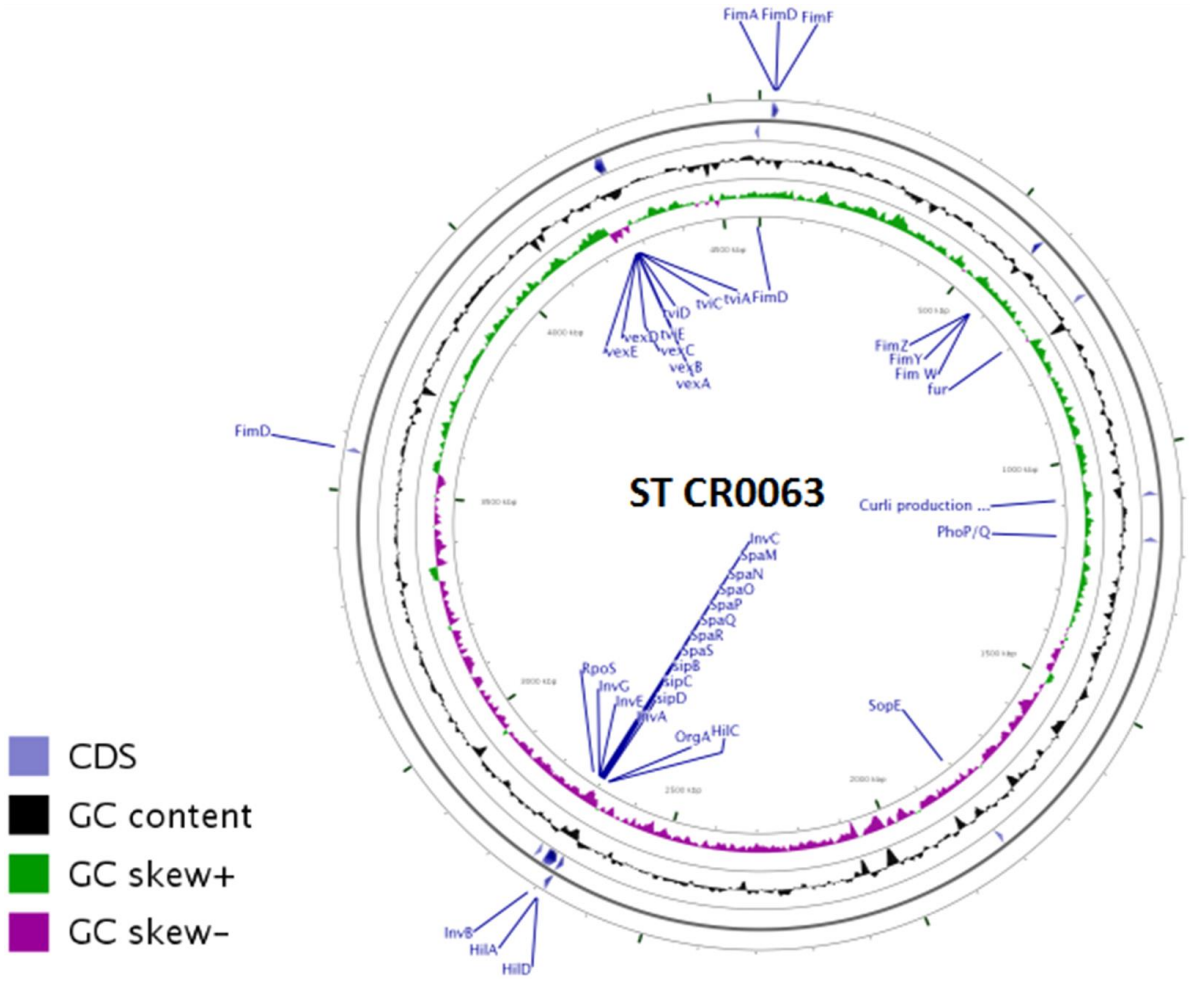


Figure 8: Circular Genome view of ST CR0063. *Positions of some of the major virulence factors and their regulators identified in ST CR0063marked in the circular genome generated using CGview (Stothard and Wishart, 2005)*

Chapter 3

Determination of the pan-genome structure and its boundaries in *Salmonella* Typhi with insights into adaptation mechanisms

---

Part of this chapter was published as:

Genome dynamics and evolution of *Salmonella* Typhi strains from the typhoid endemic zones. (2012) **Baddam R**, Kumar N, Shaik S, Lankapalli AK, Ahmed N. **Sci Rep**. 4:7457.

3.1 Introduction

The pan-genome of a species represents the complete inventory of genes in the population and is always significantly greater than the gene content of any individual (Rodriguez-Valera F, 2012). The pan-genome concept also has been applied at varied taxonomic levels like serovar, genus, kingdom etc. in previous studies (Lapierre and Gogarten, 2009). 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 (Tettelin et al., 2005). 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 (Vernikos et al., 2014). Thus the pan-genome analysis helps us to better understand the genomic diversity and provides cues about the mechanisms underlying adaptation and evolution of bacteria. Studies based on this concept using multiple whole genome sequences of other *S. enterica* serovars such as Paratyphi A and Agona have recently provided significant insights into evolution of these serovars (Zhou et al., 2014) (Zhou et al., 2013). Further this kind of analysis at species level carried out on *Salmonella enterica* could identify some novel gene families specific to certain serovars (Jacobsen et al., 2011).

In the past, various genomic studies have attributed signatures like pseudogenisation (loss of gene function) or gene deletion for host restriction in pathogenic bacteria (Bäumler and Fang, 2013). It was also reported that in human-restricted serovar *S.* Typhi, pseudogenisation is an active process compared to other generalist serovars like *S.* Typhimurium (McClelland et al., 2001). Further, the extent of this pseudogenisation also varies considerably even among host restricted serovars. Pseudogenes constitute upto 4.5% of *S.* Typhi gene pool, making them an important driver of genome reassortment over time (Holt et al., 2008). 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. In total genomes of eight strains were analyzed in this study which are associated with different clinical manifestations - outbreaks, sporadic cases, carrier strains and fatal episodes. Along with four strains described in previous chapter, four other strains from NCBI were also included. These include genomes of MDR strain CT18 from Vietnam, strain P-stx-12 isolated from a carrier individual in India, strain ST0208 associated with a sporadic case in Kuala Lumpur, Malaysia and strain CR0044 isolated from a carrier individual in 2007 in Kelantan, Malaysia (Parkhill et al., 2001) (Ong et al., 2012) (Yap et al., 2012b) (Yap et al., 2012a). The two strains CR0044 and CR0063 isolated from carrier individuals were known to share PFGE profile with the strain BL196. Few other strains from Southeast Asia which are available in NCBI could not be included in the analysis because of the low coverage and poor quality of data.

3.2 Methodology

3.2.1 Refinement of assembly and annotation

The contigs of draft genomes were further combined to scaffolds based on paired end read information. These were initially ordered according to a reference using standalone BLAST. Further high quality filtered reads of respective strains were mapped to these ordered scaffolds using BWA alignment tool (Li and Durbin, 2009). The alignment file thus generated for each genome was visualized using Tablet alignment viewer program (Milne et al., 2010) to sort and validate the scaffolds order manually based on paired-end read information. The distance between the read pairs known as insert length was utilized to confirm the order of the scaffolds; i.e if two scaffolds are in correct order then the pair at the end of the scaffold should have its mate at the start of the following scaffold as shown in Figure 9. Further this approach was also useful to resolve and place certain repetitive elements correctly. The regions with low coverage were also inspected before including them into final genome.

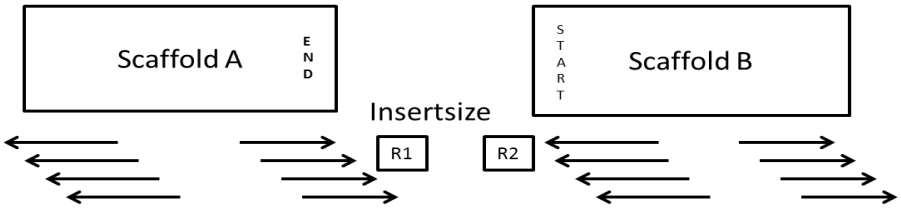


Figure 9: Schematic representation of procedure adopted for validation of scaffold order using paired-end read information.

After finalizing the order of these scaffolds 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, in order to avoid erroneous extension of annotation. The genome thus obtained was submitted to ISGA pipeline for annotation (Hemmerich et al., 2010). The two complete genomes CT-18 and P-stx-12 were also re-annotated to homogenize the data with a single annotation platform. The whole genome alignment of all these eight genomes was generated using progressiveMauve with all default settings (Darling et al., 2010).

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, an inbuilt module of PanOCT was used. The BLASTP result of query ORFs against the functional genes of *Salmonella* strains was provided as input to PanOCT (Fouts et al., 2012). 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 is considered in the final pseudogene list, so that the number will not be over represented.

3.2.2 Phylogenomic analysis

The whole genome based phylogeny was performed for all eight strains using Gegenees (version 1.1.4) (Agren et al., 2012) which employs a fragmented all-against-all comparison of the genomes and builds a distance matrix file suitable to construct a phylogenetic tree and heatmap. The BLASTN algorithm was considered for this comparison with a step size of 100. The distance matrix file produced is exported in nexus format for phylogenetic tree construction by NJ (Neighbor-joining) method using SplitsTree software (version 1.1.4) (DH, 1998).

The core genome based phylogeny was also built using core gene clusters without paralogs. The nucleotide sequences of these orthologous core genes were aligned using MAFFT (Katoh and Standley, 2013) followed by removal of alignments gaps using TrimALL (Capella-Gutiérrez et al., 2009). These pruned alignments were concatenated using PERL script and were supplied as an input to RAxML (Stamatakis, 2014) for phylogenetic analysis. The core genome based phylogenetic tree was constructed using GTR nucleotide substitution model with gamma correction. A consensus Maximum Likelihood phylogenetic tree was constructed after 1000 replicates/bootstraps.

3.2.3 Detection of Mobile elements

All of these strains included various mobile phages or phage like elements. To identify these elements, *PhiSpy* (Akhter et al., 2012), an algorithm that combines both similarity and composition based strategies, was used. These predictions were compared with results obtained from PHAST web server (A Fast Phage search tool) (You Z Karlene L, Jonathan J. D, David S. Wishart, 2011). Insertion sequence (IS) elements were identified using IS finder (Siguier et al., 2006). All these tools were run with default settings by using nucleotide fasta sequences. The genomic islands in these strains were identified using IslandViewer (Langille and Brinkman, 2009) by submitting genbank files.

3.2.4 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 (Li et al., 2003) 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 (Tettelin et al., 2008).

Initially orthologous clusters of all strains were generated with the percent match threshold of 85% and minimum protein length of 50 amino acids (aa). Later the functional genes and pseudogenes were analyzed separately as mentioned by (Liang et al., 2012), 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 50aa whereas for pseudogenes it was set to 10aa. The extrapolations of pan-genome and core genome of functional and pseudogenes was done as mentioned above, but individually for each of them.

3.2.5 COG Functional classification

The query genes were aligned against a CDD database (Marchler-Bauer et al., 2011) downloaded from the NCBI COG database using RPS blast and the error value threshold as 0.00001. The best

hit for each gene and its corresponding COG was extracted. The genes which failed to produce a significant hit were considered as hypothetical genes. The statistical two sample z-proportionality test was applied to calculate the significance for the enrichment of various functional classes.

3.2.6 Detection of SNP in core Genome

The core functional gene clusters are 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 (Larkin et al., 2007).

3.3 Results

3.3.1 Phylogenomic analysis

The whole genome based phylogenetic tree allowed us to understand the close genetic relationship among various strains as shown in Figure 10. The strain BL196 isolated during the outbreak, and the carrier strain CR0044 isolated a year later, co-clustered revealing close similarity.

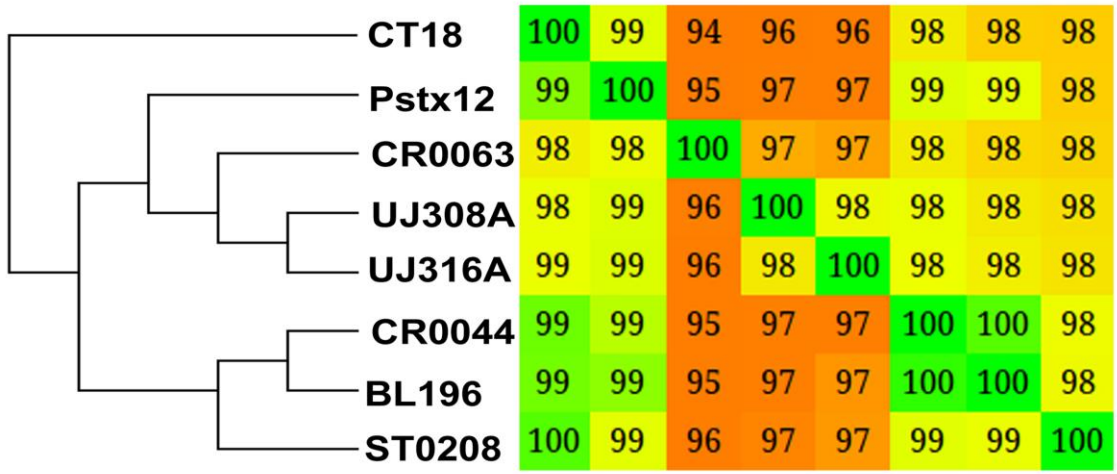


Figure 10: Phylogenomic Tree: *The whole genome information was used to build the distance matrix using Gegenees. The phylogenetic tree was developed using SplitsTree by NJ method. This revealed close similarity among genomes and also co-clustering of strains isolated from the same regions.*

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 (Thong et al., 1996). 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 (Thong et al., 1996). The strains CT18, P-stx-12, ST0208 have shown up independently in the tree.

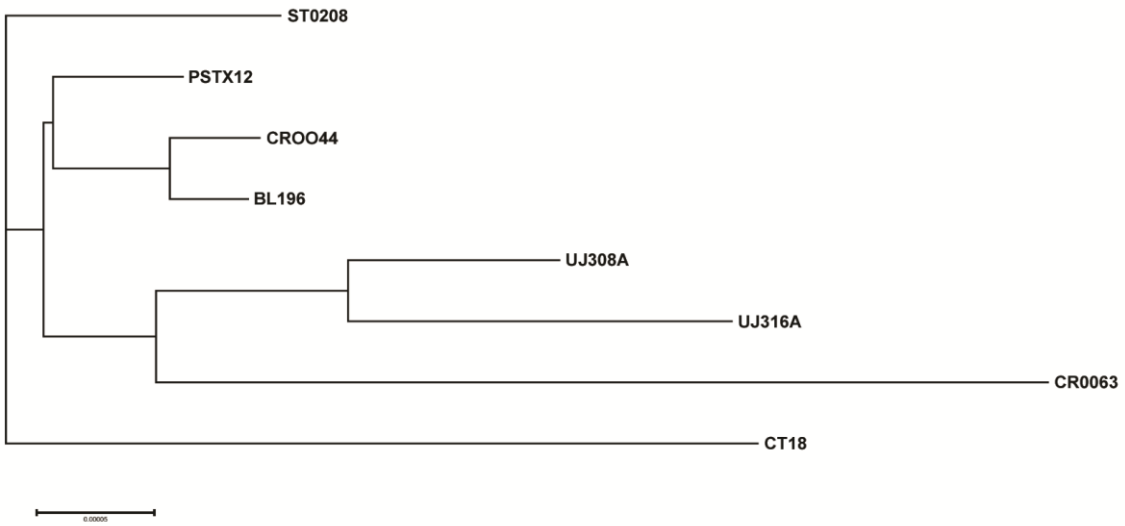


Figure 11: Core genome based phylogeny: *A core genome based consensus Maximum Likelihood phylogenetic tree constructed after 1000 replicates/bootstraps.*

A similar co-clustering pattern was also observed with Maximum Likelihood based phylogenetic tree constructed using core gene clusters without paralogs as shown in Figure 11. This analysis once again reinforces the genetically monomorphic nature of this pathogen and our observations are in concurrence with the previous findings based on MLST and other techniques (Holt et al., 2008) (Kidgell et al., 2002) . The close similarity of these genomes is also reflected in whole genome

alignment shown in Figure 12, where each single colored block represents similar sequences in the respective genomes.

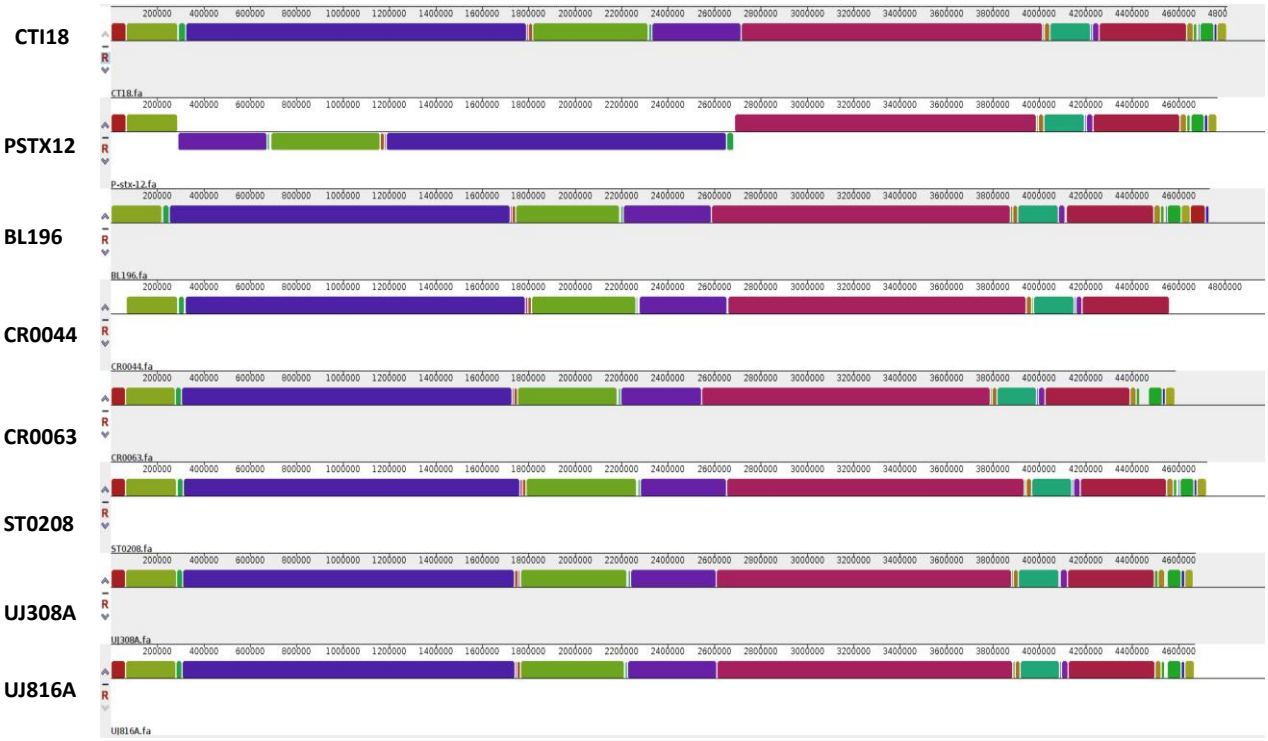


***Figure 12*: *Genome alignment*:** *The whole genome alignment of all eight genomes was generated using progressiveMauve. Each colored block represents similar sequences in the respective genomes.*

3.3.2 Mobile elements

The phages and Insertion Sequence (IS) elements of the two complete genomes CT18 and P-stx-12 have been already reported (Parkhill et al., 2001) (Ong et al., 2013). The IS elements belonging to the family IS200/605, IS3, IS256 were commonly observed in all the other 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 phages together with various phage remnants in each of the genomes as listed in Table 7. Gifsy-2 and fels-2 phages 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 was used to represent the status of major Salmonella pathogenicity islands in these genomes as shown in Figure 13 (Alikhan et al., 2011). The major variations were observed in SPI6 and SPI 10. These two islands are known to encode fimbrial operons, the changes in these surface structures are known to have implications in virulence.

Strain name	Putative phage	Completeness	size(kb)
BL196	**Gifsy_2_NC_010393**	**Intact**	**44.6**
	Entero_Fels_2_NC_010463	**Intact**	**50**
	Entero_Fels_2_NC_010463	**Intact**	**32.7**
	Cronob_vB_CsaM_GAP32_NC_019401	**Incomplete**	**29**
	Entero_cdtI_NC_009514	**Incomplete**	**6.1**
CR0044	**Gifsy_2_NC_010393**	**Intact**	**44.6**
	Entero_Fels_2_NC_010463	**Intact**	**50.1**
	Entero_Fels_2_NC_010463	**Intact**	**32.7**
	Cronob_vB_CsaM_GAP32_NC_019401	**Incomplete**	**28.9**
	Entero_cdtI_NC_009514	**Incomplete**	**6.1**
CR0063	**Gifsy_2_NC_010393**	**Intact**	**44.6**
	Entero_Fels_2_NC_010463	**Intact**	**49.6**
	Salmon_RE_2010_NC_019488	**Intact**	**46.8**
	Cronob_vB_CsaM_GAP32_NC_019401	**Incomplete**	**26.9**
	Entero_cdtI_NC_009514	**Incomplete**	**6.1**
ST0208	**Psychr_pOW20_A_NC_020841**	**Intact**	**45.1**
	Gifsy_2_NC_010393	**Intact**	**44.6**
	Entero_Fels_2_NC_010463	**Intact**	**50.7**
	Entero_Fels_2_NC_010463	**Intact**	**32.7**
	Cronob_vB_CsaM_GAP32_NC_019401	**Incomplete**	**28.8**
	Entero_cdtI_NC_009514	**Incomplete**	**6.1**
	Entero_Fels_2_NC_010463	**Incomplete**	**18.4**
UJ308A	**Gifsy_2_NC_010393**	**Intact**	**44.6**
	Entero_Fels_2_NC_010463	**Intact**	**52.3**
	Salmon_RE_2010_NC_019488	**Intact**	**51.7**
	Cronob_vB_CsaM_GAP32_NC_019401	**Incomplete**	**28.3**
	Entero_cdtI_NC_009514	**Incomplete**	**6.1**
	Entero_P4_NC_001609	**Incomplete**	**12.8**
UJ816A	**Gifsy_2_NC_010393**	**Intact**	**44.6**
	Entero_Fels_2_NC_010463	**Intact**	**50.1**
	Entero_Fels_2_NC_010463	**Intact**	**30.9**
	Entero_P4_NC_001609	**Incomplete**	**12.7**
	Entero_cdtI_NC_009514	**Incomplete**	**6.1**
	Cronob_vB_CsaM_GAP32_NC_019401	**Incomplete**	**28.9**
	Entero_P4_NC_001609	**Incomplete**	**10.6**

***Table 7*: *Phage elements*:** *The table enlists all the major phage related elements identified in the compared S. Typhi strains.*

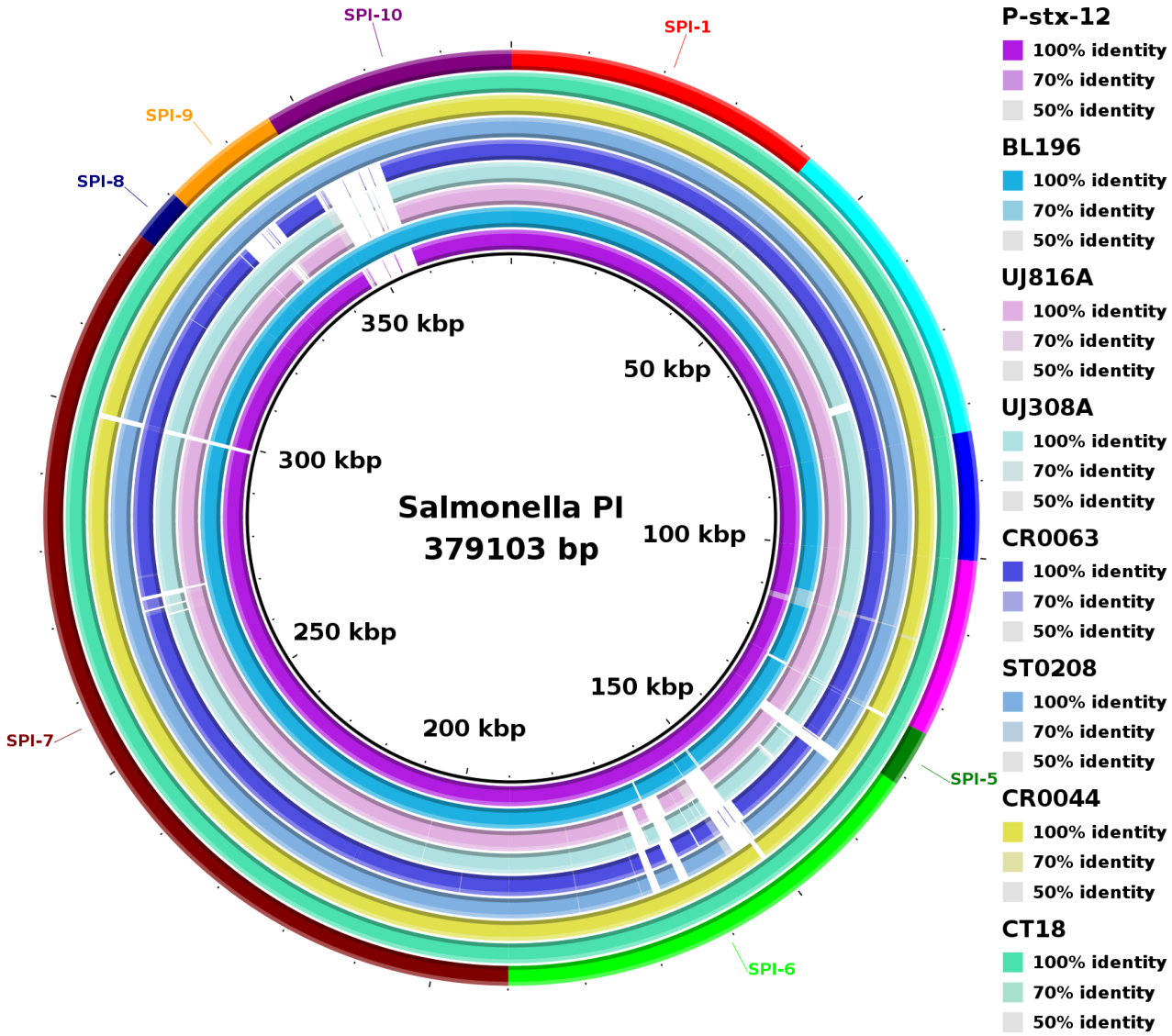


Figure 13*: *Comparative analysis of Salmonella Pathogenicity islands: *The status of major pathogenicity islands of Salmonella Typhi was identified among all the strains. Each inner ring represents an individual genome in a particular color along with their percentage identity level shown in gradation of color. The pathogenicity islands are marked in the outer ring.*

A list of genomic islands detected in these genomes as well as of the genes encoded by them is provided (Excel sheet 1). 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 strains along with the orthologous genes shared by them have already been discussed previously (Parkhill et al., 2001) (Ong et al., 2013).

3.3.3 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 (Tettelin et al., 2008). The Heap's law can be represented by the equation 1: $n = k * N^{-\alpha}$, where n is pan-genome size, N is the number of genomes and k, γ are constants for a specific curve where $\alpha = 1 - \gamma$. The exponential term α determines whether pan-genome of a bacterial species is closed or open. For α >1 (γ <0) the pan-genome is considered closed i.e. sampling more genomes will not affect the pan-genome size, whereas for α <1 (0<γ >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 *Salmonella* strains revealed an alpha value of 0.913 implying a highly conservative nature of these endemic isolates (Figure 14a).

Further to investigate the effect of pseudogenisation on gene frequency distributions of functional and pseudogenes, pan and core genomes of these 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 14c).

The values K and γ for functional gene clusters after curve fitting were determined as 4054 and 0.06 respectively, with an 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 14b, 14c), 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 14c and thus reemphasizes that pseudogenisation of functional genes is an ongoing process in *S*. Typhi.

3.3.4 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 tends 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 2 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 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 14a). The core genome of functional and pseudogenes was also determined separately and these distributions gave 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 14b) with the convergence value $tg(\theta)$ as 3495 obtained upon solving the equation.

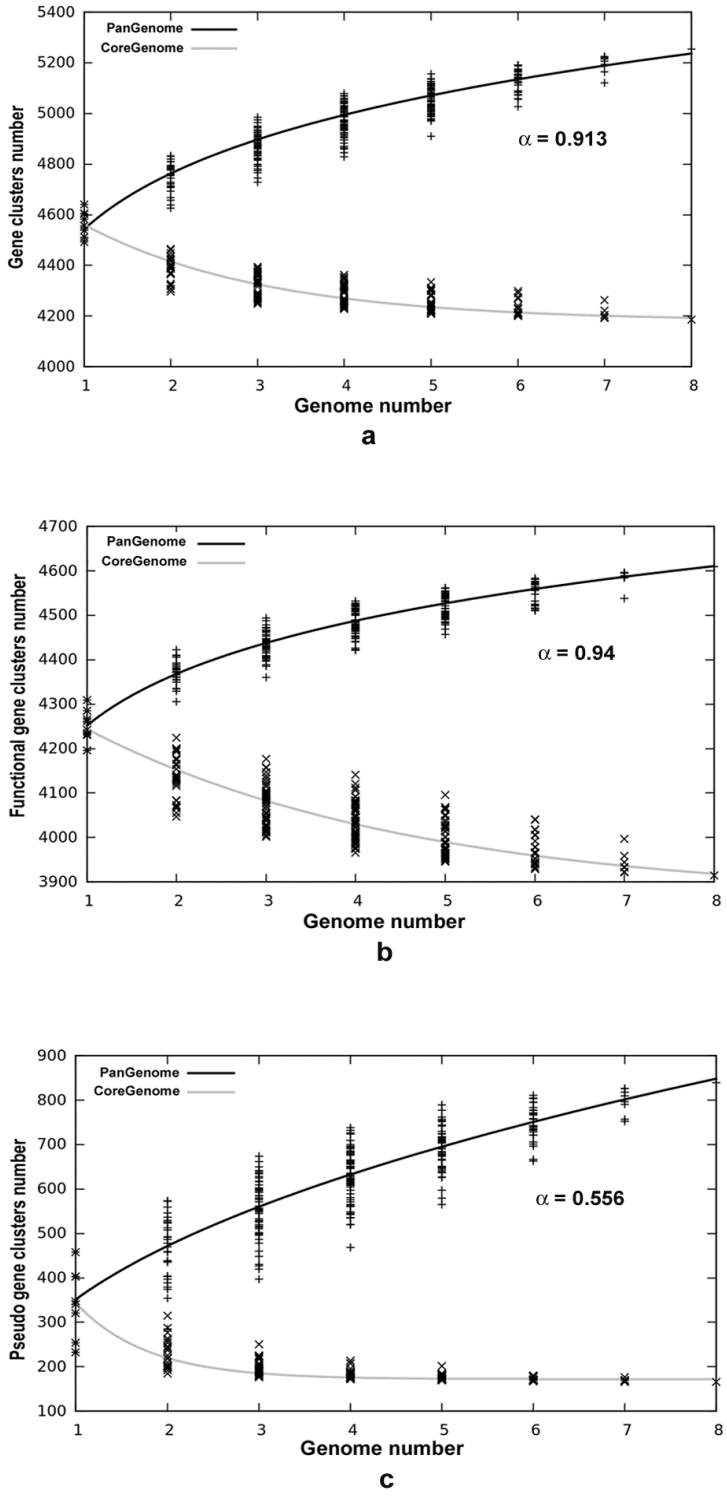


Figure 14: Pan and Core Genome Distribution: *(a). Pan and core genome developments using median values of the combinations of all eight genomes. (b). Pan and core genome developments of functional genes of these eight isolates. Here it can be observed that core genome is decreasing sharply. (c). Pan 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.*

However in the case of pseudogenes, the core genome has already reached its convergence with a $tg(\theta)$ value of 166 genes (Figure 14c). Further, the core genome profiles of functional and pseudogenes (Figure 14b, 14c), 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 included various fimbrial proteins, methyl accepting chemotaxis proteins and certain secreted effector proteins. Some of these 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 (McClelland et al., 2004) (Nuccio and Bäumler, 2014).

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, amino acid transport and metabolism, inorganic ion transport etc (Figure 15a). 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 were observed in the same functional categories related to metabolic functions as observed in case of pseudogenes (Figure 15b). This enrichment of pseudogenes observed in the metabolism related genes was also statistically significant according to the proportionality z-test with P value of 0.0001.

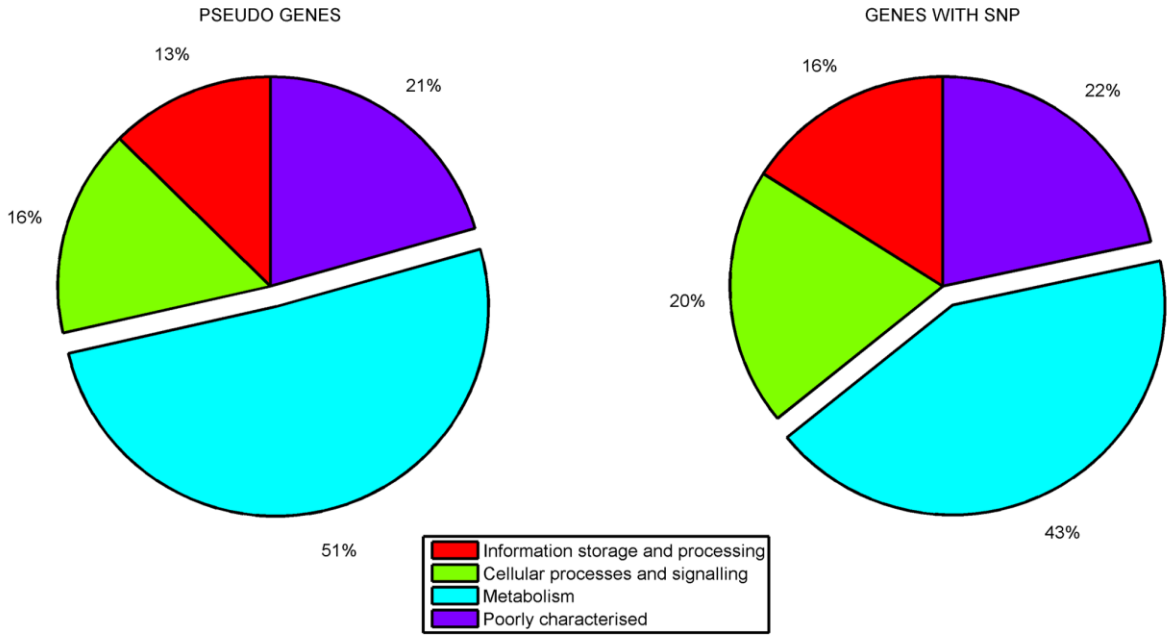


Figure 15: The proportion of functionally classified pseudogenes and the functional genes with SNPs according to COG classification: *The pie chart represents the proportion of various functional classes among the pseudogenes and the functional genes with SNPs. The figure clearly shows the enrichment of metabolism related genes in pseudogenes and the functional genes with SNPs.*

Thus, from our observations, as depicted (Figure 15a, 15b), 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.

3.3.5 Accessory pseudogene content analysis

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 16)

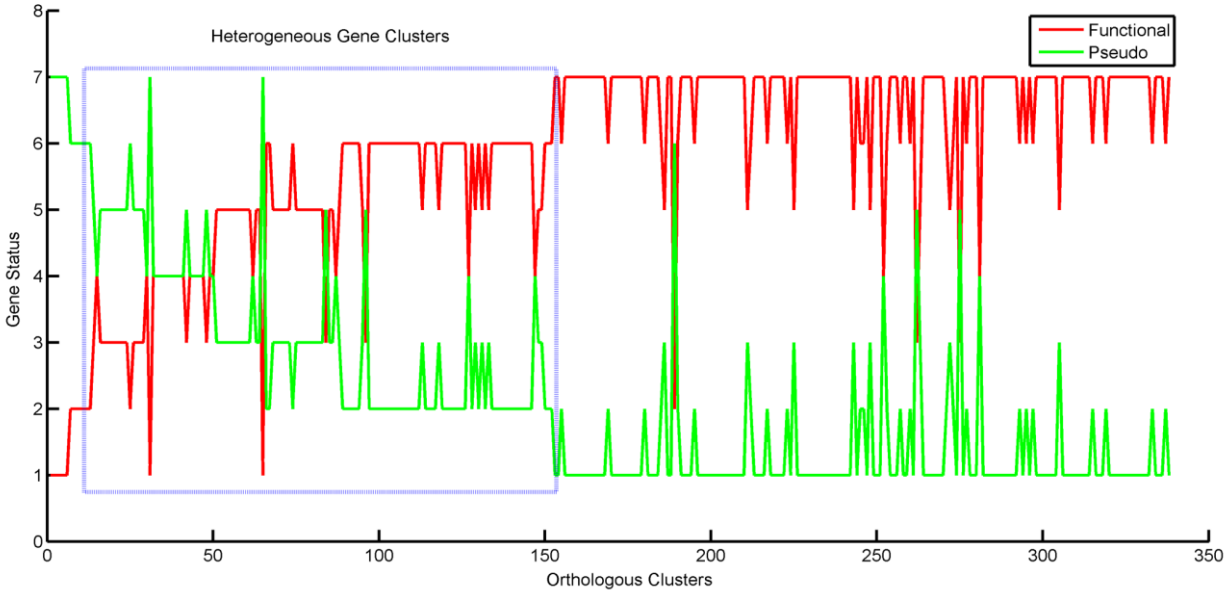


Figure 16: Accessory pseudogene clusters analysis*: The status of genes in each accessory pseudogene cluster was marked as P for pseudogene, F for functional complement and N for absence of gene. The clusters where the orthologs were present in P or F states were considered in plot. This shows the heterogeneous existence of functional and pseudogene complements in the population.*

The 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 status P or F in at least two query strains has shown that these were also enriched in metabolism related functions and this proportion was statistically significant with P value of 0.015 (Figure 17). Thus the comparison points at the existence of heterogeneous strains with varying metabolic potential and may confer an adaptive advantage for the persistence of the pathogen.

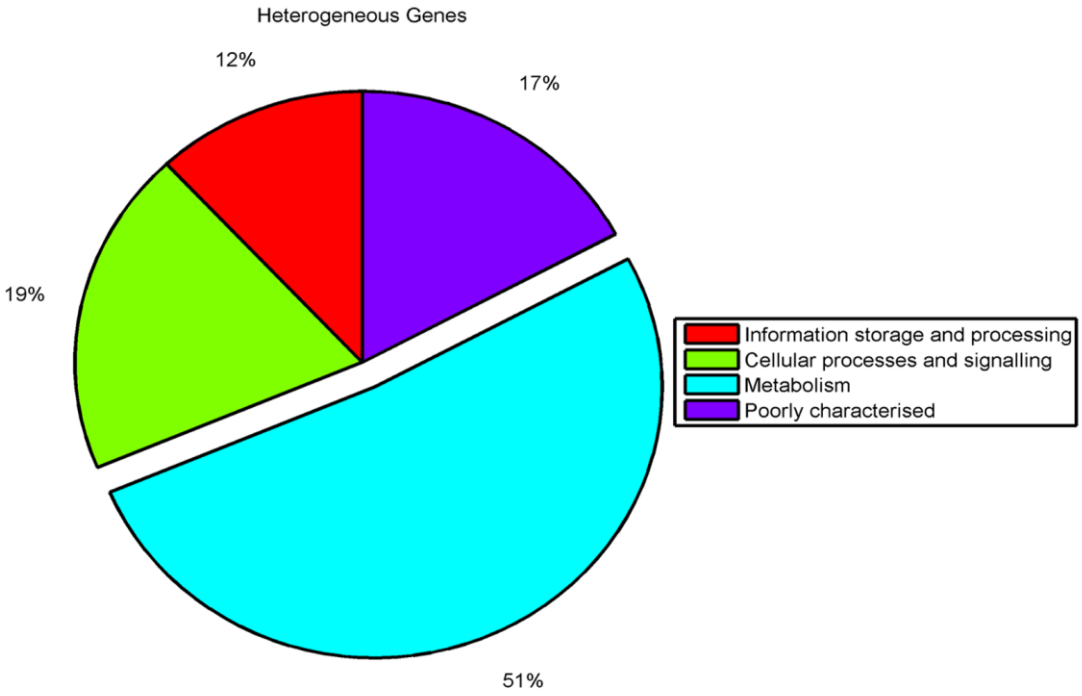


Figure 17: Proportion of heterogeneous genes classified according to COG functional categories*: The figure represents the distribution of accessory pseudogenes (those having variable functional and pseudogene status in atleast two strains) among various COG functional categories. The genes related to metabolism were clearly enriched in the accessory pseudogenes.*

3.4 Discussion

The previous whole genome based study by(Holt et al., 2008) has shown that *S*.Typhi 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 sharing similarity of 94-98 % as shown (Figure 10, 12). 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 or gain of function (Holt et al., 2008) (Parkhill et al., 2001) (McClelland et al., 2004). 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. We identified most of the pseudogenes including those caused due to frame shift mutations, as these were not detected in previous studies because of the low quality of sequence data.

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 (Kuenne et al., 2013). 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 pseudogenes recorded steep increase in pan-genome (Figure 14). 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 pseudogenization perhaps does not entail concurrent or consequent gene deletion in case of *Salmonella* Typhi in contrast to other important human pathogens such as *Mycobacterium leprae* wherein pseudogenization is followed by deletion thus downsizing the genome (Cole et al., 2001).

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 15). 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 336 clusters with mixed profile of pseudo and functional gene complements in various strains (Figure 16). Upon functional classification, even these polymorphic genes were found to be enriched in metabolism related functions (Figure 17). This could be an advantageous mechanism for the bacterium to modulate its metabolic repertoire through pseudogenisation depending on its specific local niche (Rohmer et al., 2011). Similar survival strategy is reported in other pathogenic bacteria where virulence optimization is achieved at the cost of certain metabolic genes (Touchon et al., 2009) (Maurelli et al., 1998).

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 (Holt et al., 2008). However any gain of function may be rare, or only occurring in a small number of genes through point mutations (Olson, 1999). 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. Given this situation, a definitive mechanism can be confirmed only through genetic and functional studies involving serial isolates.

*S.*Typhi encounters drastically different environments from its initial point of entry into the small intestine upto its final colonization of internal organs like gall bladder for chronic carriage (Reis and Horn, 2010). 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 are possibly tailoring a more hospitable metabolic environment. Collectively, these orchestrated genome dynamics most likely appear to aid in persistence and host adaptation.

Chapter 4

Comparative genomic analysis of strains associated with an outbreak and carrier individuals

---

Part of this chapter was published as:

Genome dynamics and evolution of *Salmonella* Typhi strains from the typhoid endemic zones. (2012) **Baddam R**, Kumar N, Shaik S, Lankapalli AK, Ahmed N. **Sci Rep**. 4:7457.

4.1 Introduction

For a host adapted strain like *S*.Typhi, survival in the host and dissemination are vital for establishing persistent infections (Ruby et al., 2012). A small percentage of infected individuals develop asymptomatic chronic carrier state thereby serving as reservoirs of infection and lead to shedding of bacterium in stools for a long period of time (Monack et al., 2004). This carrier state leads to endemicity of infection in developing countries with limited sanitation facilities and also serve as main factor responsible for periodic outbreaks (Dougan and Baker, 2014) . Therefore identification of these carriers is crucial for design and implementation of effective control measures, but none of the currently available diagnostics are completely trustworthy for this purpose. Due to intermittent shedding of bacterium in stools, identification by culturing techniques demands collection of multiple samples and is not very feasible (Gopinath et al., 2012). In endemic zones, serological tests fail due to high level of background antibodies in serum (Waddington et al., 2014). Some of the isolates from carrier individuals have been shown to exhibit similar PFGE profiles with other *S*.Typhi isolates from various regions of Southeast Asia, indicating spread of infection is probably through these carrier individuals (Thong KL Yassin RM, Sudarmono P, Padmidewi M, Soewandojo E, Handojo I, Sarasombath S, Pang T, 1995). Therefore they play an important role in understanding of transmission dynamics and effective epidemiological tracking. The underlying mechanisms which facilitate persistence of bacteria in internal organs like gall bladder are not completely clear, although mechanisms like biofilm formation on gall stones to protect from harsh environment have been implicated in previous studies (Gonzalez-Escobedo et al., 2011). The treatment options of these carriers include heavy dosage of antimicrobials for longer duration, which lead to many side effects (Bäumler et al., 2011).

A suitable collection of genomic data from *S*.Typhi strains associated with outbreak and carrier individuals will provide an opportunity to determine the microevolutionary changes, which may provide insights into the ability of bacteria to develop chronic carrier state. This analysis can also

help in identification of some important genetic loci which can be utilized in diagnosis of carriers. Therefore in this part of study, we determined the state specific SNPs of BL196 strain isolated during outbreak in 2005 and strains (CR0063, CR0044) isolated in 2007 from carrier individuals as part of surveillance program. Further we have also attempted to determine and analyze the specific functional and pseudogenes content among the pair wise comparison of this carrier and outbreak isolates.

4.2 Methodology

4.2.1 SNP analysis

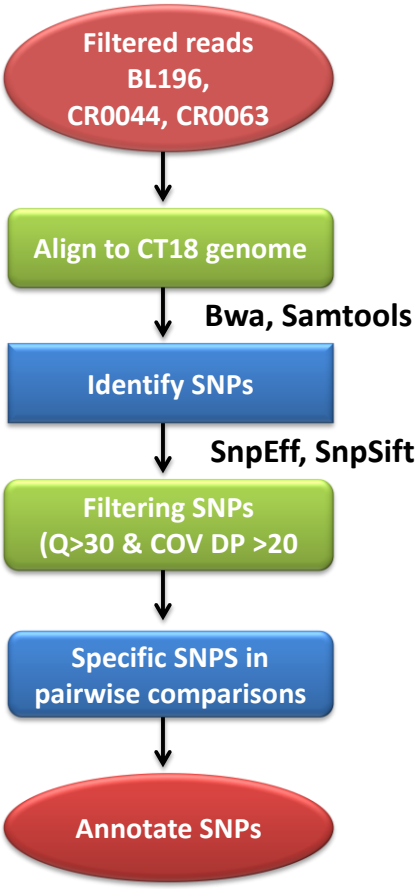


Figure 18: SNP analysis: *Methodology followed for detection of SNPs in pairwise comparisons of outbreak and carrier strains.*

The high quality filtered reads of each strain considered in pair-wise comparisons were aligned to finished reference genome CT18 individually using BWA alignment tools (Li and Durbin, 2009). SAMtools (Li et al., 2009) were applied for sorting and indexing of the above generated alignment. The variant calling was performed using mpileup option of SAMtools and output obtained in bcf (binary variant call format) is converted into vcf (Variant Call Format) using bcf tools. SnpSift and SnpEff were used for filtering of SNPs (Quality >30 and coverage depth >20) and their annotation (Cingolani et al., 2012). The following command line was used for all this analysis.

```
###### command line start #####
./bwa index -a is -p index reference_genome.fasta
./bwa aln -e 15 -I -f read1.aln index Read1_filtered.fastq
./bwa aln -e 15 -I -f read2.aln index Read2_filtered.fastq
./bwa sampe -f merged.sam index read1.aln read2.aln Read1_filtered.fastq Read2_filtered.fastq
./samtools view -o merged.bam -b -S -T reference_genome.fasta merged.sam
./samtools sort merged.bam merged_sorted.bam
./samtools index merged_sorted.bam
./samtools mpileup -d 8000 -ugf reference_genome.fasta merged_sorted.bam >final.bcf
./bcftools view -bvcg final.bcf > final_1.bcf
./bcftools view final_1.bcf >final.vcf
###### command line end #####
```

Once the high quality SNPs were identified separately for each individual strain, two pair wise comparisons were considered – outbreak strain BL196 & carrier strain CR0063 and outbreak strain BL196 & carrier strain CR0063, from Kelantan state of Malaysia. After subtracting the SNPs shared by both strains in comparison, the SNPs which are specific to certain associated state were identified.

4.2.2 Gene content variation analysis

In order to determine if there are any state specific genes that entail different clinical level phenotypes of the strains we also performed pairwise comparisons of the strains of different clinical spectrum as shown in Figure 19.

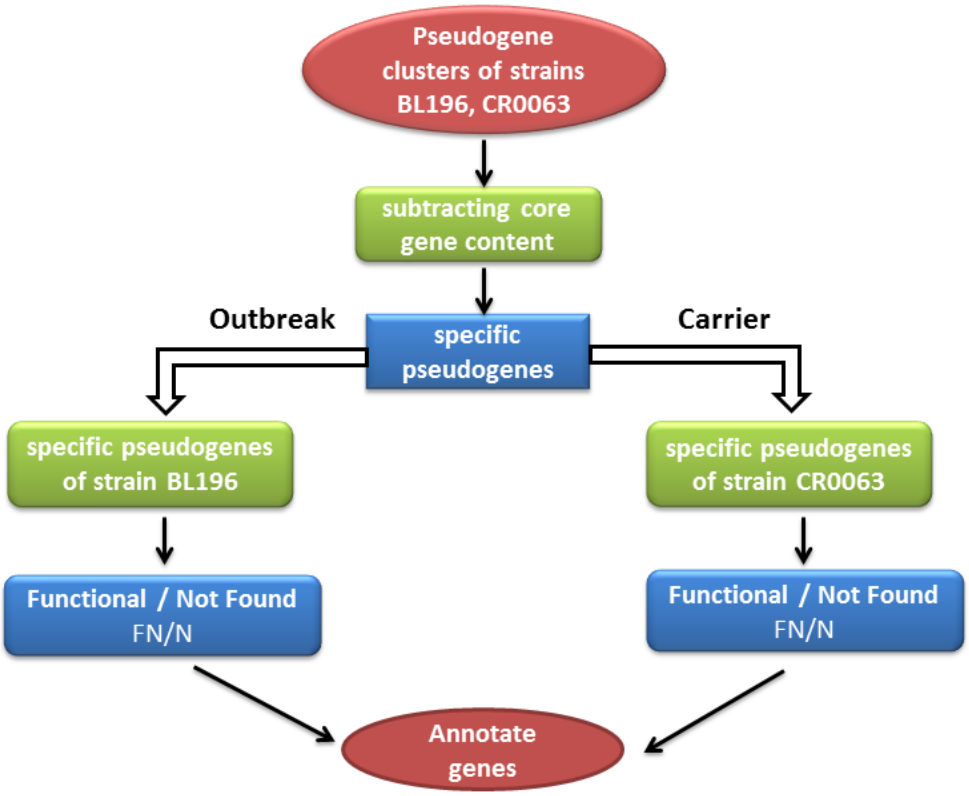


Figure19: Determination of variable gene content*: Schematic representation followed for the identification of variable pseudogene content in pairwise comparison of strains BL196 and CR0063.*

For this, we considered pair-wise comparisons in 2 sets: outbreak versus carrier strains (BL196 & CR0044 and, BL196&CR0063) from Kelantan, Malaysia. 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

corresponding given strain in comparison carried an ortholog in a different functional state (active or pseudo) in the corresponding strain, or vice versa. In this way, variation in functional and pseudogene contents of all the strains in pairwise comparisons was determined.

4.3 Results and Discussion

All the final SNPs were manually validated by loading the respective alignment file into Tablet program for visualization. In pair-wise comparison of outbreak strain BL196 with the carrier strain CR0063, 162 SNPs were identified in common for both the strains, where as in case of comparison with another carrier strain CR0044 351 SNPS were identified in common for both the strains. Further specific SNPs identified in both comparisons are shown below in Table 8

outbreak strain BL196 vs carrier strain CR0063					
Specific strain	**SNPs**	**Missense**	**Silent**	**Nonsense**	**Intergenic**
BL196	**199**	**109**	**58**	**5**	**29**
CR0063	**335**	**137**	**100**	**6**	**92**
outbreak strain BL196 vs carrier strain CR0044					
BL196	**2**	**-**	**2**	**-**	**-**
CR0044	**6**	**3**	**1**	**-**	**-**

Table 8: *The SNPs identified in observed in state specific comparisons*

The number of state specific SNPs were very low in case of strains BL196 & CR0044 when compared to that of BL196&CR0063. Further the annotation of these specific SNPs which are falling in coding regions and their functional classification was carried out. However this analysis could not reveal any potential association of specific genes to a particular clinical state. In pair-wise comparison of BL196 and CR0063, certain specific SNPs were clustered in a region that encoded a bacteriophage protein in BL196 strain and an IS element in CR0063 strain. Similarly in case of BL196 and CR0044 strains comparison, certain specific SNPs were clustered in a region that encoded a bacteriophage protein & IS element in BL196 strain and a bacteriophage protein in CR0044 strain. However the quality of alignment in these regions was very low and usually aligned

by a single read or marked by huge deletions on both sides of these regions, therefore they were not included in final list. The specific functional and pseudogene contents of the strains in two pairwise comparisons are summarized below in Table 9.

outbreak strain BL196 vs carrier strain CR0063				
Functional genes (core genes = 3778)			**Pseudogenes (core genes = 189)**	
Specific strain	**Pseudogene (P)**	**No Ortholog (N)**	**Functional gene (P)**	**No Ortholog (N)**
BL196	**80**	**35**	**57**	**52**
CR0063	**56**	**39**	**81**	**58**
outbreak strain BL196 vs carrier strain CR0044				
Functional genes (core genes = 3723)			**Pseudogenes (core genes = 235)**	
Specific strain	**Pseudogene (P)**	**No Ortholog (N)**	**Functional gene (P)**	**No Ortholog (N)**
BL196	**41**	**40**	**28**	**20**
CR0044	**28**	**30**	**42**	**29**

Table 9: *Gene content variation statistics as observed in the state specific comparisons*

After the determination of these state specific genes, they were further annotated and functionally classified. Although this analysis helped us develop pairwise inventories of complimentary active 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, like conveying an acute or a carrier stage.

Chapter 5

Summary and Outlook

Typhoid fever continues to be of a major public health concern as it still remains endemic in many developing countries with minimal resources, where substandard sanitation facilitates the transmission of pathogen through fecal-oral route. Further, the emergence and spread of MDR *S*. Typhi, lack of reliable diagnostic tests for accurate identification and failure of preventive measures like vaccination make the situation more complicated. *S*. Typhi being a monomorphic pathogen, comparison of whole genome sequences provides an opportunity to understand genetic structure of its population at a much higher resolution than with any traditional genotyping method and also provides insights as to the history of pathogen evolution. To begin with, we sequenced four *Salmonella* Typhi strains from the endemic zones of Southeast Asia and Oceania using Illumina sequencing technology. The read fragments obtained after sequencing were assembled into contigs using Velvet assembler. The improvement of assembly was made in various steps, first with the use of scaffolding tool SSPACE for combining contigs wherever possible by utilizing paired end information. Further, the scaffolds were sorted initially based on BLAST analysis with a closely related reference genome CT18. In the next step, reads were aligned to these scaffolds and their order was validated with the help of paired end read information. The draft chromosomes of all the strains obtained after the above processing steps were subjected to annotation using RAST and ISGA platforms.

The whole genome sequences of our chosen strains along with four other genomes from NCBI database were considered for a comparative genomics study. All these genomes/strains representing endemic zones were associated with different clinical manifestations - outbreaks, sporadic cases, carrier strains and fatal episodes. These genomes 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 such unique strains. The phylogenetic analysis based on whole genomes or core gene clusters revealed co-clustering of strains isolated from the same geographical

regions as well as high genetic relatedness among various strains. The close similarity of these genomes was also reflected in whole genome alignment generated using progressiveMauve. These observations once again reinforce the genetically monomorphic nature of *S.*Typhi and were in accordance with the previous findings based on MLST and other techniques. In order to understand the possible sources of variation, we have tried to identify IS elements, Phages and Salmonella pathogenicity islands encoded by these strains. However, only minor differences were observed with respect to the content of mobile elements and determination of their exact copy number was hampered due to draft status of some of the genomes.

As the comparison of whole genome sequences in previous study has pointed out that gene loss through pseudogenisation might be playing a more dominant role than gene gain by lateral gene transfer in evolution of this pathogen, we have attempted to understand the patterns that will emerge by comparison of gene inventories of these strains. For this, pan-genome analyses of these strains were performed and gene-frequency distributions were investigated among various strains with respect to functional and pseudogene complements. The pan-genome analyses of *Salmonella* strains revealed an alpha value of 0.913 implying a highly conservative nature of these endemic isolates and this was expected in view of host restriction in case of intracellular organisms as it limits opportunities for any significant intake of foreign DNA. In addition, when the same analysis was carried out for functional and pseudogenes separately, an α value of 0.94 was observed for functional genes, whereas in case of pseudogenes it was 0.556 indicating that pseudogenisation of functional genes is an ongoing process in *S.* Typhi. These findings were further supported by the observed decline in core functional gene content and a sharp increase in accessory pseudogene content. Also, the core genome of pseudogenes constituted a minor component of the total pseudogene content unlike that of core genome of functional genes. 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

has 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. Therefore, the study highlights that pattern of changes in the gene inventories of the bacterial population might be involved in redefining the metabolic complement of *S*.Typhi.

The above results prompted to us to further explore the differences in physiological capabilities which might be due to state specific genes that entail different clinical level phenotypes of the strains. For this, we performed pair-wise comparisons of the strains associated with outbreak and carrier strains in 2 sets. Although this analysis helped us develop pairwise inventories of complimentary active 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.

Comprehensive genomic analysis of this collection of strains also provided us with some significant pointers regarding host adaptation of this organism which could possibly be influenced by conserved nature of its genome. Further, inclusion of more number of genomes in the analysis may possibly enhance the quality and significance of these observations. However, deciphering definitive mechanisms of certain observations such as the heterogeneity displayed by functional and pseudogene content of these isolates, and if this could be an adaptive measure achieved by restoration of function through any mechanism or it was just a reflection of different subsequent changes in the population structure, requires further in-depth studies involving genetic and functional screens of serial isolates.

Nevertheless, with a global collection of genome sequence data, it was possible to advance the current understanding of the carrier state in *Salmonella* pathogens which is a major cause of continuous emergence and reemergence of typhoid in endemic regions. An in-depth comparative analysis with a global collection can provide a window towards identification of novel biomarkers that might aid in development of effective diagnostic tests, more importantly the ones which

distinguish acute and carrier states in endemic zones. A detailed analysis of the emerging patterns in gene inventory accompanied by changes in population structure can shed light on the exact metabolic acumen of this pathogen. But it also implies need for understanding its explicit implications *in vivo*. Finally, an improved understanding of the overall host-pathogen dynamics would facilitate better understanding of the disease state to underpin development of control strategies.

References

Achtman, M. (2008). Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. Annu. Rev. Microbiol. 62, 53–70.

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.

Akhter, S., Aziz, R. K., and Edwards, R. A. (2012). PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. Nucleic Acids Res. 40.

Akoh, J. A. (1991). Relative sensitivity of blood and bone marrow cultures in typhoid fever. Trop. Doct. 21, 174–6.

Alekshun, M. N., and Levy, S. B. (1999). The mar regulon: multiple resistance to antibiotics and other toxic chemicals. Trends Microbiol. 7, 410–3.

Alikhan, N. F., Petty, N. K., Ben Zakour, N. L., and Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics 12, 402.

Anwar, E., Goldberg, E., Fraser, A., Acosta, C. J., Paul, M., and Leibovici, L. (2014). Vaccines for preventing typhoid fever. Cochrane database Syst. Rev. 1, CD001261.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., et al. (2008). The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9, 75.

Baker, S., Favorov, M., and Dougan, G. (2010). Searching for the elusive typhoid diagnostic. BMC Infect. Dis. 10, 45.

Baker, S. J., Gunn, J. S., and Morona, R. (1999). The Salmonella typhi melittin resistance gene pqaB affects intracellular growth in PMA-differentiated U937 cells, polymyxin B resistance and lipopolysaccharide. Microbiology 145 (Pt 2, 367–78.

Baker, S., Sarwar, Y., Aziz, H., Haque, A., Ali, A., Dougan, G., Wain, J., and Haque, A. (2005). Detection of Vi-negative Salmonella enterica serovar typhi in the peripheral blood of patients with typhoid fever in the Faisalabad region of Pakistan. J. Clin. Microbiol. 43, 4418–25.

Bale, J. (2007). Kauffmann-White scheme - 2007: salmonella identification serotypes and antigen formulae. London: Centre for Infections Health Protection Agency.

Bäumler, A., and Fang, F. C. (2013). Host specificity of bacterial pathogens. Cold Spring Harb. Perspect. Med. 3, a010041.

Bäumler, A. J., Tsolis, R. M., Ficht, T. A., and Adams, L. G. (1998). Evolution of host adaptation in Salmonella enterica. Infect. Immun. 66, 4579–87.

Bäumler, A. J., Winter, S. E., Thiennimitr, P., and Casadesús, J. (2011). Intestinal and chronic infections: Salmonella lifestyles in hostile environments. Environ. Microbiol. Rep. 3, 508–17.

Bhutta, Z. A. (2006). Current concepts in the diagnosis and treatment of typhoid fever. BMJ 333, 78–82.

Brenner, F. W., Villar, R. G., Angulo, F. J., Tauxe, R., and Swaminathan, B. (2000). Salmonella nomenclature. J. Clin. Microbiol. 38, 2465–7.

Buckle, G. C., Walker, C. L. F., and Black, R. E. (2012). Typhoid fever and paratyphoid fever: Systematic review to estimate global morbidity and mortality for 2010. J. Glob. Health 2, 010401.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–3.

Chen, F., Poppe, C., Liu, G.-R., Li, Y.-G., Peng, Y.-H., Sanderson, K. E., Johnston, R. N., and Liu, S.-L. (2009). A genome map of Salmonella enterica serovar Agona: numerous insertions and deletions reflecting the evolutionary history of a human pathogen. FEMS Microbiol. Lett. 293, 188–95.

Chen, C. Y., Eckmann, L., Libby, S. J., Fang, F. C., Okamoto, S., Kagnoff, M. F., Fierer, J., and Guiney, D. G. (1996). Expression of Salmonella typhimurium rpoS and rpoS-dependent genes in the intracellular environment of eukaryotic cells. Infect. Immun. 64, 4739–43.

Choo, K. E., Davis, T. M., Ismail, A., Tuan Ibrahim, T. A., and Ghazali, W. N. (1999). Rapid and reliable serological diagnosis of enteric fever: comparative sensitivity and specificity of Typhidot and Typhidot-M tests in febrile Malaysian children. Acta Trop. 72, 175–83.

Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M., and Lu, X. (2012). Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front. Genet. 3, 35.

Cole, S. T., Eiglmeier, K., Parkhill, J., James, K. D., Thomson, N. R., Wheeler, P. R., Honore, N., Garnier, T., Churcher, C., Harris, D., et al. (2001). Massive gene decay in the leprosy bacillus. Nature 409, 1007–1011.

Collinson, S. K., Clouthier, S. C., Doran, J. L., Banser, P. A., and Kay, W. W. (1996). Salmonella enteritidis agfBAC operon encoding thin, aggregative fimbriae. J. Bacteriol. 178, 662–7.

Craig, L., Pique, M. E., and Tainer, J. A. (2004). Type IV pilus structure and bacterial pathogenicity. Nat. Rev. Microbiol. 2, 363–78.

Crawford, R. W., Gibson, D. L., Kay, W. W., and Gunn, J. S. (2008). Identification of a bile-induced exopolysaccharide required for Salmonella biofilm formation on gallstone surfaces. Infect. Immun. 76, 5341–9.

Darling, A. E., Mau, B., and Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5, e11147. doi:10.1371/journal.pone.0011147.

Darton, T. C., Blohmke, C. J., and Pollard, A. J. (2014). Typhoid epidemiology, diagnostics and the human challenge model. Curr. Opin. Gastroenterol. 30, 7–17.

Deng, W., Liou, S. R., Plunkett, G., Mayhew, G. F., Rose, D. J., Burland, V., Kodoyianni, V., Schwartz, D. C., and Blattner, F. R. (2003). Comparative genomics of Salmonella enterica serovar Typhi strains Ty2 and CT18. J. Bacteriol. 185, 2330–2337.

DH, H. (1998). SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics 14(1), 68–73.

Didelot, X., Eyre, D. W., Cule, M., Ip, C. L. C., Ansari, M. A., Griffiths, D., Vaughan, A., O'Connor, L., Golubchik, T., Batty, E. M., et al. (2012). Microevolutionary analysis of Clostridium difficile genomes to investigate transmission. Genome Biol. 13, R118.

Dougan, G., and Baker, S. (2014). Salmonella enterica serovar Typhi and the pathogenesis of typhoid fever. Annu. Rev. Microbiol. 68, 317–36.

Duncan, M. J., Mann, E. L., Cohen, M. S., Ofek, I., Sharon, N., and Abraham, S. N. (2005). The distinct binding specificities exhibited by enterobacterial type 1 fimbriae are determined by their fimbrial shafts. J Biol Chem 280, 37707–37716.

Dutta, S., Das, S., Mitra, U., Jain, P., Roy, I., Ganguly, S. S., Ray, U., Dutta, P., and Paul, D. K. (2014). Antimicrobial resistance, virulence profiles and molecular subtypes of Salmonella enterica

serovars Typhi and Paratyphi A blood isolates from Kolkata, India during 2009-2013. PLoS One 9, e101347.

Ekblom, R., and Wolf, J. B. W. (2014). A field guide to whole-genome sequencing, assembly and annotation. Evol. Appl. 7, n/a–n/a.

Emary, K., Moore, C. E., Chanpheaktra, N., An, K. P., Chheng, K., Sona, S., Duy, P. T., Nga, T. V. T., Wuthiekanun, V., Amornchai, P., et al. (2012). Enteric fever in Cambodian children is dominated by multidrug-resistant H58 Salmonella enterica serovar Typhi with intermediate susceptibility to ciprofloxacin. Trans. R. Soc. Trop. Med. Hyg. 106, 718–24.

Engels, E. A., and Lau, J. (2000). Vaccines for preventing typhoid fever. Cochrane database Syst. Rev., CD001261.

Engstrand, L. (2009). How will next-generation sequencing contribute to the knowledge concerning Helicobacter pylori? Clin. Microbiol. Infect. 15, 823–8.

Ewing, B., and Green, P. (1998). Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. Genome Res. 8, 186–194.

Flicek, P., and Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly. Nat. Methods 6, S6–S12.

Fouts, D. E., Brinkac, L., Beck, E., Inman, J., and Sutton, G. (2012). PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. Nucleic Acids Res. 40.

Gal-Mor, O., Boyle, E. C., and Grassl, G. A. (2014). Same species, different diseases: how and why typhoidal and non-typhoidal Salmonella enterica serovars differ. Front. Microbiol. 5, 391.

Gonzalez-Escobedo, G., Marshall, J. M., and Gunn, J. S. (2011). Chronic and acute infection of the gall bladder by Salmonella Typhi: understanding the carrier state. Nat. Rev. Microbiol. 9, 9–14.

Gopinath, S., Carden, S., and Monack, D. (2012a). Shedding light on Salmonella carriers. Trends Microbiol. 20, 320–327.

Gopinath, S., Carden, S., and Monack, D. (2012b). Shedding light on Salmonella carriers. Trends Microbiol. 20, 320–327.

Guo, L., Lim, K. B., Gunn, J. S., Bainbridge, B., Darveau, R. P., Hackett, M., and Miller, S. I. (1997). Regulation of lipid A modifications by Salmonella typhimurium virulence genes phoP-phoQ. Science 276, 250–3.

Guzman, C. A., Borsutzky, S., Griot-Wenk, M., Metcalfe, I. C., Pearman, J., Collioud, A., Favre, D., and Dietrich, G. (2006). Vaccines against typhoid fever. Vaccine 24, 3804–11.

Haghjoo, E., and Galán, J. E. (2004). Salmonella typhi encodes a functional cytolethal distending toxin that is delivered into host cells by a bacterial-internalization pathway. Proc. Natl. Acad. Sci. U. S. A. 101, 4614–9.

Haraga, A., Ohlson, M. B., and Miller, S. I. (2008). Salmonellae interplay with host cells. Nat. Rev. Microbiol. 6, 53–66.

Hemmerich, C., Buechlein, A., Podicheti, R., Revanna, K. V, and Dong, Q. F. (2010). An Ergatis-based prokaryotic genome annotation web server. Bioinformatics 26, 1122–1124.

Holt, K. E., Dolecek, C., Chau, T. T., Duy, P. T., La, T. T. P., Hoang, N. V. M., Nga, T. V. T., Campbell, J. I., Manh, B. H., Vinh Chau, N. Van, et al. (2011). Temporal fluctuation of multidrug resistant salmonella typhi haplotypes in the mekong river delta region of Vietnam. PLoS Negl. Trop. Dis. 5, e929.

Holt, K. E., Parkhill, J., Mazzoni, C. J., Roumagnac, P., Weill, F.-X., Goodhead, I., Rance, R., Baker, S., Maskell, D. J., Wain, J., et al. (2008). High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. Nat. Genet. 40, 987–93.

Hongo, E., Morimyo, M., Mita, K., Machida, I., Hama-Inaba, H., Tsuji, H., Ichimura, S., and Noda, Y. (1994). The methyl viologen-resistance-encoding gene smvA of Salmonella typhimurium. Gene 148, 173–4.

Jacobsen, A., Hendriksen, R. S., Aaresturp, F. M., Ussery, D. W., and Friis, C. (2011). The Salmonella enterica pan-genome. Microb. Ecol. 62, 487–504.

Kao, R. R., Haydon, D. T., Lycett, S. J., and Murcia, P. R. (2014). Supersize me: how whole-genome sequencing and big data are transforming epidemiology. Trends Microbiol. 22, 282–91.

Kariuki, S., Revathi, G., Kiiru, J., Mengo, D. M., Mwituria, J., Muyodi, J., Munyalo, A., Teo, Y. Y., Holt, K. E., Kingsley, R. A., et al. (2010). Typhoid in Kenya is associated with a dominant

multidrug-resistant Salmonella enterica serovar Typhi haplotype that is also widespread in Southeast Asia. J. Clin. Microbiol. 48, 2171–6.

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–80.

Kidgell, C., Reichard, U., Wain, J., Linz, B., Torpdahl, M., Dougan, G., and Achtman, M. (2002). Salmonella typhi, the causative agent of typhoid fever, is approximately 50,000 years old. Infect. Genet. Evol. 2, 39–45.

Kingsley, R. A., van Amsterdam, K., Kramer, N., and Bäumler, A. J. (2000). The shdA gene is restricted to serotypes of Salmonella enterica subspecies I and contributes to efficient and prolonged fecal shedding. Infect. Immun. 68, 2720–7.

Kingsley, R. A., Santos, R. L., Keestra, A. M., Adams, L. G., and Bäumler, A. J. (2002). Salmonella enterica serotype Typhimurium ShdA is an outer membrane fibronectin-binding protein that is expressed in the intestine. Mol. Microbiol. 43, 895–905.

Kuenne, C., Billion, A., Abu Mraheil, M., Strittmatter, A., Daniel, R., Goesmann, A., Barbuddhe, S., Hain, T., and Chakraborty, T. (2013). 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.

Kumar, Y., Sharma, A., and Mani, K. R. (2011). Re-emergence of susceptibility to conventionally used drugs among strains of Salmonella Typhi in central west India. J. Infect. Dev. Ctries. 5, 227–30.

Kumarasamy, K., and Krishnan, P. (2012). Report of a Salmonella enterica serovar Typhi isolate from India producing CMY-2 AmpC β-lactamase. J. Antimicrob. Chemother. 67, 775–6.

Lagesen K Rødland E, Stærfeldt HH, Ussery DW RT, H. P. F. (2000). RNammer: consistent annotation of rRNA genes in genomic sequences. Nucleic Acids Res 35, 3100–3108.

Langille, M. G. I., and Brinkman, F. S. L. (2009). IslandViewer: an integrated interface for computational identification and visualization of genomic islands. Bioinformatics 25, 664–665.

Lapierre, P., and Gogarten, J. P. (2009). Estimating the size of the bacterial pan-genome. Trends Genet. 25, 107–10.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947–2948.

Levine, M. M., Grados, O., Gilman, R. H., Woodward, W. E., Solis-Plaza, R., and Waldman, W. (1978). Diagnostic value of the Widal test in areas endemic for typhoid fever. Am. J. Trop. Med. Hyg. 27, 795–800.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–9.

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res 13, 2178–2189.

Liang, W., Zhao, Y., Chen, C., Cui, X., Yu, J., Xiao, J., and Kan, B. (2012). Pan-genomic analysis provides insights into the genomic variation and evolution of Salmonella Paratyphi A. PLoS One 7, e45346.

Lim, P. L., Tam, F. C., Cheong, Y. M., and Jegathesan, M. (1998). One-step 2-minute test to detect typhoid-specific antibodies based on particle separation in tubes. J. Clin. Microbiol. 36, 2271–8.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of next-generation sequencing systems. J. Biomed. Biotechnol. 2012, 251364.

Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., Geer, L. Y., Geer, R. C., Gonzales, N. R., et al. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. 39, D225–9.

Marcus, S. L., Brumell, J. H., Pfeifer, C. G., and Finlay, B. B. (2000). Salmonella pathogenicity islands: big virulence in small packages. Microbes Infect. 2, 145–56.

Mardis, E., McPherson, J., Martienssen, R., Wilson, R. K., and McCombie, W. R. (2002). What is finished, and why does it matter. Genome Res. 12, 669–71.

Marineli, F., Tsoucalas, G., Karamanou, M., and Androutsos, G. (2013). Mary Mallon (1869-1938) and the history of typhoid fever. Ann. Gastroenterol. Q. Publ. Hell. Soc. Gastroenterol. 26, 132–134.

Mathur, R., Oh, H., Zhang, D., Park, S.-G., Seo, J., Koblansky, A., Hayden, M. S., and Ghosh, S. (2012). A mouse model of Salmonella typhi infection. Cell 151, 590–602.

Maurelli, A. T., Fernández, R. E., Bloch, C. A., Rode, C. K., and Fasano, A. (1998). "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, 3943–8.

Maurice, J. (2012). A first step in bringing typhoid fever out of the closet. Lancet 379, 699–700.

McClelland, M., Sanderson, K. E., Clifton, S. W., Latreille, P., Porwollik, S., Sabo, A., Meyer, R., Bieri, T., Ozersky, P., McLellan, M., et al. (2004a). Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of Salmonella enterica that cause typhoid. Nat Genet 36, 1268–1274.

McClelland, M., Sanderson, K. E., Clifton, S. W., Latreille, P., Porwollik, S., Sabo, A., Meyer, R., Bieri, T., Ozersky, P., McLellan, M., et al. (2004b). Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of Salmonella enterica that cause typhoid. Nat Genet 36, 1268–1274.

McClelland, M., Sanderson, K. E., Spieth, J., Clifton, S. W., Latreille, P., Courtney, L., Porwollik, S., Ali, J., Dante, M., Du, F., et al. (2001). Complete genome sequence of Salmonella enterica serovar Typhimurium LT2. Nature 413, 852–856.

Meysman, P., Sánchez-Rodríguez, A., Fu, Q., Marchal, K., and Engelen, K. (2013). Expression divergence between Escherichia coli and Salmonella enterica serovar Typhimurium reflects their lifestyles. Mol. Biol. Evol. 30, 1302–14.

Mills, D. M., Bajaj, V., and Lee, C. A. (1995). A 40 kb chromosomal fragment encoding Salmonella typhimurium invasion genes is absent from the corresponding region of the Escherichia coli K-12 chromosome. Mol. Microbiol. 15, 749–59.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., and Marshall, D. (2010). Tablet-next generation sequence assembly visualization. Bioinformatics 26, 401–402.

Mirza, S., Kariuki, S., Mamun, K. Z., Beeching, N. J., and Hart, C. A. (2000). Analysis of plasmid and chromosomal DNA of multidrug-resistant Salmonella enterica serovar typhi from Asia. J. Clin. Microbiol. 38, 1449–52.

Monack, D. M., Mueller, A., and Falkow, S. (2004). Persistent bacterial infections: the interface of the pathogen and the host immune system. Nat. Rev. Microbiol. 2, 747–65.

Moncrief, M. B., and Maguire, M. E. (1999). Magnesium transport in prokaryotes. J. Biol. Inorg. Chem. 4, 523–7.

Mortimer, P. P. (1999). Mr N the milker, and Dr Koch's concept of the healthy carrier. Lancet 353, 1354–6.

Moshitch, S., Doll, L., Rubinfeld, B. Z., Stocker, B. A., Schoolnik, G. K., Gafni, Y., and Frankel, G. (1992). Mono- and bi-phasic Salmonella typhi: genetic homogeneity and distinguishing characteristics. Mol. Microbiol. 6, 2589–97.

Muscas, P., Rossolini, G. M., Chiesurin, A., Santucci, A., and Satta, G. (1994). Purification and characterization of type 1 fimbriae of Salmonella typhi. Microbiol. Immunol. 38, 353–8.

Nagarajan, N., and Pop, M. (2013). Sequence assembly demystified. Nat. Rev. Genet. 14, 157–67.

Nair, S., Schreiber, E., Thong, K. L., Pang, T., and 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.

Navarro, F., Llovet, T., Echeita, M. A., Coll, P., Aladueña, A., Usera, M. A., and Prats, G. (1996). Molecular typing of Salmonella enterica serovar typhi. J. Clin. Microbiol. 34, 2831–4.

Niemann, S., and Supply, P. (2014). Diversity and Evolution of Mycobacterium tuberculosis: Moving to Whole-Genome-Based Approaches. Cold Spring Harb. Perspect. Med. 4.

Nuccio, S.-P., and Bäumler, A. J. (2014). Comparative analysis of Salmonella genomes identifies a metabolic network for escalating growth in the inflamed gut. MBio 5, e00929–14.

Ochiai, R. L., Acosta, C. J., Danovaro-Holliday, M. C., Baiqing, D., Bhattacharya, S. K., Agtini, M. D., Bhutta, Z. A., Canh, D. G., Ali, M., Shin, S., et al. (2008). A study of typhoid fever in five Asian countries: disease burden and implications for controls. Bull. World Health Organ. 86, 260–8.

Ochman, H., and Groisman, E. A. (1995). The evolution of invasion by enteric bacteria. Can. J. Microbiol. 41, 555–61.

Ochman, H., Soncini, F. C., Solomon, F., and Groisman, E. A. (1996). Identification of a pathogenicity island required for Salmonella survival in host cells. Proc. Natl. Acad. Sci. U. S. A. 93, 7800–4.

Octavia, S., and Lan, R. (2009). Multiple-locus variable-number tandem-repeat analysis of Salmonella enterica serovar Typhi. J. Clin. Microbiol. 47, 2369–76.

Olopoenia, L. A., and King, A. L. (2000). Widal agglutination test - 100 years later: still plagued by controversy. Postgrad. Med. J. 76, 80–4.

Olson, M. V (1999). When less is more: gene loss as an engine of evolutionary change. Am. J. Hum. Genet. 64, 18–23.

Ong, S. Y., Pratap, C. B., Wan, X. H., Hou, S. B., Rahman, A. Y. A., Saito, J. A., Nath, G., and Alam, M. (2012). Complete Genome Sequence of Salmonella enterica subsp enterica Serovar Typhi P-stx-12. J. Bacteriol. 194, 2115–2116.

Ong, S. Y., Pratap, C. B., Wan, X., Hou, S., Rahman, A. Y., Saito, J. A., Nath, G., and Alam, M. (2013). The Genomic Blueprint of Salmonella enterica subspecies enterica serovar Typhi P-stx-12. Stand Genomic Sci 7, 483–496.

Parkhill, J., Dougan, G., James, K. D., Thomson, N. R., Pickard, D., Wain, J., Churcher, C., Mungall, K. L., Bentley, S. D., Holden, M. T., et al. (2001). Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. Nature 413, 848–852.

Parry, C. M., Hien, T. T., Dougan, G., White, N. J., and Farrar, J. J. (2002). Typhoid fever. N. Engl. J. Med. 347, 1770–1782. doi:10.1056/NEJMra020201.

Parry, C. M., Wijedoru, L., Arjyal, A., and Baker, S. (2011). The utility of diagnostic tests for enteric fever in endemic locations. Expert Rev. Anti. Infect. Ther. 9, 711–25.

Passey, M. (1995). The new problem of typhoid fever in Papua New Guinea: how do we deal with it? P. N. G. Med. J. 38, 300–4.

Pickard, D., Kingsley, R. A., Hale, C., Turner, K., Sivaraman, K., Wetter, M., Langridge, G., and Dougan, G. (2013). A genomewide mutagenesis screen identifies multiple genes contributing to Vi capsular expression in Salmonella enterica serovar Typhi. J. Bacteriol. 195, 1320–6.

Pop, M. (2009). Genome assembly reborn: recent computational challenges. Brief. Bioinform. 10, 354–66.

Reis, R. S. Dos, and Horn, F. (2010). Enteropathogenic Escherichia coli, Samonella, Shigella and Yersinia: cellular aspects of host-bacteria interactions in enteric diseases. Gut Pathog. 2, 8.

Rodriguez-Valera F, U. D. W. (2012). Is the pan-genome also a pan-selectome? F1000Research 1, 16.

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.

Roumagnac, P., Weill, F.-X., Dolecek, C., Baker, S., Brisse, S., Chinh, N. T., Le, T. A. H., Acosta, C. J., Farrar, J., Dougan, G., et al. (2006). Evolutionary history of Salmonella typhi. Science 314, 1301–4.

Ruby, T., McLaughlin, L., Gopinath, S., and Monack, D. (2012). Salmonella's long-term relationship with its host. FEMS Microbiol. Rev. 36, 600–15.

Sabat, A. J., Budimir, A., Nashev, D., Sá-Leão, R., van Dijl, J. m, Laurent, F., Grundmann, H., and Friedrich, A. W. (2013). Overview of molecular typing methods for outbreak detection and epidemiological surveillance. Euro Surveill. 18, 20380.

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.

Shea, J. E., Hensel, M., Gleeson, C., and Holden, D. W. (1996). Identification of a virulence locus encoding a second type III secretion system in Salmonella typhimurium. Proc. Natl. Acad. Sci. U. S. A. 93, 2593–7.

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., and Chandler, M. (2006). ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 34, D32–D36.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–3.

Stothard, P., and Wishart, D. S. (2005). Circular genome visualization and exploration using CGView. Bioinformatics 21, 537–9.

Sztein, M. B., Salerno-Goncalves, R., and McArthur, M. A. (2014). Complex adaptive immunity to enteric fevers in humans: lessons learned and the path forward. Front. Immunol. 5, 516.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V, Crabtree, J., Jones, A. L., Durkin, A. S., et al. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome" (vol 102, pg 13950, 2005). Proc. Natl. Acad. Sci. U. S. A. 102, 16530.

Tettelin, H., Riley, D., Cattuto, C., and Medini, D. (2008). Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 11, 472–477.

Thong, K. L., Cheong, Y. M., Puthucheary, S., Koh, C. L., and Pang, T. (1994). Epidemiologic analysis of sporadic Salmonella typhi isolates and those from outbreaks by pulsed-field gel electrophoresis. J. Clin. Microbiol. 32, 1135–41.

Thong, K. L., Passey, M., Clegg, A., Combs, B. G., Yassin, R. M., and Pang, T. (1996). Molecular analysis of isolates of Salmonella typhi obtained from patients with fatal and nonfatal typhoid fever. J. Clin. Microbiol. 34, 1029–33.

Thong KL Yassin RM, Sudarmono P, Padmidewi M, Soewandojo E, Handojo I, Sarasombath S, Pang T, P. S. (1995). Analysis of Salmonella typhi isolates from southeast Asia by pulsed-field gel electrophoresis. J Clin Micro Boil 33, 1938–1941.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., et al. (2009). Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet. 5, e1000344.

Treangen, T. J., and Messeguer, X. (2006). M-GCAT: interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species. BMC Bioinformatics 7, 433.

Turner, A. K., Nair, S., and Wain, J. (2006). The acquisition of full fluoroquinolone resistance in Salmonella Typhi by accumulation of point mutations in the topoisomerase targets. J. Antimicrob. Chemother. 58, 733–40.

Vernikos, G., Medini, D., Riley, D. R., and Tettelin, H. (2014). Ten years of pan-genome analyses. Curr. Opin. Microbiol. 23C, 148–154.

Virlogeux, I., Waxin, H., Ecobichon, C., and Popoff, M. Y. (1995). Role of the viaB locus in synthesis, transport and expression of Salmonella typhi Vi antigen. Microbiology 141 (Pt 1, 3039–47.

Waddington, C. S., Darton, T. C., and Pollard, A. J. (2014). The challenge of enteric fever. J. Infect. 68.

Wain, J., Diem Nga, L. T., Kidgell, C., James, K., Fortune, S., Song Diep, T., Ali, T., O Gaora, P., Parry, C., Parkhill, J., et al. (2003). Molecular analysis of incHI1 antimicrobial resistance plasmids from Salmonella serovar Typhi strains associated with typhoid fever. Antimicrob. Agents Chemother. 47, 2732–9.

Wain, J., Diep, T. S., Ho, V. A., Walsh, A. M., Nguyen, T. T., Parry, C. M., and White, N. J. (1998). Quantitation of bacteria in blood of typhoid fever patients and relationship between counts and clinical features, transmissibility, and antibiotic resistance. J. Clin. Microbiol. 36, 1683–7.

Wain, J., Hendriksen, R. S., Mikoleit, M. L., Keddy, K. H., and Ochiai, R. L. (2014). Typhoid fever. Lancet.

Wilson, R. P., Raffatellu, M., Chessa, D., Winter, S. E., Tükel, C., and Bäumler, A. J. (2008). The Vi-capsule prevents Toll-like receptor 4 recognition of Salmonella. Cell. Microbiol. 10, 876–90.

Yang, J., Chen, L., Sun, L., Yu, J., and Jin, Q. (2008). VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. Nucleic Acids Res. 36, D539–42.

Yap, K. P., Gan, H. M., Teh, C. S. J., Baddam, R., Chai, L. C., Kumar, N., Tiruvayipati, S. A., Ahmed, N., and Thong, K. L. (2012a). 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.

Yap, K. P., Teh, C. S. J., Baddam, R., Chai, L. C., Kumar, N., Avasthi, T. S., Ahmed, N., and Thong, K. L. (2012b). 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.

Yap, Y. F., and Puthucheary, S. D. (1998). Typhoid fever in children--a retrospective study of 54 cases from Malaysia. Singapore Med. J. 39, 260–2.

You Z Karlene L, Jonathan J. D, David S. Wishart, Y. L. (2011). "PHAST: A Fast Phage Search Tool." Nucl. Acids Res 39, 347–352.

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–9.

Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J., and Shen, B. (2011). A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. PLoS One 6, e17915.

Zhou, Z., McCann, a., Weill, F.-X., Blin, C., Nair, S., Wain, J., Dougan, G., and Achtman, M. (2014). Transient Darwinian selection in Salmonella enterica serovar Paratyphi A during 450 years of global spread of enteric fever. Proc. Natl. Acad. Sci. 111, 12199–12204.

Zhou, Z., McCann, A., Litrup, E., Murphy, R., Cormican, M., Fanning, S., Brown, D., Guttman, D. S., Brisse, S., and Achtman, M. (2013). Neutral Genomic Microevolution of a Recently Emerged Pathogen, Salmonella enterica Serovar Agona. PLoS Genet. 9.

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

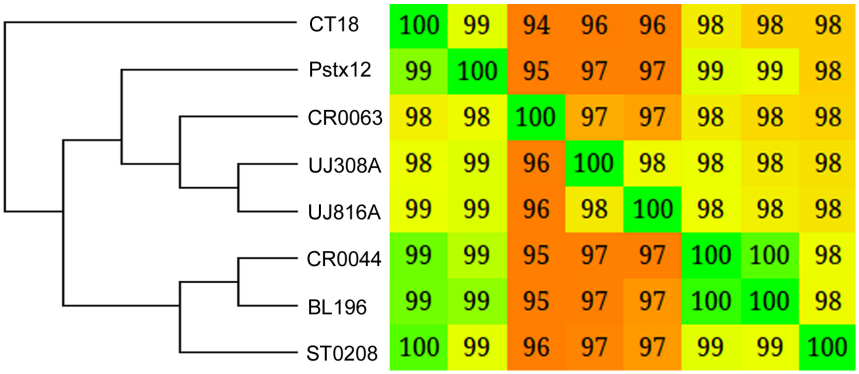


Figure 1 | Phylogenomic Tree. The whole genome information was used to build the distance matrix using Gegenees. The phylogenetic tree was developed using SplitsTree by NJ method. This revealed close similarity among genomes and also co-clustering of strains isolated from the same regions.

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).

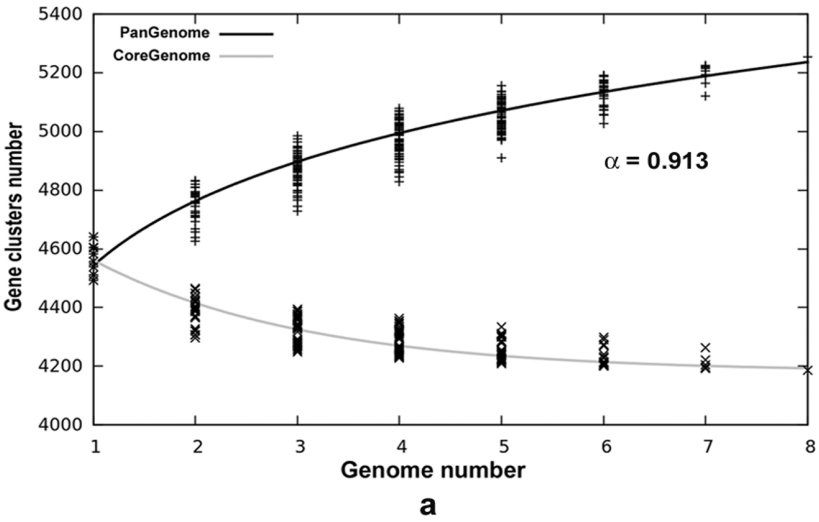
Further, to investigate the effect of pseudogenisation on gene frequency distributions of functional and pseudogenes, their pan and core genome components were determined separately. The pan-genome of functional genes contained a total of 4632 genes which was 1.03 times the average functional gene content 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).

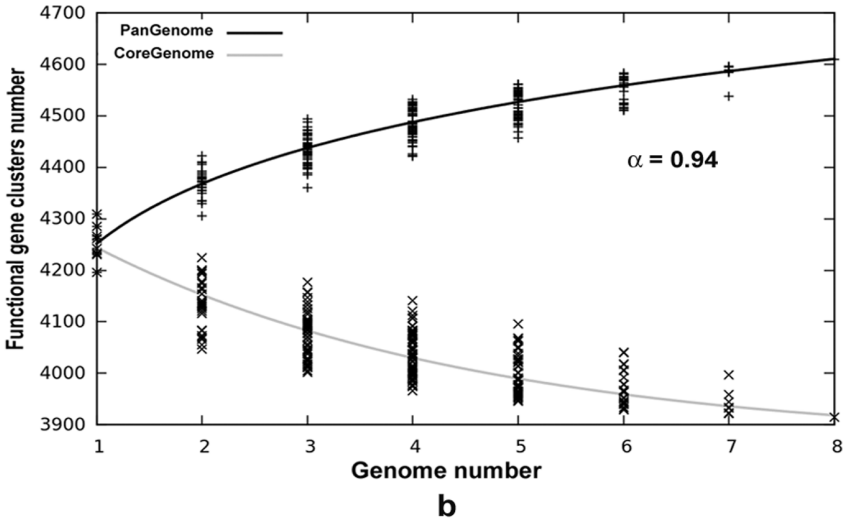
The values k and γ for functional gene clusters after curve fitting were determined as 4054 and 0.06 respectively, with a α value of 0.94. In contrast with what we observed in functional genes scenario, after curve fitting, α value for pseudogenes was 0.556 (Figure 3b, 3c), with k and γ values determined 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





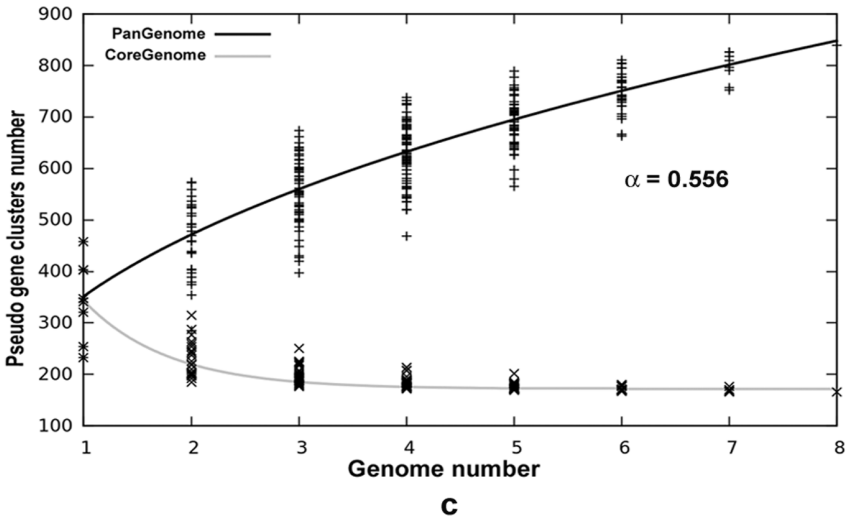


Figure 3 | Pan and Core Genome Distribution. (a). Pan and core genome developments using median values of the combinations of all eight genomes. (b). Pan and core genome developments of functional genes of these eight isolates. Here it can be observed that core genome is decreasing sharply. (c). Pan 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.

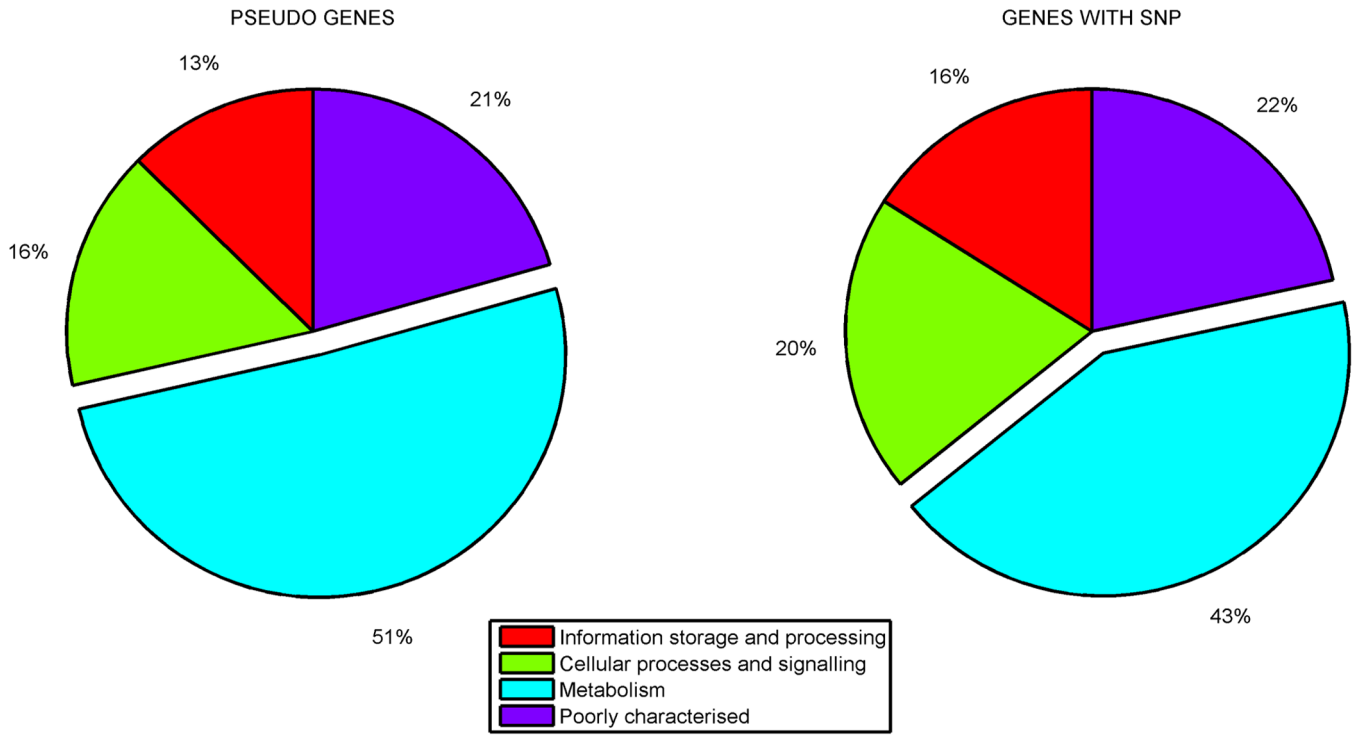


Figure 4 | The proportion of functionally classified pseudogenes and the functional genes with SNPs according to COG classification. The pie chart represents the proportion of various functional classes among the pseudogenes and the functional genes with SNPs. The figure clearly shows the enrichment of metabolism related genes in pseudogenes and the functional genes with SNPs.

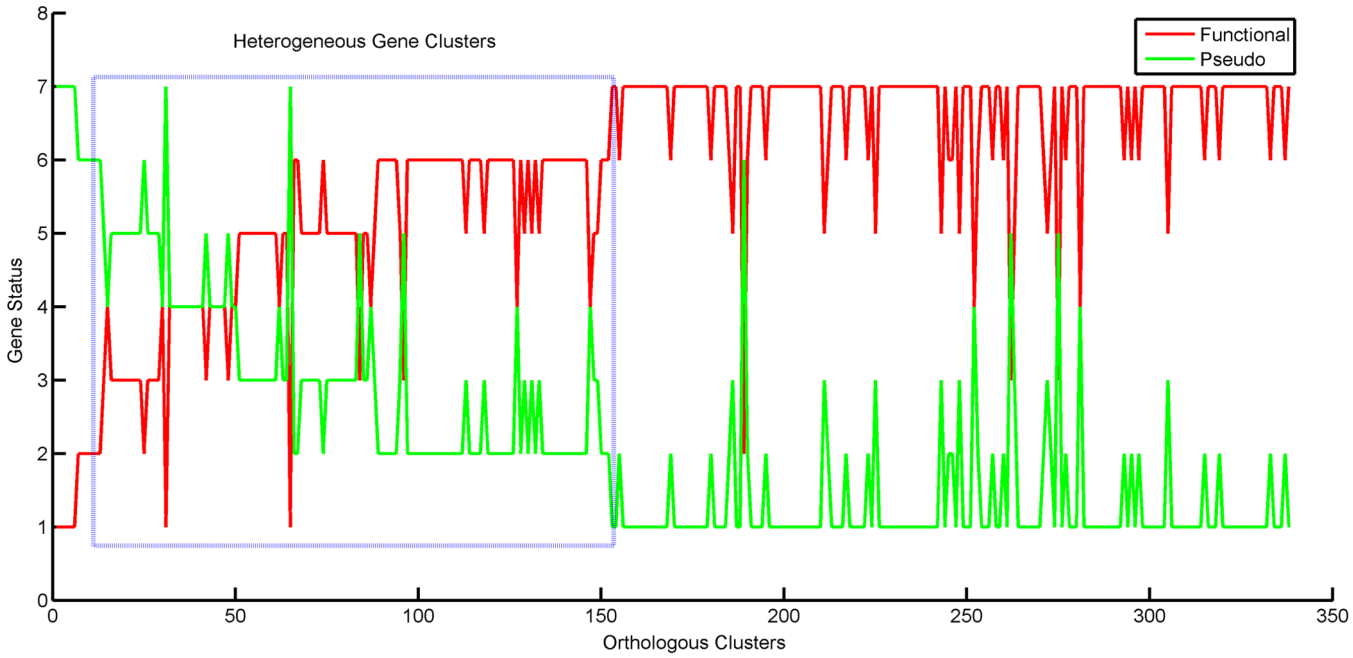


Figure 5 | Accessory pseudogene clusters analysis. The status of genes in each accessory pseudogene cluster was marked as P for pseudogene, F for functional complement and N for absence of gene. The clusters where the orthologs were present in P or F states were considered in plot. This shows the heterogeneous 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.

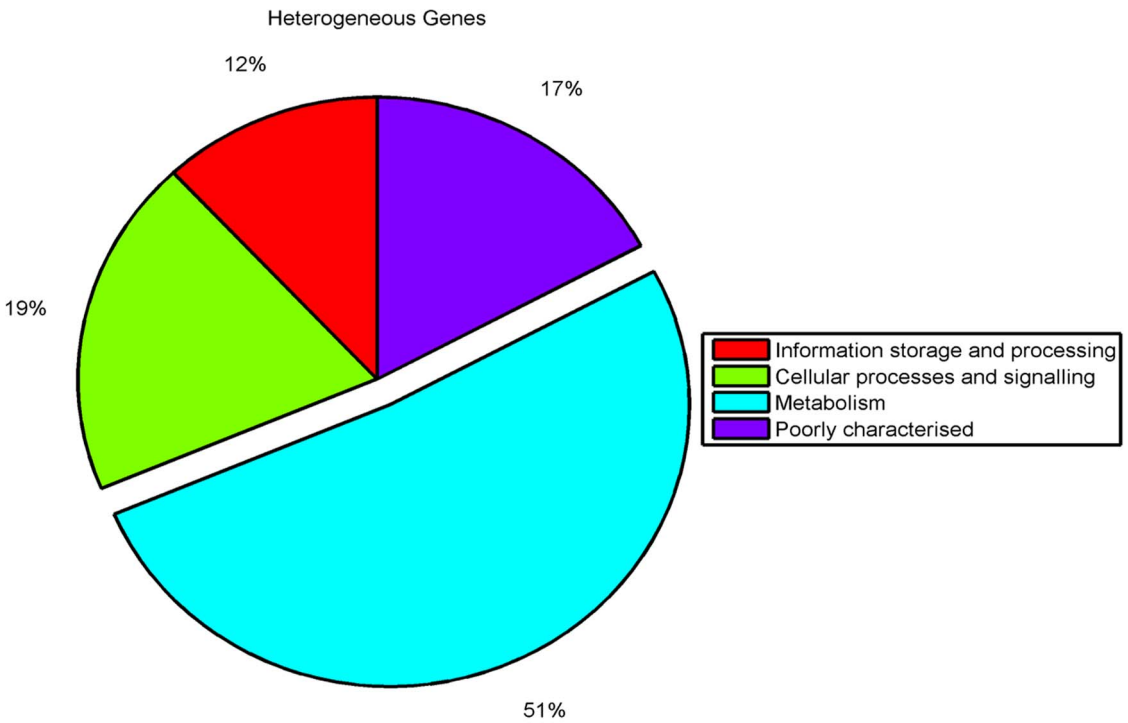


Figure 6 | Proportion of heterogeneous genes classified according to COG functional categories. The figure represents the distribution of accessory pseudogenes (those having variable functional and pseudogene status in atleast two strains) among various COG functional categories. The genes related to 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).

Gut Pathogens

GENOME ANNOUNCEMENT 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

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].

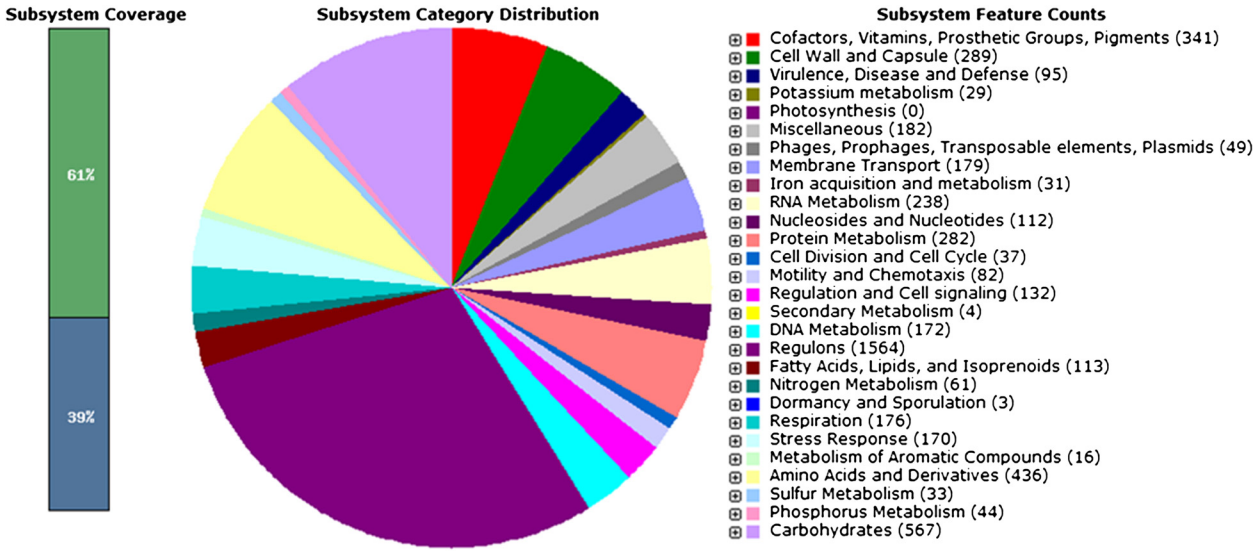


Figure 1 Subsystem distribution of ST CR0063. The subsystem statistics of ST CR0063 based on genome annotations performed according to RAST conventions.

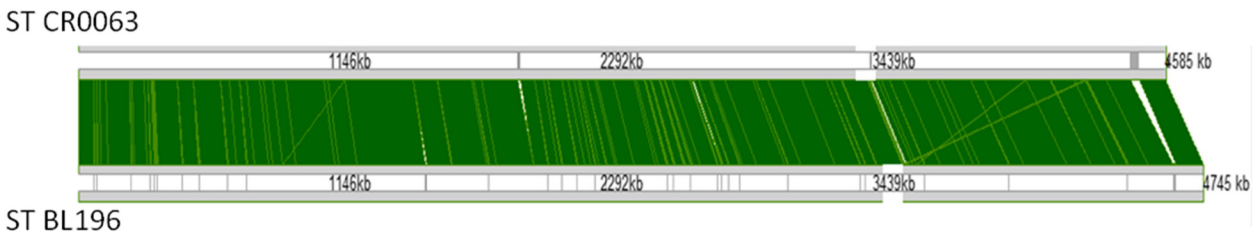


Figure 2 Comparison of *Salmonella* Typhi strains ST CR0063 and ST BL196. Comparison of whole genome sequences of *S. Typhi* strains using MG-CAT – one strain was isolated from a carrier individual (ST CR0063) and another from an infected individual (ST BL196) during a prolonged outbreak of Typhoid fever 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,

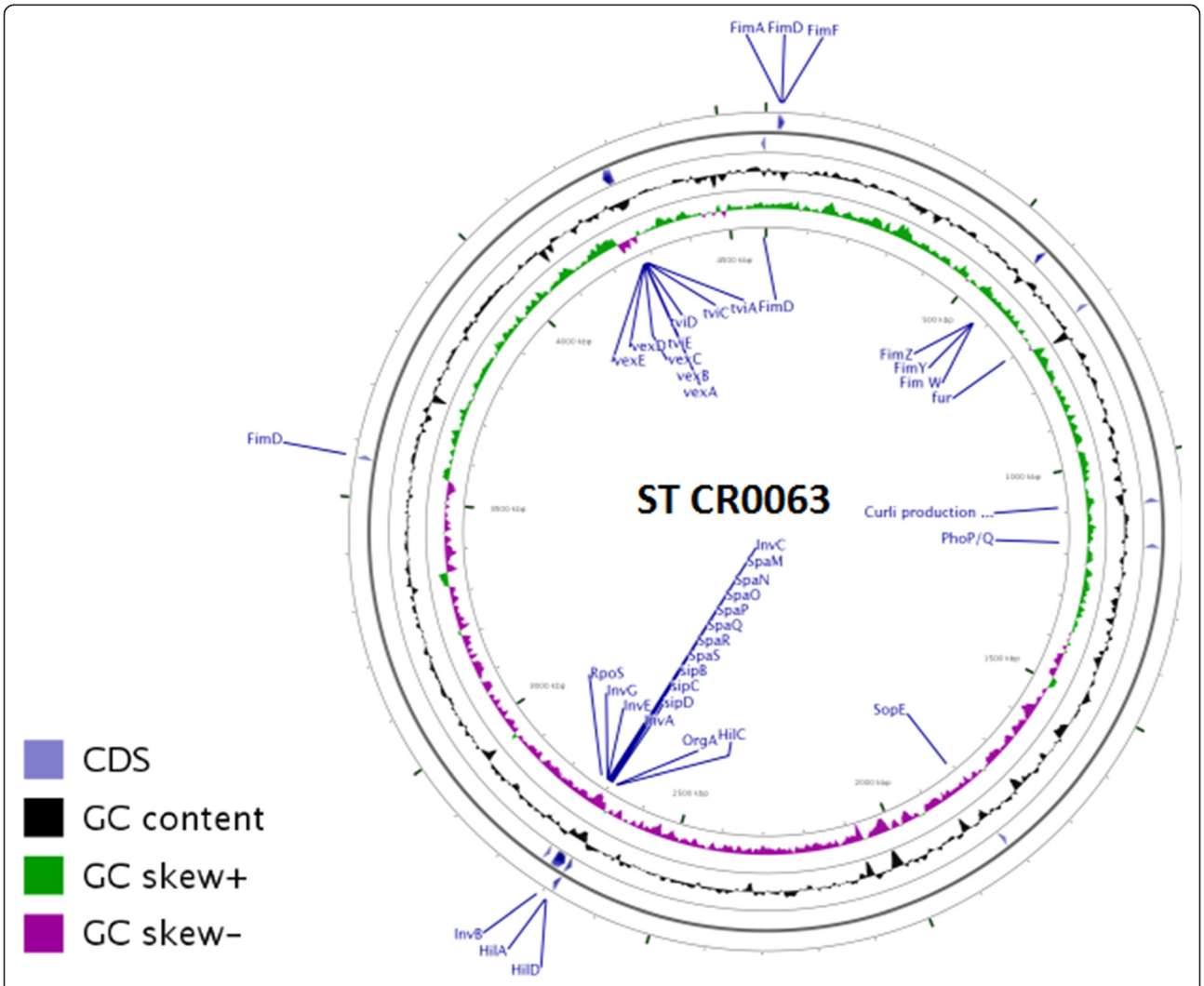


Figure 3 Circular Genome view of ST CR063. Positions of some of the major virulence factors and their regulators identified in ST CR0063 marked in the circular genome generated using CGview [26].

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.

JB Journals.ASM.org

GENOME ANNOUNCEMENT

Genome Sequence and Comparative Pathogenomics Analysis of a *Salmonella enterica* Serovar Typhi Strain Associated with a Typhoid Carrier in Malaysia

Kien-Pong Yap,[a,b] Han Ming Gan,[c] Cindy Shuan Ju Teh,[a,b] Ramani Baddam,[d] Lay-Ching Chai,[a,b] Narender Kumar,[d] Suma Avasthi Tiruvayipati,[a,d] Niyaz Ahmed,[a,d] and Kwai-Lin Thong[a,b]

Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia[a]; Laboratory of Biomedical Science and Molecular Microbiology, UMBIO Research Cluster, University of Malaya, Kuala Lumpur, Malaysia[b]; ScienceVision SB, Setia Alam, SeksyenU13, Shah Alam, Selangor, Malaysia[c]; and Pathogen Biology Laboratory, Department of Biotechnology, School of Life Sciences, University of Hyderabad, Hyderabad, India[d]

***Salmonella enterica* serovar Typhi is a human pathogen that causes typhoid fever predominantly in developing countries. In this article, we describe the whole genome sequence of the *S.* Typhi strain CR0044 isolated from a typhoid fever carrier in Kelantan, Malaysia. These data will further enhance the understanding of its host persistence and adaptive mechanism.**

Typhoid fever caused by human-specific *Salmonella enterica* serovar Typhi (*S.* Typhi) remains a major health problem that affects 21.7 million people, with 217,000 deaths worldwide annually (6). *S.* Typhi is transmitted through the oral-fecal route and sometimes persists in the body, establishing an asymptomatic chronic carrier (10, 12). The risk of developing gallbladder diseases, including carcinoma, is also higher among typhoid carriers (5, 10, 21).

Although typhoid fever is endemic in many countries, including Malaysia, little is known about the mechanism of survival and persistence of *S.* Typhi in the host. Therefore, the genome sequence and comparative pathogenomics analysis of carrier strain will provide in-depth understanding of its persistence and adaptive mechanism within its host.

S. Typhi CR0044 was isolated from stool sample of a typhoid carrier in Kelantan, Malaysia, in 2007. This strain was subtyped as ST1 by multilocus sequence typing (14) and was highly similar to the outbreak strain in 2005 by pulsed field gel electrophoresis (PFGE) (2). Genome sequencing of *S.* Typhi strain CR0044 was performed using the Ilumina Genome Analyzer (GA2x, pipeline version 1.60, insert size 300), which generated 1.0 gigabyte of data with a 90× depth coverage and a 73-bp read length. Genome assembly was constructed with Velvet (26) using the *de novo* approach, which generated 201 contigs with a minimum contig length of more than 200 bp and an average size of 23,367 bp. The open reading frames (ORFs) of the resultant contigs were predicted using RAST (1) and Prodigal (13) and subsequently annotated using Blast2GO (4), whereas tRNA and rRNA genes were identified with tRNAscan-SE (17) and RNAmmer (15), respectively. The predicted genome size is approximately 4,769,054 bp, with an average GC content of 52.1% and coding percentage of 85.8. The genome revealed approximately 4,884 coding sequences (CDS) with an average length of 825 bp. The genome also contains predicted 69 tRNA and 22 rRNA genes.

The genome revealed a type III secretion system and flagellum subsystem as reported in *S.* Typhi strains Ty2 and CT18 (7, 12). The genome contains genes reported in Ty2 and CT18, such as the gene coding for type 4 fimbrial assembly protein, the *yjbEFGH* locus, *yhjD* conserved clusters, and *wca* genes, which are related to cell wall and biofilm formation and host persistence (3, 8, 7, 12, 18, 25). It is noteworthy that the genome sequence also revealed the presence of the GGDEF family protein YeaJ, which is associated with cell surface adhesion and biofilm formation, which was not identified in Ty2 and CT18 (9, 19). The gene encoding the rhamnogalacturonide transporter RhiT for rhamnose utilization was also found adjacent to a transposase gene in CR0044 (20). Interestingly, the genome also revealed a zonular occludens toxin family protein that was not previously reported in *Salmonella* spp.

S. Typhi in Southeast Asia is genetically diverse, with genome variations and clonal expansion reported (2, 11, 15, 16, 21, 22, 23, 24). The dynamic nature of the *S.* Typhi chromosome greatly enhances its persistence and adaptation within the host, which allows the pathogen to survive and thrive in typhoid carriers. The genomic information obtained here could unveil the genome evolution and mechanism involved in carrier-state transformation.

Nucleotide sequence accession numbers. This Whole Genome Shotgun project has been deposited in GenBank under accession no. AKZO00000000. The version described in this paper is the first version, AKZO01000000. The Bioproject designation for this project is PRJNA160187.

ACKNOWLEDGMENTS

This research is supported by a University of Malaya High Impact Research Grant—Molecular Genetics (reference no. UM.C/625/1HIR/MOHE/-02 [A000002-5000 1]), for which K.-L. Thong is the grant holder for the high-impact research project under the title "Pathogenomic and Phenomic of Food-Borne Disease." We also acknowledge Indo-German International Research Training Group—Internationales Graduiertenkolleg (GRK1673)—Functional Molecular Infection Epidemiology, an initiative of the German Research Foundation (DFG) with N.A. from the University of Hyderabad (India). We are also grateful to M/s Genotypic Technology Pvt., Ltd., Bengaluru, India, for support with Illumina sequencing.

We acknowledge Safwan Jusoh from the ICT department, University of Malaya, for assisting us with computing solutions and allowing us to

Received 6 August 2012 **Accepted** 17 August 2012

Address correspondence to Kwai-Lin Thong, thongkl@um.edu.my.

doi:10.1128/JB.01416-12

use their servers and computational facilities and Soo Tein Ngoi from LBSMM, IPS, for technical assistance with DNA preparation.

REFERENCES

1. **Aziz RK, et al.** 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics **9**:75. doi:10.1186/1471-2164-9-75.
2. **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.
3. **Cano DA, Bernal GD, Tierrez A, Portillo FG, Casadesus J.** 2002. Regulation of capsule synthesis and cell motility in *Salmonella enterica* by the essential gene igaA. Genetics **162**:1513–1523.
4. **Conesa A, et al.** 2005. Blast2Go: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics **21**: 3674–3676.
5. **Crawford RW, et al.** 2010. Gallstones play a significant role in Salmonella spp. gallbladder colonization and carriage. Proc. Natl. Acad. Sci. U. S. A. **107**:4353–4358.
6. **Crump J, Mintz E.** 2010. Global trends in typhoid and paratyphoid fever. Clin. Infect. Dis. **50**:241–246.
7. **Deng W, et al.** 2003. Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18. J. Bacteriol. **185**:2330–2337.
8. **Ferrieres L, Aslam SN, Cooper RM, Clarke DJ.** 2007. The yjbEFGH locus in Escherichia coli K-12 is an operon encoding proteins involved in exopolysaccharide production. J. Microbiol. **153**:1070–1080.
9. **Garcia B, et al.** 2004. Role of the GGDEF protein family in Salmonella cellulose biosynthesis and biofilm formation. J. Mol. Microbiol. **54**:264–277.
10. **Gonzalez-Escobedo G, Marshall JM, Gunn JS.** 2011. Chronic and acute infection of the gall bladder by *Salmonella* Typhi: understanding the carrier state. Nat. Rev. Microbiol. **9**:9–14.
11. **Holt KE, et al.** 2011. Temporal fluctuation of multidrug resistant *Salmonella* Typhi haplotypes in the Mekong River delta region of Vietnam. PLoS Negl. Trop. Dis. **5**:e929. doi:10.1371/journal.pntd.0000929.
12. **Holt KE, et al.** 2008. High throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. Nat. Genet. **40**:987–993.
13. **Hyatt D, et al.** 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics **11**:119. doi: 10.1186/1471-2105-11-119.
14. **Kidgell C, et al.** 2002. Salmonella typhi, the causative agent of typhoid fever, is approximately 50,000 years old. Infect. Genet. Evol. **2**:39–45.
15. **Lagesen K, et al.** 2007. RNammer: consistent annotation of rRNA genes in genomic sequences. Nucleic Acids Res. **35**:3100–3108.
16. **Le TA, et al.** 2007. Clonal expansion and microevolution of quinolone-resistant *Salmonella enterica* serotype Typhi in Vietnam from 1996 to 2004. J. Clin. Microbiol. **45**:3485–3492.
17. **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.
18. **Mamat U, et al.** 2008. Single amino acid substitutions in either YhjD or MsbA confer viability to 3-deoxy-D-manno-oct-2-ulosonic acid-depleted Escherichia coli. J. Mol. Microbiol. **67**:633–648.
19. **Parkhill J, et al.** 2001. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. Nature **413**:848–852.
20. **Rodionov DA, Gelfand MS, Pattat NHC.** 2004. Comparative genomics of the KdgR regulon in Erwinia chrysanthemi 3937 and other gammaproteobacteria. J. Microbiol. **150**:3571–3590.
21. **Shukla VK, Singh H, Pandey M, Upadhyay SK, Nath G.** 2000. Carcinoma of the gallbladder—is it a sequel of typhoid? Dig. Dis. Sci. **45**:900–903.
22. **Thong KL, Cheong YM, Puthucheary SD, Koh CL, Pang T.** 1994. Epidemiologic analysis of sporadic *Salmonella* Typhi isolates and those from outbreaks by pulsed-field gel electrophoresis, J. Clin. Microbiol. **32**: 1135–1141.
23. **Thong KL, Puthucheary SD, Pang T.** 1996. Genome size variation among recent human isolates of *Salmonella* Typhi. Res. Microbiol. **148**:229–235.
24. **Thong KL, et al.** 1995. Analysis of *Salmonella* Typhi from Southeast Asia by pulsed-field gel electrophoresis. J. Clin. Microbiol. **33**:1938–1941.
25. **Townsend SM, et al.** 2001. *Salmonella enterica* serovar Typhi possesses a unique repertoire of fimbrial gene sequences. J. Infect. Immun. **69**:2894–2901.
26. **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

Insights from the Genome Sequence of a *Salmonella enterica* Serovar Typhi Strain Associated with a Sporadic Case of Typhoid Fever in Malaysia

Kien-Pong Yap,[a,b] Cindy Shuan Ju Teh,[a,b] Ramani Baddam,[c] Lay-Ching Chai,[a,b] Narender Kumar,[c] Tiruvayipati Suma Avasthi,[a,c] Niyaz Ahmed,[a,c] and Kwai-Lin Thong[a,b]

Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia[a]; Laboratory of Biomedical Science and Molecular Microbiology, UMBIO Research Cluster, University of Malaya, Kuala Lumpur, Malaysia[b]; and Pathogen Biology Laboratory, Department of Biotechnology, School of Life Sciences, University of Hyderabad, Hyderabad, India[c]

***Salmonella enterica* serovar Typhi is the causative agent of typhoid fever, which causes nearly 21.7 million illnesses and 217,000 deaths globally. Herein, we describe the whole-genome sequence of the *Salmonella* Typhi strain ST0208, isolated from a sporadic case of typhoid fever in Kuala Lumpur, Malaysia. The whole-genome sequence and comparative genomics allow an in-depth understanding of the genetic diversity, and its link to pathogenicity and evolutionary dynamics, of this highly clonal pathogen that is endemic to Malaysia.**

Salmonella enterica serovar Typhi (*S.* Typhi) is a human intracellular pathogen of global importance, infecting 21.7 million people and causing 217,000 deaths annually (4). The challenges of higher incidence of typhoid fever in developing countries have led to an increased health burden (4).

In 2008, 201 cases of sporadic outbreaks were reported in Malaysia (18). The genome sequence of this sporadically associated strain will provide insights into possible genetic events that would confer a fitness advantage.

S. Typhi strain ST0208 was isolated from the stool sample of a typhoid fever patient admitted to University Malaya Medical Centre (UMMC), Kuala Lumpur, Malaysia, in 2008. The strain was characterized by pulsed-field gel electrophoresis (PFGE), repetitive extragenic palindromic (REP)-PCR, and antimicrobial susceptibility profiling (25). The genome sequence of *S.* Typhi ST0208 was determined using an Illumina genome analyzer (GA2x, pipeline version 1.60) with an insert size of 300 bp, which generated 1.83 gigabytes of data with an average coverage of 165× and yielded 1,499,986 paired-end reads with a 100-bp read length. Genome assembly was constructed *de novo* using Velvet (26), which generated 222 contigs. The resultant contigs were uploaded into the RAST server (2, 17) to predict the open reading frames (ORFs) by using Glimmer3 (5) and validated with the ISGA integrated system (8). In brief, the predicted ORFs were annotated by searching against clusters of orthologous group (21) and SEED (7) databases, whereas tRNA and rRNA genes were identified by using tRNAscan-SE (15) and RNAmmer (12), respectively. The draft genome size is approximately 4,798,272 bp in length, with an average GC content of 52.0%, and is composed of 4,890 predicted coding sequences with an average length of 810 bp. A mean percentage of 83.7% of nucleotides of the genome are predicted to encode proteins. The genome reveals 71 tRNA and 22 rRNA predicted genes.

The genome contains several monosaccharide and polysaccharide metabolism-related genes, which were not reported in *S.* Typhi strains Ty2 and CT18 (6, 11, 18), such as D-galactarate permease, gluconate permease, tagatose-6-phosphate kinase, trehalase, and arabinose-proton transporter, that could be associated with host persistence (20). The genome sequence revealed four multidrug resistance clusters, *mdtABCR* and *marABCR* proteins (1, 3, 19), DNA gyrase subunit A and B, and topoisomerase subunit (IV) A and B (9), which were also identified in *S.* Typhi Ty2 and *S.* Typhi CT18 (6, 11, 19). Hypothetical proteins and pathogenicity islands were annotated.

High genetic diversity of *S.* Typhi was detected among human strains in Malaysia and Southeast Asia (10, 14, 22, 23, 24). Variation in the genome sequence revealed by the strain ST0208 is consistent with the proposed key theory of persistent adaptation and optimization of function (13, 16, 19). Genomic information from locality-specific strains associated with clinical manifestation allows genome evolution and endemicity to be studied.

Nucleotide sequence accession numbers. This whole-genome shotgun project has been deposited in DDBJ/EMBL/GenBank under the accession number AJXA00000000. The version described in this paper is the first version, AJXA01000000. The BioProject designation for this project is PRJNA160181.

ACKNOWLEDGMENTS

This research is supported by the University of Malaya High Impact Research Grant (reference UM.C/625/1HIR/MOHE/-02 [A000002-50001]), Molecular Genetics, in which K.-L. Thong is the grant holder for the high-impact research project under the title of Pathogenomics and Phenomics of Food-borne Bacterial Diseases.

We are also grateful to M/s Genotypic Technology Pvt. Ltd., Bengaluru, India, for their assistance with Illumina sequencing. We acknowledge Safwan Jusoh from the ICT department, University of Malaya, for assisting us with computing solutions and allowing us to use the servers and computational facilities and S. T. Ngoi from LBSMM, IPS, for technical assistance with DNA preparation.

REFERENCES

1. **Alekshun MN, Levy SB.** 1999. The mar regulon: multiple resistance to antibiotics and other toxic chemicals. Trends Microbiol. 7:410–413.

Received 15 June 2012 **Accepted** 29 June 2012

Address correspondence to Kwai-Lin Thong, thongkl@um.edu.my.

doi:10.1128/JB.01062-12

2. **Aziz RK, et al.** 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics **9**:75.
3. **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.
4. **Crump J, Mintz E.** 2010. Global trends in typhoid and paratyphoid fever. Clin. Infect. Dis. **50**:241–246.
5. **Delcher AL, Bratke KA, Powers EC, Salzberg SL.** 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics **23**:673–679.
6. **Deng W, et al.** 2003. Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18. J. Bacteriol. **185**:2330–2337.
7. **Disz T, et al.** 2010. Accessing the SEED genome database via web services API: tools for programmers. BMC Bioinformatics **11**:319.
8. **Hemmerich C, Buechlein A, Podicheti R, Revanna KV, Dong Q.** 2010. An Ergatis-based prokaryotic genome annotation web server. Bioinformatics **26**:1122–1124.
9. **Hirose K, et al.** 2002. DNA sequence analysis of DNA gyrase and DNA topoisomerase IV quinolone resistance-determining regions of *Salmonella enterica* serovar Typhi and serovar Paratyphi A. Antimicrob. Agents Chemother. **46**:3249–3252.
10. **Holt KE, et al.** 2011. Temporal fluctuation of multidrug resistant *Salmonella* Typhi haplotypes in the Mekong River delta region of Vietnam. PLoS Negl. Trop. Dis. **5**:e929.
11. **Holt KE, et al.** 2008. High throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. Nat. Genet. **40**:987–993.
12. **Lagesen K, et al.** 2007. RNAmmer: consistent annotation of rRNA genes in genomic sequences. Nucleic Acids Res. **35**:3100–3108.
13. **Lapierre P, Gogarten JP.** 2009. Estimating the size of the bacterial pangenome. Trends Genet. **25**:107–110.
14. **Le TA, et al.** 2007. Clonal expansion and microevolution of quinolone-resistant *Salmonella enterica* serotype Typhi in Vietnam from 1996 to 2004. J. Clin. Microbiol. **45**:3485–3492.
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.
16. **Matthews TD, Rabsch W, Maloy S.** 2011. Chromosomal rearrangements in *Salmonella enterica* serovar Typhi strains isolated from asymptomatic human carriers. mBio **2**:e00060-11. doi:10.1128/mBio.00060-11.
17. **Meyer F, et al.** 2008. The metagenomic RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics **9**:386.
18. **Ministry of Health Malaysia.** 2008. Epidemiology of foodborne diseases in Malaysia, p 89. Director General Ministry of Health Malaysia technical report. Ministry of Health Malaysia, Putrajaya, Malaysia.
19. **Parkhill J, et al.** 2001. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. Nature **413**:848–852.
20. **Shelburne SA, et al.** 2008. A direct link between carbohydrate utilization and virulence in the major human pathogen group A Streptococcus. Proc. Natl. Acad. Sci. U. S. A. **105**:1698–1703.
21. **Tatusov RL, Koonin EV, Lipman DJ.** 1997. A genomic perspective on protein families. Science **278**:631–637.
22. **Thong KL, Cheong YM, Puthucheary SD, Koh CL, Pang T.** 1994. Epidemiologic analysis of sporadic *Salmonella* Typhi isolates and those from outbreaks by pulsed-field gel electrophoresis. J. Clin. Microbiol. **32**:1135–1141.
23. **Thong KL, Puthucheary SD, Pang T.** 1996. Genome size variation among recent human isolates of *Salmonella* Typhi. Res. Microbiol. **148**:229–235.
24. **Thong KL, et al.** 1995. Analysis of *Salmonella* Typhi from Southeast Asia by pulsed-field gel electrophoresis. J. Clin. Microbiol. **33**:1938–1941.
25. **Tiong V, et al.** 2010. Macrorestriction analysis and antimicrobial susceptibility profiling of *Salmonella enterica* at a university teaching hospital, Kuala Lumpur. Jpn. J. Infect. Dis. **63**:317–322.
26. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18**:821–829.

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, Kwai-Lin Thong, Tiruvayipati Suma Avasthi, Sabiha Shaik, Kien-Pong Yap, Cindy Shuan Ju Teh, Lay-Ching Chai, Narender Kumar and Niyaz Ahmed

J. Bacteriol. **2012, 194(18):5122. DOI: 10.1128/JB.01051-12.**

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.

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.

Journal of Bacteriology

Genetic Fine Structure of a Salmonella enterica Serovar Typhi Strain Associated with the 2005 Outbreak of Typhoid Fever in Kelantan, Malaysia

Ramani Baddam, Narender Kumar, Kwai-Lin Thong, Soo-Tein Ngoi, Cindy Shuan Ju Teh, Kien-Pong Yap, Lay-Ching Chai, Tiruvayipati Suma Avasthi and Niyaz Ahmed

J. Bacteriol. **2012, 194(13):3565. DOI: 10.1128/JB.00581-12.**

JB Journals.ASM.org

GENOME ANNOUNCEMENT

Genetic Fine Structure of a *Salmonella enterica* Serovar Typhi Strain Associated with the 2005 Outbreak of Typhoid Fever in Kelantan, Malaysia

Ramani Baddam,[a] Narender Kumar,[a] Kwai-Lin Thong,[b,c] Soo-Tein Ngoi,[b,c] Cindy Shuan Ju Teh,[c] Kien-Pong Yap,[b,c] Lay-Ching Chai,[b,c] Tiruvayipati Suma Avasthi,[a,b] and Niyaz Ahmed[a,b]

Pathogen Biology Laboratory, Department of Biotechnology, School of Life Sciences, University of Hyderabad, Hyderabad, India[a]; Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia[b]; and Laboratory of Biomedical Science and Molecular Microbiology, UMBIO Research Cluster, University of Malaya, Kuala Lumpur, Malaysia[c]

Among enteric pathogens, *Salmonella enterica* serovar Typhi is responsible for the largest number of food-borne outbreaks and fatalities. The ability of the pathogen to cause systemic infection for extended durations leads to a high cost of disease control. Chronic carriers play important roles in the evolution of *Salmonella* Typhi; therefore, identification and in-depth characterization of isolates from clinical cases and carriers, especially those from zones of endemicity where the pathogen has not been extensively studied, are necessary. Here, we describe the genome sequence of the highly virulent *Salmonella* Typhi strain BL196/05 isolated during the outbreak of typhoid in Kelantan, Malaysia, in 2005. The whole-genome sequence and comparative genomics of this strain should enable us to understand the virulence mechanisms and evolutionary dynamics of this pathogen in Malaysia and elsewhere.

Salmonella enterica serovar Typhi and other pathogenic salmonellae are endemic in some countries (8, 15), and outbreaks occur due to unhygienic conditions (16), leading to alarming morbidity and mortality figures (9, 18). This results in a huge burden on the public health machinery. *Salmonella* Typhi persists for an extended duration in its host (18) due to modulation of the host immune responses together with genetic rearrangements that confer fitness advantages (12, 13, 17).

The incidence of typhoid in Kelantan had always been higher than in other states of Malaysia. In the 2005 outbreak (April to June 2005), 735 cases and 2 deaths occurred (18). We hypothesized that the genome sequences of the underlying strains would provide more insights to enhance understanding of endemicity or persistence of typhoid in Kelantan.

Salmonella Typhi BL196/05 was isolated from blood samples of a severe typhoid case in Kelantan during the notorious outbreak of 2005. The strain was characterized by PCR to determine the presence of many different virulence genes. The genome sequence was analyzed and annotated exactly as described previously (4, 5, 11, 21). Briefly, the 73-bp paired-end Illumina sequence reads, amounting to 1.7 gigabytes of data (insert size, 300 bp), were generated with 80× genome coverage. Velvet (26), with the hash length set to 37, was used to assemble sequence reads into 191 contigs; these were further assembled into a draft genome and were submitted to RAST (6) to determine the following data. The size of the single chromosome was approximately 4,744,056 bp, with a G+C content of 53.21% and a coding percentage of 87.1. There were 4,875 protein coding sequences found, with an average length of 875 bp. The genome revealed 76 tRNA and 22 rRNA genes. Our strain did not harbor any plasmid, and its phage typing revealed Vi phage type B1. The genome sequence revealed two *mar* regulons, *marRAB* and *marC*, as reported also in *Salmonella* Typhi strains CT18 and Ty2, and these were homologous to *Escherichia coli mar* (multiple antibiotic resistance) regulon members (2, 3, 10, 20, 25). The genes encoding a melittin resistance protein, PqaB, and a polymyxin resistance protein, PmrD (7), were located in the genome as also seen in the *Salmonella* Typhi CT18 and Ty2 genomes. The methyl viologen resistance gene (14) *smvA* was also identified. Several pathogenicity islands as well as a pool of hypothetical proteins were annotated.

Considerable genetic diversity exists among human isolates of *Salmonella* Typhi in Malaysia and Southeast Asia (22, 23), and the heterogeneity in genome sizes (24) points to a high degree of plasticity in the *Salmonella* pan-genome. The genome sequence of strain BL196/05 also revealed rearrangements putatively relevant to virulence optimization, persistence, and adaptation within the host. Such information would be harnessed to improve understanding of genome evolution and adaptation dynamics of *Salmonella* in Malaysia and to develop point-of-care diagnostics. Further, future efforts are needed to evaluate the significance of comparative genomics/genotypic data in juxtaposition with host genetic polymorphisms to herald the beginning of the "functional molecular infection epidemiology" (1) of typhoid. Finally, we recommend greater use of data determined in studies of clinical isolates and strains specific to outbreaks and countries rather than use of the reference (type) strains alone.

Nucleotide sequence accession numbers. This genome project has been deposited at GenBank under accession no. AJGK00000000. The version described herein is the first version, AJGK01000000. The Bioproject designation for this project is PRJNA85621.

Received 9 April 2012 **Accepted** 10 April 2012

Address correspondence to Niyaz Ahmed, niyazSL@uohyd.ernet.in.

doi:10.1128/JB.00581-12

ACKNOWLEDGMENTS

We thankfully acknowledge support received from the University of Malaya (High Impact Research Grant; reference 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 at the Institute of Biological Sciences at the Universiti Malaya, Kuala Lumpur, Malaysia, and a Visiting Adjunct Professor at the Universiti Brunei—Darussalam, Brunei. N.K. acknowledges the Junior Research Fellowship received from the Council of Scientific and Industrial Research (CSIR), India.

We are also grateful to M/s Genotypic Technology Pvt. Ltd., Bengaluru, India, for their measureless efforts with respect to Illumina sequencing. We acknowledge the Bioinformatics Facility (BIF) at the Department of Biotechnology, University of Hyderabad, for the use of servers and computational infrastructure. We thank Dipshikha Chakravortty for critical reading of the manuscript. All other members of the Ahmed and Thong laboratories are gratefully acknowledged for their help and support.

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. **Alekshun MN, Levy SB.** 1997. Regulation of chromosomally mediated multiple antibiotic resistance: the mar regulon. Antimicrob. Agents Chemother. **41**:2067–2075.
3. **Alekshun MN, Levy SB.** 1999. The mar regulon: multiple resistance to antibiotics and other toxic chemicals. Trends Microbiol. **7**:410–413.
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. **Avasthi TS, et al.** 2011. Genome of multidrug-resistant uropathogenic Escherichia coli strain NA114 from India. J. Bacteriol. **193**:4272–4273.
6. **Aziz RK, et al.** 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics **9**:75. doi:10.1186/1471-2164-9-75.
7. **Baker SJ, Gunn JS, Morona R.** 1999. The *Salmonella typhi* melittin resistance gene *pqaB* affects intracellular growth in PMA-differentiated U937 cells, polymyxin B resistance and lipopolysaccharide. Microbiology **145**(Pt. 2):367–378.
8. **Crump JA, Mintz ED.** 2010. Global trends in typhoid and paratyphoid fever. Clin. Infect. Dis. **50**:241–246.
9. **Crump JA, Luby SP, Mintz ED.** 2004. The global burden of typhoid fever. Bull. World Health Organ. **82**:346–353.
10. **Deng W, et al.** 2003. Comparative genomics of Salmonella enterica serovar Typhi strains Ty2 and CT18. J. Bacteriol. **185**:2330–2337.
11. **Devi SH, et al.** 2010. Genome of Helicobacter pylori strain 908. J. Bacteriol. **192**:6488–6489.
12. **Everest P, Wain J, Roberts M, Rook G, Dougan G.** 2001. The molecular mechanisms of severe typhoid fever. Trends Microbiol. **9**:316–320.
13. **Holt KE, et al.** 2008. High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. Nat. Genet. **40**:987–993.
14. **Hongo E, et al.** 1994. The methyl viologen-resistance-encoding gene smvA of Salmonella typhimurium. Gene **148**:173–174.
15. **House D, Bishop A, Parry C, Dougan G, Wain J.** 2001. Typhoid fever: pathogenesis and disease. Curr. Opin. Infect. Dis. **14**:573–578.
16. **Jain S, et al.** 2009. Multistate outbreak of Salmonella Typhimurium and Saintpaul infections associated with unpasteurized orange juice—United States, 2005. Clin. Infect. Dis. **48**:1065–1071.
17. **Jones BD, Falkow S.** 1996. Salmonellosis: host immune responses and bacterial virulence determinants. Annu. Rev. Immunol. **14**:533–561.
18. **Ministry of Health Malaysia.** 2006. Epidemiology of foodborne diseases in Malaysia, chapter 2, p 86–87. Director General Ministry of Health Malaysia Technical Report. Ministry of Health Malaysia, Putrajaya, Malaysia.
19. **Monack DM.** 2012. Salmonella persistence and transmission strategies. Curr. Opin. Microbiol. **15**:100–107.
20. **Parkhill J, et al.** 2001. Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. Nature **413**:848–852.
21. **Siddavattam D, et al.** 2011. Genome of a novel isolate of Paracoccus denitrificans capable of degrading N,N-dimethylformamide. J. Bacteriol. **193**:5598–5599.
22. **Thong KL, Cheong YM, Puthucheary SD, Koh CL, Pang T.** 1994. Epidemiologic analysis of sporadic Salmonella Typhi isolates and those from outbreaks by pulsed-field gel electrophoresis. J. Clin. Microbiol. **32**:1135–1141.
23. **Thong KL, et al.** 1995. Analysis of *Salmonella* Typhi from Southeast Asia by pulsed-field gel electrophoresis. J. Clin. Microbiol. **33**:1938–1941.
24. **Thong KL, Puthucheary SD, Pang T.** 1996. Genome size variation among recent human isolates of *Salmonella typhi.* Res. Microbiol. **148**:229–235.
25. **Winfield MD, Groisman EA.** 2004. Phenotypic differences between Salmonella and Escherichia coli resulting from the disparate regulation of homologous genes. Proc. Natl. Acad. Sci. U. S. A. **101**:17162–17167.
26. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18**:821–829.

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

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).

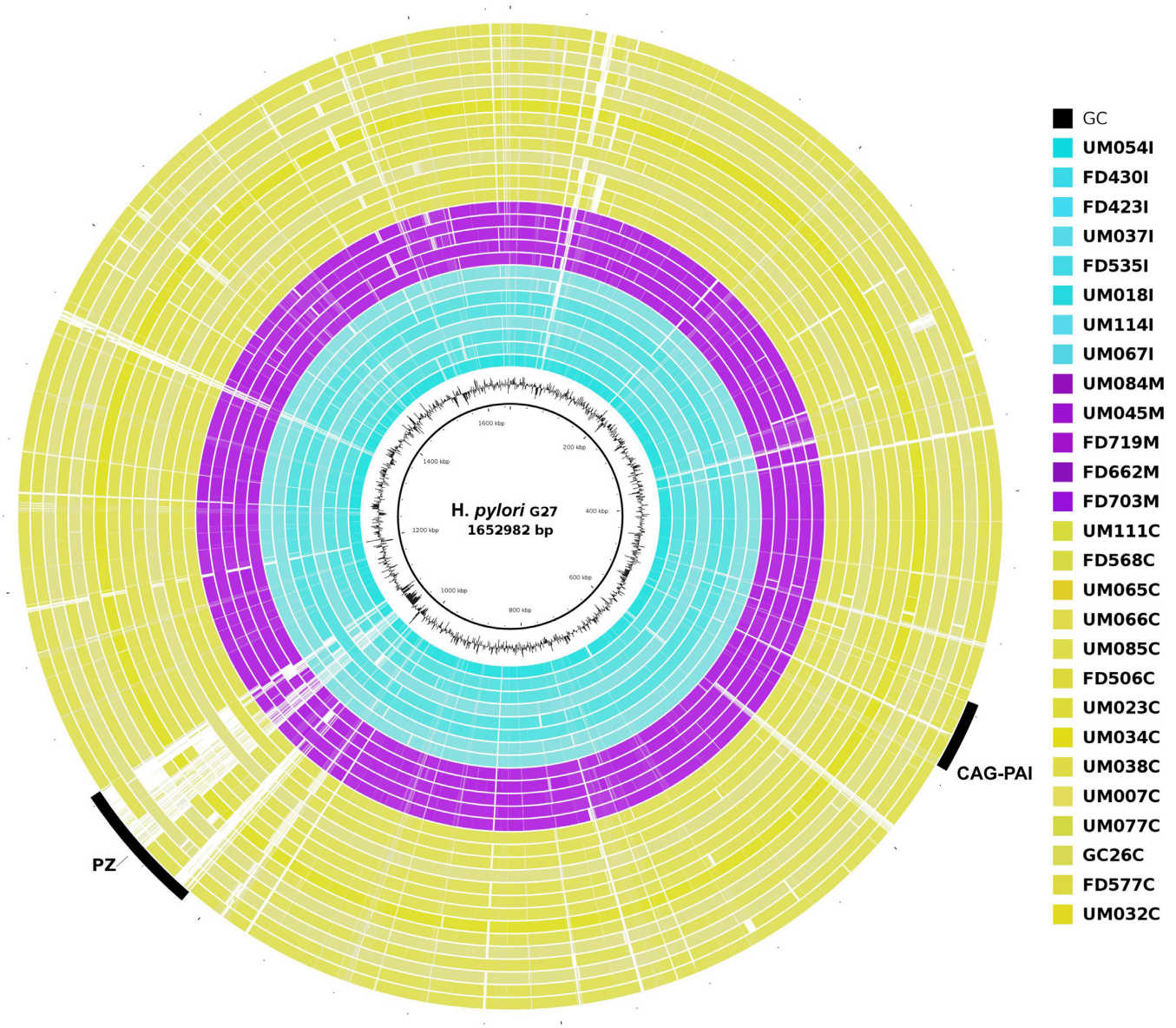


Figure 1. A circular representation of the genomes of Malaysian isolates: the draft genomes of 27 strains were aligned against the genome of reference strain *H. pylori* G27. Each genome is represented by a ring. The yellow rings represent *H. pylori* genomes from Malaysian-Chinese, purple represents 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).

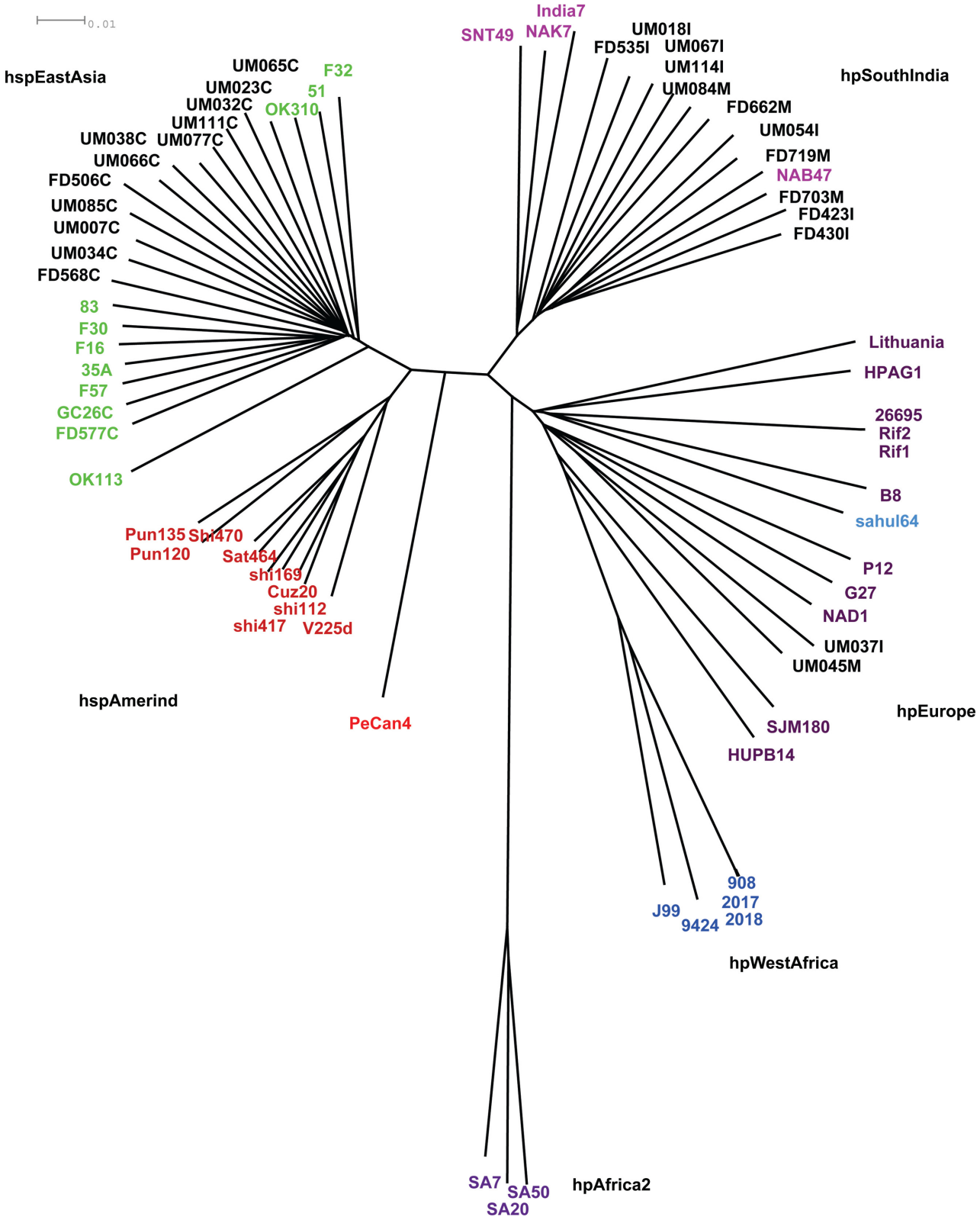


Figure 2. The whole genome phylogenetic analysis: the figure represents a whole genome comparison-based phylogenetic tree of various complete and draft *H. pylori* genomes from different geographical regions. The tree was constructed based on neighbor joining algorithm using SplitsTree. The Malaysian 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)

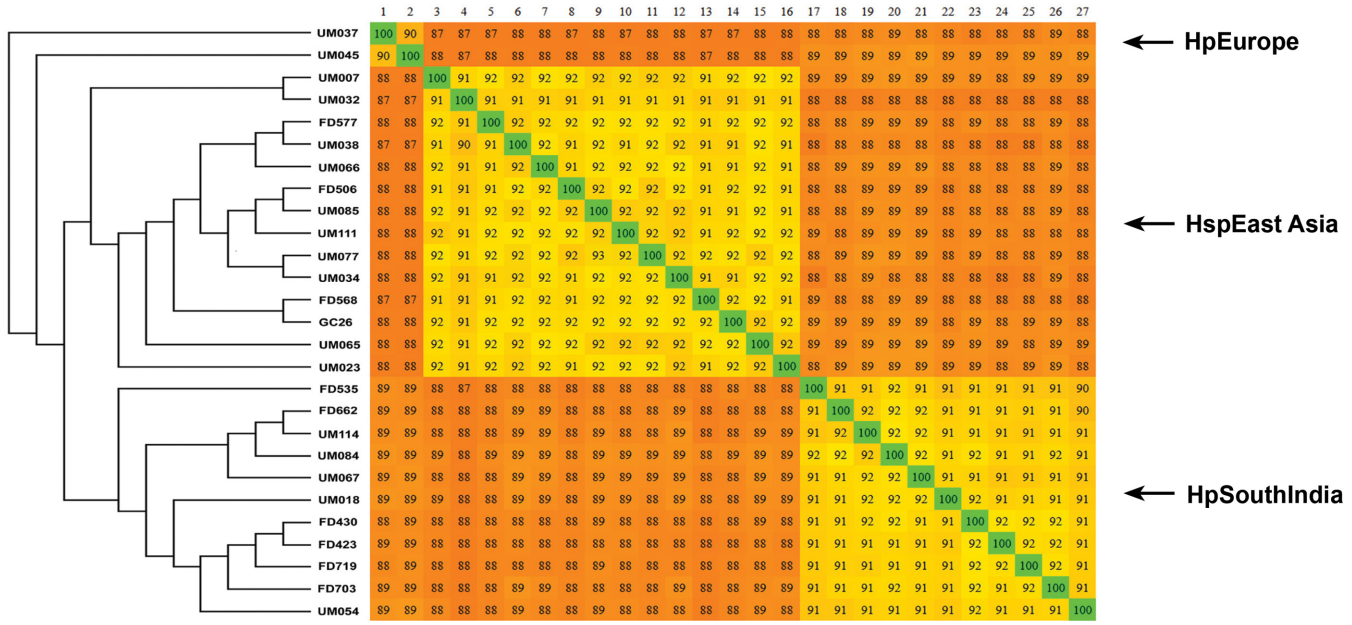


Figure 3. The analysis of 27 Malaysian genomes: the Neighbor-Joining phylogenetic tree constructed after the alignment of 27 Malaysian *H. pylori* isolates representing various ethnic groups. The heat plot shows average similarity values among the strains.

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).

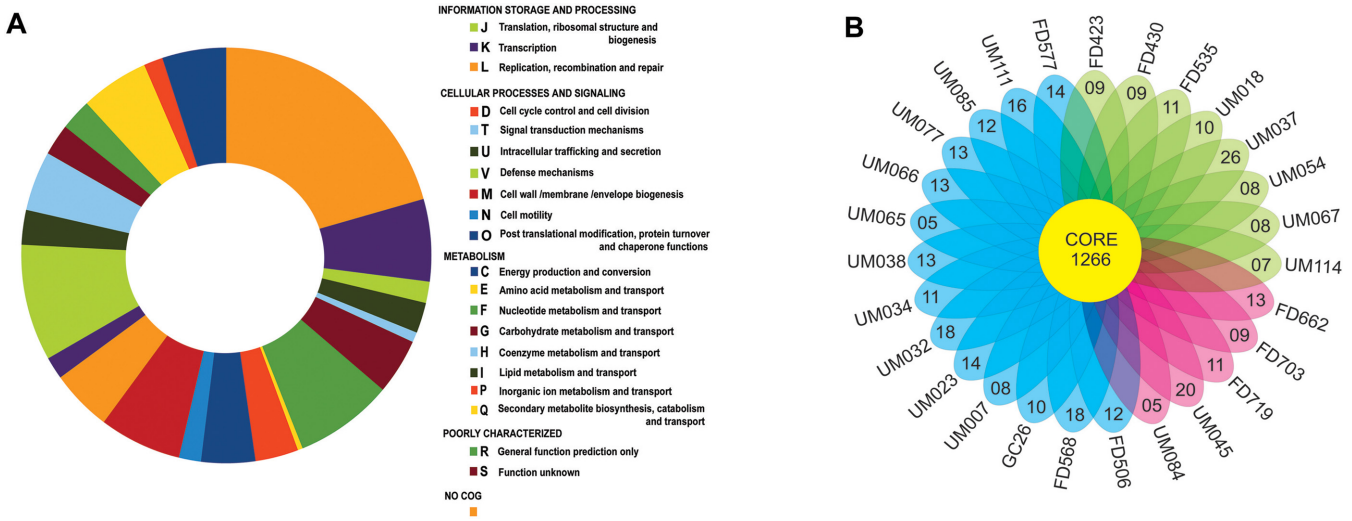


Figure 4. The functional COG classification of genes: (**A**) the COG functional classification representing core genome of the Malaysian *H. pylori* isolates. (**B**) Core and specific gene content observed among various strains compared in the study.

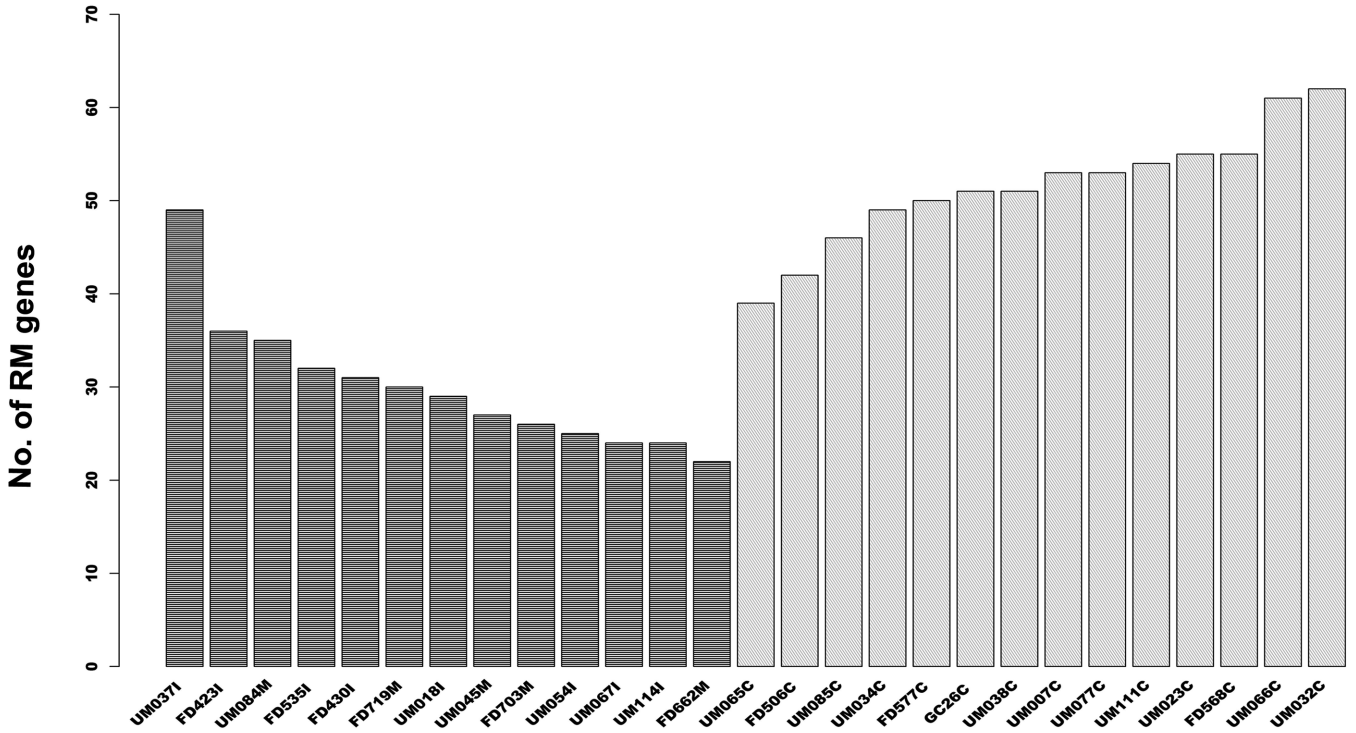


Figure 5. The distribution of RM genes in various strains: the graph shows the number of genes annotated to encode putative RM functions in each strain. The Y-axis represents the number of genes and the X-axis denotes strain names.

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-

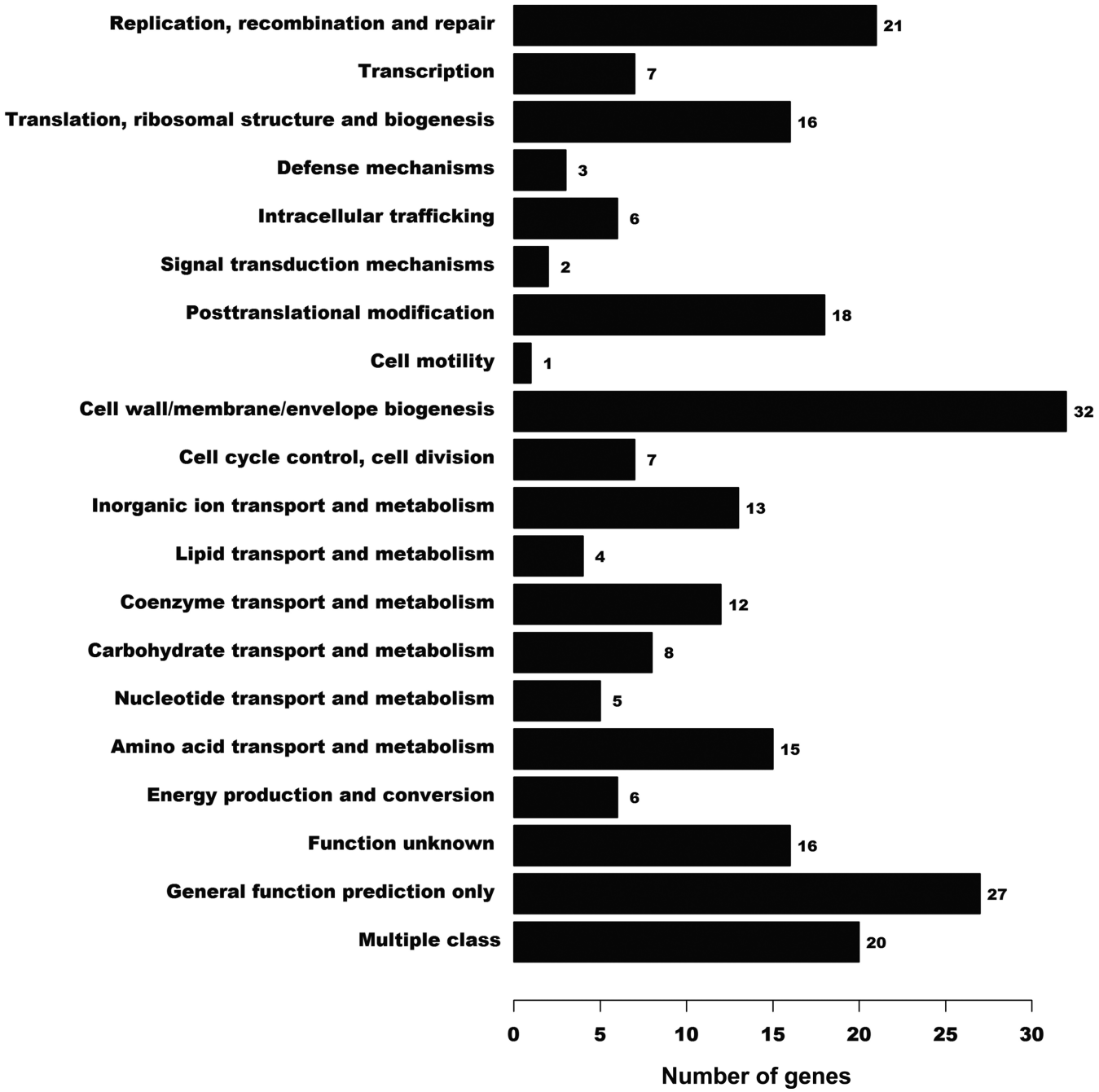


Figure 6. The functional classification of differentially evolving genes: the graph shows the COG functional classification of various core genes with signs of differential evolution among East-Asian and non-East-Asian strains. The Y-axis denotes the functional category and the X-axis represents the 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.

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

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

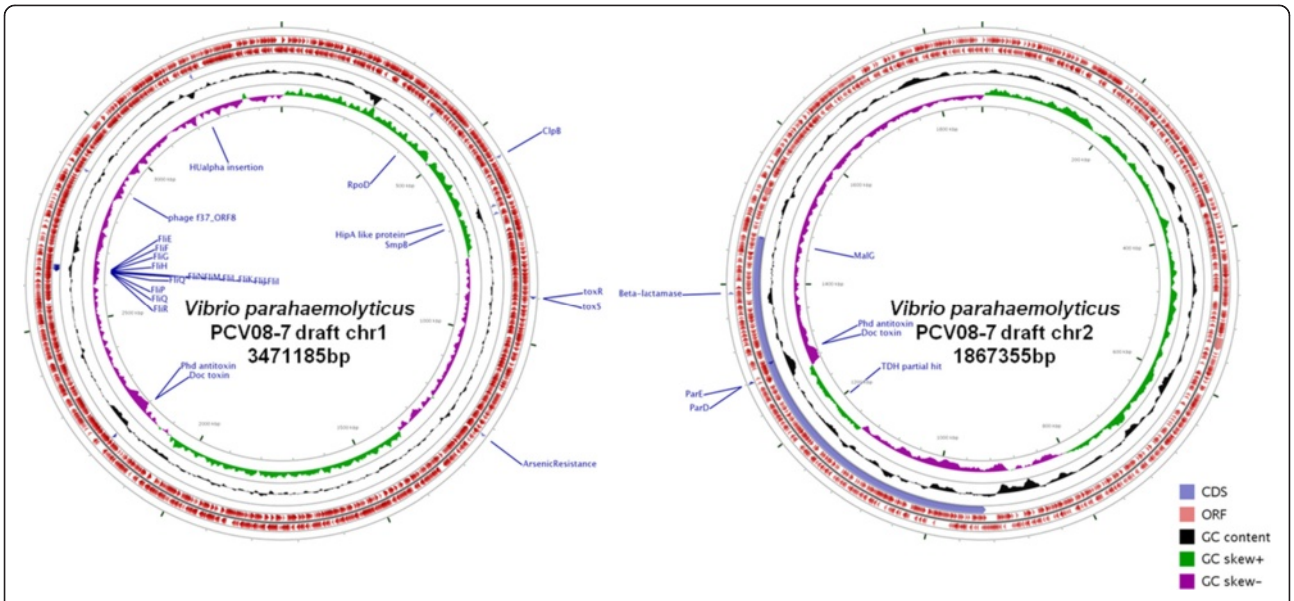


Figure 1 Circular view of *Vibrio parahaemolyticus* PCV08-7 draft genome. Diagrammatic representation of major genes carried by the two chromosomes of *Vibrio parahaemolyticus* PCV08-7 genome using CGview [25].

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

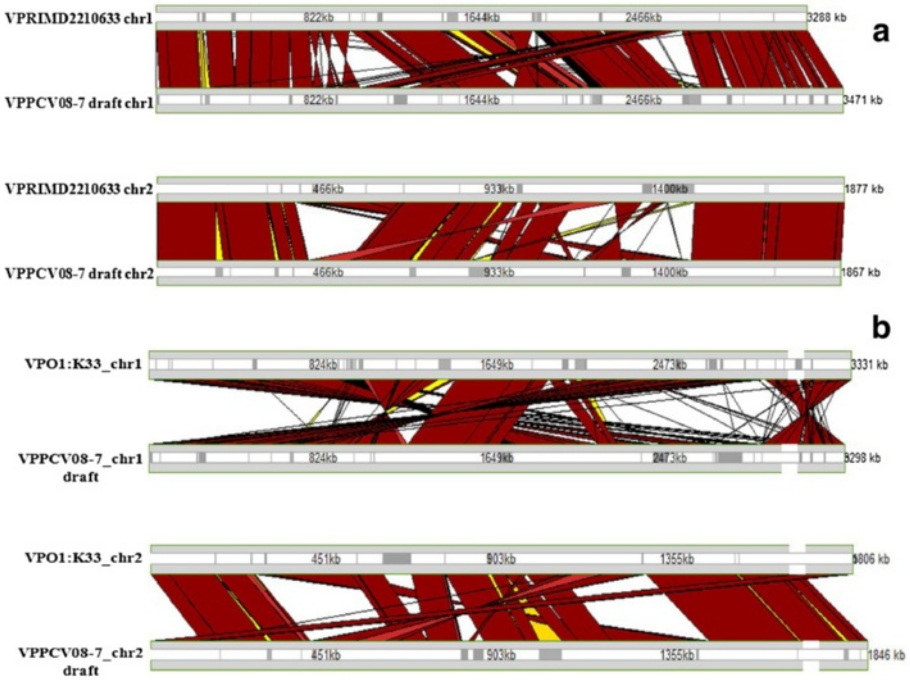


Figure 2 Alignment of the genome of strain RIMD2210633 against that of isolate PCV08-7 and strain O1:K33. (a) Comparison of chromosomes of strain RIMD2210633 (VPRIMD2210633 chr1, VPRIMD2210633 chr2) with the draft chromosomes of PCV08-7 (VPPCV08-7 draft chr1, VPPCV08-7 draft chr2) using M-GCAT. **(b)** 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.

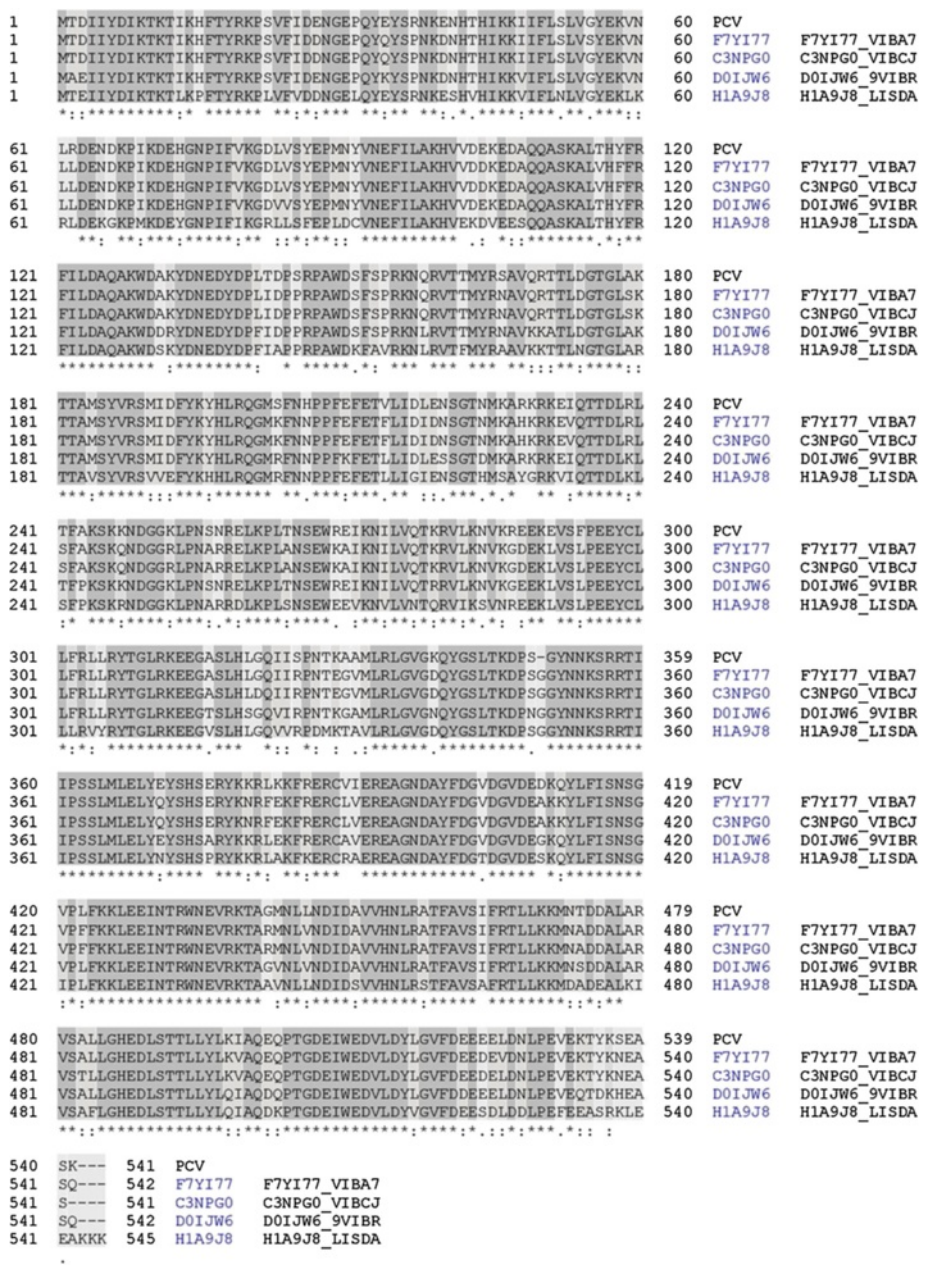


Figure 3 Alignment of a unique PCV08-7 protein sequence similar to *Photobacterium damselae subsps. damselae*. A unique sequence from PCV08-7 genome showed similarity with putative uncharacterized proteins of *V. anguillarum* 775 (F7YI77), *V. cholera* MJ-1236 (C3NPG0) and *Vibrio sp.* RC586 (D0IJW6) 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

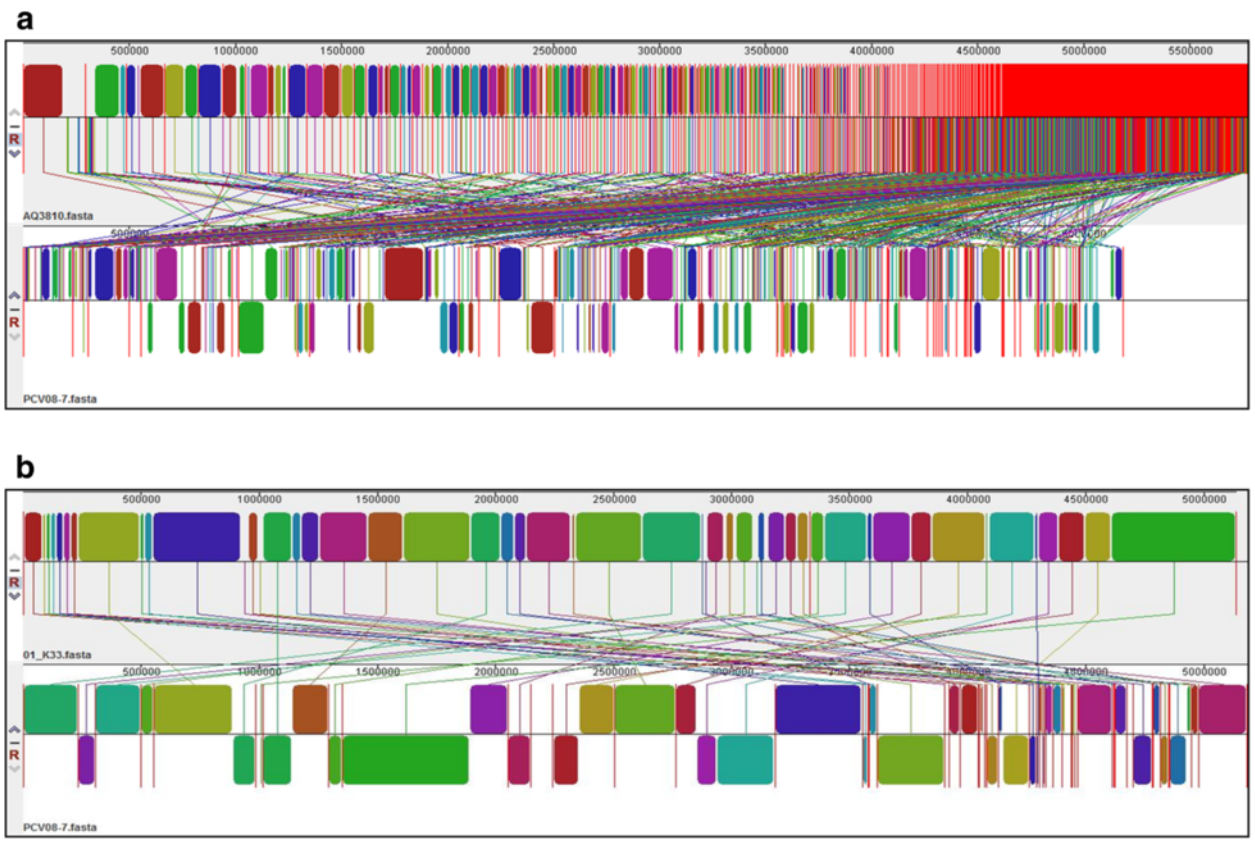


Figure 4 (a) ***V. parahaemolyticus*** **AQ3810 alignment against PCV08-7 genome: Concatenated chromosome 1 and chromosome 2 of** ***V. parahaemolyticus*** **AQ3810 (AQ3810.fasta) against** ***V. parahaemolyticus*** **PCV08-7 (PCV08-7.fasta) (b)** ***V. parahaemolyticus*** **O1:K33 alignment against PCV08-7.** Concatenated chromosome 1 and chromosome 2 of *V. parahaemolyticus* O1:K33 (O1_K33.fasta) against *V. parahaemolyticus* PCV08-7 (PCV08-7.fasta).

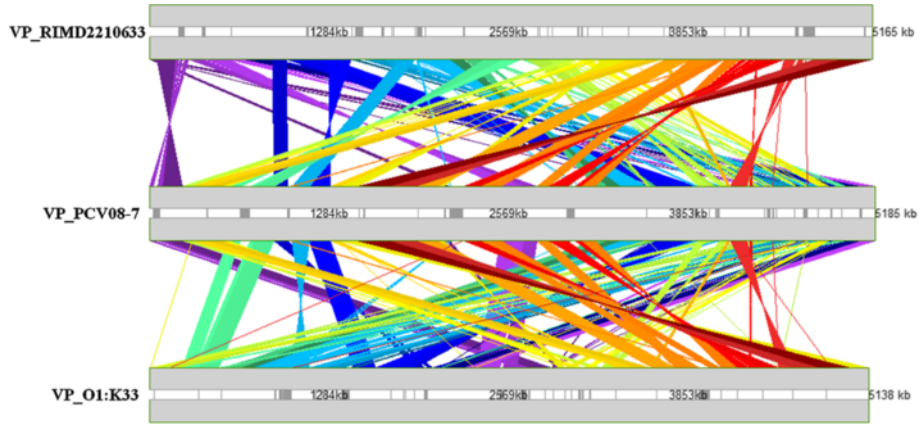


Figure 5 Comparison of whole genome sequences of strains RIMD2210633, PCV08-7 and O1:K33. Alignment of complete genomes of *V. parahaemolyticus* RIMD2210633, *V. parahaemolyticus* PCV08-7 and *V. parahaemolyticus* O1:K33, showing PCV08-7 being more similar to RIMD2210633.

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.

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.

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.

JOURNAL OF BACTERIOLOGY, Oct. 2011, p. 5598–5599 Vol. 193, No. 19
0021-9193/11/$12.00 doi:10.1128/JB.05667-11

Genome of a Novel Isolate of *Paracoccus denitrificans* Capable of Degrading *N*,*N*-Dimethylformamide

Dayananda Siddavattam,[1]* Timmanagouda B. Karegoudar,[2] Santosh Kumar Mudde,[2] Narender Kumar,[3] Ramani Baddam,[3] Tiruvayipati Suma Avasthi,[3] and Niyaz Ahmed[3,4,5]*

Department of Animal Sciences, School of Life Sciences, University of Hyderabad, Hyderabad, India[1]; Department of Biochemistry, Gulbarga University, Gulbarga, Karnataka, India[2]; Pathogen Biology Laboratory, Department of Biotechnology, School of Life Sciences, University of Hyderabad, Hyderabad, India[3]; Institute of Life Sciences, University of Hyderabad, Hyderabad, India[4]; and Institute of Biological Sciences, University of Malaya, Kuala Lumpur, Malaysia[5]

Received 27 June 2011/Accepted 8 July 2011

The bacterial genus *Paracoccus* is comprised of metabolically versatile organisms having diverse degradative capabilities and potential industrial and environmental applications for bioremediation in particular. We report a *de novo*-assembled sequence and annotation of the genome of a novel isolate of *Paracoccus denitrificans* originally sourced from coal mine tailings in India. The isolate was capable of utilizing *N*,*N*-dimethylformamide (DMF) as a source of carbon and nitrogen and therefore holds potential for bioremediation and mineralization of industrial pollutants. The genome sequence and biological circuitry revealed thereupon will be invaluable in understanding the metabolic capabilities, functioning, and evolution of this important bacterial organism.

Paracoccus denitrificans is a Gram-negative coccoid bacterium capable of thriving in soil under either aerobic or anaerobic conditions. An important characteristic of this bacterium is its ability to single-handedly convert nitrate to dinitrogen via a process called denitrification (2). *Paracoccus* bacteria have gained significant attention as model organisms due to the overlap with some of the important features specific to mitochondria and that they presumably constitute the ancestors of eukaryotic mitochondria (7, 15). Given this, genomics-enabled insights into the evolution of this taxon are very important.

The Indian isolate of *P. denitrificans* described herein was obtained from coal mine leftovers and was cultured in mineral salts medium (MM1) devoid of any traditional carbon and nitrogen source but fortified with *N*,*N*-dimethylformamide (DMF) (0.5% [vol/vol]). The isolate was previously identified as *Ochrobactrum* species based on partial 16S rRNA gene typing (14); however, the genome sequence provides a clear basis for its identity as *P. denitrificans*.

The genome sequence was determined by Illumina genome analyzer (GA2x, pipeline version l.6) and comprised of sequence traces equivalent to 890 megabytes of data encompassing 72-bp paired-end reads with insert size of 300 bp, and the genome coverage achieved was about 60 times. In addition to the chromosomal complement, our *P. denitrificans* isolate also carried a 60-kb plasmid which has yet to be analyzed with respect to the genes that it carries. The chromosomal sequence was assembled *de novo* into contigs using Velvet (16) assembly tool with a hash length of 39. The draft genome was annotated with the help of the RAST server (1) comparing outputs from GLIMMER (6), GeneMark (3), and EasyGene (11). Artemis (12) was used to obtain the following statistics of the genome. In addition, tRNAscan-SE (13) was used to scan for the total number of tRNAs. The size of the *P. denitrificans* Indian isolate was approximately 3 Mb (2985589 bp) with a G+C content of 65.65% and a coding percentage of 82.4 with 3,744 protein-coding sequences of an average length of 662 bp. The genome revealed 34 tRNA genes (one gene encoding a selenocyteine) and 3 rRNA genes. Analyses for tracing important genes were carried out using the alignment tools Mauve (4, 5), BLAT (9), and Mummer (10). The BioCyc (8) database was used for coordinates of the genes in comparison with the *P. denitrificans* PD1222 genome. Further, NCBI BLAST was used for manual curation. Regions containing genes encoding nitrite transporter, nitrite reductase, nitrate reductase, and ferredoxin were readily located. A few RuBisCO genes along with genes conferring resistance to fosmidomycin, ethidium bromide viologen, tellurium, bicyclomycin, arsenic, and acriflavin were found. In addition, enzymes involved in mineralization of DMF (dimethyl formidase, dimethyl amine dehydrogenase, and monomethyl amine dehydrogenase), cytochrome *c* oxidases, and various proteins related to the cytochrome family were identified.

These observations and the ensuing comparative genomic analyses shall be extremely useful in both furthering fundamental understanding of bioremediation mechanisms encoded by this and other *Paracoccus* species and working toward identification of the scientific basis to weigh the beneficial and harmful effects of such organisms in the environment.

Nucleotide sequence accession number. The genome sequence is deposited in GenBank under accession number CP002897.

* Corresponding author. Mailing address for Dayananda Siddavattam: Department of Animal Sciences, School of Life Sciences, University of Hyderabad, Professor CR Rao Road, Gachibowli, Hyderabad 500 046, India. Phone: 91 40 23134578. Fax: 91 40 23010120. E-mail: sdSL@uohyd.ernet.in. Mailing address for Niyaz Ahmed: Pathogen Biology Laboratory, Department of Biotechnology, School of Life Sciences, University of Hyderabad, Professor CR Rao Road, Gachibowli, Hyderabad 500 046, India. Phone: 91 40 23134585. Fax: 91 40 66794585. E-mail: niyazSL@uohyd.ernet.in.

This genome program was supported by the University of Hyderabad through an interim grant to N.A. as a part 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/University Grants Commission India) of which N.A. is a Speaker. D.S. and T.B.K. acknowledge support from the Department of Biotechnology of the Indian Government for undertaking research on DMF degradation.

We are grateful to M/s Genotypic Technology Pvt. Ltd., Bengaluru, India, for their help with Illumina sequencing. We acknowledge project management/laboratory facility support rendered by Gutti Navamallika.

REFERENCES

1. **Aziz, R. K., et al.** 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics **9:**75.
2. **Baumann, B., M. Snozzi, A. J. Zehnder, and J. R. Van Der Meer.** 1996. Dynamics of denitrification activity of *Paracoccus denitrificans* in continuous culture during aerobic-anaerobic changes. J. Bacteriol. **178:**4367–4374.
3. **Besemer, J., and M. Borodovsky.** 2005. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res. **33:**W451–W454.
4. **Darling, A. C., B. Mau, F. R. Blattner, and N. T. Perna.** 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. **14:**1394–1403.
5. **Darling, A. E., B. Mau, and N. T. Perna.** 2010. Progressive Mauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One **5:**e11147.
6. **Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg.** 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. **27:**4636–4641.
7. **John, P., and F. R. Whatley.** 1975. *Paracoccus denitrificans* and the evolutionary origin of the mitochondrion. Nature **254:**495–498.
8. **Karp, P. D., et al.** 2005. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Res. **3:**6083–6089.
9. **Kent, W.** 2002. BLAT—the BLAST-like alignment tool. Genome Res. **12:**656–664.
10. **Kurtz, S., et al.** 2004. Versatile and open software for comparing large genomes. Genome Biol. **5:**R12.
11. **Larsen, T. S., and A. Krogh.** 2003. EasyGene–a prokaryotic gene finder that ranks ORFs by statistical significance. BMC Bioinformatics **4:**21.
12. **Rutherford, K., et al.** 2000. Artemis: sequence visualization and annotation. Bioinformatics **16:**944–945.
13. **Schattner, P., A. N. Brooks, and T. M. Lowe.** 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. **33:**W686–W689.
14. **Veeranagouda, Y., P. V. Emmanuel Paul, P. Gorla, D. Siddavattam, and T. B. Karegoudar.** 2006. Complete mineralisation of dimethylformamide by *Ochrobactrum* sp. DGVK1 isolated from the soil samples collected from the coalmine leftovers. Appl. Microbiol. Biotechnol. **71:**369–375.
15. **Yip, C. Y., M. E. Harbour, K. Jayawardena, I. M. Fearnley, and L. A. Sazanov.** 2011. Evolution of respiratory complex I: "supernumerary" subunits are present in the alpha-proteobacterial enzyme. J. Biol. Chem. **286:**5023–5033.
16. **Zerbino, D. R., and E. Birney.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18:**821–829.

JOURNAL OF BACTERIOLOGY, Aug. 2011, p. 4272–4273 Vol. 193, No. 16
0021-9193/11/$12.00 doi:10.1128/JB.05413-11

Genome of Multidrug-Resistant Uropathogenic *Escherichia coli* Strain NA114 from India▽

Tiruvayipati Suma Avasthi,[1]† Narender Kumar,[1]† Ramani Baddam,[1]† Arif Hussain,[1]*
Nishant Nandanwar,[1] Savita Jadhav,[2] and Niyaz Ahmed[1,3,4]*

Pathogen Biology Laboratory, Department of Biotechnology, School of Life Sciences, University of Hyderabad, Hyderabad, India[1]; Department of Microbiology, Dr. D. Y. Patil Medical College, Dr. D. Y. Patil University, Pimpri, Pune, India[2]; Institute of Life Sciences, University of Hyderabad Campus, Hyderabad, India[3]; and Institute of Biological Sciences, University of Malaya, Kuala Lumpur, Malaysia[4]

Received 31 May 2011/Accepted 6 June 2011

Uropathogenic *Escherichia coli* (UPEC) causes serious infections in people at risk and has a significant environmental prevalence due to contamination by human and animal excreta. In developing countries, UPEC assumes importance in certain dwellings because of poor community/personal hygiene and exposure to contaminated water or soil. We report the complete genome sequence of *E. coli* strain NA114 from India, a UPEC strain with a multidrug resistance phenotype and the capacity to produce extended-spectrum beta-lactamase. The genome sequence and comparative genomics emanating from it will be significant in understanding the genetic makeup of diverse UPEC strains and in boosting the development of new diagnostics/vaccines.

Pathogenic *Escherichia coli* constitutes a significant threat to public health, and the emergence of extended-spectrum beta-lactamase (ESBL)-producing *E. coli* with high virulence potential is alarming (14, 16, 20, 22, 23). Comparative genomics holds significant promise in understanding the genome organization of such bacteria and thereby identifying coordinates highly relevant in the development of intervention strategies (1). Our group has recently studied uropathogenic *E. coli* (UPEC) from the western Indian city of Pune (11), whereupon strain NA114 emerged as an ideal representative of the entire Pune collection. The three major characteristics of strain NA114 that make it epidemiologically and clinically significant are its affiliation with serogroup O25, its placement in phylogenetic group B2, and its sequence type, ST131 (19). The latter denotes a pandemic clone frequently associated with community-acquired antimicrobial-resistant infections (23). Motivated by these facts, we performed complete in-depth sequencing, annotation, and analysis of the genome of UPEC strain NA114, which was originally obtained from the urine of a 70-year-old male patient with prostatitis from Pune. Antibiotic sensitivity tests revealed that it was a multidrug-resistant strain refractory to several common antibiotics and was an ESBL producer (11).

(This work constitutes part of the unpublished doctoral work of Arif Hussain.)

The genome sequence was determined by Illumina Genome Analyzer (GA2x, pipeline ver. l.6) and consisted of sequence traces equivalent to 8 gigabytes of data, encompassing 54-bp paired-end reads with an insert size of 300 bp, and the genome coverage achieved was 500×. The sequence was assembled using Velvet (26), and the contigs were ordered with respect to the best-aligned positions compared to the reference genome of *E. coli* SE15 (25) using Mauve (5, 6). The genome alignment tools BLAT (15) and MUMmer (17) were also used to validate the aligned contigs. The genome was annotated with the help of the RAST server (2), and putative CDSs were identified by comparing outputs from Glimmer (7), Genemark (4), and EasyGene (18). Artemis (24) was used to glean the following details of the genome.

The size of the NA114 chromosome was 4,935,666 bp with a G+C content of 51.16% and a coding percentage of 88.4% with 4,875 protein coding sequences with an average length of 901 bp. The genome revealed 67 tRNA and 3 rRNA genes. We also found several virulence genes, including *iha*, *sat*, *fimH*, *kpsM*, *iutA*, and *malX*, which correspond to the genes of *E. coli* CFT073 (9). In addition, genes corresponding to another UPEC strain, UTI89, such as *fyuA* and *usp* etc., were located. PCR-based analysis showed that this strain carried multiple virulence genes infrequently described in a clone of this type, including *sfa*, *aer*, *cnf*, and an intact polyketide synthase (*pks*) island (12). *E. coli* NA114 also contains other virulence factors, such as *pap*, *fim*, and genes for iron uptake systems such as the hemin uptake system and the yersiniabactin siderophore (*ybt*). In addition to a 4.935-Mb chromosomal genome, strain NA114 also harbored a single plasmid of 3.5 kb which has yet to be analyzed with regard to its replicon type and resistance gene profiles, if it has any.

These observations and the comparative genomic studies emanating therefrom could be extremely useful both in improving our fundamental understanding of multidrug resistance mechanisms encoded by UPEC and in the design of effective drugs to control and manage the alarming health hazards caused by ESBL-producing bacteria in both the developing and developed parts of the world.

* Corresponding author. Mailing address: Pathogen Biology Laboratory, Department of Biotechnology, School of Life Sciences, University of Hyderabad, Professor CR Rao Road, Gachibowli, Hyderabad 500 046, India. Phone: 91 40 23134585. Fax: 91 40 66794585. E-mail for Niyaz Ahmed: niyazSL@uohyd.ernet.in. E-mail for Arif Hussain: pbl_uoh@gmail.com.
† These authors contributed equally.
▽ Published ahead of print on 17 June 2011.

Nucleotide sequence accession number. The genome sequence of *E. coli* NA114 has been deposited in GenBank under accession no. CP002797.

This genome program was supported by the University of Hyderabad through an interim grant to N.A. as a part 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/University Grants Commission 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 thankful to Seyed E. Hasnain, Joerg Hacker, Lothar H. Wieler, and Christa Ewers for helpful advice and suggestions. Akash Ranjan is gratefully acknowledged for timely support and discussions. We are grateful to M/s Genotypic Technology Pvt. Ltd., Bengaluru, India, for their help with Illumina sequencing. We are also thankful to the authorities at Dr. D. Y. Patil University, Pune, India, for their cooperation and facilitation of this study. We acknowledge project management/laboratory facilitation support rendered by Gutti Navamallika.

REFERENCES

1. **Ahmed, N., U. Dobrindt, J. Hacker, and S. E. Hasnain.** 2008. Genomic fluidity and pathogenic bacteria: applications in diagnostics, epidemiology and intervention. Nat. Rev. Microbiol. **6:**387–394.
2. **Aziz, R. K., et al.** 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics **9:**75.
3. Reference deleted.
4. **Besemer, J., and M. Borodovsky.** 2005. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res. **33:**W451–W454.
5. **Darling, A. C., B. Mau, F. R. Blattner, and N. T. Perna.** 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. **14:**1394–1403.
6. **Darling, A. E., B. Mau, and N. T. Perna.** 2010. Progressive Mauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One **5:**e11147.
7. **Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg.** 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. **27:**4636–4641.
8. Reference deleted.
9. **Guyer, D. M., J. S. Kao, and H. L. Mobley.** 1998. Genomic analysis of a pathogenicity island in uropathogenic *Escherichia coli* CFT073: distribution of homologous sequences among isolates from patients with pyelonephritis, cystitis, and catheter-associated bacteriuria and from fecal samples. Infect. Immun. **66:**4411–4417.
10. Reference deleted.
11. **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:**e18063.
12. **Johnson, J. R., B. Johnston, M. A. Kuskowski, J. P. Nougayrede, and E. Oswald.** 2008. Molecular epidemiology and phylogenetic distribution of the *Escherichia coli pks* genomic island. J. Clin. Microbiol. **46:**3906–3911.
13. Reference deleted.
14. **Kaper, J. B., J. P. Nataro, and H. L. T. Mobley.** 2004. Pathogenic *Escherichia coli*. Nat. Rev. Microbiol. **2:**123.
15. **Kent, W.** 2002. BLAT—the BLAST-like alignment tool. Genome Res. **12:**656–664.
16. **Kumarasamy, K. K., M. A. Toleman, T. R. Walsh, J. Bagaria, F. Butt, et al.** 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.
17. **Kurtz, S., et al.** 2004. Versatile and open software for comparing large genomes. Genome Biol. **5:**R12.
18. **Larsen, T. S., and A. Krogh.** 2003. EasyGene—a prokaryotic gene finder that ranks ORFs by statistical significance. BMC Bioinformatics **4:**21.
19. **Lau, S. H., et al.** 2008. UK epidemic *Escherichia coli* strains A-E, with CTX-M-15 b-lactamase, all belong to the international O25:H4-ST131 clone. J. Antimicrob. Chemother. **62:**1241–1244.
20. **Livermore, D. M.** 2009. Beta-lactamases—the threat renews. Curr. Protein Pept. Sci. **10:**397–400.
21. Reference deleted.
22. **Miriagou, V., et al.** 2010. Acquired carbapenemases in Gram-negative bacterial pathogens: detection and surveillance issues. Clin. Microbiol. Infect. **16:**112–122.
23. **Rogers, B. A., H. E. Sidjabat, and D. L. Paterson.** 2011. *Escherichia coli* O25b-ST131: a pandemic, multiresistant, community-associated strain. J. Antimicrob. Chemother. **66:**1–14.
24. **Rutherford, K., et al.** 2000. Artemis: sequence visualization and annotation. Bioinformatics **16:**944–945.
25. **Toh, H., et al.** 2010. Complete genome sequence of the wild-type commensal *Escherichia coli* strain SE15, belonging to phylogenetic group B2. J. Bacteriol. **192:**1165–1166.
26. **Zerbino, D. R., and E. Birney.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18:**821–829.

JOURNAL OF BACTERIOLOGY, July 2011, p. 3385–3386 Vol. 193, No. 13
0021-9193/11/$12.00 doi:10.1128/JB.05006-11

Genomes of Two Chronological Isolates (*Helicobacter pylori* 2017 and 2018) of the West African *Helicobacter pylori* Strain 908 Obtained from a Single Patient▽

Tiruvayipati Suma Avasthi,[1]† Singamaneni Haritha Devi,[1]† Todd D. Taylor,[2] Narender Kumar,[1] Ramani Baddam,[1] Shinji Kondo,[2] Yutaka Suzuki,[3] Hervé Lamouliatte,[4] Francis Mégraud,[5,6] and Niyaz Ahmed[1,7,8]*

Pathogen Biology Laboratory, Department of Biotechnology, School of Life Sciences, University of Hyderabad, Hyderabad, India[1]; Quantitative Biology Center, RIKEN, 1-7-22 Suehiro, Tsurumi, Yokohama, Kanagawa 230-0045, Japan[2]; Graduate School of Frontier Sciences, University of Tokyo, Tokyo, Japan[3]; CHU de Bordeaux, Hospital Saint André, 33077 Bordeaux, France[4]; Université de Bordeaux, Laboratoire de Bactériologie, 33000 Bordeaux, France[5]; INSERM U853, 33076 Bordeaux Cedex, France[6]; Institute of Life Sciences, University of Hyderabad Campus, Hyderabad, India[7]; and Institute of Biological Sciences, University of Malaya, Kuala Lumpur, Malaysia[8]

Received 4 April 2011/Accepted 12 April 2011

The diverse clinical outcomes of colonization by *Helicobacter pylori* reflect the need to understand the genomic rearrangements enabling the bacterium to adapt to host niches and exhibit varied colonization/virulence potential. We describe the genome sequences of the two serial isolates, *H. pylori* 2017 and 2018 (the chronological subclones of *H. pylori* 908), cultured in 2003 from the antrum and corpus, respectively, of an African patient who suffered from recrudescent duodenal ulcer disease. When compared with the genome of the parent strain, 908 (isolated from the antrum of the same patient in 1994), the genome sequences revealed genomic alterations relevant to virulence optimization or host-specific adaptation.

The high genetic variability of *Helicobacter pylori* (1–4) points to its capacity toward adaptive evolution (6, 9, 12, 16, 17). DNA profiling reveals minor differences in clinical or related strains, suggestive of microevolution (5, 18, 19, 22, 25, 28), although such methods do not explain the underlying rearrangements. Multiple genome sequences of *H. pylori* better explain its lifestyle and evolution (7, 11, 15, 21, 24, 27). However, chronological isolates, especially those obtained from single patients (18, 19, 25) or single families (28), have not been sequenced.

H. pylori isolates 2017 and 2018 represent chronological subclones of strain 908 recovered after a decade of the original isolation of the parent strain from a West African duodenal ulcer disease patient in France (8, 25). Recently, strain 908 was completely sequenced by our group (15). Herein, we report full genome sequences of the subsequent isolates, 2017 and 2018.

Genomes were determined by Illumina Genome Analyzer (GA2x, pipeline version l.6) and comprised of sequence reads equivalent to 60 Mb for each isolate, encompassing 101-bp paired-end reads with an insert size of 300 bp, and the genome coverage achieved was 50X (15). The sequence reads were assembled using Velvet (29) with the hash length set to 21 (15). In view of the phylogenetic relatedness of 908 to *H. pylori* J99 (8, 15), the assembled contigs were ordered with respect to the best-aligned positions when compared to the genome of reference strain J99 using BLAT (20). The genomes were annotated with the help of the RAST server (10), and putative coding sequences (CDSs) were identified by comparing outputs from Glimmer (14), Genemark (13), and EasyGene (23). Finally, manual curation was carried out. Artemis (26) was used to glean the following details of the two genomes.

The sizes of the 2017 and 2018 draft genomes were 1,548,238 and 1,562,832 bp, respectively, with G+C contents of 39.3 and 39.29%, respectively. The genomes of 2017 and 2018 revealed coding percentages of 91.5 and 91.6, respectively, and contained 1,593 and 1,603 protein coding sequences, respectively, with average lengths of 894 and 896 bp, respectively. Each of the genomes had 36 tRNA and 3 rRNA genes, and a few pseudogenes and putative phagelike products were identified. Both the genomes displayed a conserved repertoire of housekeeping genes corresponding to various metabolic pathways, a largely intact *cag*PAI, the genomic island *tfs3*, and virulence-associated alleles of *vacA*, as also described earlier (25), and revealed the presence of several plasticity zone open reading frames (ORFs) and putative virulence factors. Comparative genomic analysis of the 2017 and 2018 genomes revealed that they are almost identical and descended from that of strain 908 (15).

In conclusion, the genome sequences prove the clonal origin of the three isolates (908, 2017, and 2018) and thus reinforce our stance that the patient under study did in fact harbor only a single strain which survived eradication therapy (25) and that the subclones, 2017 and 2018, did not represent exogenous reinfection by a new source.

Nucleotide sequence accession numbers. The genome sequences for 2017 and 2018 are deposited in GenBank under accession numbers CP002571 and CP002572, respectively.

* Corresponding author. Mailing address: Pathogen Biology Laboratory, Department of Biotechnology, School of Life Sciences, University of Hyderabad, Professor CR Rao Road, Gachibowli, Hyderabad 500 046, India. Phone: 91 40 23134585. Fax: 91 40 66794585. E-mail: niyazSL@uohyd.ernet.in.

† These authors contributed equally to this study.

▽ Published ahead of print on 22 April 2011.

The genome program was carried out under the wider umbrella of the European *Helicobacter* Study Group (EHSG), of which N.A. and F.M. are fellows. Functional analysis of these genomes in the context of chronological evolution is part of the Indo-German International Research Training Group—Internationales Graduiertencolleg (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. S.H.D. received her postdoctoral fellowship under the UoH-DBT/CREBB program of the University of Hyderabad and the Indian Department of Biotechnology of the Ministry of Science and Technology.

We are thankful to Barry Marshall, Leonardo A. Sechi, and Ramy K. Aziz for helpful advice and suggestions. We are also grateful to M/s Genotypic Technology Pvt. Ltd. Bengaluru, India, for their unqualified efforts in training T.S.A. and N.K. in next-generation sequencing platforms and data analysis; our specific thanks are due to Raja Mugasimangalam and Sudha Narayana Rao and Vidya Niranjan of Genotypic for assistance with resequencing of the genome of strain 908.

REFERENCES

1. **Ahmed, N.** 2010. Replicative genomics can help *Helicobacter* fraternity usher in good times. Gut Pathog. **2:**25.
2. **Ahmed, N.** 2009. A flood of microbial genomes—do we need more? PLoS One **4:**e5831.
3. **Ahmed, N., S. Tenguria, and N. Nandanwar.** 2009. *Helicobacter pylori*—a seasoned pathogen by any other name. Gut Pathog. **1:**24.
4. **Ahmed, N., U. Dobrindt, J. Hacker, and S. E. Hasnain.** 2008. Genomic fluidity and pathogenic bacteria: applications in diagnostics, epidemiology and intervention. Nat. Rev. Microbiol. **6:**387–394.
5. **Akopyanz, N. S., N. O. Bukanov, T. U. Westblom, S. Kresovich, and D. E. Berg.** 1992. DNA diversity among clinical isolates of *Helicobacter pylori* detected by PCR-based RAPD fingerprinting. Nucleic Acids Res. **20:**5137–5142.
6. **Akopyants, N. S., K. A. Eaton, and D. E. Berg.** 1995. Adaptive mutation and cocolonization during *Helicobacter pylori* infection of gnotobiotic piglets. Infect. Immun. **63:**116–121.
7. **Alm, R. A., et al.** 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature **397:**176–180.
8. **Alvi, A., et al.** 2007. Microevolution of *Helicobacter pylori* type IV secretion systems in an ulcer disease patient over a ten-year period. J. Clin. Microbiol. **45:**4039–4043.
9. **Atherton, J. C., and M. J. Blaser.** 2009. Coadaptation of *Helicobacter pylori* and humans: ancient history, modern implications. J. Clin. Invest. **119:**2475–2487.
10. **Aziz, R. K., et al.** 2008. The RAST server: Rapid Annotations using Subsystems Technology. BMC Genomics **9:**75.
11. **Baltrus, D. A., et al.** 2009. The complete genome sequence of *Helicobacter pylori* strain G27. J. Bacteriol. **191:**447–448.
12. **Baltrus, D. A., K. Guillemin, and P. C. Phillips.** 2008. Natural transformation increases the rate of adaptation in the human pathogen *Helicobacter pylori*. Evolution **62:**39–49.
13. **Besemer, J., and M. Borodovsky.** 2005. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res. **33:**W451–W454.
14. **Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg.** 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. **27:**4636–4641.
15. **Devi, S. H., et al.** 2010. Genome of *Helicobacter pylori* strain 908. J. Bacteriol. **192:**6488–6489.
16. **Falush, D., et al.** 2001. Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. Proc. Natl. Acad. Sci. U. S. A. **98:**15056–15061.
17. **Gressmann, H., et al.** 2005. Gain and loss of multiple genes during the evolution of *Helicobacter pylori*. PLoS Genet. **1:**e43.
18. **Gustavsson, A., M. Unemo, B. Blomberg, and D. Danielsson.** 2005. Genotypic and phenotypic stability of *Helicobacter pylori* markers in a nine-year follow-up study of patients with noneradicated infection. Dig. Dis. Sci. **50:**375–380.
19. **Israel, D. A., et al.** 2001. *Helicobacter pylori* genetic diversity within the gastric niche of a single human host. Proc. Natl. Acad. Sci. U. S. A. **98:**14625–14630.
20. **Kent, W.** 2002. BLAT—the BLAST-like alignment tool. Genome Res. **12:**656–664.
21. **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.
22. **Kersulyte, D., H. Chalkauskas, and D. E. Berg.** 1999. Emergence of recombinant strains of *Helicobacter pylori* during human infection. Mol. Microbiol. **31:**31–43.
23. **Larsen, T. S., and A. Krogh.** 2003. EasyGene—a prokaryotic gene finder that ranks ORFs by statistical significance. BMC Bioinformatics **4:**21.
24. **Oh, J. D., et al.** 2006. The complete genome sequence of a chronic atrophic gastritis *Helicobacter pylori* strain: evolution during disease progression. Proc. Natl. Acad. Sci. U. S. A. **103:**9999–10004.
25. **Prouzet-Mauleon, V., et al.** 2005. Pathogen evolution in vivo: genome dynamics of two isolates obtained 9 years apart from a duodenal ulcer patient infected with a single *Helicobacter pylori* strain. J. Clin. Microbiol. **43:**4237–4241.
26. **Rutherford, K., et al.** 2000. Artemis: sequence visualization and annotation. Bioinformatics **16:**944–945.
27. **Tomb, J. F., et al.** 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature **388:**539–547.
28. **Van der Ende, A., et al.** 1996. Heterogeneous *Helicobacter pylori* isolates from members of a family with a history of peptic ulcer disease. Gastroenterology **111:**638–647.
29. **Zerbino, D. R., and E. Birney.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18:**821–829.