

**Genome dynamics of human gastric pathogen,
*Helicobacter pylori***

**Thesis Submitted to the University of Hyderabad
For the Degree of**

DOCTOR OF PHILOSOPHY

By

NARENDER KUMAR

(Reg. No: 11LTPH01)



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

July, 2015

University of Hyderabad

(A Central University by an Act of Parliament)
Department of Biotechnology and Bioinformatics
School of Life Sciences



DECLARATION

The research work presented in the thesis entitled “**Genome dynamics of human gastric pathogen, *Helicobacter pylori***”, 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 : Narender Kumar

Reg. No. : 11LTPH01

University of Hyderabad

(A Central University by an Act of Parliament)
Department of Biotechnology and Bioinformatics
School of Life Sciences



CERTIFICATE

This is to certify that Mr. Narender Kumar 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 dynamics of human gastric pathogen, *Helicobacter pylori*”** for submission for the degree of Doctor of Philosophy of this University.

Date:

Dr. Niyaz Ahmed
Research supervisor & Head
Dept. of Biotechnology &
Bioinformatics

Date:

Prof. P. Reddanna
Dean
School of Life Sciences

Dedicated to my Parents & Family members
For their infinite love and support

ACKNOWLEDGEMENT

I would like to acknowledge my supervisor and mentor Dr. Niyaz Ahmed for his constant support and guidance throughout the tenure of this research work. I am really thankful for his encouragement and confidence he showed in me. I would like to heart fully acknowledge his openness to new ideas that allowed me to venture into the field of Bioinformatics. I see him as my motivation and would continue improve upon and learn.

I would also like to thank my doctoral research committee members for their constant review and comments on my research work. The help extended by the Department of Biotechnology & Bioinformatics and support extended by various Heads of the department is greatly acknowledged. I would like to thank various Deans of School of Life Sciences for their continued assistance during my research tenure. All the faculties of Life Sciences from whom I gained the motivation for research are also acknowledged. I am also thankful to the non-teaching staff of School of Life Sciences, Dept. of Biotechnology & Bioinformatics and Pathogen Biology Laboratory for their support during the official work and cooperation. The funding support of CSIR, DBT, ICMR, UM - HIR, IRTG – GRK 1673, EUROMATHS-FP7, DBT-BIF, DBT-CREBB, DST-FIST, UGC-SAP, UPE are also greatly acknowledged.

I would like to thank Prof. Barry James Marshal (The Nobel Laureate) for the enriching discussions and guidance and source of motivation during my research work. The help from Prof. Jamuna Vadivelu for providing strains and Dr. Vanitha Mariappan for help during my work on Malaysian strains is also acknowledged. I would like to Prof. Carmen Molina Paris for her support and guidance during our visit to University of Leeds.

I would like to express my thanks to the members of NGS team Ramani, Sabiha, Aditya, Kishore, Suma, Kshitij, Majid, Anusha and Dr. Haritha. The discussions with them on the white board, working together late night, the

endeavors of constant learning and support from them both professionally and personally definitely needs a mention. I have made great friends and colleagues here like Ramani who have been a source of constant encouragement for me. I would like to thank Ramani for her critical analysis and comments on my work and for the lively discussions we had. I greatly acknowledge Aditya for being a source of alternatives whenever I had problems during my work. I would also like to thank Sabiha from I have learnt a lot and have shared amazing moments as part of NGS team.

I cannot imagine the completion of my research work without the help and support of all the present and former members Pathogen Biology Laboratory. I would like to thank Savita, Dr. Nishant, Amit and Kishore for their friendship and all the amazing moments we had together.

I would also like to thank Neha for her constant support, patience and encouragement throughout the journey of my Ph.D. work.

I cannot end without the acknowledgement of the love and support of my Family members, especially my Parents without whom I cannot imagine myself.

INDEX

	Page No
Abbreviations	I-II
List of Figures	III-VI
List of Tables	VII-VIII
 Chapter I	 1-15
General Introduction	
 <i>Helicobacter pylori</i>	 2
Epidemiology	2-4
Diagnosis	4-5
Treatment regimens	5-6
Clinical Manifestations & virulence determinants	6-8
Evolution	8-12
Genome characteristics	12-13
Rationale and objectives of the study	14-15
 Chapter II	 16-31
Establishing an in house pipeline for the whole genome sequencing and assembly of NGS data	
 Introduction	 17-20
Methodology	21-24
Genomic DNA quality check and sequencing	21
Quality assessment of the sequence data	22
Contig assembly	23-24
Gene prediction and annotation	24
Genome submission	24
Results and discussion	25-30

Quality assessment	25-26
<i>de novo</i> contig assembly	27-28
Genome annotation	29-30
Conclusions	31
Chapter III	32-47
Understanding the microevolution of <i>H. pylori</i> in a single stomach	
Introduction	33-35
Methodology	36-38
Patient information	36
Genome sequencing, assemble and annotation	36
Whole genome comparison and phylogenetic analysis	36-37
Analyzing the core and accessory genome content	37
Analyzing the virulence associated regions	37
Virulence gene content	38
Results and discussion	39-46
Assembly comparison of the whole genomes	39-40
Phylogenetic analysis	40-41
Core and accessory gene pool	41-42
Variations in the virulence apparatus	42-43
Variations in CagPAI	44-45
Outer membrane proteins	45-46
Conclusions	47
Chapter IV	48-68
Comparative genomic analysis of <i>H. pylori</i> to understand the population genetic structure in multiethnic societies such as Malaysia	
Introduction	49-51
Methodology	52-54

Strain collection and ethical approval	52
Bacterial culturing and DNA isolation	52
Functional annotation, calculation of core and specific gene content	53
Phylogenetic analysis	53-54
Virulence genes and phage detection	54
Results and discussion	55-67
Genome statistics	55-56
Phylogeny determination	57-59
Virulence potential	59-61
The core genome	61-62
Lineage specific genes	63-64
Restriction-Modification genes	64-66
Differentially evolving genes	66-67
Conclusions	68
 Chapter V	 69-83
Understanding the population genetic structure of <i>H. pylori</i> in India	
 Introduction	 70-72
Methodology	73-75
Strain collection	73
Population structure analysis	74
Phylogenetic analysis	74
Virulence content analysis	75
Results and discussion	76-81
Population structure analysis	76-78
Virulence potential	78-80
CagA-EPIYA motif	80-81
Conclusions	82-83

Chapter VI	84-90
Summary and outlook	85-90
References	91-110
Publications	111-187

List of abbreviations

<i>H. pylori</i>	<i>Helicobacter pylori</i>
PCR	Polymerase Chain Reaction
RUT	Rapid Urease Test
UBT	Urea Breath Test
BMI	Basic Metabolic Index
MALT	Mucosa Associated Lymphoid Tissue
IARC	International Agency for Research on Cancer
OMP	Outer Membrane Protein
CagPAI	Cytotoxin associated gene Pathogenicity Associated Island
CagA	Cytotoxin associated gene A
VacA	Vacuolation associated cytotoxin A
<i>dupA</i>	Duodenal Ulcer Promoting gene A
ATP	Adenosine Tri-phosphate
DNA	Deoxyribonucleic Acid
RAPD	Randomly Amplified Polymorphic DNA
MLST	Multi Locus Sequence Typing
NGS	Next Generation Sequencing
NCBI	National Centre for Biotechnology Information
ERIC	Enterobacterial Repetitive Intergenic Consensus

REP	Repetitive Extragenic Palindromic Sequences
QC	Quality Control
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
NJ Algorithm	Neighbour Joining Algorithm
MCL Algorithm	Markov Cluster Algorithm
BLAST	Basic Local Alignment Search Tool
VFDB	Virulence Factor Database
COG	Cluster of Orthologous Groups
ANI	Average Nucleotide Identity
NUD	Non Ulcer Dyspepsia
RPS BLAST	Reversed Position Specific BLAST
RM	Restriction-Modification
BAPS	Bayesian Analysis for Population Structure
WHO	World Health Organization

List of figures

Figure 1: Prevalence of *H. pylori* infection.

Figure 2: population structure of *H. pylori* obtained from STRUCTURE analysis of six housekeeping genes amplified from 761 global isolates.

Figure 3: The map of human migration as predicted by the studies on *H. pylori* and emergence of various lineages dominant in different geographical regions of the world.

Figure 4: Circular representation of *H. pylori* P12 genome showing its characteristics such as Plasticity Zones (PZ), CagPAI, comB cluster genes. The circles starting from inside represents GC skew plot, G+C content, location strain specific genes, genes encoded on the plus strand and those encoded on the minus strand.

Figure 5: Various NGS platforms together with their sequencing principles.

Figure 6: A typical fastq format representing its components.

Figure 7: A read quality graph obtained after the quality filter of NAK7 reads. The red and green dots represent the quality of the bases before and after filtering, respectively. Black line indicates the threshold score for filtering the reads.

Figure 8: The subsystems distribution of the functions encoded by genes of NAK7 genome.

Figure 9: The pipeline standardized for the whole genome assembly.

Figure 10: Whole genome alignment of 908, 2017 and 2018 generated using MGCAT. The green color represents the similar regions whereas the white color represents absence in other genomes.

Figure 11: Whole genome phylogenetic tree of 33 *H. pylori* strains including the strains used in this study. The branches of the tree are colored differently to highlight various lineages.

Figure 12: Venn diagram showing the comparative gene pool structure of *H. pylori* strains.

Figure 13: The comparison of Cag-PAI from three isolates. The brown color represents the gene in forward strand and those in blue represent genes encoded on reverse strand.

Figure 14: ClustalW alignment of *cag7* gene from 908, 2017 and 2018 with insertions and deletions in the repetitive regions.

Figure 15: Various outer membrane proteins encoded by the 908 genome. The rings starting from inside represent GC skew, GC content, 2017, 2018, outer membrane genes in forward strand and reverse strands.

Figure 16: 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 where each strain is represented by a ring. The yellow rings represent Malaysian-Chinese, purple represents Malay and light blue represents Malaysian-Indian strains. The variable regions such as plasticity zones (PZ) and Cag-PAI are compared across all genomes using BRIG image generator.

Figure 17: 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.

Figure 18: 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 the average similarity values among the strains.

Figure 19: 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 20: The distribution of RM genes in various strains: The graph shows number of genes annotated to encode putative RM functions in each strain. The Y axis represents number of genes and the X axis denotes strain names.

Figure 21: 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 X axis represents the number of genes in a particular functional category.

Figure 22: Admixture analysis of MLST sequences from 101 strains using BAPS. The Blue, Red and green colors represent clusters of strains belonging to HpEurope, HpSouthIndia and HpEastAsia.

Figure 23: Whole genome based phylogenetic tree generated using Gegenees tool with UPGMA algorithm. The colors represent clusters belonging to a particular lineage. The black arrows indicate the clusters containing *H. pylori* isolates from India.

Figure 24: *cagA* gene based phylogenetic tree constructed using MEGA tool with Neighbour-Joining algorithm after 1000 bootstraps. The tree branch confidence values are above 0.75 dividing the genomes into two major clusters, one formed by EastAsian *cagA* and other by European and Indian isolates.

List of tables

Table 1: Quality control statistics obtained from NGSQC tool kit for various *H. pylori* strains

Table 2: Contig assembly runs for NAK7 using velvet with various hash lengths. The optimum hash length is colored red

Table 3: Contig assembly statistics obtained for various *H. pylori* strains

Table 4: Genome characteristics of various *H. pylori* strains sequenced and assembled in this study

Table 5: Submitted genomes and their GenBank Accession IDs

Table 6: Genome characteristics of the three strains involved in the study.

Table 7: Comparative virulence profile of 908, 2017 and 2018 isolates. (+) represents the presence while (-) represents absence or partial presence.

Table 8: Genome characteristics of the sequenced *H. pylori* strains.

Table 9: The genes showing lineage specificity: The table shows the identified genes having high prevalence in any one of East-Asian and non-East-Asian genomes.

Table 10: Strain collection for this study.

Table 11: Status of 58 virulence genes among five sequenced genomes from India. (+) represents the presence while (-) shows absence of a gene.

Chapter I

General Introduction

Helicobacter pylori

Helicobacter pylori (*H. pylori*) is a Gram negative, microaerophilic and spiral shaped bacterium that colonizes the human stomach (Marshall and Warren 1984). The bacterium was first identified in 1882 by two Australian scientists, Prof. Robin Warren and Prof. Barry James Marshall while examining the gastric biopsies of the patients. The curved bacilli were later isolated and cultured to prove Koch's postulates and thus establishing it as a casual organism for gastritis and ulcers, which were then considered to be caused due to acid (Marshall et al. 1985). This landmark discovery brought two major turnarounds; first, it negated the long prevailing enigma that the stomach is a sterile environment and the second, the patients no longer had to take lifelong medication and could be easily treated with a short regimen of few antibiotics. For their remarkable discovery, the scientists were awarded Nobel Prize in Physiology or Medicine in 2005. Since then, *H. pylori* has become the epicenter of many interventions aimed at elucidation of its epidemiology and mechanisms underlying its evolution and transmission.

Epidemiology

The discovery of a bacterial cause for some of the highly prevalent gastric diseases underpinned a number of studies worldwide to understand its epidemiology and associated health burden. Almost half of the world population is colonized by *H. pylori* making it the most dominant infection around the globe (Graham 2015). Recent epidemiological studies have revealed a decline in peptic ulcer disease and gastric cancer throughout the world. This could be attributed to increased use of antibiotics and rise in the healthcare systems (Eusebi et al. 2014). The infection is potentially contracted in the childhood and can persist for decades if not eradicated by administering antibiotics (Kuipers et al. 2000). The exact mode of *H. pylori* transmission is still

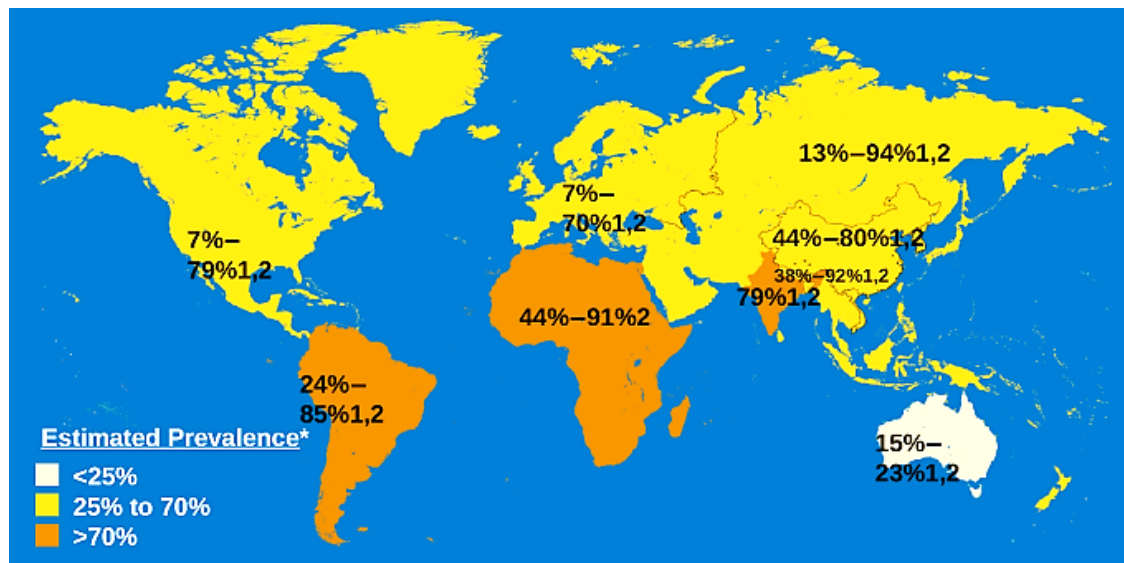


Figure 1: Prevalence of *H. pylori* infection. (Adapted from Khalifa et al. 2010 & Goh et al. 2011)

unclear, but several studies on isolates from family members exhibited intra familial acquisition as the dominant force of transmission (Didelot et al. 2013). The infection burden is higher in case of developing nations with 70-80% prevalence in contrast to 30-40% observed in developed nations (Figure 1) (Eusebi et al. 2014). In a study from Europe, the prevalence was shown to be lower (20-30%) in Northern Europe (den Hollander et al. 2013) compared to Southern countries like Portugal (60-80%) (Bastos et al. 2013). Moreover, in another study from Netherlands ethnicity emerged as a major factor for the observed *H. pylori* prevalence (den Hollander et al. 2013). In Asia, the studies revealed a higher prevalence of *H. pylori*, upto 60-80% (Sodhi et al. 2013). A study from Korea reported *H. pylori* seroprevalence of 54.4% (Lim et al. 2013) whereas in China it was reported to be 63.4% (Zhu et al. 2014). In India, a latest study recorded a prevalence of approximately 60 % similar to its neighboring countries like Bhutan (Vilaichone et al. 2013). Epidemiological studies from African countries like Morocco and Ethiopia revealed 75.5 and 65.7 *H. pylori* seroprevalence in the population (Benajah et al. 2013; Mathewos et al.

2013). Taken together, it appears that predominance of *H. pylori* could be best described as per the socioeconomic status of the regions concerned, although ethnicity and geography are potential risk factors as well (Eusebi et al. 2014).

Diagnosis

The approaches in vogue for the detection and diagnosis of *H. pylori* can be divided in two major categories; invasive and non-invasive (Garza-Gonzalez et al. 2014). The invasive methods include histology examination, bacterial culture, PCR, rapid urease test (RUT) of the biopsies of gastric samples and serological examination by use of immuno-assays. The histological approach relies on the Hematoxylin and Eosin staining to visualize the gastric morphology and the counter staining by Giemsa to observe the bacterium (Fallone et al. 1997). The sensitivity and specificity is dependent on the experience of the histologist, intensity of colonization and biopsy collection. Culturing the bacteria directly from the biopsy on selective plates offers 100 % specificity and >80% sensitivity of the results, but is labor intensive and time consuming (Hirschl and Makristathis 2007). PCR promises great advantage as it can detect even low bacterial loads and can be easily applied to range of samples from stool to biopsies. The method also aids in identification of antibiotic resistance but, suffers from disadvantage of even detecting dead bacteria and sometimes leading to false positive results (Rimbara et al. 2013). RUT is highly specific test that utilizes the activity of urease enzyme produced by *H. pylori* causing a color change of the pH sensitive media (Garza-Gonzalez et al. 2014). More accurate results are obtained from various serological methods that involve detection of antibodies like IgG. Recently, six major virulence factors such as CagA, VacA, GroEl, gGT, HcpC and UreA have been listed as promising targets for the detection of *H. pylori* by ELISA (Erzin et al. 2006; Malfertheiner et al. 2007). Serological detection provides information on the active immune

response and inflammation in the biopsies. On the other hand, non-invasive methods include approaches such as Urea Breath Test (UBT) which is a preliminary test for the presence of *H. pylori*. UBT relies on the urease production capability of *H. pylori* that converts ^{13}C or ^{14}C labelled Urea into CO_2 and ammonia (Goddard and Logan 1997). The CO_2 is circulated through the blood vessels and is exhaled from the lungs and is recognized by the detector to confirm *H. pylori*'s presence. Because of its high sensitivity, the method is also used to confirm the eradication of *H. pylori* after drug therapy (Shirin et al. 2005).

Treatment regimens

Various treatment regimens have been developed that utilize a varied combination of drugs. Most widely used regimen involves triple drug therapy that includes a proton pump inhibitor (PPI), amoxicillin and clarithromycin for two weeks (Garza-Gonzalez et al. 2014). This forms the first line therapy which is supplemented with second or third line therapies upon failure of the earlier. Recent years have witnessed a rise in clarithromycin resistance leading to a decline in the eradication to 65% (Megraud 2013). Triple drug regimen with metronidazole as replacement of clarithromycin achieved an eradication rate of 84% (Sasaki et al. 2013). In a study from Korea, factors such as age, female gender, BMI, amoxicillin and/or clarithromycin resistance were found to significantly influence the outcome of treatment (J. Y. Lee et al. 2014). Therefore, a tailored therapy based on the drug susceptibility of the patient is gaining greater acceptance. The second line of drug therapy involves use of PPI, amoxicillin and metronidazole after the failure of first line of drugs to eradicate *H. pylori* (O'Connor et al. 2014). The time duration between the first and the second line appears to have a strong influence on the success in eradication therapy (Nishizawa et al. 2013). In addition, use of bismuth like fluoroquinolones in combination with triple drugs (quadruple drug therapy) is an effective strategy to combat even resistant strains.

Hence, the choice and combination of drugs used for *H. pylori* eradication needs a careful survey and knowledge of the dominant genotype in the region (O'Connor et al. 2014).

Clinical Manifestations and virulence determinants

H. pylori infection has been associated with a myriad of upper gastrointestinal diseases such as chronic gastritis, peptic ulcer disease, mucosa-associated lymphoid tissue (MALT) lymphoma and gastric adenocarcinoma (Kusters et al. 2006). Although nearly half of the world population is colonized by the bacterium, only a minority develop clinical symptoms. A number of studies have linked *H. pylori* infection to enhanced risk of developing gastric cancer and therefore, the International Agency for Research on Cancer (IARC) regarded *H. pylori* as class I carcinogen in 1994 (Roesler et al. 2014).

From the recent reports, it is conceivable that *H. pylori* harbors a plethora of genetic factors underpinning the infection outcome (Yamaoka 2008). These virulence genes can be grouped into three categories. First category consists of strain specific genes that are present only few strains of *H. pylori*. Second category comprises of the phase-variable genes that remain in a dynamic on off status under various physiological conditions. This category is majorly enriched in outer membrane proteins (OMPs) such as OipA, SabA, SabB, BabA, BabC and HopZ (Oh et al. 2006). The third category comprises of genes that exhibit genotypic and structural polymorphism such as *vacA* and *cagA* (Yamaoka 2008). Some of the major virulence genes and their functions are discussed below.

Cytotoxin associated gene pathogenicity associated island (CagPAI) is the most studied region because of its correlation with severe clinical outcomes. The region is nearly 40 kb in size and encodes nearly 30 proteins that form a TypeIV secretion system (Censini et al. 1996). The pilus

like typeIV secretion system allows the transfer of CagA oncoprotein directly into the gastric epithelial cells. Once inside, CagA triggers the perturbation of a myriad of signaling pathways leading to cellular and morphological alterations (Higashi et al. 2002; Odenbreit et al. 2000). The *cagA* gene exhibits high genetic polymorphism in the C terminal repeat region among EastAsian and Western isolates (Hatakeyama 2009). The region harbors EPIYA repeats containing the sites of phosphorylation and depending upon the sequence composition flanking these repeats, four different types have been recognized; EPIYA-A, EPIYA-B, EPIYA-C and EPIYA-D (Hatakeyama 2004). The Western *H. pylori* isolates carry EPIYA-A, EPIYA-B & EPIYA-C motifs in the *cagA* whereas the EastAsian strains comprise EPIYA-A, EPIYA-B and EPIYA-D repeat motifs at the C-termini. The strains containing EPIYA-D motifs have been shown to exhibit higher pro-inflammatory potential (Hatakeyama 2009) and thus correlate more frequently with the severe disease like gastric cancer. Therefore, EastAsian strains that contain the same motifs pose a greater risk of gastric disease compared to Western isolates (Yamaoka 2010).

Another widely studied virulence factor VacA that induces vacuolation in the gastric epithelial cells upon delivery and leads to cell death (Leunk et al. 1988). The gene *vacA* exhibits high genetic polymorphism among different isolates of *H. pylori*. The sites of variability is confined to three main regions the s (signal sequence), i (intermediate region) and the m (middle region). *H. pylori* harbors different combinations such as s1m1, s1m2, s2m1 and s2m2 in *vacA* genes and display consequent alteration in the severity of vacuolation (Wroblewski et al. 2010). The EastAsian strains having high virulence potential mostly encompass s1m1 or s1m2 *vacA* whereas Western *H. pylori* strains mostly retain the s2m1 or s2m2 genotypes of *vacA* (Atherton et al. 1995). Many epidemiological studies have found strong association of s1m1 with severe disease outcomes like peptic ulcer and gastric cancer (Lopez-Vidal et al. 2008).

In addition, a variety of other virulence factors such as *dupA*, Urease and several outer membrane proteins have been highlighted in the recent studies. Duodenal Ulcer Promoting gene A (*dupA*) is positively correlated with duodenal ulcers and having a protective effect against gastric cancer (Schmidt et al. 2009). The gene is also found to induce elevated IL-8 secretion in both *in vivo* and *in vitro* experiments. The *dupA* knockout strains also revealed higher susceptibility to low acid compared with the wild type strains illustrating its significance in the pathogenesis (H. Lu et al. 2005). Another important factor for the resistance of *H. pylori* to low acid is because of the enzyme Urease (Montecucco and Rappuoli 2001). The enzyme converts urea into ammonia and CO₂, both of these aids in forming a protective cloud guarding the bacterium from low acid stress. Apart from this ammonia, produced by the enzyme also induces alteration in the vesicular membrane transport, reduction in ATP generation and consequent decline in protein synthesis of the gastric epithelial cells (Montecucco and Rappuoli 2001). *H. pylori* also harbor a large repository of OMPs that play an important role in adaptation and host pathogen interaction (Tomb et al. 1997). These include adhesins like Lewis b blood group antigen binding adhesin (BabA) and sialyl Lewis X binding adhesin (SabA). These adhesins have strong affinity towards sugar moieties present on the epithelial cells thus aiding in colonization and inducing pro-inflammatory secretions (Ilver et al. 1998). Additionally, many other proteins like HopC, HopB and HopZ have also been shown to be important mediators of gastric inflammation (Odenbreit et al. 2009).

Evolution

The high prevalence of *H. pylori* instigated a greater interest towards understanding the genetic characteristics. A myriad of approaches such as DNA fingerprinting, Restriction Amplified Fragment Length Polymorphism (RFLD), Fluorescently Amplified Fragment Length

Polymorphisms (FAFLP) and REP-PCR were employed to understand the diversity and transmission of *H. pylori* (Suzuki et al. 2012). *H. pylori* strains were found to be highly diverse which can be attributed to the increased recombination and mutation rates (Achtman et al. 1999; Bjorkholm et al. 2001). The studies on chronological isolates from the same patients revealed existence of a single dominant genotype though some cases of mixed infections have also been reported (Prouzet-Mauleon et al. 2005). The isolates from various regions of the world revealed a high diversity but the strains from the same geographical area exhibited higher genetic similarity (Achtman et al. 1999). Therefore, it was important to decipher the population structure to better understand *H. pylori*'s evolution and transmission dynamics. Multi Locus Sequence Typing analysis of highly conserved genes provides greater resolution and is much more accurate in providing evolutionary insights (Maiden et al. 1998). MLST analysis of six housekeeping genes from various isolates around the globe revealed six major lineages of *H. pylori*: HpEurope, HpAfrica1, HpAfrica2, HpEastAsia, HpAsia2 and HpNEAfrica shown in figure 1 (Falush et al. 2003; Linz et al. 2007). These lineages were classified in accordance with their geographical predominance (Achtman et al. 1999). Further, lineages such as HpEastAsia could be divided into three sub lineages: HspEastAsia majorly found in Japan, China and Korea, HspAmerind distributed among Inuits and Amerinds of North and South America, and HspMaori found among NewZealanders and Moari people from Polynesia (Linz et al. 2007). HpAfrica1 was classified into HspSAfrica and HspWAfrica. *H. pylori* strains consisting of HspSAfrica and HpAfrica2 were predominant in SouthAfrica whereas those ascribing to HspWAfrica predominate in West Africa and form a minor proportion in SouthAfrica. HpAsia2 is limited to Malaysia, India and parts of South Asia. The isolates from Europeans form HpEurope which are also found in Israel, Turkey Bangladesh and Sudan. The outcomes also resonate in the recent

reports based on whole genome sequencing exhibiting the accuracy of MLST based typing approach.

The evolutionary history of *H. pylori* and its origins can also render important insights into human evolution and migration (Ahmed et al. 2009). *H. pylori* shares a long co-evolution history along with their human host and their association even predates any event of human migration out of Africa approximately 60,000 years ago (Moodley et al. 2012).

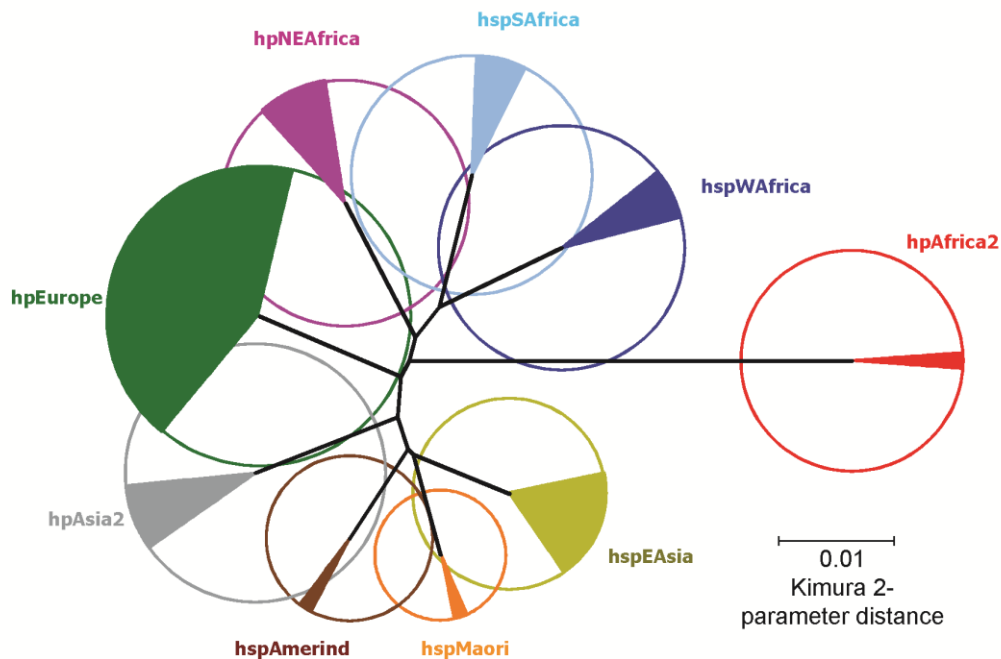


Figure 2: Population structure of *H. pylori* obtained from STRUCTURE analysis of six housekeeping genes amplified from 761 global isolates. Figure courtesy (Falush et al. 2003)

H. pylori possess characteristics that make it a surrogate marker for studying their human hosts: (i) humans are the only natural host for *H. pylori* (Kusters et al. 2006), (ii) the bacterium has been associated with the host for a very long time (iii) the transmission within the families helps

maintain the genetic signatures (iv) The pathogen exhibits high genetic diversity compared to humans that might be more informative of the population structure. The study carried out by Linz et al. observed decrease in genetic diversity as we branch out from EastAfrica in juxtaposition with humans (Kusters et al. 2006).

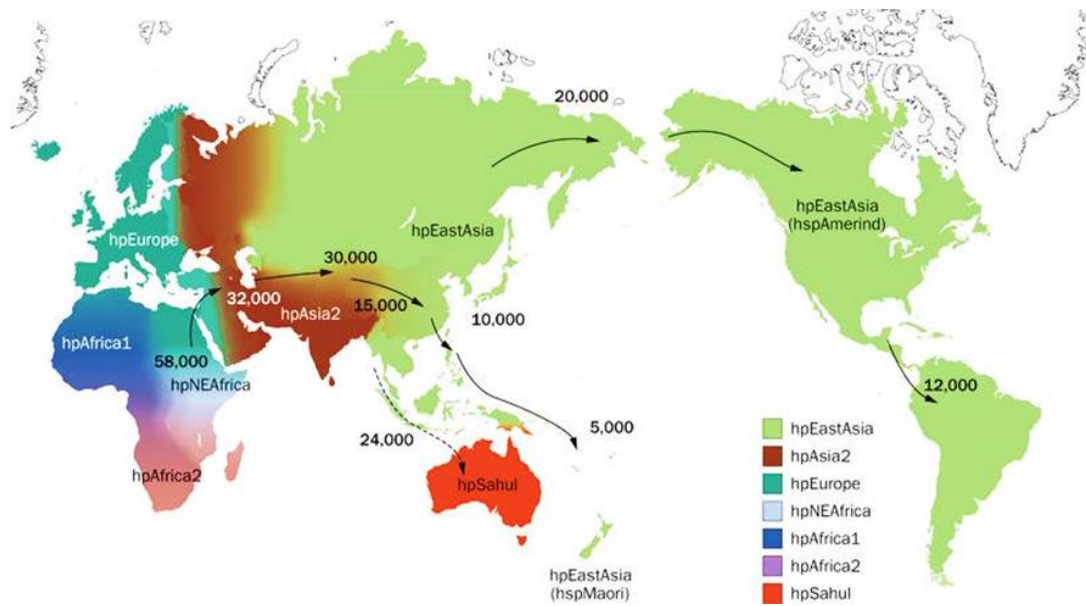


Figure 3: The map of human migration as predicted by the studies on *H. pylori* and emergence of various lineages dominant in different geographical regions of the world. Figure courtesy (Yamaoka et al. 2010)

The population structure deciphered from the global collection could coincide with human evolution and could explain human migration out of Africa, the cradle of modern human population. Hence, it can be surmised that as humans migrated out of Africa to populate different geographical regions, there occurred a diversification of *H. pylori* into distinct lineages by virtue of host specific adaptations. Therefore, studying evolution and diversity of *H. pylori* might provide significant cues about the riddles of human migration and evolution with greater

detail (Ahmed et al. 2009). Recent studies based on the whole genome sequences promises more robust population structure analysis with ultimate resolution allowing a better observations on population admixture and coevolution dynamics.

Genome characteristics

The importance of the organism and the high variability it exhibited, made it mandatory to perform whole genome sequencing to unravel its genomic characteristics. The first whole genome sequencing of the strain 26695 in 1997 provided important genomic insights (figure 4) (Tomb et al. 1997). This was followed by genome sequencing of a number of isolates from various parts of the globe. As of today there are more than 51 whole genome sequences of

H. pylori strains representing a diverse collection at global level. *H. pylori* have an average genome size of 1.6mb with a G+C content of 39%. There are certain regions having a sharp decline in the G+C content such as CagPAI and plasticity zones. The genome encodes nearly 1600 genes having an average length of 945bp. *H. pylori* genome harbors ~36 tRNAs and 6 rRNAs along with several repeat regions. These repeats are mainly concentrated in the OMPs and dictate their phase-variation. The comparative genome studies have revealed a core genome content of ~1200 genes encoding the basic functional machinery of *H. pylori*. The genome also encodes two types of putative secretory systems, the sec pathway and another is type IV secretion system located in CagPAI. The comparative analyses of the whole genomes have revealed a high proportion of strains specific genes. Nearly half of these are localized in the plasticity zones that exhibit high strain to strain genetic variability (Kersulyte et al. 2009). The region also harbors some of the recently studied virulence factors highlighting the importance of strain specific genes (S. Devi et al. 2014).

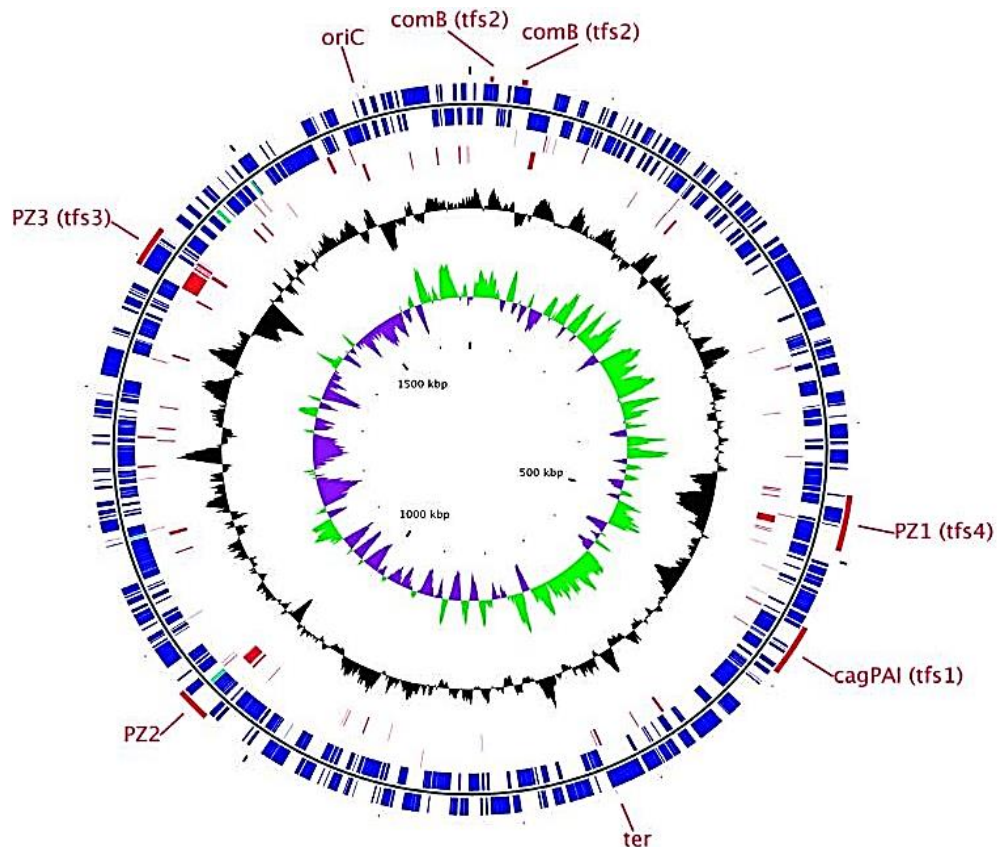


Figure 4: Circular representation of *H. pylori* P12 genome showing its characteristics such as Plasticity Zones (PZ), CagPAI, comB cluster genes. The circles starting from inside represents GC skew plot, G+C content, location strain specific genes, genes encoded on the plus strand and those encoded on the minus strand.

Figure courtesy (Fischer, W., et al. 2010)

Another striking feature of *H. pylori* genome is the increased proportion of restriction modification genes (Vasu and Nagaraja 2013). *H. pylori* on an average contain 20-30 restriction modification genes and few of them like *iceA* have already been associated with increased risk of peptic ulcer (Peek et al. 1998). Therefore, whole genomes could provide invaluable information about the organism and their comparative analysis would uncover novel mechanisms and identify new targets for designing effective treatment and control regimens.

Rationale and objectives of the study

All the above information clearly depicted the importance of *H. pylori*. Several MLST based studies have sought to understand its epidemiology and population structure but failed to provide further insights into its evolutionary mechanisms. In comparison, the whole genome sequencing offers the ultimate resolution for the comparison of genomes (Walker et al. 2013). Even for the monomorphic organisms that exhibit high genetic similarity and could be difficult to characterize by traditional approaches, the whole genome sequencing offers solution and can be applied to understand their evolution. The ease of sequencing has made it possible to establish WGS approach for tracking the spread and transmission of clinically important pathogens (Harris et al. 2013; Walker et al. 2013). In addition, genome sequences also provide information on the genetic capabilities, making it possible to narrow down on key coordinates underlying the functional phenotypic differences (Holden et al. 2013). Banking on these advantages, in this study, we have attempted to undertake whole genome approach towards understanding the dynamics of *H. pylori*.

It is apparent from various studies that once acquired at the early age *H. pylori* establishes persistent infection for decades. Several studies on isolates from single patients report a single dominant genotype though there are minor reports of mixed infection (Covacci et al. 1999). Given this, it would be necessary to understand the changes that this bacterium undergoes in a single host, overtime. The analysis would allow us to gain significant insights into the persistence mechanisms employed by *H. pylori* to colonize and cause infection. This would be necessary to narrow down on key virulence and colonization factors and would foster the development of effective control strategies. Another important issue that needs investigation is to understand the evolutionary dynamics and population structure of *H. pylori* in mixed population setup such as

Malaysia, where people of distinct ethnicities have coexisted for a long time (Alfizah et al. 2012). The host specific nature of this gastric pathogen along with the other factors has been implicated in dictating the infection outcome and severity. The comparative analysis of the isolates from a mixed population setup could shed light on the impact of co-evolution of various genotypes and their population behavior. The study promises to provide a clear understanding of the discordant infection rates and associated disease outcomes among people of distinct ethnic backgrounds. A clear understanding of these would be very important for the formulation and application of effective control regimens entailing regions like Asia with population comprising of people from different ethnic, religious backgrounds. And the last part of the thesis addresses the population structure of *H. pylori* in India and the cues it could provide about the prevalence and severity of *H. pylori* associated gastric diseases. The study aimed to gain insights into the virulence potential and a comparison with EastAsian genotype. All of the above mentioned initiatives required whole genome sequencing and assembly of isolates from regions of interest. Therefore, the thesis work has been divided into four chapters with the following objectives:

1. Establishing an in-house pipeline for whole genome sequencing and assembly of NGS data
2. Understanding the microevolution of *H. pylori* in a single stomach
3. Comparative genomic analysis of *H. pylori* to understand the population genetic structure in multiethnic societies such as Malaysia
4. Understanding the population genetic structure of *H. pylori* in India

CHAPTER II

Establishing an in house pipeline for the whole genome sequencing and assembly of NGS data

Part of this chapter was published as:

- 1) 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. **Kumar N**, Mukhopadhyay AK, Patra R, De R, Baddam R, Shaik S, Alam J, Tiruvayipati S, Ahmed N. **J Bacteriol.** 2012 Nov;194(21):5963-4.
- 2) Genomes of two chronological isolates (*Helicobacter pylori* 2017 and 2018) of the West African *Helicobacter pylori* strain 908 obtained from a single patient. Avasthi TS, Devi SH, Taylor TD, **Kumar N**, Baddam R, Kondo S, Suzuki Y, Lamouliatte H, Mégraud F, Ahmed N. **J Bacteriol.** 2011 Jul;193(13):3385-6.

Introduction

The emergence of NGS has triggered a revolution in the way we study evolution and seek answers of biological problems (Schuster 2008). Consequent technological advances fueled a surge in comparative genomic analysis and spurred many new areas of research such as metagenomics (Buermans and den Dunnen 2014). Whole genome sequences not only provide better resolution at a base scale, they also enable researchers to map genome wide alterations taking place over generations in response to selection/adaptation requirements. Therefore, it is heralding a paradigm shift in our ability to understand various mechanisms of adaptation, transmission and evolution of bacterial pathogens (Toft and Andersson 2010).

Whole genome sequencing offers certain advantages over limited sequencing approaches like MLST, ribotyping etc. Comparing whole genome sequences allows access to a plethora of variable loci as compared to relatively less proportion measured using traditional approaches. This ensures higher discriminatory power even when comparing clonal bacteria like *Mycobacterium tuberculosis* where traditional fingerprinting techniques sometimes remain inconclusive (Harris et al. 2010). Whole genome sequencing could reveal a lot of clinically relevant information in considerably short time span and provides more robust information with minimal errors compared to any traditional typing techniques. Therefore, epidemiological studies to track evolution and transmission of bacterial pathogens could be best understood using whole genome sequence based comparisons (Koboldt et al. 2013). In addition, understanding the population structure of a bacterial pathogen using MLST might be skewed because of high level of similarities among the housekeeping gene loci, in contrast, the whole genome sequences with higher discriminatory power offers a better alternative (Parkhill 2013).

The advent of NGS dates back to 2001 when 454 sequencing was first introduced (Ronaghi 2001). This was followed by Solexa sequencing/Illumina in 2008 and several others in the recent past (Bentley et al. 2008). As a result public databases such as NCBI have experienced a steady rise in the number of genome sequences over the past decade due to the emergence of time efficient and cost-effective DNA sequencing technologies. The daunting task of library preparation that demands huge amount of time and manpower has now been replaced by much faster and streamlined approaches. Further, it also leads to a consequent reduction in the time required and a consequent reduction of errors caused due to manual interventions (van Dijk et al. 2014).

Depending upon the sequencing chemistry employed, there are different platforms available (figure 5) for whole genome sequencing. Illumina platform that produces shorter reads of 300-250 bp relies on reversible dye terminator technology (Bentley et al. 2008), while Roche platform which generates much longer reads when compared to Illumina, is dependent upon the amplicon enrichment followed by pyro-sequencing approach (Margulies et al. 2005). The other platforms like SOLiD are based on the property of DNA ligases to read the bases in the amplified templates obtained after emulsion PCR (Mardis 2013). However, each of these platforms have their own pro and cons, but we chose Illumina because of the following four main reasons: 1) it generates paired-end information which is important in downstream assembly process, 2) the per base cost is much lower compared to other platforms, 3) it generated high coverage depth that aids in mapping variations and 4) having prokaryotic genomes (less complicated and smaller compared to eukaryotic genomes) to work with.

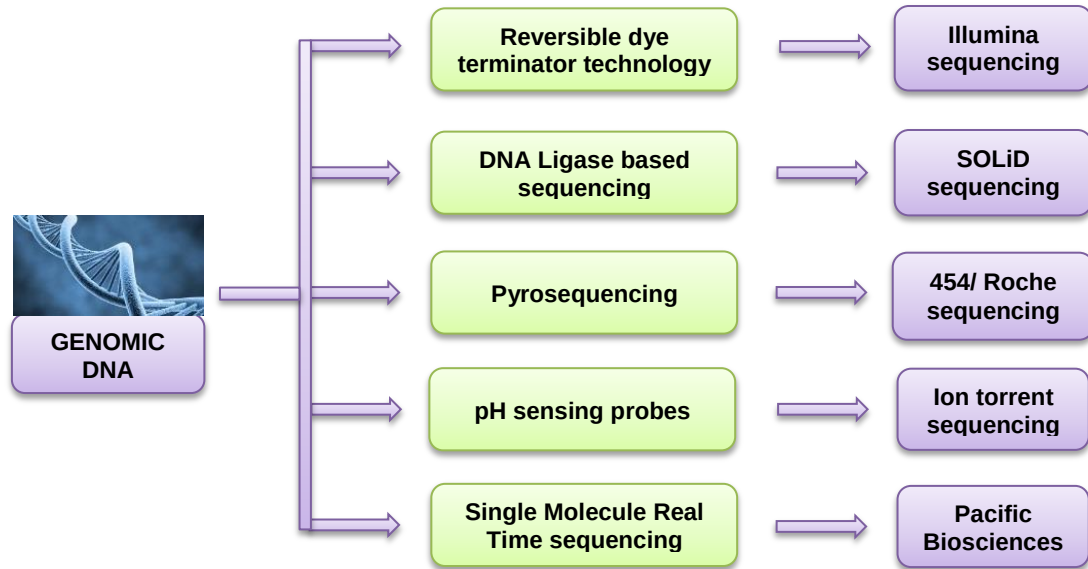


Figure 5: Various NGS platforms together with their sequencing principles

Illumina sequencing workflow mainly comprises of library preparation followed by bridge amplification and sequencing. The library preparation starts with random shearing of the genomic DNA and selection of desired sized fragments using gel electrophoresis.

Further, Illumina specific adapters are then ligated at the ends of these fragments using DNA ligase. These adapter ligated fragments are then hybridized to an array of nucleotides complementary to the ligated adapters on a glass slide known as flow cell. The library fragments are then amplified to a million times using bridge amplification method forming clusters of identical fragments also known as polony. The bases contained in the fragments comprising these polonies are determined by the sequencer in a systematic manner using a complex of enzymes and nucleotides in cycles as depicted in figure 2 (Bentley et al. 2008).

Therefore, NGS platforms end up generating millions of short fragments called reads from which reconstructing the genome could be a challenging task. The approaches utilized for

assembling a draft/complete genome broadly can be divided into two categories, a reference based approach and *de novo* approach. The reference based approach relies entirely on the chosen reference genome whereas; the *de novo* method is independent of any reference sequence. Both the approaches suffers from demerits such as losing strain specific information in case of the former while in the latter case repeat regions could be a daunting task to resolve. In this chapter we discuss the development of a hybrid approach that filters benefits of both the methods and therefore, enables a better genome assembly. The reads are first assembled *de novo* into contigs which are then aligned against a reference genome (preferably the same species). The paired reads are mapped back to the reference sorted contigs and the alignment information is utilized to validate their order in the genome. This approach helps to retain maximum strain specific information and also aid in resolving repetitive elements to a greater extent and has been applied to carry out assembly of more than 20 bacterial genomes. In addition, this method is also efficient in detecting possible contamination or even plasmids in the genome sequence data and therefore helps in achieving better genome assemblies.

Methodology

Genomic DNA quality check and sequencing

The genomic DNA of strains isolated from patients with different disease conditions as mentioned in following chapters of the thesis was subjected to quality check prior to sequencing. The DNA quality was analyzed using NanoDrop to estimate the concentration (~200-350 ng/μl) and quality (230/260 ratio for organic impurities). The integrity of isolated DNA was analyzed on a 0.8% agarose gel electrophoresis prior to library preparation for sequencing. The samples for which a single band was observed with minor smearing were selected for sequencing. Nearly ~5-10 μg of high quality genomic DNA was shipped to the company at Bangalore, for sequencing. The paired-end sequencing was carried out on Illumina GAIIX Next Generation Sequencing platform.

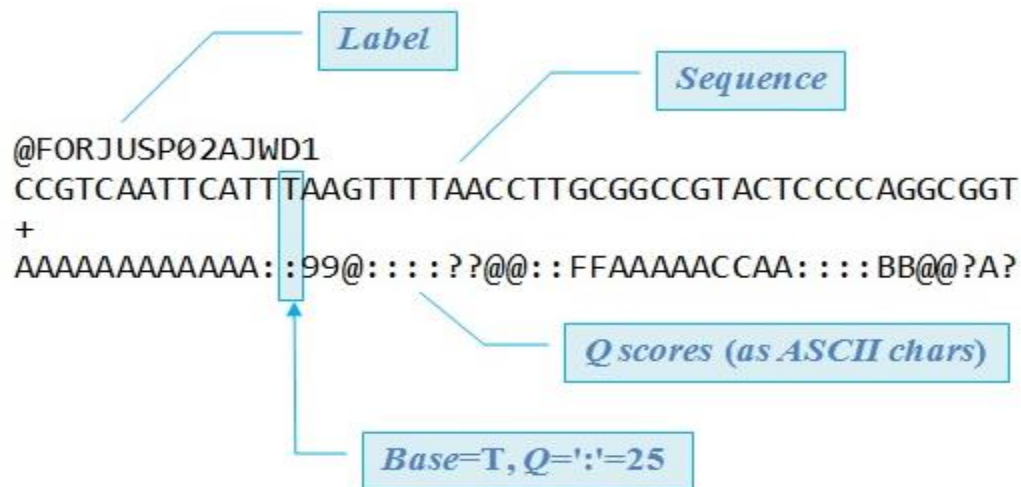


Figure 6: A typical fastq format representing its components.

Quality assessment of the sequence data

The sequence data were received in the form of .fastq files. A typical representation of a .fastq format is mentioned in the figure 6. These raw files were subjected to quality check using NGSQC toolkit (Patel and Jain 2012). The reads were initially assessed for the average quality throughout the read length. The reads were then trimmed if the quality of the bases dropped below a threshold Phred quality score of 20 as shown in the figure 7.

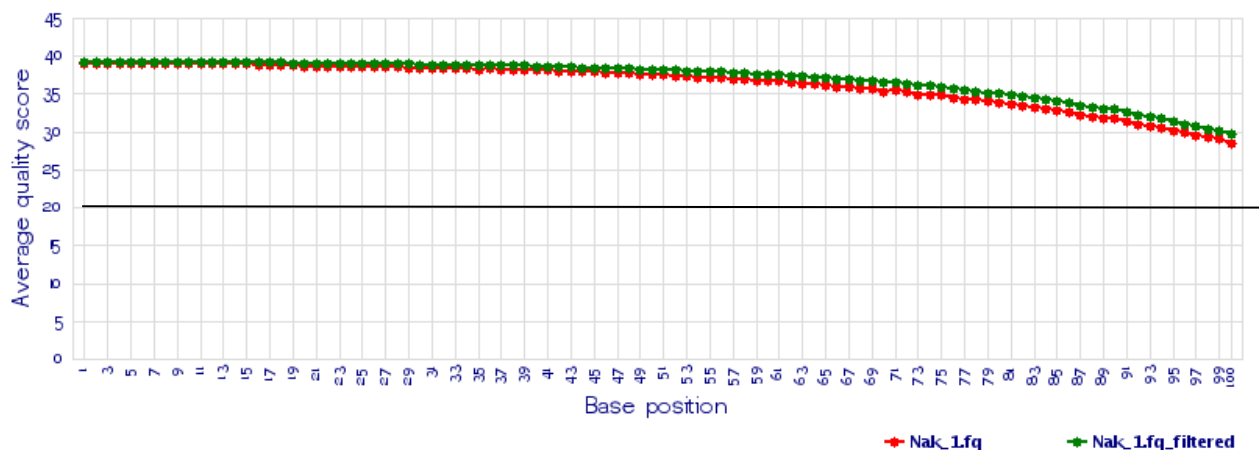


Figure 7: A read quality graph obtained after the quality filter of NAK7 reads. The red and green dots represent the quality of the bases before and after filtering, respectively. Black line indicates the threshold score for filtering the reads.

In addition the reads which carried more 30% of their bases below the threshold quality score were also removed. The filtered and trimmed high quality reads were taken for the downstream contig assembly.

Contig assembly

The high quality filtered reads (.fastq files) were assembled into contigs using velvet *denovo* assembler (Zerbino and Birney 2008). The reads contained in the two .fastq files were merged into a single .fastq file by using a perl script (shufflesequences_fastq.pl). The merged file was used as input for the velvet assembler. The assembler runs in two stages, first velveth executes the hashing of reads and creates a roadmap file. The output files together with various error correction modules are employed by velvetg to assemble the contigs. The exact command lines and the parameters for one representative assembly are mentioned below:

```
$/shuffleSequences_fastq.pl read1 read2 shuffled.fq
```

```
$/velvethoutput_folder K1,K2,2 -shortPaired -fastqshuffled.fq
```

```
$/velvetgoutput_folder -ins_length_sd 20 -ins_length 200 -exp_cov auto -  
min_contig_lgth 200 -cov_cutoff auto -scaffolding no -unused_reads no
```

Here, K1 and K2 refer to the chosen start and end of the hash length and must be odd numbers.

In this manner the assembly was repeated for various hash lengths. A comparison of assembly statistics was carried out to select the optimal hash length that provided minimum number of contigs, higher N50, approximate genome size and utilized maximum amount of reads. The assembled contigs with the optimum hash length were selected for carrying out scaffolding using SSPACE tool (Boetzer et al. 2011). This tool merges contigs into scaffolds (by placing N's in between) using the paired-end information of aligned reads. The gaps in the scaffolds were filled as much possible using Gapfiller tool (Nadalin et al. 2012). The output was then manually

screened for errors after aligning the reads back to contigs and followed by visualization of the alignment file in tablet tool (Milne et al. 2010).

Gene prediction and annotation

The draft genomes were submitted to online annotation server RAST (Rapid Annotation using Subsystems technologies) for gene prediction and annotation (Aziz et al. 2008). The .gbk file from RAST server was loaded in Artemis to harvest the genome characteristics of the assembled genomes. The draft genomes were also screened by tRNAscan-SE (Schattner et al. 2005) and RNAmmer (Lagesen et al. 2007) online servers to predict tRNAs and rRNAs respectively.

Genome submission

The submission of the draft genomes to NCBI involves a series of steps as mentioned below.

1. Create an online account in NCBI /GenBank.
2. Under Bioproject portal generate a Bioproject ID for the submission by providing the information about the strains and purpose of sequencing.
3. Create a template file (.sbt) online: <http://www.ncbi.nlm.nih.gov/WebSub/template.cgi> with the details of submitter and other authors.
4. Generate a (.fsa) file of the contigs obtained after assembly. The file must contain all the contigs above 200bp of length. The headers of the contigs should contain the following information in the same manner as mentioned below.

>contig001 [organism=Helicobacter pylori] [strain=NAK7] [host=human] [gcode=11]
5. The final .sqn file is created using the tool tbl2asn downloaded from NCBI which requires both .fsa and .sbt file as input.

Results and discussion

Quality assessment

The paired-end raw reads generated after sequencing of different *H. pylori* strains were analyzed for quality. Each dataset contained two separate file representing both the read pairs. Both the read files were analyzed separately for read quality. The filtering statistics generated for both the files corresponding to each genome are mentioned in table 1. As observed from the table all the sequencing runs generated ~90% high quality (hq) reads with almost negligible primer adapter contamination. Number of non-ATGC bases that gives an indication of the unambiguous base calls during sequencing also remained very low signifying low proportion of errors during the sequencing. The base composition analysis revealed a low G+C content which was in accordance the characteristics of *H. pylori*.

Table 1: Quality control statistics obtained from NGSQC tool kit for various *H. pylori* strains.

	NAK7		NAD1		NAB47		UM018	
QC parameters	Read1	Read2	Read1	Read2	Read1	Read2	Read1	Read2
Total No. of reads (x10 ⁶)	5.4	5.4	3.6	3.6	6.24	6.24	1.56	1.56
Total No. of hq reads (x10 ⁶)	5	5	3.4	3.4	6.17	6.17	1.43	1.43
Percent hq reads	92.32	92.32	95	95	98.88	98.88	91.6	91.6
Total No. of bases (x10 ⁶)	545.2	545.2	264.3	264.3	449.2	449.2	156.1	156.1
Total No. of bases in hq reads (x10 ⁶)	503.3	503.3	251.3	251.3	444.2	444.2	143	143
Percent hq bases in hq reads	98.11	97.47	249.0	248.6	99.12	99.08	98.27	98.56

	UM054		UM007		UM034		UM045	
Total No. of hq bases in hq reads (x10 ⁶)	493.8	490.6	99.09	98.04	440.3	440.2	140.5	140.9
Percent A's	32.84	27.83	31.79	28.95	32.84	27.83	33.08	26.74
Percent G's	20.64	19.35	20.23	19.55	20.64	19.35	20.95	19.32
Percent C's	19.24	20.61	19.54	20.22	19.24	20.61	19.42	20.98
Percent T's	27.28	32.20	28.44	31.28	27.28	32.2	26.55	32.96
QC parameters	Read1	Read2	Read1	Read2	Read1	Read2	Read1	Read2
Percent non-ATGC bases	0.01	0.01	0.01	0.03	0.29	0.03	0.17	0.98
primer/adapter contaminated reads	4	10	18	42	1	0	1	0
Total No. of reads (x10 ⁶)	1.55	1.55	1.4	1.4	1.71	1.71	1.86	1.55
Total No. of hq reads (x10 ⁶)	1.43	1.43	1.28	1.28	1.56	1.56	1.69	1.43
Percent hq reads	92.02	92.02	91.37	91.37	91.1	91.1	91.21	92.02
Total No. of bases (x10 ⁶)	155.4	155.4	140.3	140.3	171.9	171.9	186.1	155.4
Total No. of bases in hq reads (x10 ⁶)	143	143	128.2	128.2	156.6	156.6	169.8	143
Percent hq bases in hq reads	98.09	98.12	98.3	98.5	98.3	98.58	98.25	98.09
Total No. of hq bases in hq reads (x10 ⁶)	140.2	140.3	126	126.3	153.9	154.4	166.8	140.2
Percent A's	33.42	25.68	32.6	28.02	32.71	27.57	32.18	33.42
Percent G's	21.43	19.55	20.33	19.13	20.63	19.22	20.57	21.43
Percent C's	19.59	21.47	19.23	20.33	19.26	20.64	19.42	19.59
Percent T's	25.55	33.3	27.84	32.51	27.41	32.57	27.83	25.55
Percent non-ATGC bases	0.18	0.02	0.17	0.93	0.16	0.96	0.17	0.18
No. of primer/adapter contaminated reads	0	2	2	0	3	0	2	0

de novo contig assembly

The high quality filtered reads were utilized to carry out assembly of shorter read fragments into larger contiguous set of sequences called contigs. The assembly was carried out using velvet *de novo* assembler, that relies on user defined Kmer (hash length) to assemble contigs. The statistics mentioned in the table 2 illustrates a significant variation in various assembly parameters with varying hash lengths. A hash length is considered optimum if it generates minimum number of contigs, higher N50 value and utilizes maximum proportion of the read information. Therefore, choosing an optimum hash length forms an important step towards the assembly of the read data. For all the *H. pylori* strains the assembly was carried out at different hash lengths and an optimum hash length was chosen as described above. The statistics of genome assembly of different strains are shown in the table 3. Thereafter, the assembled contigs were sorted according to a reference genome using an inhouse written perl script. The sorted contigs were used as a reference to align the paired-end reads using bwa (H. Li and Durbin 2009) and samtools (H. Li et al. 2009). The alignment file was visualized in tablet where the paired-end information of the reads was utilized to validate the order of contigs. These contigs were later joined to form scaffolds using SSPACE tool to formulate a draft genome.

Table 2: Contig assembly runs for NAK7 using velvet with various hash lengths. The optimum hash length is colored red.

Hash length	No. of contigs	N50	Max. contig length	Approx. genome size	No. reads utilized
59	168	43166	114885	1577894	9799461
61	160	43192	114884	1578834	9755774
63	143	62931	125013	1579711	9711311
65	139	62931	125013	1579585	9667649
67	130	55593	125013	1579384	9621889

69	129	62931	125013	1579830	9576928
71	125	64085	125013	1580185	9528618
73	124	62931	125013	1579858	9473856
75	117	62931	125013	1580135	9419345
77	115	67610	125013	1580342	9361413
79	113	67612	125013	1580922	9299885
81	113	67614	125013	1580581	9236807
83	108	67616	125013	1581642	9175262
85	105	65522	125013	1581674	9108254
87	95	67620	140938	1581927	9038310
89	97	65895	178669	1582084	8963733
91	99	64112	239607	1581680	8882743
93	110	47598	224628	1581051	8793092
95	229	15531	50592	1577551	8691108
97	1353	1578	7481	1466673	8246455

Table 3: Contig assembly statistics obtained for various *H. pylori* strains.

Strains	No. of contigs	N50	Max. contig size	Approx. genome size	Reads utilized	Total No. of reads
UM018	72	64430	232496	1611122	2774351	2860370
UM054	81	39527	169147	1591048	2772269	2860186
UM007	72	60721	121584	1572401	2483668	2565436
UM034	65	60609	141393	1604159	3009014	3132784
UM045	58	63348	162928	1590413	3237000	3396122
NAD1	103	37513	96405	1603433	6958655	7061678
NAB47	107	30166	88644	1585150	12381629	12480382
NAK7	95	67620	140938	1581927	9038310	10067448

Genome annotation

The draft genomes were submitted to RAST annotation server for gene prediction and annotation. The genome characteristics as gleaned from Artemis are mentioned in table 4. All the sequenced genomes harbored on an average 1600 coding sequences with an approximate CDS length of 903 bp. The genomes revealed a high gene density with a coding percentage of nearly 91%. The strains also encoded 3 rRNA operons and 36 tRNA sequences. The subsystems distribution of genes predicted for one of the representative *H. pylori* strain NAK7 is shown in figure 8.

Table 4: Genome characteristics of various *H. pylori* strains sequenced and assembled during this study.

Features	Origin	Genome size (bp)	G+C content	CDS	Av. Length CDS (bp)	Coding %	rRNAs	tRNAs
UM018	Malaysia	1617433	39.05	1579	937	91.5	3	36
UM054	Malaysia	1603218	39.12	1585	926	91.6	4	36
UM007	Malaysia	1568678	38.84	1557	924	91.7	3	36
UM034	Malaysia	1714278	38.59	1669	940	91.5	4	38
UM045	Malaysia	1623875	38.96	1595	933	91.7	3	36
NAK7	India	1590891	39.05	1575	926	91.6	4	36
NAD1	India	1594674	38.98	1580	922	91.3	4	36
NAB47	India	1593604	39.15	1577	926	91.6	4	36

The assembled contigs/draft genomes were processed as per the guidelines of NCBI whole genome shotgun submission and were submitted with the accession numbers as mentioned in table 5.

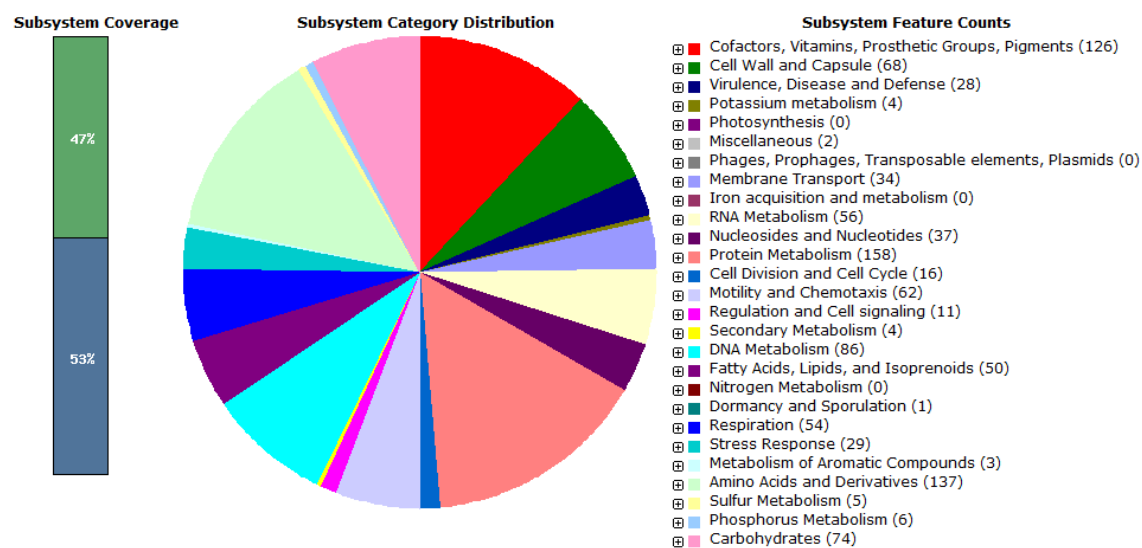


Figure 8: The subsystems distribution of the functions encoded by genes of NAK7 genome.

Table 5: Submitted genomes and their GenBank Accession IDs.

Genomes	Accession ID
NAK7	AONJ00000000
NAD1	AJGJ02000000
NAB47	AJFA02000000
UM018/I	AONK00000000
UM054/I	AONL00000000
UM007/C	AONM00000000
UM034/C	AONN00000000
UM045/M	AONO00000000
HP2017	CP002571
HP2018	CP002572

Conclusions

The advent of NGS has triggered a revolution in the field of comparative genomics in the recent past. In this chapter, we harnessed the power of high throughput sequencing technologies to carrying out the sequencing and whole genome assembly of various *H. pylori* strains. We established a concise and accurate pipeline (figure 9) to sequence and assemble bacterial genomes. The method utilizes the paired-read information to validate the sort order of contigs obtained after reference based alignment. Therefore, it employs a hybrid approach consisting of *de novo* (paired-end information) and reference guided method to achieve a better genome assembly. In addition, assembly of various *H. pylori* genomes also served as starting point for different whole genome comparative studies discussed in other chapters of the thesis.

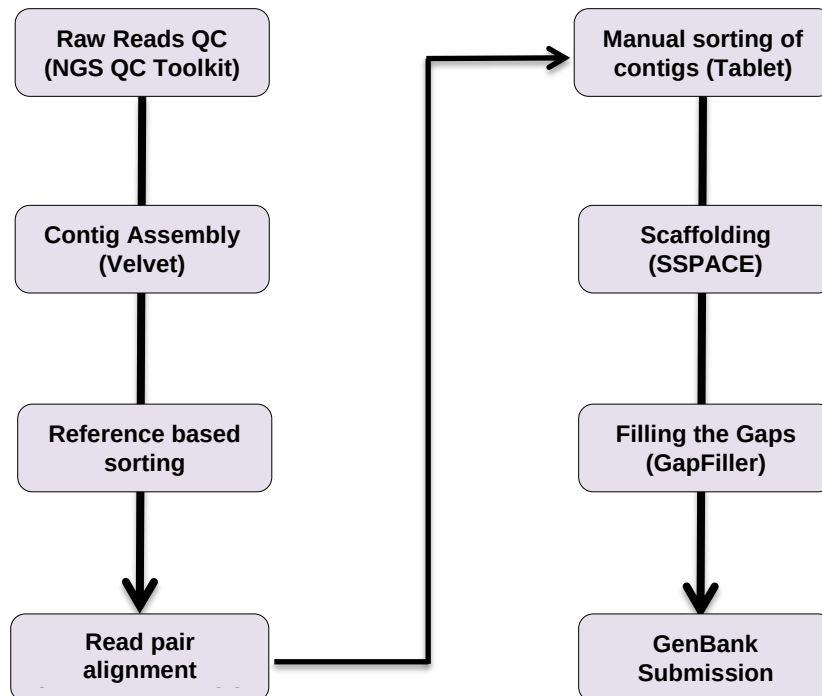


Figure 9: The pipeline standardized for the whole genome assembly.

CHAPTER III

Understanding the microevolution of *H. pylori* in a single stomach

Part of this chapter was published as:

- 1) Genomes of two chronological isolates (*Helicobacter pylori* 2017 and 2018) of the West African *Helicobacter pylori* strain 908 obtained from a single patient. Avasthi TS, Devi SH, Taylor TD, **Kumar N**, Baddam R, Kondo S, Suzuki Y, Lamouliatte H, Mégraud F, Ahmed N. **J Bacteriol.** 2011 Jul;193(13):3385-6.
- 2) Genome of *Helicobacter pylori* strain 908. Devi SH1, Taylor TD, Avasthi TS, Kondo S, Suzuki Y, Megraud F, Ahmed N. **J Bacteriol.** 2010 Dec;192(24):6488-9.

Introduction

H. pylori has co-evolved with humans for more than 60,000 years (Moodley et al. 2012) and therefore, its genome carries imprints that parallels with the migration and diversification history of humans (Ghose et al. 2002). The bacterium is usually transmitted in families as apparent from the existence of low genetic diversity among the strains isolated from family members. The infection is usually acquired during the childhood and can persist lifelong if not eradicated. Though majority of studies have concentrated on the elucidation of various pathogenic mechanisms, a detailed understanding of its adaptation and microevolution in a single host remains elusive. Recently, Linz et al demonstrated that *H. pylori* undergo a burst of spontaneous mutations during initial phase of colonization in the host (Linz et al. 2014). The analysis of multiple chronological isolates from a single patient has also revealed a higher recombination and mutation rates for *H. pylori* (Israel et al. 2001). A detailed understanding of the microevolution in a single host would facilitate development of better and effective control and eradication regimens (Kennemann et al. 2011). The knowledge would also further our understanding of various adaptive mechanisms employed by *H. pylori* to persist in the dynamic gut environment.

The availability of whole genome sequences enables better resolution for understanding the genomic perturbations that the pathogen might have occurred in a single host when compared to limited sequencing approaches (van El et al. 2013). The genome sequences not only aid in typing but also serve as a source to investigate the evolutionary changes and to map transmission routes. In addition, it provides a more accurate means to predict virulence potential and therefore would aid in designing effective epidemiological tracking and control regimens (Toft and Andersson 2010).

In this chapter, we have used whole genome comparative analysis to understand the microevolution in chronological isolates from a single patient who suffered a recrudescence gastric ulcer. In 1994, an African patient living in France was reported with duodenal ulcer (Prouzet-Mauleon et al. 2005). The endoscopic examination of the patient revealed a severe inflammation and *H. pylori* strain HP908 was cultured from the gastric biopsy specimen. The patient was administered a 10 day triple drug therapy comprising of lansoprazole, amoxicillin and clarithromycin. The patient was pronounced clear of infection after the eradication therapy through follow up endoscopic examination. In 2003, after a gap of 9 years the same patient reported similar symptoms and was diagnosed with recrudescence duodenal ulcer upon endoscopic examination. Two *H. pylori* strains HP2017 and HP2018 were cultured from the biopsies collected from the antrum and corpus regions. The patient was again administered similar treatment and was cleared of infection and symptoms.

At the end of the above exercise we had a collection of three strains; HP908 isolated in 1993 and HP2017 & HP2018 both isolated in 2003. The three strains revealed high genetic similarity as observed by fingerprinting assays such as enterobacterial repetitive intergenic consensus (ERIC) and repetitive extragenic palindrome sequences (REP) PCR. The high genetic similarity also suggested a possible chronological relationship among these strains. Therefore, we carried out whole genome sequencing and comparative analysis to understand the variations among the strains. The study could render important insights into the microevolutionary mechanisms of *H. pylori*. The whole genome comparison also revealed a closed similarity among the strains. The phylogenetic analysis revealed an African ancestry for the strains as they clustered together with J99 which was also isolated from a patient of African origin. The comparative analysis of their genetic composition revealed similar virulence capabilities and the variations were mainly concentrated in plasticity regions, CagPAI and OMPs. The outer membrane proteins

experienced elevated proportion of small variations that could have provided adaptive advantage to *H. pylori*. The existence of strain specific genes even in these chronological isolates is in line with the high mutation and recombination rates of *H. pylori*. The analysis lays emphasis on the role of variations in OMPs which are known to facilitate better adaptation and persistence of *H. pylori* in a single host.

Methodology

Patient information

The strains involved in this study were isolated from a single African patient living in France at two different time points. First isolation of strain 908 took place in 1994 when the patient was diagnosed with duodenal ulcer. It was followed by isolation of two strains 2017 and 2018 in 2003 upon recrudescence infection reported by the same patient. In total, 3 strains were isolated at a gap of nearly 10 years. The strains were cultured from the biopsy as mentioned by Valérie Prouzet-Mauléon et al. The isolated strains were maintained at -70 °C.

Genome sequencing, assembly and annotation

The genomic DNA from single colony isolate was sequenced using Illumina GAIIx yielding 60 Mb of sequence data. The library size of 300bp was read up to 101 bp from both sides generating paired end sequence data giving an approximate genome coverage of 50X. The read data was filtered using NGSQC Toolkit and was assembled into contigs using velvet *de novo* assembler. The contigs were ordered according to a close reference genome J99 using BLASTn. The ordered contigs were processed in a similar manner as mentioned in chapter 1 and formulated into draft genomes. The genomes were annotated using RAST server and the .gbk file was analyzed in Artemis to glean genome characteristics (Rutherford et al. 2000).

Whole genome comparison and phylogenetic analysis

The draft genome sequences of the three isolates were aligned using MGCAT with default parameters (Treangen and Messeguer 2006). To analyze the phylogenetic relationship, a whole genome based phylogenetic tree was constructed after pooling these isolates with 30 other complete genomes from the NCBI database representing strains from different geographical regions. The phylogenetic tree was

constructed using Gegenees tool (Agren et al. 2012) that relies on fragmented BLAST approach for calculating similarity among the genomes. An accurate all vs all fragment based alignment was carried as part of the tool with parameters of fragment size and step size set to 200 and 100 respectively. The distance matrix generated by Gegenees tool was used as an input to SplitsTree to visualize the phylogenetic tree using UPGMA algorithm (Loewenstein et al. 2008).

Analyzing the core and accessory genome content

To calculate the shared content of the three isolates, OrthoMCL program was used (L. Li et al. 2003). The tool carries out an all vs all BLASTp of the amino acid sequences encoded by predicted genes. The output is sorted and analyzed using MCL algorithm to identify orthologous genes. A similarity threshold of 70 percent was used to determine the orthologous genes. The output of OrthoMCL was further utilized to decipher the core genome or the gene clusters having orthologs in all three genomes. The orthologous gene cluster which had orthologs in only one or two genomes constituted the accessory gene clusters.

Analyzing virulence associated regions

CagPAI and plasticity zones are the two horizontally acquired regions and are related to virulence because of the genes they encode (Occhialini et al. 2000). The annotated files of the genome were manually scanned for the presence of these regions. The CagPAI region is located at the 3' end of the glutamate racemase gene and encodes nearly 26 proteins. Both gene wise and nucleotide comparisons of the whole CagPAI region were made to identify the variations among the three isolates. The other important region also known as plasticity zone was identified using Island viewer (Langille and Brinkman 2009) that employs four different algorithms to identify genomic islands. Their comparison was also made in parallel to Cag-PAI.

Virulence gene content

All the virulence genes of *H. pylori* reported in the Virulence Factor Database (VFDB) (Chen et al. 2005) were downloaded. Their status in the three isolates was determined by carrying out a BLASTp search of predicted genes in strains against those reported in VFDB. The presence and absence of a certain gene was calculated based on the query coverage and identity threshold of 70 and 85 percent respectively. The changes in the outer membrane proteins SabA and BabA were analyzed manually using ClustalW (Thompson et al. 2002).

Results and discussion

Assembly comparison of whole genomes

The three isolates 908, 2017 and 2018 were sequenced, assembled and submitted to NCBI database as draft genomes. The genome characteristics of the three strains are given in table 6.

Table 6: Genome characteristics of the three chronological isolates involved in the study.

	HP908	HP2017	HP2018
Size	1549666	15488238	1562832
GC%	39.3	39.3	39.29
Coding %	91.5	91.5	91.6
Avg. gene length	883	893	896
No. of genes	1607	1608	1616
Predicted ORFs	1595	1593	1603
Pseudogenes	12	15	13
IS elements	IS605, IS606	IS605, IS606	IS605, IS606

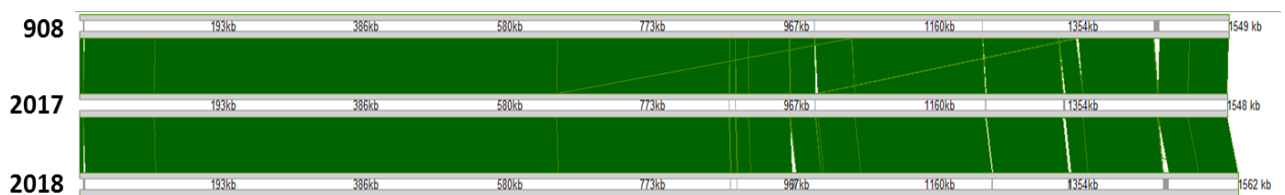


Figure 10: Whole genome alignment of 908, 2017 and 2018 generated using MGCAT. The green color represents the similar regions whereas the white color represents absence in other genomes.

The three strains possessed highly similar genetic characteristics even with regard to the composition of mobile genetic elements. All the strains encoded nearly 1600 genes with an average CDS length of 890 bp. The whole genome alignment of the three strains revealed nearly 98 % ANI (Average Nucleotide Identity) which was also evident in the figure 10. These observations dictated that the genomes were highly similar and confirmed the observations made by Alvi et.al. using PCR and other fingerprinting techniques.

Phylogenetic analysis

In order to assess the genetic relatedness among the strains together in comparison to other strains from different lineages, we constructed a whole genome phylogenetic tree as shown in the figure 11. All 33 strains grouped into phylogenetically distinct clusters representing particular lineage they belong to. The pattern of clustering was also in accordance with various MLST based studies (Falush et al. 2003). The strains of our interest (908, 2017 and 2018) clustered close to J99 revealing an African lineage for them. The observation also reflected upon the African ethnicity of the patient and therefore, was in consent with the fact that *H. pylori* are very specific to the host population.

The close clustering of the three strains as evident from the phylogenetic tree also supported the notion that strains were clonal and indeed 2017 and 2018 were descendants of the ancestral strain 908 which lend a strong support to the inferences made on the basis of various PCR and fingerprinting assays reported previously (Prouzet-Mauleon et al. 2005). The observation also ruled out the possibility of two different infections on separate occasions. Together, the data confirmed our model of chronological study.

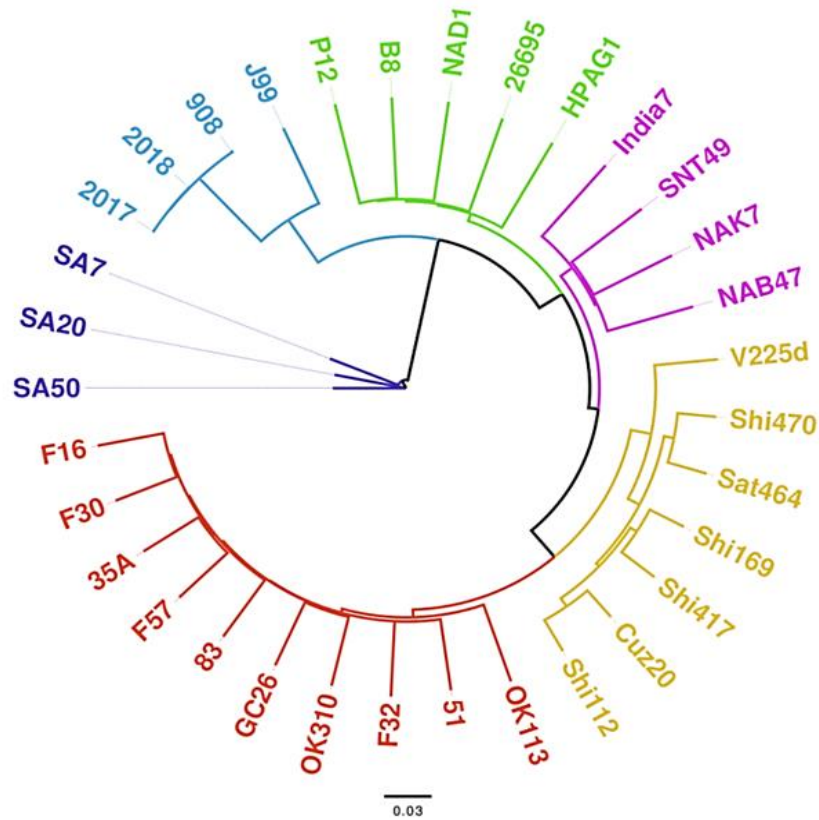


Figure 11: Whole genome phylogenetic tree of 33 *H. pylori* strains including the strains used in this study. The branches of the tree are colored differently to highlight various lineages.

Core and accessory gene pool

To determine the difference in the gene pool of these *H. pylori* strains, we analyzed their core and accessory gene content. A total of 1548 genes formed the core encoding various basic functions, while the accessory content varied from 23-30 genes. The strains also harbored approximately 20 strains specific genes per strain (figure 12). The core genome also harbored genes related to outer membrane functions and other virulence genes. The data also showed a comparable virulence potential of the three strains.

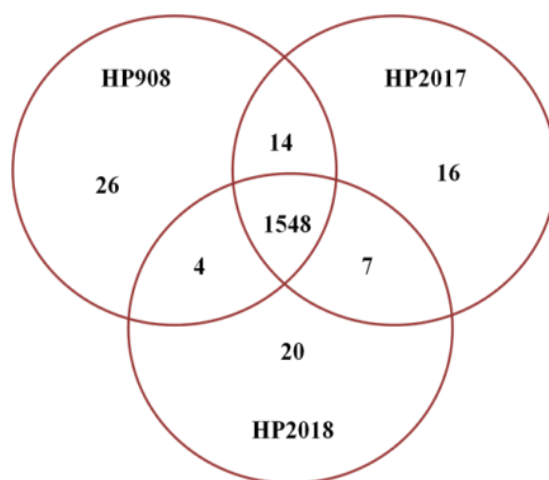


Figure 12: Venn diagram showing the comparative gene pool structure of *H. pylori* strains.

Variations in virulence apparatus

As 2017 and 2018 were found to be the descendants of 908, we were interested in deciphering the evolutionary changes they had undergone over a period of 10 years. First, we analyzed the variations within the two major virulence factors CagA and VacA that may have occurred during the decade long colonization. EPIYA motifs located at the C-terminus of CagA gets phosphorylated in the host cytoplasm thus dictating lot of alterations in different signaling cascades. Based on the amino acid sequences that flank EPIYA sequence, different types of EPIYA motifs have been identified. The strain 908 contained A, B and C type EPIYA motifs which were also found to be conserved among 2017 and 2018. The gene was found to be conserved among the three isolates with minimal changes signifying the importance of this gene in the pathogenesis of *H. pylori*. The comparison of VacA protein from the three genomes revealed a non-synonymous mutation making it functional in 2017 and 2018 compared to 908 in which the gene was non-functional due to an internal stop codon. Further, we carried out comparative analysis of all the virulence genes reported in VFDB that illustrated a comparable virulence potential (table 7) among the strains and therefore, provide support for observation of similar inflammatory damage after a period of 10 years.

Table 7: Comparative virulence profile of 908, 2017 and 2018 isolates. (+) represents the presence while (-) represents absence or partial presence.

	HP908	HP2017	HP2018
Adherence factors			
<i>babA</i>	+	+	+
<i>babB</i>	-	-	-
<i>hopZ</i>	-	-	-
<i>sabA</i>	+	+	+
Lewis antigens			
<i>fucT</i>	+	+	+
<i>fucU</i>	+	+	+
<i>neuA</i>	+	+	+
<i>neuB</i>	+	+	+
Endotoxin (LPS)			
<i>gluE</i>	+	+	+
<i>gluP</i>	+	+	+
<i>kdtB</i>	+	+	+
<i>lpxB</i>	+	+	+
<i>rfaC</i>	+	+	+
<i>rfaJ</i>	+	+	+
<i>rfbD</i>	+	+	+
<i>rfbM</i>	+	+	+
<i>wbcJ</i>	+	+	+
<i>wbpB</i>	+	+	+
<i>Urease</i>	+	+	+
<i>ureA</i>	+	+	+
<i>ureB</i>	+	+	+
<i>ureE</i>	+	+	+
<i>ureF</i>	+	+	+
<i>ureG</i>	+	+	+
<i>ureH</i>	+	+	+
<i>ureI</i>	+	+	+
Motility (flagella)			
<i>flaA</i>	+	+	+
<i>flaB</i>	+	+	+
<i>flgG</i>	+	+	+
Proinflammatory effect			
HP-NAP(<i>napA</i>)	+	+	+
<i>oipA</i>	+	+	+

Variations in CagPAI

The presence of Cag-PAI has been shown to have high a correlation with sever disease outcomes and thus is an important factor for pathogenicity (Brandt et al. 2005). The island measures around 32-40 kb in size and harbors nearly 30 genes that encode proteins required to construct a Type IV secretion system apparatus which is utilized to inject virulence proteins such as CagA directly into the host epithelial cells to bring about inflammatory changes. We compared the cag-PAI from three isolates to identify the variations it might have undergone during adaptation in a single host as shown in figure 13.

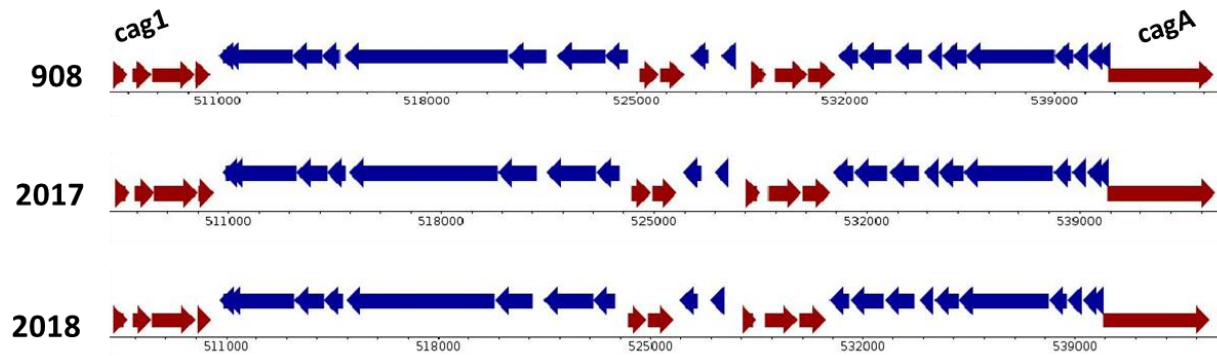


Figure 13: The comparison of Cag-PAI from three isolates. The brown color represents the gene in forward strand and those in blue represent genes encoded on reverse strand.

908_cag7	KESVKAYKDCVSRARNEKEKQECEKLLTPEARKLLEEKKSVKAYLDCVSKAKNEAERKE	900
2017_cag7	KESVKAYKDCVSRARNEKEKQ-----	839
2018_cag7	KESVKAYKDCVSRARNEKEKK----- *****:	839
908_cag7	CEKLLTPEAKKLLEAKESVKAYKDCVSRARNEKEKQECEKLLTPEAKKLLENQALDCLK	960
2017_cag7	-----	
2018_cag7	-----	
908_cag7	NAKTEAEKKRCVKDLPKDLQKKVLAKESVKAYLDCVSKAKNETERKECEKLLTPEAKKLL	1020
2017_cag7	-----	
2018_cag7	-----	
908_cag7	EAKESVKAYKDCVSRARNEKEKQECEKLLTPEARKLLEQEVKKSVKAYLDCVSRARNEK	1080
2017_cag7	-----ECEKLLTPEARKLLEQEVKKSVKAYLDCVSRARNEK	875
2018_cag7	-----ECEKLLTPEARKLLE-ESKKSVKAYLDCVSKAKNEA ***** * *****:*,**	874
908_cag7	EKQECEKLLTPEAKKFLEN----QALDCLKNAKTEAEKKRCVKDLPKDLQKKVLAKKS	1134
2017_cag7	EKQECEKLLTPEAKKFLEN----QALDCLKNAKTEAEKKRCVKDLPKDLQKKVLAKKS	929
2018_cag7	ERKECEKLLTPEAKKLLEAKESVKAYKDCVSRARNEKEKQECEKLLTPEAKKFLENQA- *:*****:*: *:.*. *. *:. * * . : *: :	933
908_cag7	VKAYLDCVSRARNEKEKQECEKLLTPEARKLLEEAKESVKAYKDCVSRARNEKEKQECEK	1194
2017_cag7	VKAYLDCVSRARNEKEKQECEKLLTPEARKLLEEAKESVKAYKDCVSRARNEKEKQECEK	989
2018_cag7	---LDCLKNAKTEAEKKRCVKDLPKDLQKK-VLAKKSVKAYLDCVSRARNETERKECEK ***** * *****:*,**	988

Figure 14: ClustalW alignment of *cag7* gene from 908, 2017 and 2018 with insertions and deletions in the repetitive regions.

Though number of genes in Cag-PAI remained conserved, we could observe insertions and deletions in *cag7/cagY* gene as shown in figure 14. The gene encodes a protein that coats the entire length of TypeIV secretion system and is reported to undergo changes by virtue of internal repeat sequences in the gene. This could be one of the mechanisms exploited by *H. pylori* to generate variations in TypeIV system's pilus length in order to adapt to the dynamic environment.

Outer membrane proteins

Outer membrane proteins allow pathogens to interact with the host but at same time these molecules are readily detected by the host to mount a counter defense. As a result, these proteins remain under constant fine-tuning for adaptation and persistence of the pathogen (Odenbreit et al. 2009). *H. pylori* encodes nearly 42 different outer membrane proteins (figure 15) belonging to three major families hop, hor and hom. SabA and BabA are the two key adhesins that are important for colonization and

persistence of *H. pylori*. The comparative analysis of these genes revealed high proportion of small nucleotide variations like SNPs and small Indels. The comparison of other adhesins also displayed similar phenomenon. Thus it was apparent from the data that outer membrane proteins were under constant fine tuning, allowing the bacterium to adapt to the dynamic environment of the stomach.

Figure 15: Various outer membrane proteins encoded by the 908 genome. The rings starting from inside represent GC skew, GC content, 2017, 2018, outer membrane genes in forward strand and reverse strands.

Conclusions

H. pylori has co-evolved with humans for a very long time that predates any event of human migration out of Africa. The bacterium is generally acquired in the childhood and may persist for long time if not eradicated by medication. The mechanisms underpinning this long persistence within a single host has been the focus of many studies. Here, we have tried to understand the variations that this bacterium might have undergone with time by means of genome wide comparison of isolates from a single patient over time or otherwise called chronological isolates. Our study together with other evidences proved that the first isolate HP908 was ancestral source for strains isolated ten years later from the same patient. The phylogenetic tree revealed an African lineage for the three strains and showed genetic relatedness to *H. pylori* J99 strain. This finding is in congruence with the ethnic background of the patient and therefore, reiterates that *H. pylori* are very specific to their host. Upon confirmation of the model of chronological evolution, we studied genome wide variations that might have occurred as a consequence of adaptations in ten years. The analysis of virulence genes and associated regions clearly depicted that the strains underwent small nucleotide variations in *cag7* gene and other outer membrane proteins. We could not find any significant recombination and horizontal gene transfer events among the three chronological isolates. In a recent study it was shown that *H. pylori* undergoes a burst of mutations i. e., SNPs rather than recombination to mediate quick adaptation (Linz et al. 2014). We observed a significant proportion of single nucleotide variations (SNVs) which indicated a strong selective pressure on the outer membrane proteins. Our conclusions were also in accordance with other studies on chronological isolates from the same patients. Hence, our analysis revealed that generation of genetically and phenotypically mosaic population in a quick succession having variants of OMP repertoires could be the key for *H. pylori*'s long persistent colonization in the human host.

CHAPTER IV

Comparative genomic analysis of *H. pylori* to understand the population genetic structure in multiethnic societies such as Malaysia

Part of this chapter was published as:

Comparative genomic analysis of *Helicobacter pylori* from Malaysia identifies three distinct lineages suggestive of differential evolution. **Kumar N**, Mariappan V, Baddam R, Lankapalli AK, Shaik S, Goh KL, Loke MF, Perkins T, Benghezal M, Hasnain SE, Vadivelu J, Marshall BJ, Ahmed N. **Nucleic Acids Res.** 2015 Jan;43(1):324-35.

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) (Perez-Perez et al. 2004). 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 (MALT) lymphoma (Khalifa et al. 2010). Generally acquired during the childhood (intra familial transfer), *H.pylori* establishes a lifelong persistent infection unless cleared by antibiotics (Kuipers et al. 2000).

The analysis of *H. pylori* strains has revealed existence of populations which are geographically localized. These populations have been classified into seven major lineages or genotypes depending on their regional prevalence, namely; HpEurope, HpSahul, HpEastAsia, HpAfrica1, HpAfrica2, HpWestAfrica and HpAsia2, based on MLST typing (Linz et al. 2007). The ability to undergo frequent mutation and recombination serves as one of the major contributors for the observed polymorphism among various *H. pylori* isolates. It also allows the bacterium to quickly adapt to the changing niche and establish a persistent gastric infection. Certain countries where people of different ethnicities, culture, lifestyle and religions reside present a pertinent model to examine the effect of migration and co-evolution. Studies entailing such settings would provide a better understanding of the evolutionary and adaptive strategies employed by *H. pylori* (Stein et al. 2013).

Malaysia is one such multiracial developing nation with a population comprising of four major ethnic groups: Malay/Bumiputera (67.4%), Chinese (24.6%), Indians (7.3%) and others (0.7%) (<http://www.statistic.gov.my>, 2010). In general, the Malays are the natives while Malaysian-Chinese were documented to have migrated from Southern China, and the Malaysian-Indians

from Southern India (Hatin et al. 2011). 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 (Hill et al. 2006).

Previous reports have shown a high prevalence of *H. pylori* infection amongst the Malaysian-Indians (69% - 75%), followed by Malaysian-Chinese (45% - 60%) and Malays (8% - 43%), and a minuscule number of inter-racial/ inter-community and inter- religion marriages resulting in a putatively reduced chance of cross-infection occurring between ethnic groups. The majority of the *H. pylori* isolates from Malays and Malaysian-Indians were found to share a recent common origin, while those from Malaysian-Chinese exhibited EastAsian ancestry. Generally, the *H. pylori* isolate collection from the Malaysian population is composed of strains from HpEastAsia, HpEurope and HpAsia2. Around 50% of the Malay isolates were similar to the Indian ones, suggesting a possible acquisition of *H. pylori* from Indians, although there is no clear evidence to support this interpretation yet. The reason for low prevalence of *H. pylori* infection in Malay population remains unclear and is likely to involve a number of environmental, genetic and host related factors (Tay et al. 2009).

Recent sequencing efforts have reported multiple genome sequences of *H. pylori* isolates from patients of different ethnicities and various disease manifestations from Malaysia (Khosravi et al. 2013; Y. Y. Lee et al. 2013). To date, there are 29 Malaysian *H. pylori* genomes (27 clinical strains and 2 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 ethnic backgrounds. The status of various virulence genes and outer membrane proteins 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 regimens.

Methodology

Strain collection and ethical approval

Biopsy samples were obtained from five non-ulcer dyspepsia (NUD) patients of different ethnicities [2 Malaysian-Chinese (UM007 and UM034), 2 Malaysian-Indians (UM018 and UM054) and 1 Malay (UM045)] at University of Malaya Medical Center (UMMC). All biopsies were obtained with written informed consent of the patients attending the Endoscopy Unit, at the UMMC. This study was approved by the Human Ethics Committees of the University of Malaya, Kuala Lumpur, Malaysia (Ref No- 943.2).

Bacterial culturing and DNA isolation

The *H. pylori* isolates were cultured by inoculating the biopsies on chocolate agar containing 4% blood base agar No. 2 (Oxoid), defibrinated horse blood (Oxoid), 2.5 mg trimethoprim, 5.0 mg vancomycin, and 1250 IU polymyxin B. Primary cultures were incubated at 37°C with 10% CO₂ for up to 10 days (with daily observation). For isolation of pure culture, a single colony was picked and sub-cultured onto the chocolate agar for 3 - 5 days. Identification of *H. pylori* isolate was performed based on microscopic morphology and biochemical testing (urease, oxidase and catalase). A plateful of *H. pylori* culture was suspended into 500µl of Tris buffer. The suspension was centrifuged (7, 5000 rpm for 10 mins) 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. The strains were sequenced assembled as mentioned in the chapter1.

Functional annotation, calculation of core and specific gene 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 blast alignment. The other 22 draft genomes reported from Malaysia were downloaded and their gene prediction was also performed as mentioned in the previous section. The core genome was calculated by identifying orthologs in every genome by applying MCL algorithm included in the orthomcl program. The parameters for deciding orthology such as identity and blast score cut-off 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 functional database (Tatusov et al. 2000) using rpsBLAST program (Natale et al. 2000) and followed by manual curation of these results.

Phylogenetic analysis

A whole genome alignment of 27 Malaysian isolates was carried with 39 others (draft or complete) from NCBI database using Gegenees tool (36). The tool utilizes a fragmented alignment algorithm to calculate average similarity among the compared genomes using blastn or blastx. 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 is displayed in terms of a heatmap matrix and is further used to deduce phylogenetic relationship which can be exported in the form of .nexus file. This nexus tree file was supplied as an input to

SplitsTree program (Huson 1998) for building an un-rooted phylogenetic tree employing Neighbor-Joining algorithm.

Virulence genes and phage detection

The whole genomes of these strains were screened for presence of virulence genes enlisted in the Virulence Factor Database (VFDB) using BLASTp program (38). The cut-off identity and query coverage was set to 70 and 60 percent, respectively. Further comparison of the amino acid sequences of CagA and VacA was carried out using clustalw. Phage related sequences in the genome were identified using PHAST server (Zhou et al. 2011) which integrates the analysis against various phage databases and compares key phage attributes to detect similar phage sequences in the query genome sequence.

Results and Discussion

Genome statistics

The whole genome sequencing of five Malaysian isolates (figure 16) revealed their chromosome sizes ranging from 1.56 to 1.62Mb. The genomes also revealed a low G+C content of 39% which is characteristic of *Helicobacter pylori*. The draft genomes were predicted to encode ~1600 genes with an average CDS measuring upto 930 amino acids. All the sequenced genomes harbored 3 rRNA operons as well as 36 tRNA genes. Two out five sequenced strains (UM045 and UM054) also harbored phage sequences encoding 136 and 12 putative genes respectively. A detailed genome statistics of these five isolates has been mentioned in table 8. The genomes were also compared using BLASTn against a reference strain G27 shown in figure 1.

Table 8: Genome characteristics of the sequenced *H. pylori* strains.

	UM018	UM054	UM007	UM034	UM045
Origin	M-Indian*	M-Indian	M-Chinese*	M-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	1,579	1,585	1,557	1,669	1,595
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)

*M-Indian: Malaysian Indian, M-Chinese: Malaysian Chinese

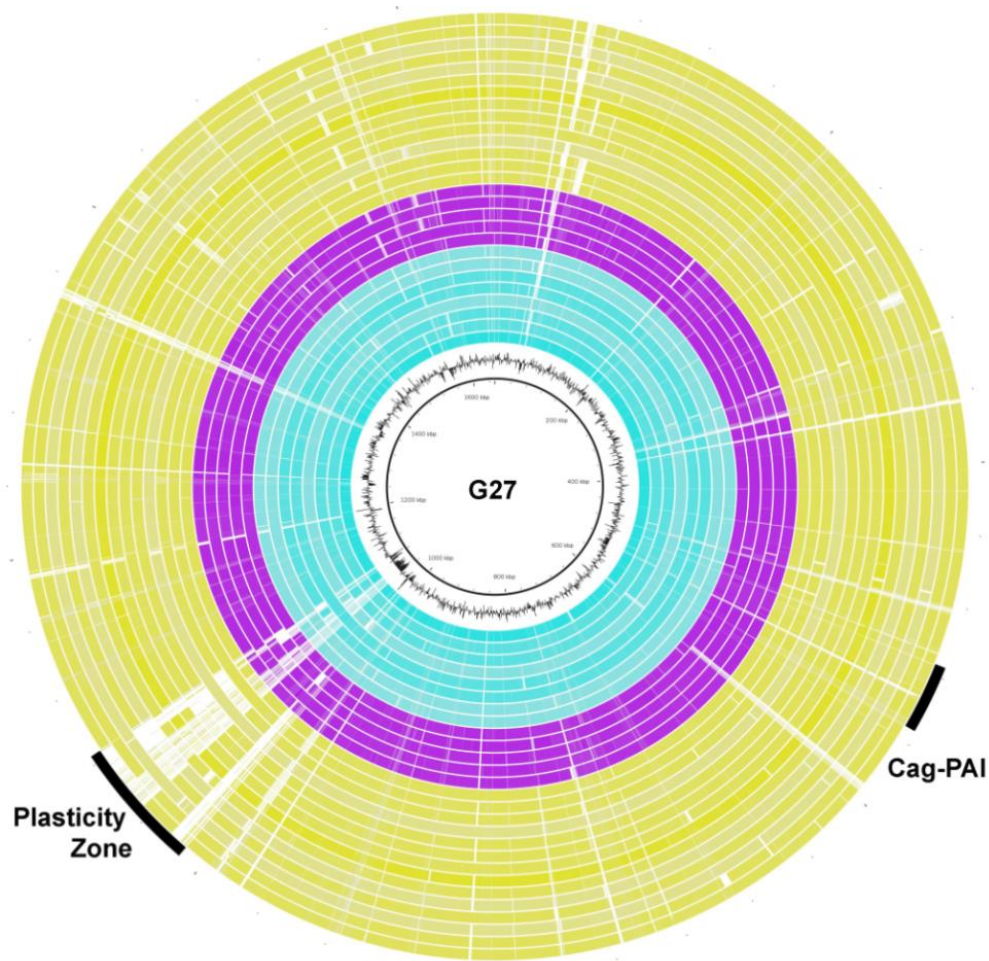


Figure 16: A circular representation of the genomes of Malaysian isolates: The draft genomes of 27 strains were aligned against the genome of reference strain *H. pylori* G27 where each strain is represented by a ring. The yellow rings represent Malaysian-Chinese, purple represents Malay and light blue represents Malaysian-Indian strains. The variable regions such as plasticity zones (PZ) and Cag-PAI are compared across all genomes using BRIG image generator.

Phylogeny determination

The sequenced isolates were pooled together with 38 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 tree. 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 17.

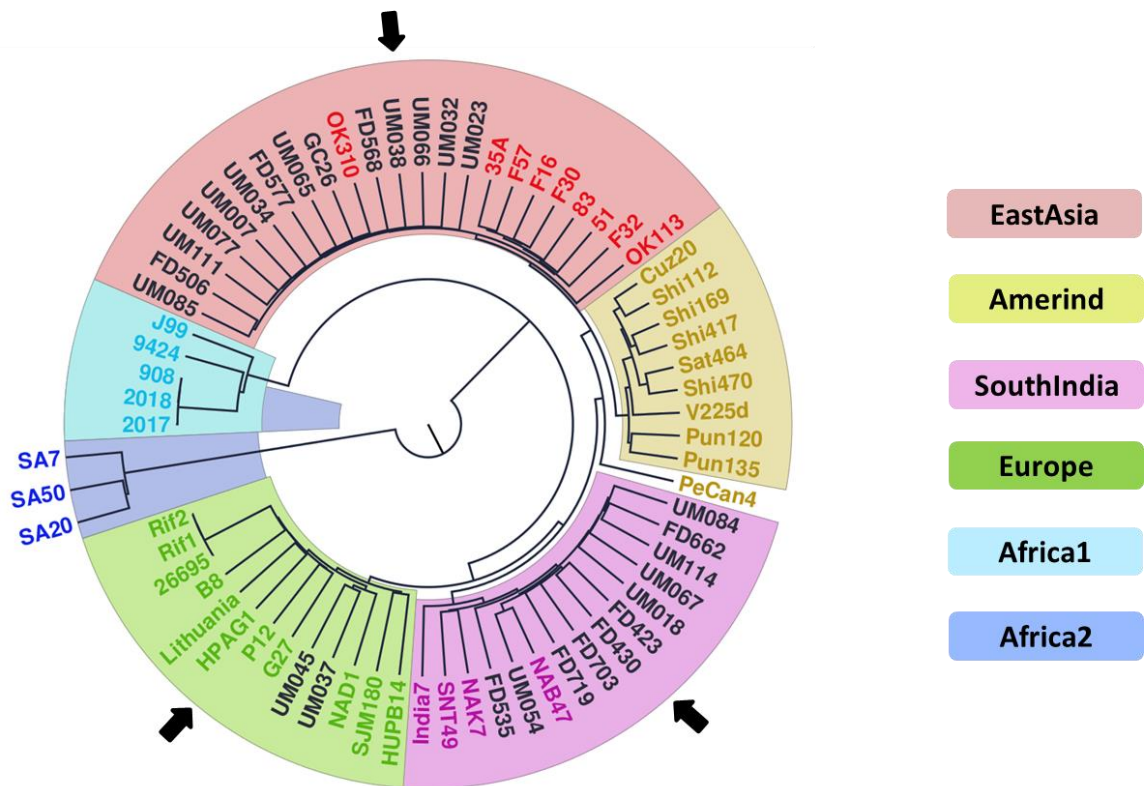


Figure 17: The whole genome phylogenetic analysis: The figure represents a whole genome comparison based phylogenetic tree of various complete and draft *H. pylori* genomes from different geographical regions. The tree was constructed based on neighbor joining algorithm using SplitsTree. The Malaysian strains used in the analysis are labeled in black whereas the other genomes are colored to represent their genotypes.

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) (Falush et al. 2003). Although few studies have reported existence of three lineages in South Asia; HpEurope, HpAsia2 and HpEastAsia, a clear understanding of their phylogeny could not be achieved because of scarcity of available genome sequences. The rapid increase in the number of genome sequences being available provided a better opportunity to achieve greater resolution in categorizing *H. pylori* strains by whole genome comparative studies. The isolates from South India and Malaysian-Indians clustered tightly forming a group named as HpSouthIndia. This close phylogenetic association among Indian and Malaysian-Indian strains is in accordance with the findings based on MLST typing (Tay et al. 2009). Moreover, UM045 and UM037 isolated from Malay and Malaysian-Indian patients respectively, clustered with HpEurope genotype suggesting their European ancestry. Further analysis of only Malaysian *H. pylori* genomes also revealed a bipartite clustering which was supplemented by the similarity score matrix shown in figure 18. All the strains of Malay and Malaysian-Indian (European) origin exhibited more similarity to each other allowing them to distinctly cluster from Malaysian-Chinese (East-Asian) strains. These findings also supported the hypothesis of ancient human migration ancestral Indians and their subsequent colonization in Malaysia. Similarly the intimacy of Malaysian-Chinese strains with HspEastAsia reiterated their common ancestry. Together the phylogenetic analysis explained a mixed population genetic structure of *H. pylori* existing in Malaysian population. These differential genotypes might underpin the observed discrepancy in the disease outcome among various ethnic groups.

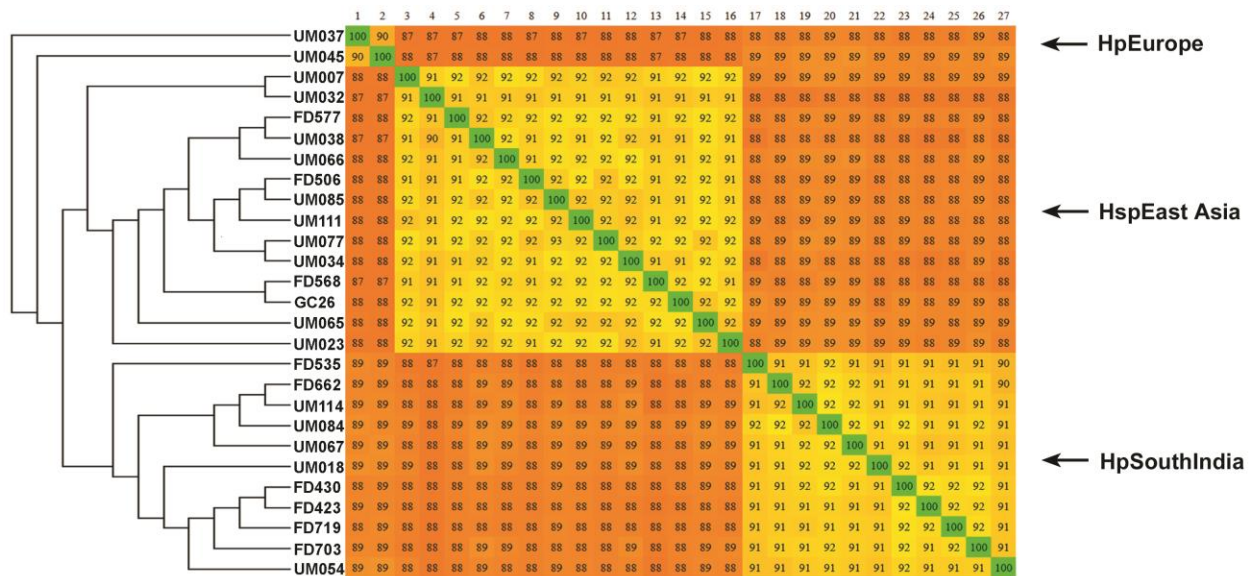


Figure 18: The analysis of 27 Malaysian genomes: The Neighbor-Joining phylogenetic tree constructed after the alignment of 27 Malaysian *H. pylori* isolates representing various ethnic groups. The heat plot shows the average similarity values among the strains.

Virulence potential

The observed phylogenetic distinction between the Malaysian isolates was further investigated for presence of differential virulence related gene content. The major virulence determinants which the bacterium uses for first host cell contact is outer membrane proteins (OMPs). These OMPs play important role in adhesion and are reported to be associated with pro-inflammatory secretions (Peek and Crabtree 2006). Outer membrane proteins in *H. pylori* have been categorized into five major different families based on their structural composition. These proteins carry out various functions ranging from host-surface interactions to non-selective porins for import of ions (Ahmed et al. 2013). 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 strains revealed a conserved

nature for most of the 62 OMPs with minor exceptions. The BLASTn similarity percentage for these proteins varied from 84 to 100 representing their polymorphic nature. Few of the genes such as *bopZ*, *bopMN*, *bopQ* (*sabB*) showed variable presence among the genomes but we failed to detect a lineage/group specific pattern among the East Asian and other groups of strains. The genes like *bomA*, *bomB* revealed a variable presence among the genomes.

The high recombination capability of *H. pylori* 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 their pathogenesis of this gastric inhabitant. The virulence database enlists 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 genomes using BLASTp . All the strains harbored intact Cag Pathogenicity Island (Cag-PAI) including CagA conserved among them. Other virulence markers like *oipA*, *vacA* and *flgG* genes which have been associated with severe disease phenotypes were also consistently present. All the genomes also possessed components of the urease cluster which allows *H. pylori* to survive in the low pH conditions (Dorer et al. 2010). The analysis thus revealed a high virulence potential of the genomes irrespective of the ethnic groups that they are isolated from.

However analysis of gene polymorphisms in CagA revealed lineage specific patterns. CagA protein encoded by the Cag-PAI is highly correlated with severe gastric outcomes. The enhanced virulence potential of CagA has earned its name as a bacterial “oncoprotein”(Ohnishi et al. 2008). Phylogenetic analysis of CagA could clearly differentiate East-Asian (Malaysian-Chinese) strains from their non-East-Asian (Malay and Malaysian Indian) counterparts. The analysis of alignment revealed a lineage specific variation not only at C-terminal encoding EPIYA phosphorylation domains 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 suggesting a differential evolution of this protein among different genotypes and its probable role in the observed disparity of the disease outcomes.

The core genome

The gene content analysis of Malaysian isolates was carried out by calculating the core and accessory genome content. The genes which 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 27 strains formed 1993 orthologous gene clusters. Among them, 1266 clusters possessed 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 in accordance with the previous reports (W. Lu et al. 2014). 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 19 and the rest 261 gene clusters remained uncategorized. Among these 261 gene clusters majority were found to be encoding putative hypothetical proteins based on their comparison with other *H. pylori* genes. Out of 1005 functionally categorized gene clusters, 115 clusters encoded proteins involved in translation, ribosomal structure and biogenesis. Other than performing housekeeping functions, studies have shown that some genes like Pol I also aids in generating genome plasticity. The core genome was also enriched in the genes related to cell wall biosynthesis, amino acid and 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 the long association with its host. Interestingly, 81 core gene clusters were identified as belonging to

multiple functional classes. These genes could represent the proteins involved in multiple pathways (Kersulyte et al. 2009). Multifunctional proteins could also be advantageous to *H. pylori* with a small genome and limited coding potential but requires further functional validation. Moreover, a high proportion of hypothetical proteins in the core genome also warrant their functional characterization to ascertain their role in the biology and thus pathogenesis of this gastric pathogen.

H. pylori have been reported to possess a high strain specific gene content majorly localized in hypervariable region known as plasticity zones (Erzin et al. 2006). Few genes from plasticity regions like hp0986 have been associated with increased pro-inflammatory secretion in cell culture studies. The strain specific content of Malaysian genomes varied from 10 to 12 genes per genome (figure 19). A high proportion of these strain specific genes were predicted to encode hypothetical proteins while few of them encoded putative restriction-modification systems (RM) in some strains.

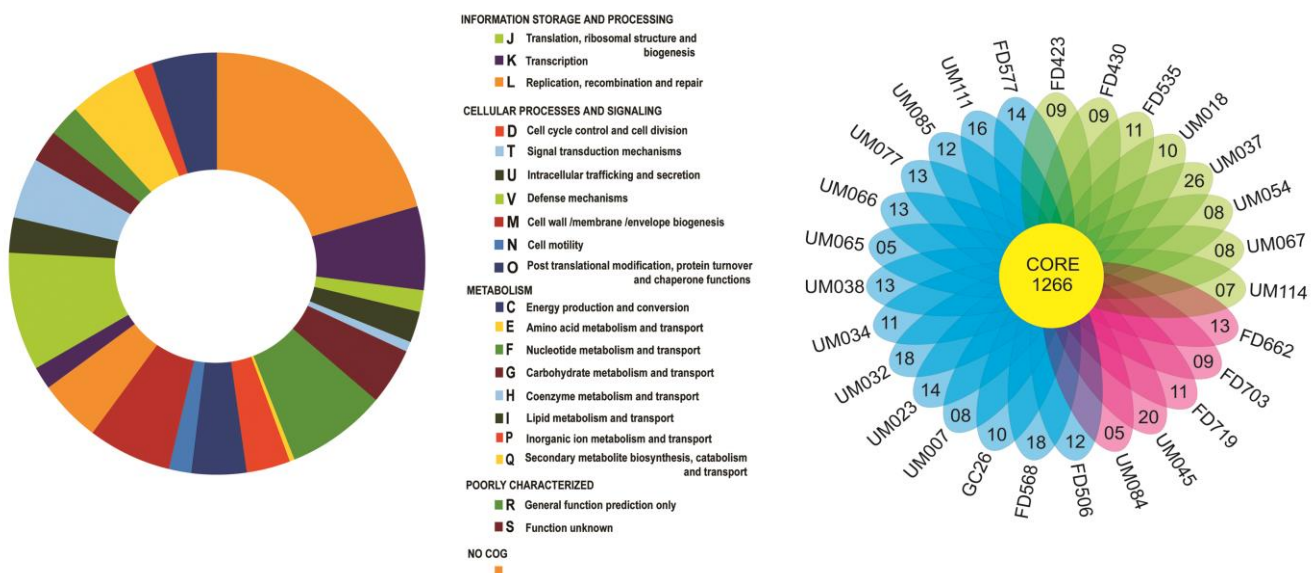


Figure 19: The functional COG classification of genes: The COG functional classification representing core genome of the Malaysian *H. pylori* isolates. Core and specific gene content observed among various strains compared in the study.

Lineage specific genes

The phylogenetic analysis classified the Malaysian isolates into three distinct genotypes. The strains belonging to HpEurope shared a close similarity to HpSouthIndia strains compared to HspEastAsia strains. We divided the strains into two groups in accordance with their phylogenetic clustering and genomic similarity. In total, 14 Malaysian-Chinese represented 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 showed differential presence among non-East Asian genomes. Out of these 33 gene clusters, four gene clusters did not have ortholog in any of the non-East Asian strains, one of which was predicted to encode putative lysozyme like protein (table 9). These proteins have been observed to be up regulated during DNA damage induced stress in *H. pylori*. The other three encoded hypothetical proteins and awaits further functional characterization. A clear functional understanding of these genes could provide significant insights into the underlying distinction between East Asian and non-East Asian genotypes and associated disease outcomes (Ahmed et al. 2008).

Table 9: The genes showing lineage specificity: The table shows the identified genes having high prevalence in any one of East-Asian and non-East-Asian genomes.

Cluster ID	Status in <i>H. pylori</i> strains		Predicted functions	Orthologs in 26695
	EA (n=14)	Non-EA (n=13)		
2426	14	0	Lysozyme family protein	HP0339
2425	14	0	Hypothetical protein	HP0344
2424	14	0	Hypothetical protein	HP0346
2423	14	0	Hypothetical protein	N.A*
2403	14	2	Lipopolysacchride biosynthesis protein	N.A*
2376	14	4	Type II Restriction Endonuclease	HP1537
2375	14	4	Type II Methylase	N.A*
2370	14	5	Glycosyltransferase	N.A*
2364	14	4	DNA Methyltransferase	HP0051
2402	1	13	Hypothetical protein	N.A*

*Represents absence of corresponding homolog in the sequenced strain 26695

Restriction-Modification genes

Previous studies on *H. pylori* genomes have revealed a high number of genes encoding RM systems (Lin et al. 2001). In this 27 Malaysian strains, a total of 1077 genes were predicted to encode RM related genes. Further clustering by UCLUST with an identity of 80%, these genes were arranged into 149 clusters (Edgar 2010). 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 the figure 20. The analysis clearly outlines the extent of diversity both in terms of numbers and allelic types of RM genes present in *H. pylori*. This might explain the observed strain to strain diversity in *H. pylori*. Higher proportion of RM gene 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 greatly pursued.

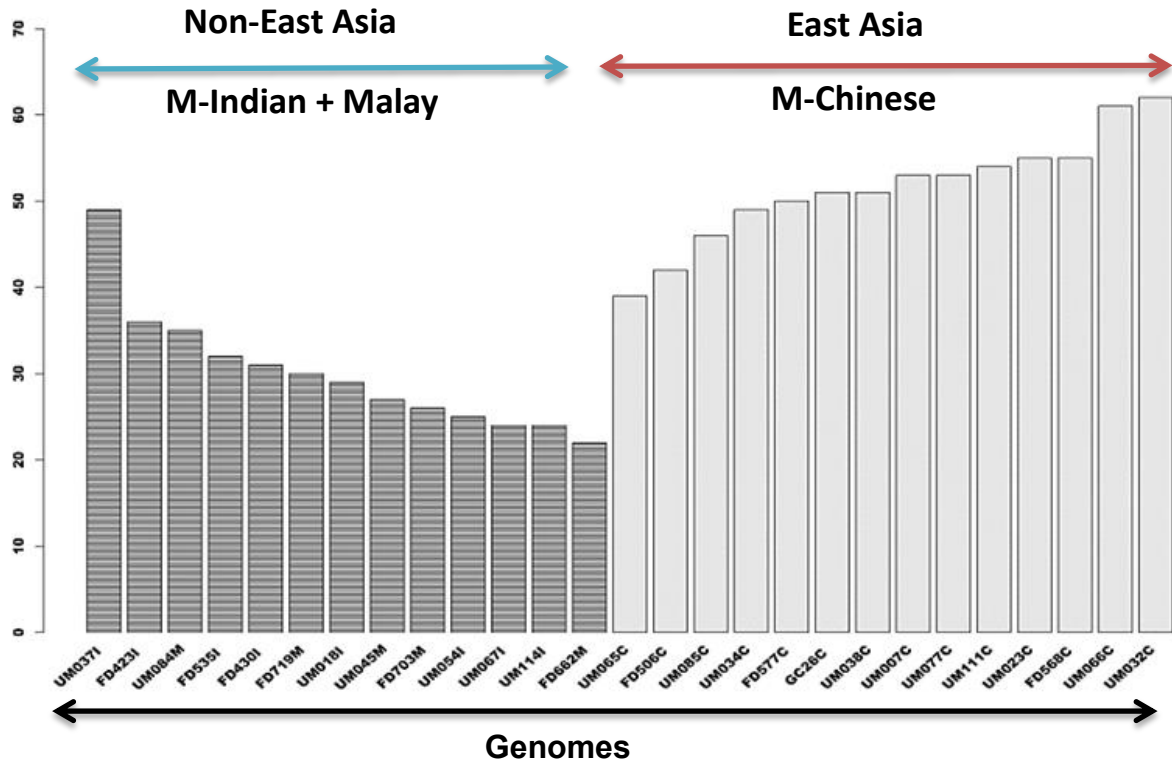


Figure 20: The distribution of RM genes in various strains: The graph shows number of genes annotated to encode putative RM functions in each strain. The Y axis represents number of genes and the X axis denotes strain names.

Recent studies have proved that inactivation of these R-M genes lead to changes in the expression of several genes in *H. pylori* (Furuta et al. 2014). Moreover, these RM genes have also been shown to exhibit phase variation (Krebes et al. 2014). Therefore a clear understanding of the role played by these RM genes is necessary to better understand the mechanisms of genetic polymorphism in *H. pylori*.

Differentially evolving genes

It has been proposed that pathogenic bacteria that resort to long term adaptation to a particular niche modulate its core gene repertoire in synchrony with their virulence complement to gain fitness advantage (Rohmer et al. 2011). Therefore we attempted to identify core gene clusters which showed 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 genes 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 could be assigned to functional categories while the rest did not find a significant hit with the COG database. These functional categories were enriched in genes with functions related to cell-wall/membrane biogenesis, recombination repair, and some poorly characterized genes (figure 21). These included various outer membrane proteins that have been proven to be differentially evolving among East-Asian and non-East-Asian genomes. Even *cagA* and *vacA* whose polymorphism is well established among the East-Asian type and non-East-Asian type strains were identified in the analysis. The core genome thus possesses a significant number of differentially evolving genes. This also mirrors the differential adaptive and evolutionary pressure experienced by these isolates.

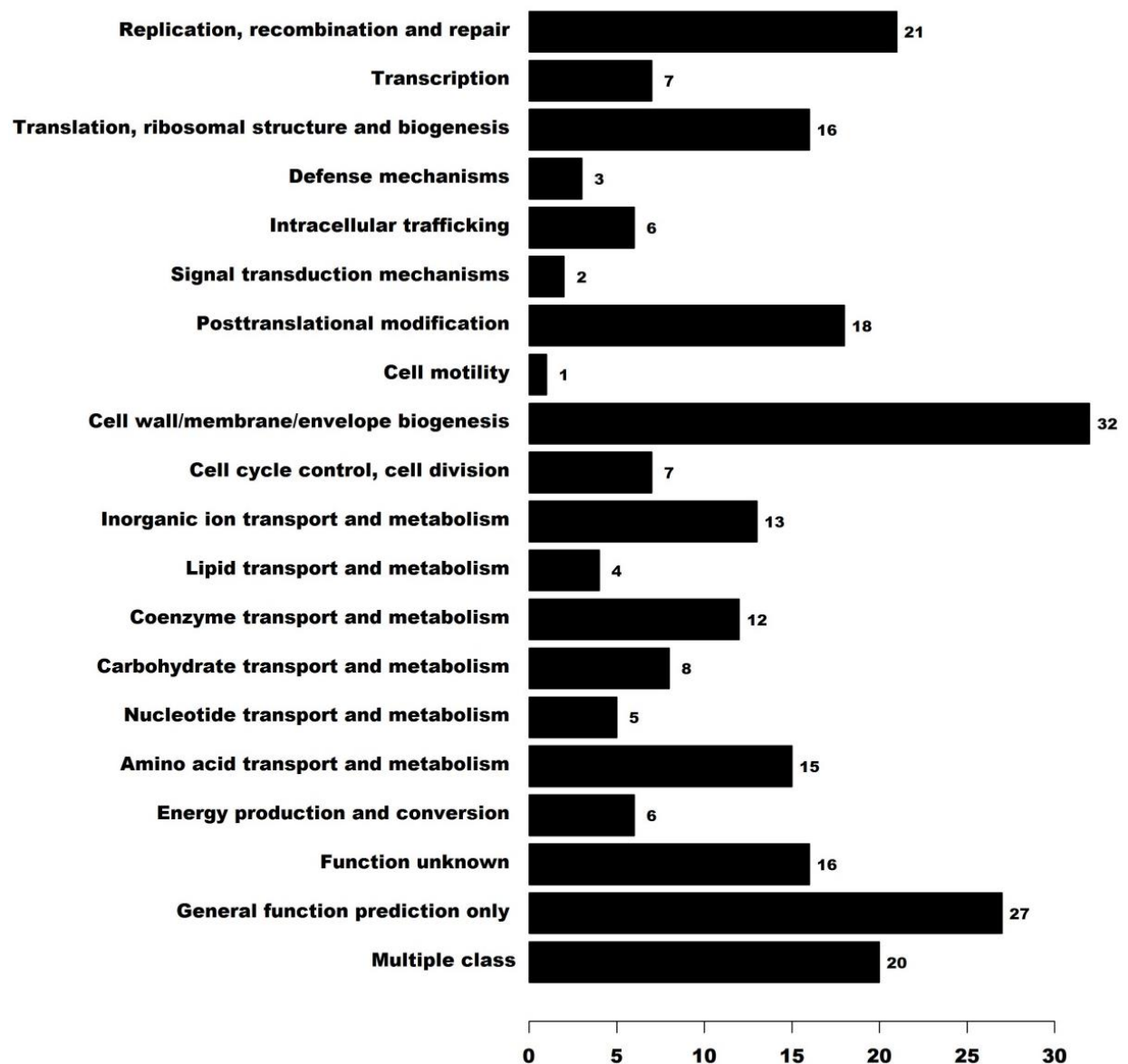


Figure 21: The functional classification of differentially evolving genes: The graph shows the COG functional classification of various core genes with signs of differential evolution among East-Asian and non-East-Asian strains. The Y axis denotes the functional category and X axis represents the number of genes in a particular functional category.

Conclusions

This study was aimed at understanding the population genetic structure of *H. pylori* and cues they could provide for the observed variation in disease outcomes. The whole genome phylogenetic analysis resolved the strains into three lineages representing the patients from various ethnic groups. 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 compared to others but remains questionable considering the draft status of the genomes. In light of these findings, further investigation of core gene pool revealed a significant proportion of genes under differential stress among East-Asian from non-East-Asian strains. These differentially evolving genes which included restriction modification genes and outer-membrane proteins different lineages should be investigated for their possible role in manipulation of the virulence repertoire. Therefore we believe that *H. pylori* employs 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 regimens and drug intervention for *H. pylori*. A better understanding of the interplay between the virulence and core gene complements would allow us to gain a better understanding of the pathogen biology and host-pathogen interactions.

CHAPTER V

Understanding the population
genetic structure of *H. pylori* in India

Introduction

H. pylori colonizes almost half of the world population and is considered a primary risk factor for a variety of gastric ailments ranging from gastritis to gastric cancer (Perez-Perez et al. 2004). *H. pylori*'s long association with humans fostered genome diversification leading to emergence of host specific genotypes (Zhu et al. 2014). Globally, *H. pylori* can be divided into 7 major lineages and sub lineages in accordance with their prevalence in various geographical regions (Breurec et al. 2011). The *H. pylori* infection and associated disease burden varies among different geographical regions and can be best depicted by the socioeconomic status of the prevailing population. The burden is higher in developing countries as opposed to the developed nations. One of the lineages, HpEastAsia has been associated with increased risk for gastric cancer as evident from higher burden of gastric cancer in East Asian countries such as Japan, China and Korea. The human populations in South Asian countries are comprised of people from different ethnic backgrounds. A similar scenario is observed with regards to *H. pylori* and its associated diseases, since different genotypes have been observed to prevail in these populations (N. Kumar et al. 2015). Therefore, a careful epidemiological survey of prevailing genotypes is necessary to assess the associated health risk it poses and to design effective control regimens.

H. pylori transmission occurs in families and this makes it a surrogate to understand ancestral human migrations and evolution. Various MLST based studies have revealed that *H. pylori* differentiation parallels human migration out of Africa. Wirth et al reported the existence of a heterogeneous mix of three independent lineages (HpEastAsia, HpEurope and HpAisa2) (Wirth et al. 2004) reflecting different waves of ancient human migrations that shaped the population of Ladakh in the northern region of India. Subsequent analysis of strains isolated from patients of different ethnic backgrounds mainly from Hyderabad, revealed a European ancestry for the *H.*

pylori (S. M. Devi et al. 2007). However, these studies were based on the isolates collected from few regions which might render inaccurate estimation of the existing population structure of *H. pylori* in India. On the other hand, studies on human migration based upon the mitochondrial haplotype sequences and Y chromosome mapping have provided evidence for two distinct waves (Reich et al. 2009) of human migration that populated India in addition to sparse mixing from the east. Hence, owing to the host specific spread of *H. pylori* genotypes, the population structure of *H. pylori* should also reflect ancestral human migrations and therefore should be comprised of more than one ancestral lineage in India. Therefore, in this study we set out to determine the population genetic structure of *H. pylori* in India and to understand their implication on the observed infection burden.

We pooled together the strains from earlier studies along with the whole genome sequences of five *H. pylori* strains reported from India and those from Malaysia. The Bayesian analysis of the MLST sequences of 101 strains revealed three major lineages: HpSouthIndia, HpEurope and HpEastAsia. Only Ladakh strains had a minor population of HpEastAsia while the rest strains belonged to the other two major lineages. These findings revealed a population structure that is in accordance with the human migrations in the region and represent the dominance of two lineages HpEurope and HpSouthIndia. The whole genome phylogenetic analysis of the five strains also resonated similar relationship by clustering along with two genotypes: HpEurope and HpSouthIndia. The investigation of their genetic composition revealed a higher virulence potential for the five sequenced strains. The analysis of *cagA* gene revealed western type signature of the EPIYA motifs in all the strains from India. Taken together, our results revealed two major lineages that prevail in the region and their comparative virulence profile and western type CagA EPIYA motifs provided strong evidence for the low occurrence of gastric cancer in

spite of having higher infection burden. The study provided cues as to how the genotype structure of *H. pylori* shapes the associated disease burden in the region.

Methodology

Strain collection

The MLST sequences of seven housekeeping genes (*atpA*, *efp*, *mutY*, *ppa*, *trpC*, *ureI*, *yphC*) of the strains reported from India and Malaysia were collected. Allele sequences of 50 isolates from Ladakh region were downloaded from PubMLST database. 23 others were obtained from the study published by Devi SM, et al and data of 83 Malaysian isolates were obtained from two other publications analyzing the strains from Malaysia. In addition, whole genome sequences of *H. pylori* from India were also mined to extract MLST alleles for the analysis. Together, MLST sequences of 7 housekeeping genes from 101 different isolates were utilized to carry out this study. The details of the strain collection are also mentioned in table 10.

Table 10: Strain collection for this study.

No. of strains	Geographical region	Reference
MLST Sequences		
50	Ladakh	MLST database
23	Hyderabad	Devi SM <i>et. al.</i>
83	Malaysia	Tay CY, <i>et. al.</i>
Whole genome sequence		
NAK7	Kerala	AONJ00000000
NAD1	Delhi	AJGJ02000000
NAB47	Bangalore	AJFA02000000
SNT49	Orissa	CP002983
India7	India	CP002331

Population structure analysis

The population structure was determined using Bayesian Analysis for Population Structure tools (BAPS). Seven different files were prepared after concatenation of individual allele sequences from all the strains. These files were provided as input to BAPS software (Corander et al. 2004) along with strain information file for the analysis of population structure. The process was carried out using initial population size (k-value) of 1-10 with a linked loci model to generate cluster of related strains. The use of admixture model employed to account for mixing of alleles that could have occurred as a consequence of long coevolution between strains of various lineages. The results of the analysis were visualized using admixture provided by BAPS tool.

Phylogenetic analysis

The draft genomes of Indian isolates (NAK7, NAD1, NAB47, SNT49 and INDIA7) as well as 39 other genomes from NCBI database were downloaded and compared using Gegenees tool. This tool was used to break the genome into fragments which are then aligned against all using BLASTn or BLASTx program. The similarities between the genomes were assessed based on the fragmented alignments which are represented in the form of a heat map. The analysis was run using the fragment length of 200 and the anchor size of 100. The heat plot displaying the similarity percentage among the genomes was exported as nexus file with the threshold set to 25%. The nexus file was supplied as an input to SplitsTree program for building an un-rooted phylogenetic tree employing UPGMA method. The newick tree file was visualized in FigTree for proper labelling and analysis (<http://tree.bio.ed.ac.uk/software/figtree/>).

Virulence content analysis

The whole genomes of the strains were screened for virulence genes of *H. pylori* that are enlisted in the Virulence Factor Database (VFDB) using BLASTn program. The cut-off for identity and query coverage was set to 85 and 70 percent, respectively. ClustalW was used for comparing the amino acid sequences of various pathogenicity related genes. The *cagA* gene sequence from 14 *H. pylori* genomes representing strains from India and Malaysia was extracted and aligned using ClustalW in MEGA software (S. Kumar et al. 1994). The alignment was used to construct *cagA* gene based tree using Neighbour-Joining algorithm with 1000 bootstraps. The tree was visualized in FigTree by loading newick file.

Results and discussions

Population structure analysis

To determine the population structure of *H. pylori* in India, we performed a BAPS analysis involving allele sequences of 7 housekeeping genes from strains reported from India and Malaysia. Since Malaysia harbors a significant proportion of Malaysian-Indian population, involving the Malaysian *H. pylori* strains in the analysis would provide higher resolution and also strengthen our conclusions. The BAPS analysis was run with initial population approximation of 1-10. An optimal clustering of the data could be obtained for the k-value of 3 after carrying out three separate runs. The admixture analysis was further performed on these clusters to decipher the extent of admixture among different clusters. The individual clusters were then classified into lineages depending on the lineage ancestry of the strains they were comprised of. The analysis revealed three major lineages HpEurope, HpSouthIndia and HpEastAsia as shown in the figure 22.



Figure 22: Admixture analysis of MLST sequences from 101 strains using BAPS. The Blue, Red and green colors represent clusters of strains belonging to HpEurope, HpSouthIndia and HpEastAsia.

The strains ascribing to Ladakh region comprised of all three lineages while those from other regions of India were comprised of two major lineages HpEurope and HpSouthAsia. The

existence of even the EastAsian lineage in Ladakh isolates lends support to ancestral human migrations from the east in the region (Majumder and Basu 2015). Our findings are also in accordance to the previous study on *H. pylori* population structure of isolates obtained from Ladakh (Wirth et al. 2004). These findings also illustrate the potential of analyzing *H. pylori* population structure to unravel the complex mysteries of human migration.

Previous study reported that Indian *H. pylori* has European roots (S. M. Devi et al. 2007). Their conclusions were based on the analysis of isolates from Hyderabad and Ladakh. Even in our data isolates from Hyderabad predominantly clustered in HpEurope and this could be the reason for the proposed European ancestry by the authors. Only when we pooled the isolates from southern India and Malaysian-Indians we could observe a second lineage HpSouthIndia as well. Therefore, it is tempting to believe that Indian *H. pylori* population is mainly comprised of two lineages HpEurope and HpSouthIndia and this corroborates very well with previous studies on ancestral human migration in India showing two distinct waves of human migrations witnessed in the region.

To glean further support to our hypothesis, we performed phylogenetic analysis of five *H. pylori* genomes available at the time of the study, together with 39 others from the database to decipher their phylogenetic relationship (figure 23). The phylogenetic tree formed clusters ascribing to different lineages of *H. pylori*. Five *H. pylori* strains from India were found to cluster in two distinct lineages HpSouthIndia and HpEurope. The observation provides a strong support for our conclusions based on MLST sequence data even at the whole genome level.

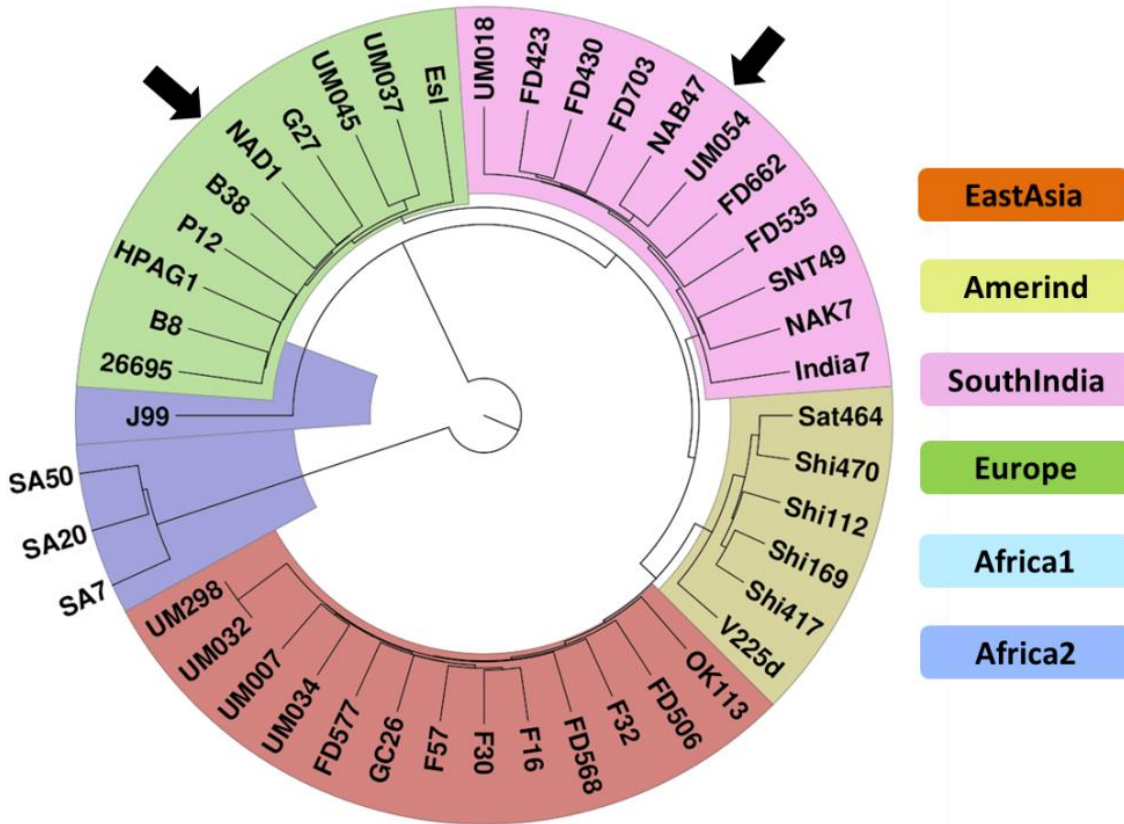


Figure 23: Whole genome based phylogenetic tree generated using Gegenees tool with UPGMA algorithm. The colors represent clusters belonging to a particular lineage. The black arrows indicate the clusters containing *H. pylori* isolates from India.

Virulence potential

Having deciphered the population structure composition of *H. pylori* strains, we were next interested to investigate the virulence potential of those fully sequenced strains. The comparison of gene content revealed a high virulence capability of the strains based on the observation that most of the 58 virulence genes were present in the strains (table 11). The high virulence capability could be one of the factors underpinning high proportion of gastritis and ulcer disease burden observed in India.

Table 11: Status of 58 virulence genes among five sequenced genomes from India. (+) represents the presence while (-) shows absence of a gene.

Genes	NAK7	NAD1	SNT49	NAB47	INDIA7
<i>ureA</i>	+	+	+	+	+
<i>ureB</i>	+	+	+	+	+
<i>ureI</i>	+	+	+	+	+
<i>ureE</i>	+	+	+	+	+
<i>ureF</i>	+	+	+	+	+
<i>ureG</i>	+	+	+	+	+
<i>ureH</i>	+	+	+	+	+
<i>flaA</i>	+	+	+	+	+
<i>flaB</i>	+	+	+	+	+
<i>flgG</i>	+	+	+	+	+
<i>vacA</i>	+	+	-	+	+
<i>cag1</i>	+	+	+	+	+
<i>cag3</i>	+	+	+	+	+
<i>cag4</i>	+	+	+	+	+
<i>cag5</i>	+	+	+	+	+
<i>virB11</i>	+	+	+	+	+
<i>cag6</i>	+	+	+	+	+
<i>cag7</i>	+	+	-	-	+
<i>cag8</i>	+	+	-	-	+
<i>cag9</i>	+	+	+	+	+
<i>cag10</i>	+	+	+	+	+
<i>cag11</i>	+	+	+	+	+
<i>cag12</i>	+	+	+	+	+
<i>cag13</i>	+	+	+	+	+
<i>cag14</i>	+	+	+	+	+
<i>cag15</i>	+	+	+	+	+
<i>cag16</i>	+	+	+	+	+
<i>cag17</i>	+	+	+	+	+
<i>cag18/cagL</i>	+	+	+	+	+
<i>cag19</i>	+	+	+	+	+
<i>cag20</i>	+	+	+	+	+
<i>cag21</i>	+	+	+	+	+
<i>cag22</i>	+	+	+	+	+
<i>fucU</i>	+	+	+	+	+
<i>fucT</i>	+	+	+	+	+
<i>neuA</i>	+	+	+	+	+
<i>neuB</i>	+	+	+	+	+
<i>sabA</i>	+	+	+	+	+
<i>gluP</i>	+	+	+	+	+
<i>gluE</i>	+	+	+	+	+
<i>rfbM</i>	+	+	+	+	+
<i>rfaC</i>	+	+	+	+	+
<i>rfaJ</i>	+	+	+	+	+
<i>rfaJ</i>	+	+	+	+	-

Genes	NAK7	NAD1	SNT49	NAB47	INDIA7
<i>wbpB</i>	+	+	+	+	+
<i>kdtB</i>	+	+	+	+	+
<i>rfbD</i>	+	+	+	+	+
<i>wbcJ</i>	+	+	+	+	+
<i>lpxB</i>	+	+	+	+	+
<i>babA</i>	+	-	+	-	+
<i>babB</i>	+	-	+	+	+
<i>hopZ</i>	+	+	-	+	-
<i>napA</i>	+	+	+	+	+
<i>cag23</i>	+	+	+	+	+
<i>cag24</i>	+	+	+	+	+
<i>cag25</i>	+	+	+	+	+
<i>cagA</i>	+	+	+	+	+

CagA EPIYA motif

CagA gene encoding an “oncoprotein” is one of the most studied virulence factors and has been associated with increased risk of gastric cancer (Prouzet-Mauleon et al. 2005). The type of EPIYA motif present also influences pro-inflammatory capability of the protein. The EastAsian strains mostly harbor A, B and D type motifs and exhibit enhanced pro-inflammatory response as compared to European strains that contain A, B and C type EPIYA motifs (Stein et al. 2013). We carried out *cagA* gene comparison of Indian strains along with other genomes to decipher their EPIYA motif composition. The multiple gene alignment and further gene based phylogeny revealed a European type EPIYA motif in all the strains as shown in the figure 24. The existence of European or less virulent EPIYA motif composition of *cagA* could be the reason for low proportion of gastric cancer cases in India.

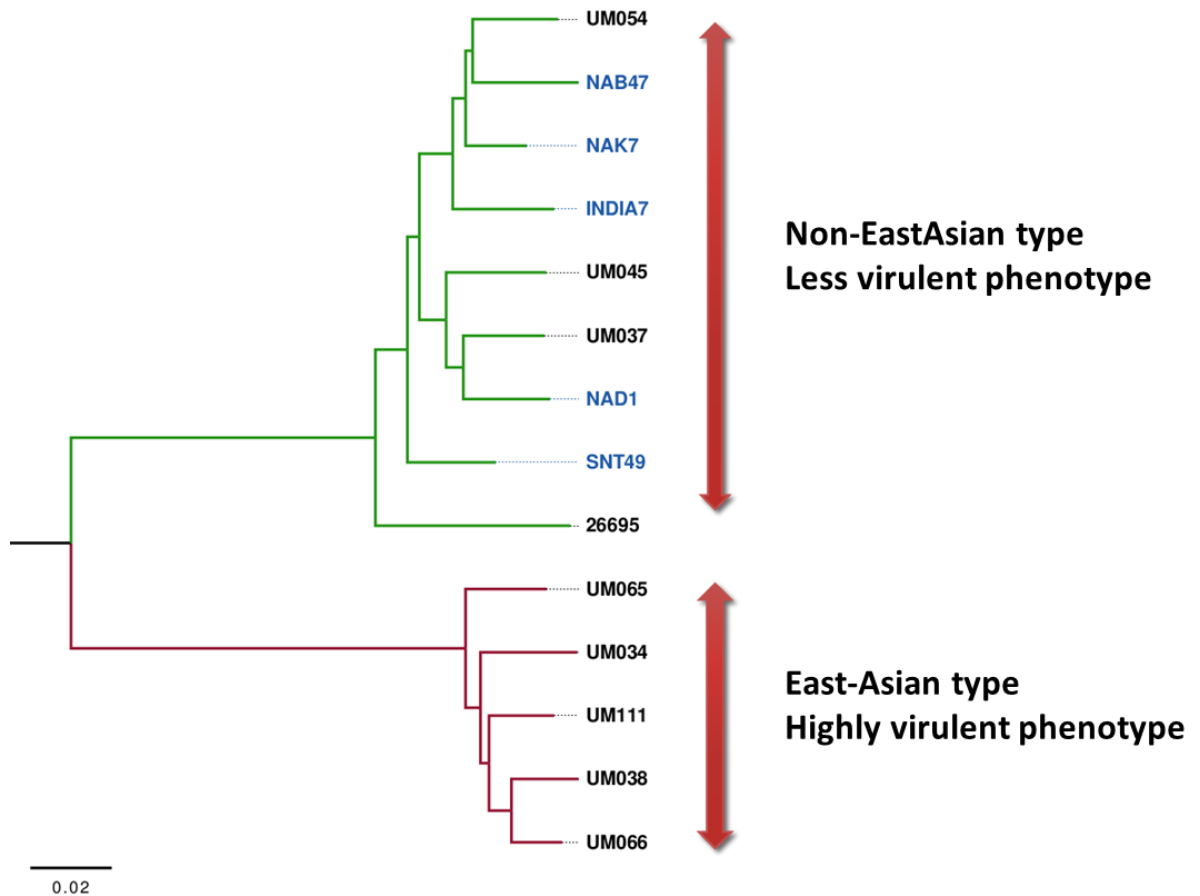


Figure 24: *cagA* gene based phylogenetic tree constructed using MEGA tool with Neighbour-Joining algorithm after 1000 bootstraps. The tree branch confidence values are above 0.75 dividing the genomes into two major clusters, one formed by EastAsian *cagA* and other by European and Indian isolates.

Conclusions

Long coevolution history together with host specificity makes *H. pylori* an ideal organism to investigate ancestral human migrations. Although a high burden of *H. pylori* infection is recorded in India, a minuscule number of attempts were made to understand the population genetic structure of *H. pylori* in the region. One of the studies reported existence of three major lineages based on the analysis of strains limited to Ladakh regions (Wirth et al. 2004). The other study reported European ancestry for the Indian *H. pylori* strains but analysis were majorly confined to Hyderabad region. Therefore, in this study we pooled the samples from north and south regions and aimed to delineate the population composition of *H. pylori* in India.

The BAPS analysis of the MLST sequence alleles of 7 seven housekeeping genes was carried out and further extent of admixture was determined. The strains from the Ladakh region harbored strains from all three lineages HpEurope, HpSouthAsia and HpEastAsia. The analysis was in accordance with the conclusions made by Wirth et al. that Ladakh region holds slightly different *H. pylori* composition which might be due to the complex history of human migration from south and east of India which is also evident from the human haplotype studies. The strains from rest of India consisted of strains mainly belonging to two lineages HpEurope and HpSouthIndia. The strains belonging to other regions revealed bipartite ancestry which also supports the hypothesis of two waves of human migration leading to peopling of India. The phylogenetic analysis of five whole genomes available at the time of the study together with 39 others from the database resonated the findings of BAPS analysis.

Having determined lineage structure of *H. pylori* in India, we also determined the possible reasons for the low occurrence of gastric cancer despite high rate of infection burden. The five Indian strains retained most of the 58 virulence genes describing a high pathogenic potential of

the strains. The analysis of EPIYA motifs structure in CagA oncoprotein revealed presence of western type EPIYA motifs which are less frequently associated with gastric cancer and was in stark contrast to EastAsian strains. These observations provided a glimpse of genetic differences that could underpin the variations observed at phenotypic level.

The study provided a clear understanding of the existing population structure of *H. pylori* in India and provided the putative factors behind the observed disease burden in the country. Addition of more *H. pylori* genome sequences from various regions representing different disease conditions would definitely provide better resolution and increased understanding of *H. pylori* related pathogenesis in the region.

CHAPTER VI

Summary and outlook

H. pylori is a major risk factor for the development of a myriad of gastric diseases such as gastritis, duodenal ulcer and gastric carcinoma. The association can also be realized from the fact that it has been recognized by WHO as a class I carcinogen. *H. pylori* has evolved variety of mechanisms to thrive in the harsh environment of the stomach. The bacterium possesses urease enzyme to dwell in the acidic environment of the stomach. The outer membrane proteins allow the bacterium to adhere and colonize the epithelial surface. The flagellar apparatus aids to quickly swim across the highly viscous mucus layer to reach a more favorable niche at the epithelial lining. Once there, it activates its virulence machinery triggering inflammation that transcends to the development of disease. The comparison of strains from different geographical regions has revealed a high genetic diversity. In addition, there is a strong geographical impact on the population structure as revealed by the observation of distinct genotypes of *H. pylori* in various regions of the world. The dissimilarities among the genotypes also corroborates with the severity of associated disease revealing different disease potential of these genotypes. For example, East Asian countries where HpEastAsia genotype predominates record elevated proportion of gastric cancer. Therefore, it is necessary to understand the the extent to which these genotypes differ in order to comprehend the discrepancies at phenotypic level.

The emergence of high throughput and cost effective technologies has equipped individual labs with whole genome sequencing capability. This has also opened avenues for many new research areas such as metagenomics widening the applicability of Next Generation Sequencing technologies. The whole genome not only provides a holistic insight of the organism's characteristics, their compositional differences can be understood at a resolution unmatched by any of the other typing methods. This allows us to monitor pathogen evolution and its epidemiology and narrow down on key targets for discovery. The approach can also be used to

track transmission and therefore help come up with better and effective control measures. Given this, we adopted a whole genome approach towards understanding *H. pylori* evolution in a single host and host communities.

The primary step forward was to setup a methodology to sequence and assemble complete genomes. We preferred Illumina platform over the others because it provided paired end reads, high coverage and being cost effective. The reads obtained after sequencing require a lot of downstream processing to reconstruct the genome. The reads were quality filtered and assembled using velvet *de novo* assembler. The contigs obtained can usually be sorted to formulate the chromosome based on two key principles. The reference based approach relies solely on the closely related reference genome to arrange the contigs into a chromosome. Though the approach is most appropriate for monomorphic genomes, it is not recommended in case of highly recombining organisms as one can lose out strains specific information. The *de novo* approach does not take into account any reference genome and relies on the paired end information generated during sequencing. While the latter approach helps to retain maximum strain specific regions, formulating a near complete genome still remains problematic because of repeat regions and gaps in the genome. Therefore, we sought to utilize the benefits of both the approaches towards obtaining better assemblies. The velvet assembled contigs were first sorted after a chosen reference genome using BLASTn. These set of reference sorted contigs were then utilized as reference to map the paired reads. The alignment information of the paired reads mapped at the start and end of the contigs were extracted to evaluate links between the contigs. This was carried out manually for each contig using the visualization tool Tablet (Milne et al. 2010). The contigs were reorganized in accordance with the paired end information while the contigs where read pair information was inconclusive were arranged as per the reference

genome. More than twenty bacterial genomes were assembled using this methodology that seemed to provide better genome assemblies.

After having learnt to sequence and assemble whole genomes, we were interested to undertake the whole genome based comparative analyses to gain insights into the evolutionary mechanisms employed by *H. pylori*. First analysis dealt with elucidating the variations among chronological isolates from a single patient. *H. pylori* infection is thought to be usually acquired in the childhood and can persist for decades. The variations that a single strain undergoes over a period of time in a single stomach therefore could render insights into the adaptive mechanisms. The chronological isolated shared a high similarity among themselves ruling out the possibility of extraneous infection by different strains. This was further confirmed by co-clustering of the isolates in the whole genome based phylogenetic tree. The genome comparison depicted a core genome content comprising of ~95% of the genes which was in line with the phylogenetic tree. Earlier, studies on sequential *H. pylori* isolates from patients have revealed a high recombination and mutation rate that seemed to subside as the time between the strains increased. Our analysis showed that much of the variation is enriched in the OMPs together with minor events of recombination. The strains were similar in terms of their virulence potential as revealed by homogeneous content of virulence genes. The results were also in accordance with the recent article describing a burst of mutations during the colonization of *H. pylori* (Linz et al. 2014). The variations in OMPs could enable the bacterium to generate an array of variants essential to mediate host pathogen interaction and effectively evade detection. The study revealed alterations in the OMPs could be the major factor that assists *H. pylori* to adapt and persist in a single gut for long time.

Another study was carried out to understand the population genetic structure of *H. pylori* in a mixed population setup such as Malaysia (Kumar et al. 2014). The comparative analysis also aimed at finding the cues that underpin the differential prevalence of *H. pylori* related infections among the three ethnic groups that constitute the Malaysian population. Among the three major ethnic groups, Malaysian Indians record the highest infection burden although record low proportion of gastric cancer when compared to the Malaysian Chinese population. The comparison involved 27 genomes comprising of isolates from all three ethnic backgrounds. The isolates belonging to the same ethnic group shared a high similarity among themselves compared to the strains of other ethnic groups. The phylogenetic analysis also revealed a similar trend as the strains clustered as per the ethnic groups they belong to. The core genome content of the 27 strains revealed a comparable virulence potential as we failed to detect any group specific virulence genes. This could be because of the long coevolution causing a homogenous spread of these virulence coordinates by virtue of frequent recombination. A significant proportion of the core genes were observed to be differentially evolving and therefore highlights the host specific adaptations. These genes could underpin the host specific nature of *H. pylori* and their investigation would further our understanding of *H. pylori* adaptive mechanisms. Another important observation was with regards to RM systems and their role in maintaining the host specific barrier. EastAsian strains were found to harbor increased proportion of RM genes when compared to non-EastAsian strains isolated from Malaysian –Indians and Malays. The observation provided strong evidence for the close clustering of all EastAsian strains in phylogenetic tree as RM genes can provide resistance to genetic exchange from other lineages. The comparison of the RM gene pool also revealed a high genetic polymorphism with strain specific RM genes. Recent reports on additional functions methyl transferases in particular have suggested their role in the epigenetic regulation of gene expression. Therefore, further detailed

investigation of their role would be necessary to understand the pathogen dynamics and regulatory networks in *H. pylori*.

In the final chapter, we sought to determine the population genetic structure of *H. pylori* in India. In spite of high infection burden, a minuscule proportion of gastric cancer incidences are recorded in the region. Therefore, it is necessary to determine the genotype structure in order to better understand the epidemiology of *H. pylori* associated diseases and accurately calculate the health risks it poses. Previous reports based on the MLST analysis have failed to provide a definitive genotype structure and this might be because of the samples representing few geographical regions. We sequenced and assembled three isolates from different regions of India. Since the number of whole genome sequences was not sufficient, we sought to analyze the MLST sequences from previous studies along with others reported in the PubMLST database. The BAPS analysis clearly revealed a bipartite population structure comprising of two major genotypes HpEurope and HpSouthIndia. The results were also in line with the human migration studies that showed India witnessed two distinct waves of human migration that populated the region. The finding also corroborated with the intimate association of *H. pylori* with its human host and provided strong support for the use of *H. pylori* to determine human ancestry. The analysis of the strains revealed conservation of a majority of reported virulence markers justifying the increased proportion of *H. pylori* related gastric diseases. The strains also harbored western type CagA EPIYA AB-C motifs that are less frequently associated with gastric cancer as compared to AB-D type observed in EastAsian strains (Yamaoka Y 2010). The comparative analysis therefore provided some cues about the existing enigma of having less cases of gastric cancer versus high infection burden. The genotype structure is quite distinct from EastAsian strains and resembles more with western lineages and even the CagA EPIYA motif resembles

the western type, both these observations were strong indicators for the observed epidemiology of gastric cancer in the region.

In conclusion, the work presented in the thesis provided a better sequencing and assembly pipeline for bacterial genomes. The work also revealed importance of whole genome sequencing and comparative genomics in understanding the evolution and epidemiology of bacterial pathogens. The comparative study of the chronological isolates revealed that *H. pylori* exercises constant outer membrane fine-tuning to adapt and persist in the stomach, over time. The analysis of isolates from a mixed population setup provided insights into the evolution and host specific adaptations leading to differential infection outcomes. The study also laid emphasis on the possible role of RM genes in the epigenetic regulation and consequent impact on the virulence of this gastric pathogen. The work would also serve as a guide to understand the population genetic structure of bacterial pathogens. Further, comparison of RM genes at the global level together with functional investigations would be very important to understand their regulatory potential and impact on *H. pylori*. In future, it would be important to understand the emergence and evolution of virulent genotypes such as HpEastAsia to gain deeper insights into biology of this gastric pathogen.

References

Achtman, M., et al. (1999), 'Recombination and clonal groupings within *Helicobacter pylori* from different geographical regions', *Mol Microbiol*, 32 (3), 459-70.

Agren, J., et al. (2012), 'Gegenes: fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups', *PLoS One*, 7 (6), e39107.

Ahmed, N., Tenguria, S., and Nandanwar, N. (2009), 'Helicobacter pylori--a seasoned pathogen by any other name', *Gut Pathog*, 1, 24.

Ahmed, N., et al. (2008), 'Genomic fluidity and pathogenic bacteria: applications in diagnostics, epidemiology and intervention', *Nat Rev Microbiol*, 6 (5), 387-94.

Ahmed, N., et al. (2013), 'Helicobacter pylori in 2013: multiplying genomes, emerging insights', *Helicobacter*, 18 Suppl 1, 1-4.

Alfizah, H., et al. (2012), 'Association of Malaysian *Helicobacter pylori* virulence polymorphisms with severity of gastritis and patients' ethnicity', *Helicobacter*, 17 (5), 340-9.

Atherton, J. C., et al. (1995), 'Mosaicism in vacuolating cytotoxin alleles of *Helicobacter pylori*. Association of specific *vacA* types with cytotoxin production and peptic ulceration', *J Biol Chem*, 270 (30), 17771-7.

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

Bastos, J., et al. (2013), 'Sociodemographic determinants of prevalence and incidence of *Helicobacter pylori* infection in Portuguese adults', *Helicobacter*, 18 (6), 413-22.

Bentley, D. R., et al. (2008), 'Accurate whole human genome sequencing using reversible terminator chemistry', *Nature*, 456 (7218), 53-9.

Benajah, D. A., et al. (2013), 'Prevalence of *Helicobacter pylori* and its recurrence after successful eradication in a developing nation (Morocco)', *Clin Res Hepatol Gastroenterol*, 37 (5), 519-26.

Bjorkholm, B., et al. (2001), 'Mutation frequency and biological cost of antibiotic resistance in *Helicobacter pylori*', *Proc Natl Acad Sci U S A*, 98 (25), 14607-12.

Boetzer, M., et al. (2011), 'Scaffolding pre-assembled contigs using SSPACE', *Bioinformatics*, 27 (4), 578-9.

Brandt, S., et al. (2005), 'NF-kappaB activation and potentiation of proinflammatory responses by the *Helicobacter pylori* CagA protein', *Proc Natl Acad Sci U S A*, 102 (26), 9300-5.

Breurec, S., et al. (2011), 'Evolutionary history of *Helicobacter pylori* sequences reflect past human migrations in Southeast Asia', *PLoS One*, 6 (7), e22058.

Buermans, H. P. and den Dunnen, J. T. (2014), 'Next generation sequencing technology: Advances and applications', *Biochim Biophys Acta*, 1842 (10), 1932-41.

Censini, S., et al. (1996), 'cag, a pathogenicity island of *Helicobacter pylori*, encodes type I-specific and disease-associated virulence factors', *Proc Natl Acad Sci U S A*, 93 (25), 14648-53.

Chen, L., et al. (2005), 'VFDB: a reference database for bacterial virulence factors', *Nucleic Acids Res*, 33 (Database issue), D325-8.

Corander, J., et al. (2004), 'BAPS 2: enhanced possibilities for the analysis of genetic population structure', *Bioinformatics*, 20 (15), 2363-9.

Covacci, A., et al. (1999), '*Helicobacter pylori* virulence and genetic geography', *Science*, 284 (5418), 1328-33.

den Hollander, W. J., et al. (2013), 'Ethnicity is a strong predictor for *Helicobacter pylori* infection in young women in a multi-ethnic European city', *J Gastroenterol Hepatol*, 28 (11), 1705-11.

Devi, S., et al. (2014), 'Helicobacter pylori antigen HP0986 (TieA) interacts with cultured gastric epithelial cells and induces IL8 secretion via NF-kappaB mediated pathway', *Helicobacter*, 19 (1), 26-36.

Devi, S. M., et al. (2007), 'Ancestral European roots of Helicobacter pylori in India', *BMC Genomics*, 8, 184.

Didelot, X., et al. (2013), 'Genomic evolution and transmission of Helicobacter pylori in two South African families', *Proc Natl Acad Sci U S A*, 110 (34), 13880-5.

Dorer, M. S., Fero, J., and Salama, N. R. (2010), 'DNA damage triggers genetic exchange in Helicobacter pylori', *PLoS Pathog*, 6 (7), e1001026.

Edgar, R. C. (2010), 'Search and clustering orders of magnitude faster than BLAST', *Bioinformatics*, 26 (19), 2460-1.

Erzin, Y., et al. (2006), 'Analysis of serum antibody profile against H pylori VacA and CagA antigens in Turkish patients with duodenal ulcer', *World J Gastroenterol*, 12 (42), 6869-73.

Eusebi, L. H., Zagari, R. M., and Bazzoli, F. (2014), 'Epidemiology of Helicobacter pylori infection', *Helicobacter*, 19 Suppl 1, 1-5.

Fallone, C. A., et al. (1997), 'Hematoxylin and eosin staining of gastric tissue for the detection of *Helicobacter pylori*', *Helicobacter*, 2 (1), 32-5.

Falush, D., et al. (2003), 'Traces of human migrations in *Helicobacter pylori* populations', *Science*, 299 (5612), 1582-5.

Furuta, Y., et al. (2014), 'Methylome diversification through changes in DNA methyltransferase sequence specificity', *PLoS Genet*, 10 (4), e1004272.

Garza-Gonzalez, E., et al. (2014), 'A review of *Helicobacter pylori* diagnosis, treatment, and methods to detect eradication', *World J Gastroenterol*, 20 (6), 1438-49.

Ghose, C., et al. (2002), 'East Asian genotypes of *Helicobacter pylori* strains in Amerindians provide evidence for its ancient human carriage', *Proc Natl Acad Sci U S A*, 99 (23), 15107-11.

Goddard, A. F. and Logan, R. P. (1997), 'Review article: urea breath tests for detecting *Helicobacter pylori*', *Aliment Pharmacol Ther*, 11 (4), 641-9.

Goh, K. L., et al. (2011), 'Epidemiology of *Helicobacter pylori* infection and public health implications', *Helicobacter*, 16 Suppl 1, 1-9.

Graham, D. Y. (2015), '*Helicobacter pylori* update: gastric cancer, reliable therapy, and possible benefits', *Gastroenterology*, 148 (4), 719-31 e3.

Harris, S. R., et al. (2013), 'Whole-genome sequencing for analysis of an outbreak of methicillin-resistant *Staphylococcus aureus*: a descriptive study', *Lancet Infect Dis*, 13 (2), 130-6.

Harris, S. R., et al. (2010), 'Evolution of MRSA during hospital transmission and intercontinental spread', *Science*, 327 (5964), 469-74.

Hatakeyama, M. (2004), 'Oncogenic mechanisms of the *Helicobacter pylori* CagA protein', *Nat Rev Cancer*, 4 (9), 688-94.

Hatakeyama, M (2009), '*Helicobacter pylori* and gastric carcinogenesis', *J Gastroenterol*, 44 (4), 239-48.

Hatin, W. I., et al. (2011), 'Population genetic structure of peninsular Malaysia Malay sub-ethnic groups', *PLoS One*, 6 (4), e18312.

Higashi, H., et al. (2002), 'SHP-2 tyrosine phosphatase as an intracellular target of *Helicobacter pylori* CagA protein', *Science*, 295 (5555), 683-6.

Hill, C., et al. (2006), 'Phylogeography and ethnogenesis of aboriginal Southeast Asians', *Mol Biol Evol*, 23 (12), 2480-91.

Hirschl, A. M. and Makristathis, A. (2007), 'Methods to detect *Helicobacter pylori*: from culture to molecular biology', *Helicobacter*, 12 Suppl 2, 6-11.

Holden, M. T., et al. (2013), 'A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic', *Genome Res*, 23 (4), 653-64.

Huson, D. H. (1998), 'SplitsTree: analyzing and visualizing evolutionary data', *Bioinformatics*, 14 (1), 68-73.

Ilver, D., et al. (1998), '*Helicobacter pylori* adhesin binding fucosylated histo-blood group antigens revealed by retagging', *Science*, 279 (5349), 373-7.

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 (25), 14625-30.

Kennemann, L., et al. (2011), '*Helicobacter pylori* genome evolution during human infection', *Proc Natl Acad Sci U S A*, 108 (12), 5033-8.

Kersulyte, D., et al. (2009), '*Helicobacter Pylori*'s plasticity zones are novel transposable elements', *PLoS One*, 4 (9), e6859.

Khalifa, M. M., Sharaf, R. R., and Aziz, R. K. (2010), '*Helicobacter pylori*: a poor man's gut pathogen?', *Gut Pathog*, 2 (1), 2.

Khosravi, Y., et al. (2013), 'Comparing the genomes of *Helicobacter pylori* clinical strain UM032 and Mice-adapted derivatives', *Gut Pathog*, 5, 25.

Koboldt, D. C., et al. (2013), 'The next-generation sequencing revolution and its impact on genomics', *Cell*, 155 (1), 27-38.

Krebes, J., et al. (2014), 'The complex methylome of the human gastric pathogen *Helicobacter pylori*', *Nucleic Acids Res*, 42 (4), 2415-32.

Kuipers, E. J., et al. (2000), 'Quasispecies development of *Helicobacter pylori* observed in paired isolates obtained years apart from the same host', *J Infect Dis*, 181 (1), 273-82.

Kumar, N., et al. (2015), 'Comparative genomic analysis of *Helicobacter pylori* from Malaysia identifies three distinct lineages suggestive of differential evolution', *Nucleic Acids Res*, 43 (1), 324-35.

Kumar, S., Tamura, K., and Nei, M. (1994), 'MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers', *Comput Appl Biosci*, 10 (2), 189-91.

Kusters, J. G., van Vliet, A. H., and Kuipers, E. J. (2006), 'Pathogenesis of *Helicobacter pylori* infection', *Clin Microbiol Rev*, 19 (3), 449-90.

Lagesen, K., et al. (2007), 'RNAmmer: consistent and rapid annotation of ribosomal RNA genes', *Nucleic Acids Res*, 35 (9), 3100-8.

Langille, M. G. and Brinkman, F. S. (2009), 'IslandViewer: an integrated interface for computational identification and visualization of genomic islands', *Bioinformatics*, 25 (5), 664-5.

Lee, J. Y., et al. (2014), 'Factors affecting first-line triple therapy of *Helicobacter pylori* including CYP2C19 genotype and antibiotic resistance', *Dig Dis Sci*, 59 (6), 1235-43.

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 (5), 338-46.

Leunk, R. D., et al. (1988), 'Cytotoxic activity in broth-culture filtrates of *Campylobacter pylori*', *J Med Microbiol*, 26 (2), 93-9.

Li, H. and Durbin, R. (2009), 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*, 25 (14), 1754-60.

Li, H., et al. (2009), 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25 (16), 2078-9.

- Li, L., Stoeckert, C. J., Jr., and Roos, D. S. (2003), 'OrthoMCL: identification of ortholog groups for eukaryotic genomes', *Genome Res*, 13 (9), 2178-89.
- Lim, S. H., et al. (2013), 'Prevalence and risk factors of *Helicobacter pylori* infection in Korea: nationwide multicenter study over 13 years', *BMC Gastroenterol*, 13, 104.
- Lin, L. F., et al. (2001), 'Comparative genomics of the restriction-modification systems in *Helicobacter pylori*', *Proc Natl Acad Sci U S A*, 98 (5), 2740-5.
- Linz, B., et al. (2014), 'A mutation burst during the acute phase of *Helicobacter pylori* infection in humans and rhesus macaques', *Nat Commun*, 5, 4165.
- Linz, B., et al. (2007), 'An African origin for the intimate association between humans and *Helicobacter pylori*', *Nature*, 445 (7130), 915-8.
- Loewenstein, Y., et al. (2008), 'Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space', *Bioinformatics*, 24 (13), i41-9.
- Lopez-Vidal, Y., et al. (2008), 'High diversity of *vacA* and *cagA* *Helicobacter pylori* genotypes in patients with and without gastric cancer', *PLoS One*, 3 (12), e3849.
- Lu, H., et al. (2005), 'Duodenal ulcer promoting gene of *Helicobacter pylori*', *Gastroenterology*, 128 (4), 833-48.

Lu, W., et al. (2014), 'Comparative analysis of the full genome of *Helicobacter pylori* isolate Sahul64 identifies genes of high divergence', *J Bacteriol*, 196 (5), 1073-83.

Maiden, M. C., et al. (1998), 'Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms', *Proc Natl Acad Sci U S A*, 95 (6), 3140-5.

Majumder, P. P. and Basu, A. (2015), 'A genomic view of the peopling and population structure of India', *Cold Spring Harb Perspect Biol*, 7 (4), a008540.

Malfertheiner, P., et al. (2007), 'Current concepts in the management of *Helicobacter pylori* infection: the Maastricht III Consensus Report', *Gut*, 56 (6), 772-81.

Mardis, E. R. (2013), 'Next-generation sequencing platforms', *Annu Rev Anal Chem (Palo Alto Calif)*, 6, 287-303.

Margulies, M., et al. (2005), 'Genome sequencing in microfabricated high-density picolitre reactors', *Nature*, 437 (7057), 376-80.

Marshall, B. J. and Warren, J. R. (1984), 'Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration', *Lancet*, 1 (8390), 1311-5.

Marshall, B. J., et al. (1985), 'Attempt to fulfil Koch's postulates for pyloric *Campylobacter*', *Med J Aust*, 142 (8), 436-9.

Mathewos, B., Moges, B., and Dagneu, M. (2013), 'Seroprevalence and trend of *Helicobacter pylori* infection in Gondar University Hospital among dyspeptic patients, Gondar, North West Ethiopia', *BMC Res Notes*, 6, 346.

Megraud, F. (2013), 'Current recommendations for *Helicobacter pylori* therapies in a world of evolving resistance', *Gut Microbes*, 4 (6), 541-8.

Milne, I., et al. (2010), 'Tablet--next generation sequence assembly visualization', *Bioinformatics*, 26 (3), 401-2.

Montecucco, C. and Rappuoli, R. (2001), 'Living dangerously: how *Helicobacter pylori* survives in the human stomach', *Nat Rev Mol Cell Biol*, 2 (6), 457-66.

Moodley, Y., et al. (2012), 'Age of the association between *Helicobacter pylori* and man', *PLoS Pathog*, 8 (5), e1002693.

Nadalin, F., Vezzi, F., and Policriti, A. (2012), 'GapFiller: a de novo assembly approach to fill the gap within paired reads', *BMC Bioinformatics*, 13 Suppl 14, S8.

Natale, D. A., et al. (2000), 'Using the COG database to improve gene recognition in complete genomes', *Genetica*, 108 (1), 9-17.

Nishizawa, T., et al. (2013), 'Delay of second-line eradication therapy for *Helicobacter pylori* can increase eradication failure', *J Gastroenterol Hepatol*, 28 (10), 1608-10.

O'Connor, A., et al. (2014), 'Treatment of *Helicobacter pylori* infection 2014', *Helicobacter*, 19 Suppl 1, 38-45.

Occhialini, A., et al. (2000), 'Distribution of open reading frames of plasticity region of strain J99 in *Helicobacter pylori* strains isolated from gastric carcinoma and gastritis patients in Costa Rica', *Infect Immun*, 68 (11), 6240-9.

Odenbreit, S., et al. (2000), 'Translocation of *Helicobacter pylori* CagA into gastric epithelial cells by type IV secretion', *Science*, 287 (5457), 1497-500.

Odenbreit, S., et al. (2009), 'Outer membrane protein expression profile in *Helicobacter pylori* clinical isolates', *Infect Immun*, 77 (9), 3782-90.

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 (26), 9999-10004.

Ohnishi, N., et al. (2008), 'Transgenic expression of *Helicobacter pylori* CagA induces gastrointestinal and hematopoietic neoplasms in mouse', *Proc Natl Acad Sci U S A*, 105 (3), 1003-8.

Parkhill, J. (2013), 'What has high-throughput sequencing ever done for us?', *Nat Rev Microbiol*, 11 (10), 664-5.

Patel, R. K. and Jain, M. (2012), 'NGS QC Toolkit: a toolkit for quality control of next generation sequencing data', *PLoS One*, 7 (2), e30619.

Peek, R. M., Jr. and Crabtree, J. E. (2006), 'Helicobacter infection and gastric neoplasia', *J Pathol*, 208 (2), 233-48.

Peek, R. M., Jr., et al. (1998), 'Adherence to gastric epithelial cells induces expression of a Helicobacter pylori gene, iceA, that is associated with clinical outcome', *Proc Assoc Am Physicians*, 110 (6), 531-44.

Perez-Perez, G. I., Rothenbacher, D., and Brenner, H. (2004), 'Epidemiology of Helicobacter pylori infection', *Helicobacter*, 9 Suppl 1, 1-6.

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 (8), 4237-41.

Reich, D., et al. (2009), 'Reconstructing Indian population history', *Nature*, 461 (7263), 489-94.

Rimbara, E., Sasatsu, M., and Graham, D. Y. (2013), 'PCR detection of *Helicobacter pylori* in clinical samples', *Methods Mol Biol*, 943, 279-87.

Roesler, B. M., Rabelo-Goncalves, E. M., and Zeitune, J. M. (2014), 'Virulence Factors of *Helicobacter pylori*: A Review', *Clin Med Insights Gastroenterol*, 7, 9-17.

Rohmer, L., Hocquet, D., and Miller, S. I. (2011), 'Are pathogenic bacteria just looking for food? Metabolism and microbial pathogenesis', *Trends Microbiol*, 19 (7), 341-8.

Ronaghi, M. (2001), 'Pyrosequencing sheds light on DNA sequencing', *Genome Res*, 11 (1), 3-11.

Rutherford, K., et al. (2000), 'Artemis: sequence visualization and annotation', *Bioinformatics*, 16 (10), 944-5.

Sasaki, H., et al. (2013), 'Ten-year trend of the cumulative *Helicobacter pylori* eradication rate for the 'Japanese eradication strategy'', *Digestion*, 88 (4), 272-8.

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 (Web Server issue), W686-9.

Schmidt, H. M., et al. (2009), 'The prevalence of the duodenal ulcer promoting gene (dupA) in *Helicobacter pylori* isolates varies by ethnic group and is not universally associated with disease development: a case-control study', *Gut Pathog*, 1 (1), 5.

Schuster, S. C. (2008), 'Next-generation sequencing transforms today's biology', *Nat Methods*, 5 (1), 16-8.

Shirin, H., et al. (2005), 'Eradication of *Helicobacter pylori* can be accurately confirmed 14 days after termination of triple therapy using a high-dose citric acid-based ¹³C urea breath test', *Digestion*, 71 (4), 208-12.

Sodhi, J. S., et al. (2013), 'Prevalence of *Helicobacter pylori* infection and the effect of its eradication on symptoms of functional dyspepsia in Kashmir, India', *J Gastroenterol Hepatol*, 28 (5), 808-13.

Stein, M., et al. (2013), 'Helicobacter pylori CagA: From Pathogenic Mechanisms to Its Use as an Anti-Cancer Vaccine', *Front Immunol*, 4, 328.

Suzuki, R., Shiota, S., and Yamaoka, Y. (2012), 'Molecular epidemiology, population genetics, and pathogenic role of *Helicobacter pylori*', *Infect Genet Evol*, 12 (2), 203-13.

Tatusov, R. L., et al. (2000), 'The COG database: a tool for genome-scale analysis of protein functions and evolution', *Nucleic Acids Res*, 28 (1), 33-6.

Tay, C. Y., et al. (2009), 'Population structure of *Helicobacter pylori* among ethnic groups in Malaysia: recent acquisition of the bacterium by the Malay population', *BMC Microbiol*, 9, 126.

Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2002), 'Multiple sequence alignment using ClustalW and ClustalX', *Curr Protoc Bioinformatics*, Chapter 2, Unit 2 3.

Toft, C. and Andersson, S. G. (2010), 'Evolutionary microbial genomics: insights into bacterial host adaptation', *Nat Rev Genet*, 11 (7), 465-75.

Tomb, J. F., et al. (1997), 'The complete genome sequence of the gastric pathogen *Helicobacter pylori*', *Nature*, 388 (6642), 539-47.

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.

van Dijk, E. L., et al. (2014), 'Ten years of next-generation sequencing technology', *Trends Genet*, 30 (9), 418-26.

van El, C. G., et al. (2013), 'Whole-genome sequencing in health care. Recommendations of the European Society of Human Genetics', *Eur J Hum Genet*, 21 Suppl 1, S1-5.

Vasu, K. and Nagaraja, V. (2013), 'Diverse functions of restriction-modification systems in addition to cellular defense', *Microbiol Mol Biol Rev*, 77 (1), 53-72.

Vilaichone, R. K., et al. (2013), 'Extremely high prevalence of *Helicobacter pylori* infection in Bhutan', *World J Gastroenterol*, 19 (18), 2806-10.

Walker, T. M., et al. (2013), 'Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study', *Lancet Infect Dis*, 13 (2), 137-46.

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 (14), 4746-51.

Wroblewski, L. E., Peek, R. M., Jr., and Wilson, K. T. (2010), '*Helicobacter pylori* and gastric cancer: factors that modulate disease risk', *Clin Microbiol Rev*, 23 (4), 713-39.

Yamaoka, Y. (2008), 'Roles of the plasticity regions of *Helicobacter pylori* in gastroduodenal pathogenesis', *J Med Microbiol*, 57 (Pt 5), 545-53.

Yamaoka, Y. (2010), 'Mechanisms of disease: *Helicobacter pylori* virulence factors', *Nat Rev Gastroenterol Hepatol*, 7 (11), 629-41.

Zerbino, D. R. and Birney, E. (2008), 'Velvet: algorithms for de novo short read assembly using de Bruijn graphs', *Genome Res*, 18 (5), 821-9.

Zhou, Y., et al. (2011), 'PHAST: a fast phage search tool', *Nucleic Acids Res*, 39 (Web Server issue), W347-52.

Zhu, Y., et al. (2014), 'Risk Factors and Prevalence of Helicobacter pylori Infection in Persistent High Incidence Area of Gastric Carcinoma in Yangzhong City', *Gastroenterol Res Pract*, 2014, 481365.

Publications

Comparative genomic analysis of *Helicobacter pylori* from Malaysia identifies three distinct lineages suggestive of differential evolution

Narender Kumar¹, Vanitha Mariappan², Ramani Baddam¹, Aditya K. Lankapalli¹, Sabiha Shaik¹, Khean-Lee Goh³, Mun Fai Loke², Tim Perkins⁴, Mohammed Benghezal⁴, Seyed E. Hasnain⁵, Jamuna Vadivelu², Barry J. Marshall⁴ and Niyaz Ahmed^{1,6,*}

¹Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, University of Hyderabad, Gachibowli, Hyderabad, 500046, India, ²Department of Medical Microbiology, Faculty of Medicine, University of Malaya, 50603, Kuala Lumpur, Malaysia, ³Department of Medicine, Faculty of Medicine, University of Malaya, 50603, Kuala Lumpur, Malaysia, ⁴School of Pathology and Laboratory Medicine, University of Western Australia, Nedlands 6009, Western Australia, Australia, ⁵Kusuma School of Biological Sciences, Indian Institute of Technology, Hauz Khas, New Delhi, 110016, India and ⁶Institute 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, hpWest-Africa 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% CO₂. 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 (NNNNNCACACTTAATT 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).

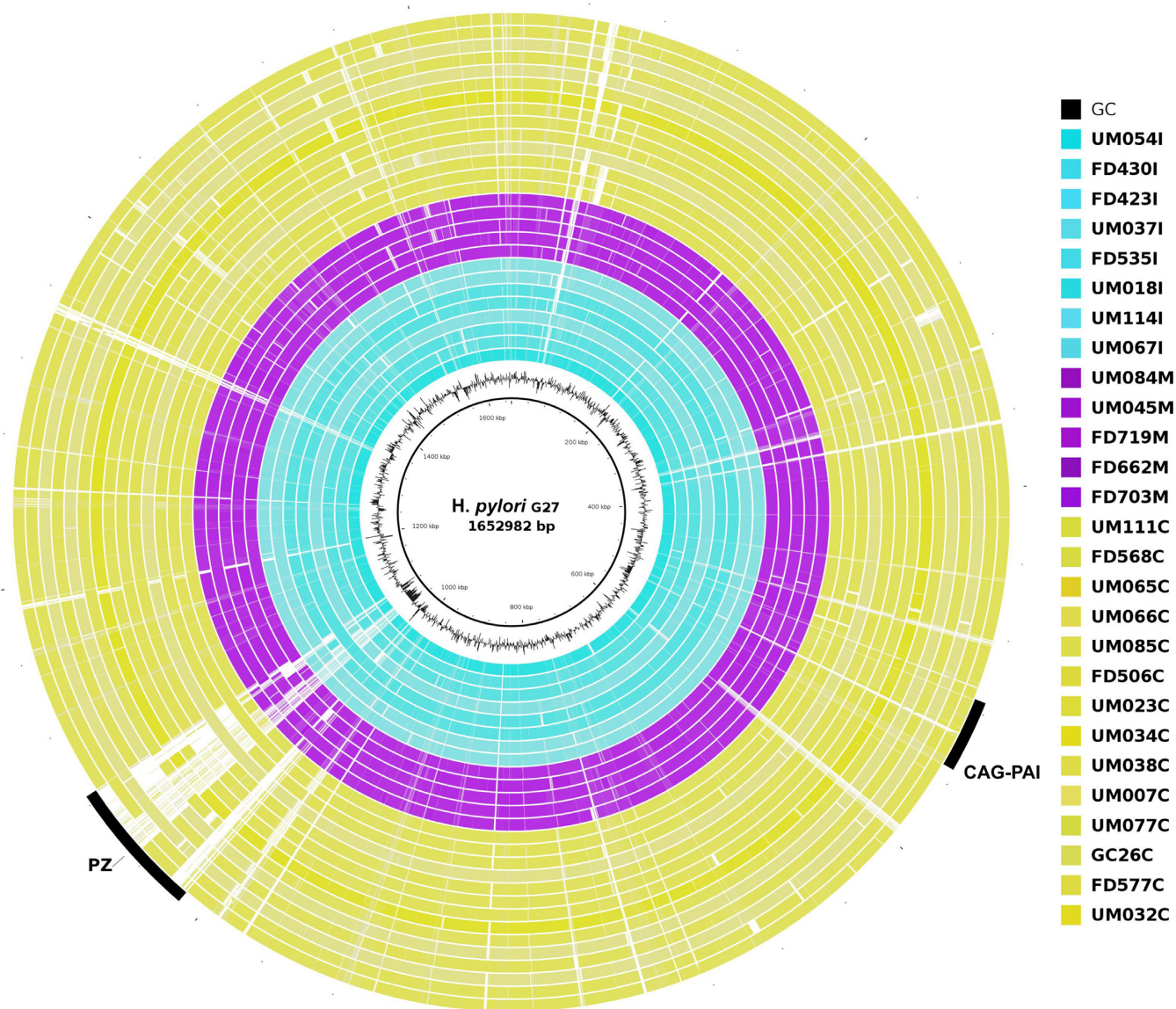


Figure 1. A circular representation of the genomes of Malaysian isolates: the draft genomes of 27 strains were aligned against the genome of reference strain *H. pylori* G27. Each genome is represented by a ring. The yellow rings represent *H. pylori* genomes from Malaysian-Chinese, purple represents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

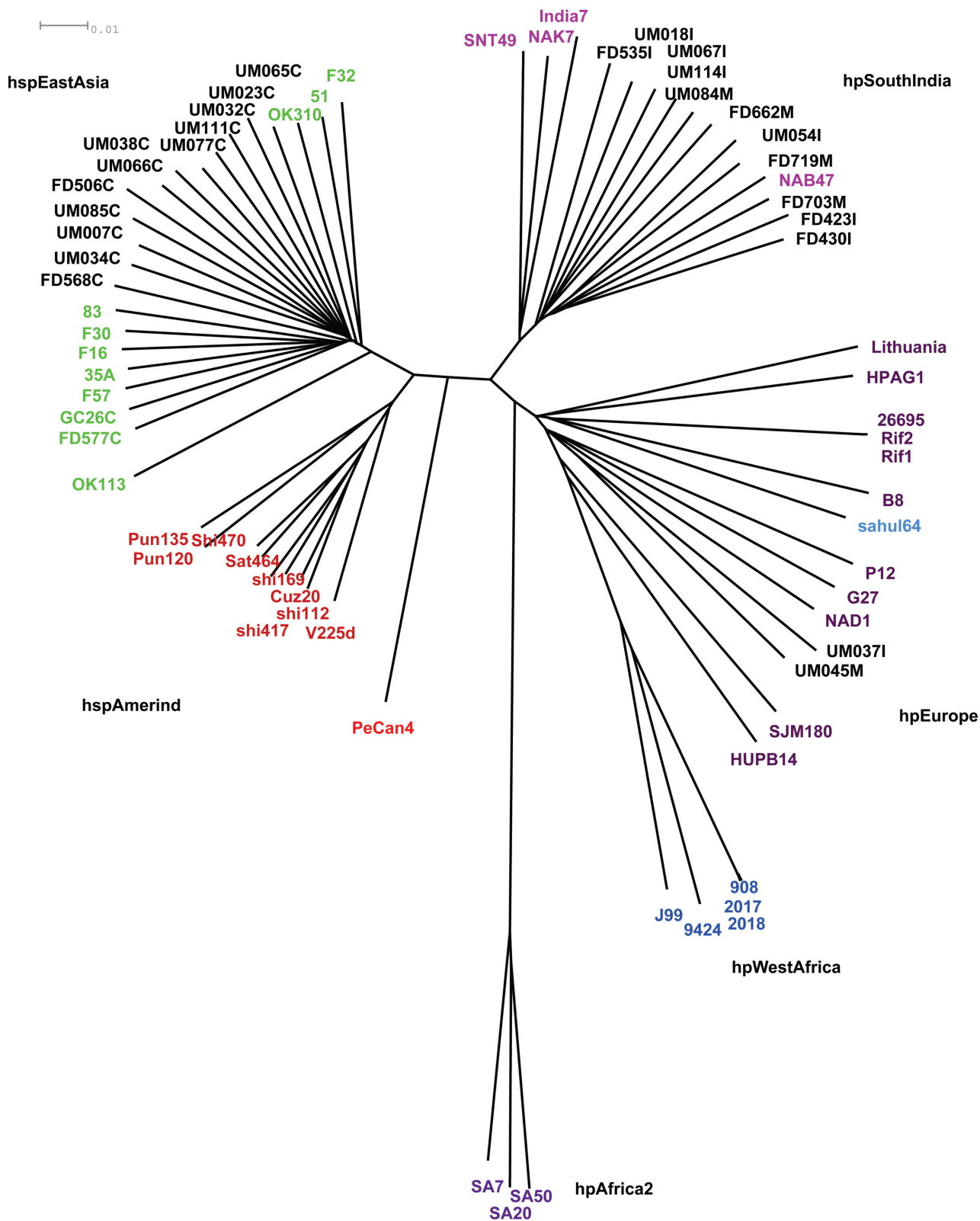
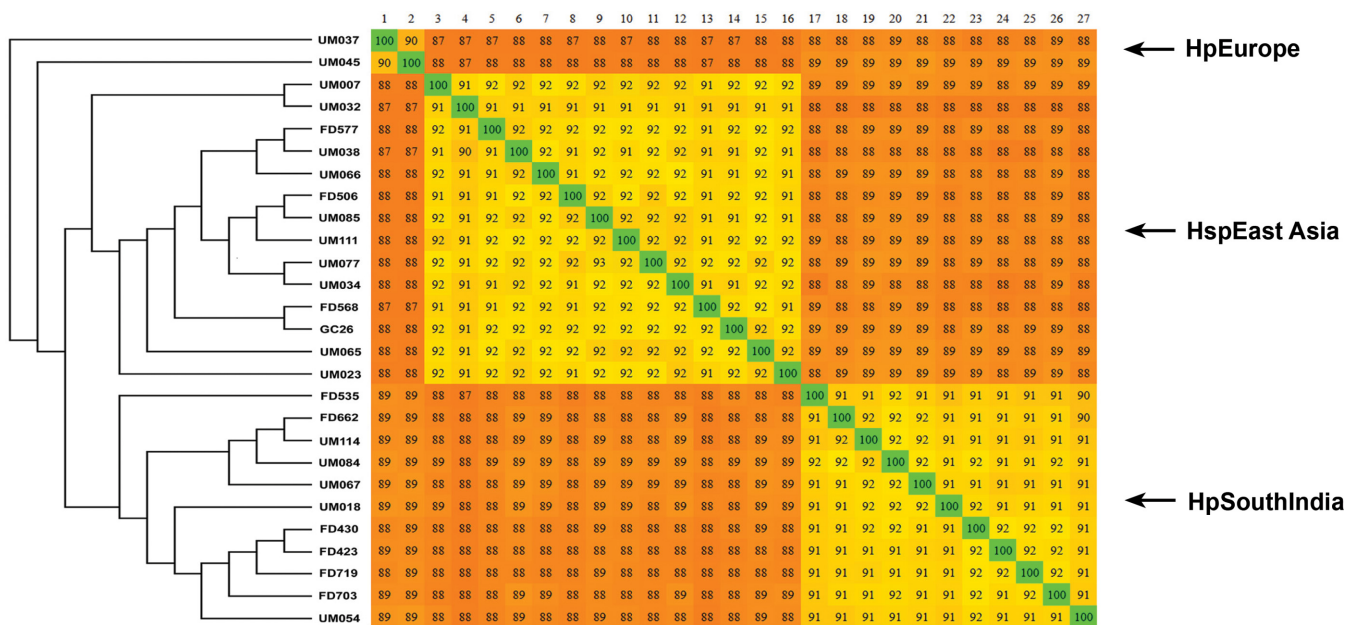


Figure 2. The whole genome phylogenetic analysis: the figure represents a whole genome comparison-based phylogenetic tree of various complete and draft *H. pylori* genomes from different geographical regions. The tree was constructed based on neighbor joining algorithm using SplitsTree. The Malaysian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

revealed lineage-specific patterns. CagA protein encoded by the *cagPAI* 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).

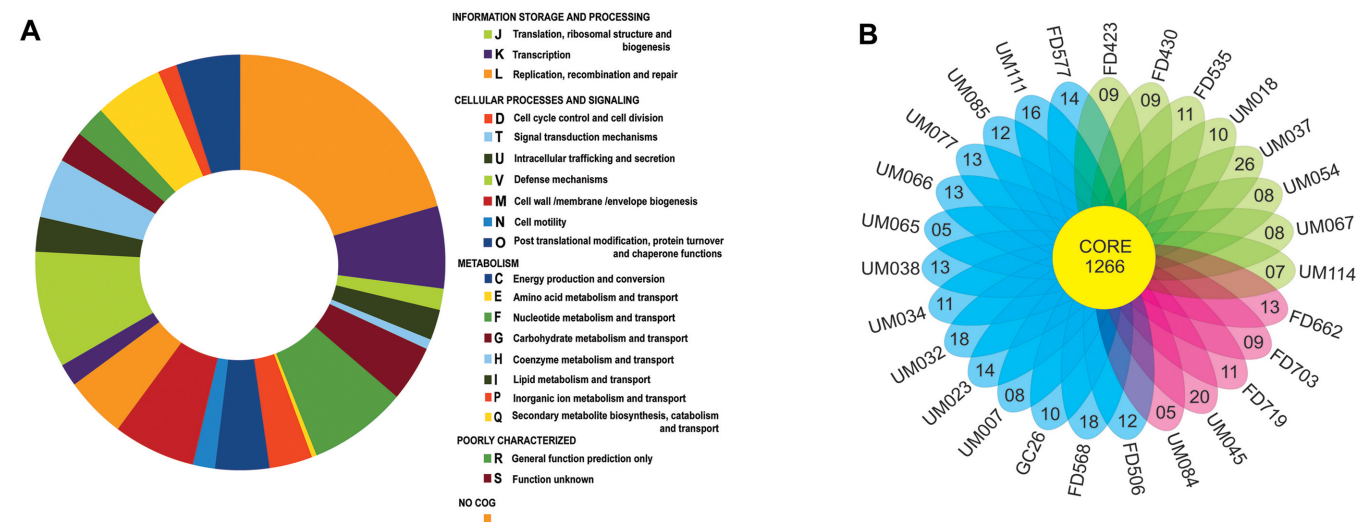


Figure 4. The functional COG classification of genes: (A) the COG functional classification representing core genome of the Malaysian *H. pylori* isolates. (B) Core and specific gene content observed among various strains compared in the study.

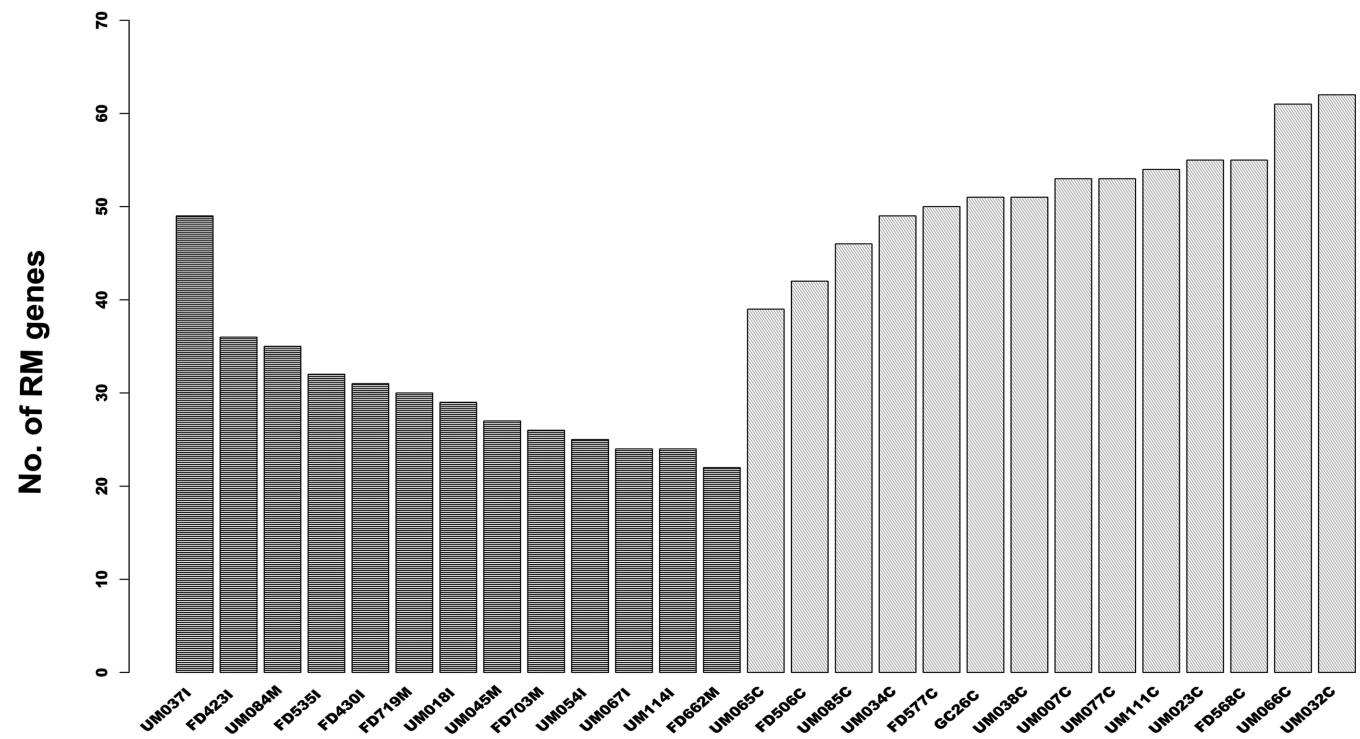


Figure 5. The distribution of RM genes in various strains: the graph shows the number of genes annotated to encode putative RM functions in each strain. The Y-axis represents the number of genes and the X-axis denotes strain names.

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 represent-

ing 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-

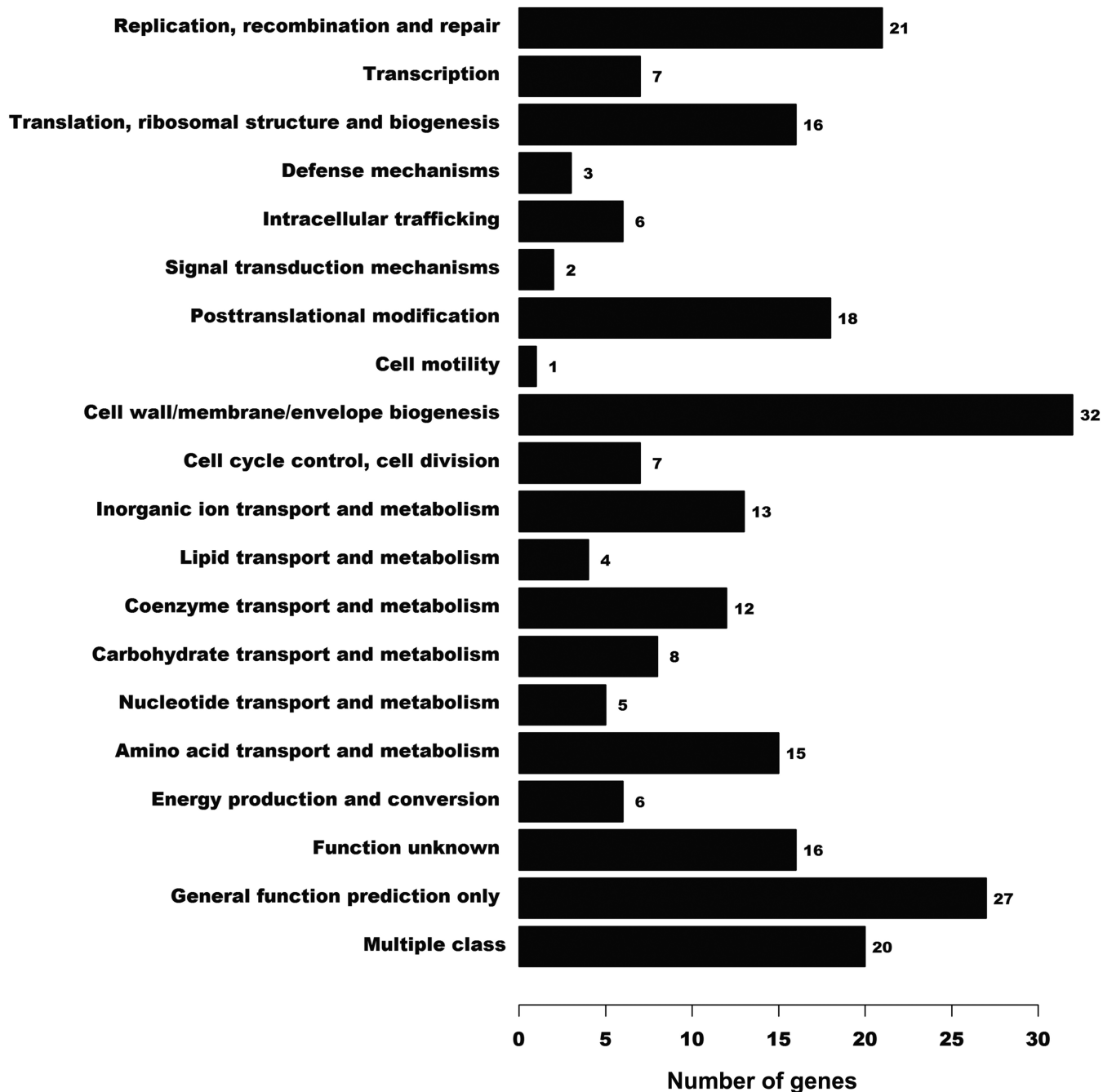


Figure 6. The functional classification of differentially evolving genes: the graph shows the COG functional classification of various core genes with signs of differential evolution among East-Asian and non-East-Asian strains. The Y-axis denotes the functional category and the X-axis represents the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

Cluster ID	Status in <i>H. pylori</i> strains		Predicted functions	Orthologs in 26695
	EA (n = 14)	Non-EA (n = 13)		
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 (AONK000000000), UM054 (AONL000000000), UM007 (AONM000000000), UM034 (AONN000000000) and UM045 (AONO000000000).

SUPPLEMENTARY DATA

[Supplementary Data](#) are available at NAR Online.

ACKNOWLEDGMENTS

We would like to thank members of the Ahmed lab, Vadi-velu 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

- Perez-Perez, G.I., Rothenbacher, D. and Brenner, H. (2004) Epidemiology of *Helicobacter pylori* infection. *Helicobacter*, **9**(Suppl. 1), 1–6.
- Khalifa, M.M., Sharaf, R.R. and Aziz, R.K. (2010) *Helicobacter pylori*: a poor man's gut pathogen? *Gut Pathog.*, **2**, 2.
- Amieva, M.R. and El-Omar, E.M. (2008) Host-bacterial interactions in *Helicobacter pylori* infection. *Gastroenterology*, **134**, 306–323.
- Cover, T.L. and Blaser, M.J. (2009) *Helicobacter pylori* in health and disease. *Gastroenterology*, **136**, 1863–1873.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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. Rehvaith, 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., Rehvaith, 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) GENE: 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. Krebs, 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.



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 serovars¹. However, serovar *S. Typhi* of *Salmonella enterica* subspecies *enterica* infects only humans resulting in a systemic infection, typhoid fever². 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 year³. 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 poor⁴. 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 mortality⁵. 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 surveillance¹⁰. 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 individuals¹¹.

In the past, various genomic studies have attributed signatures like pseudogenisation (loss of gene function) or gene deletion for host restriction in pathogenic bacteria¹². 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

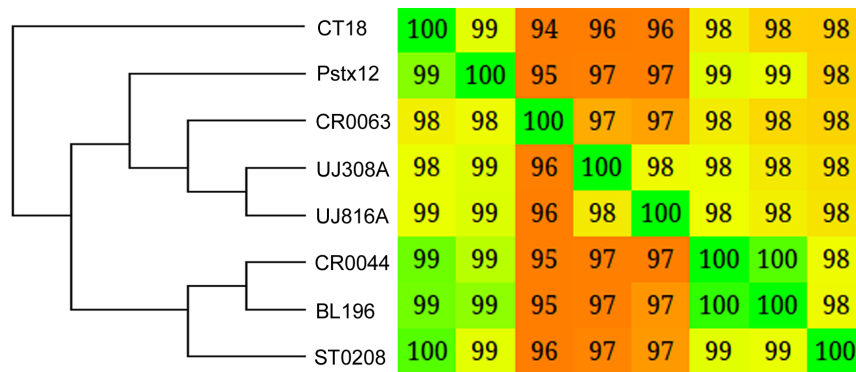


Figure 1 | Phylogenomic Tree. The whole genome information was used to build the distance matrix using Gegendes. The phylogenetic tree was developed using SplitsTree by NJ method. This revealed close similarity among genomes and also co-clustering of strains isolated from the same regions.

re-assortment over time¹⁷. 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 individual¹⁸. 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 survival¹⁸. 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 2005²³. 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, Malaysia²⁶. The previous findings have also recorded shared PFGE patterns among the isolates from Southeast Asia¹¹. Strains UJ308A and UJ816A were isolated in Papua New Guinea from fatal and non-fatal cases, respectively, in 1998²⁷. 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 strains³⁰. 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 strain³⁰. 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). BRIG³³ 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 progressiveMauve⁵⁰. 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 law³⁴. The Heap's law can be represented by the following equation:

$n = k \cdot N^{-\alpha}$, 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:

$n = k \cdot \exp\left[-\frac{N}{\tau}\right] + tg(\theta)$, 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 $k \cdot \exp\left[-\frac{N}{\tau}\right]$ 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

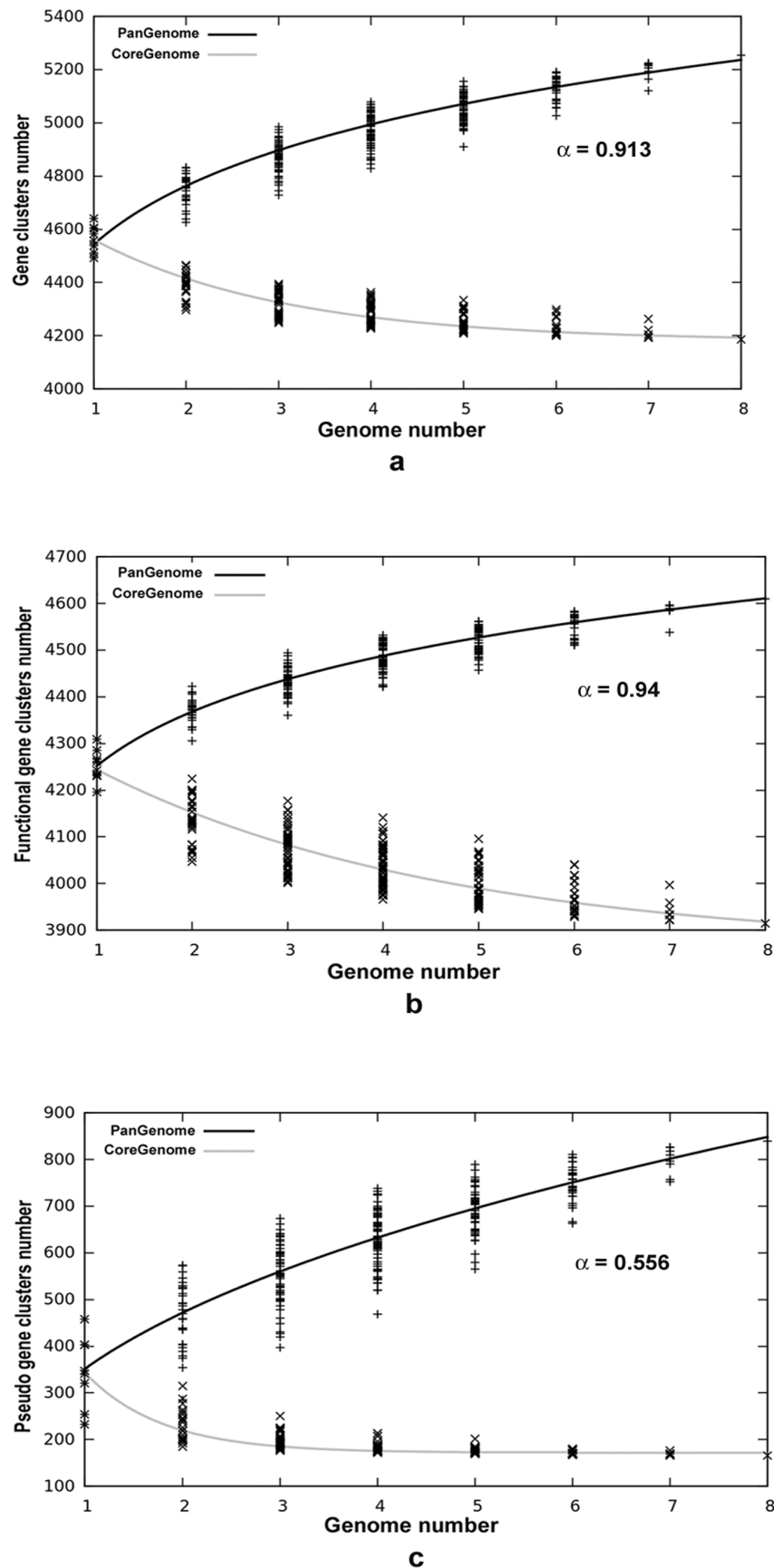


Figure 3 | Pan and Core Genome Distribution. (a). Pan and core genome developments using median values of the combinations of all eight genomes. (b). Pan and core genome developments of functional genes of these eight isolates. Here it can be observed that core genome is decreasing sharply. (c). Pan 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 vitB₁₂ 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 (604 clusters 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 complementary 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.

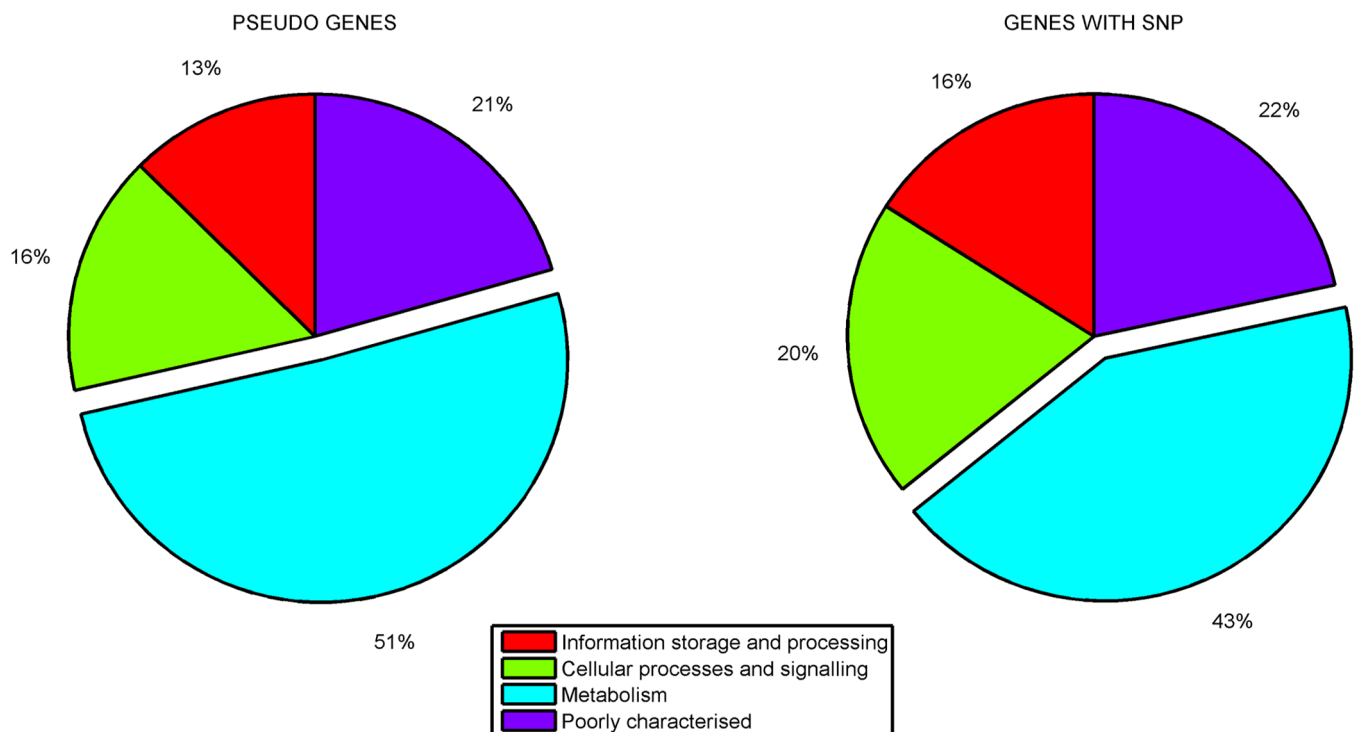


Figure 4 | The proportion of functionally classified pseudogenes and the functional genes with SNPs according to COG classification. The pie chart represents the proportion of various functional classes among the pseudogenes and the functional genes with SNPs. The figure clearly shows the enrichment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

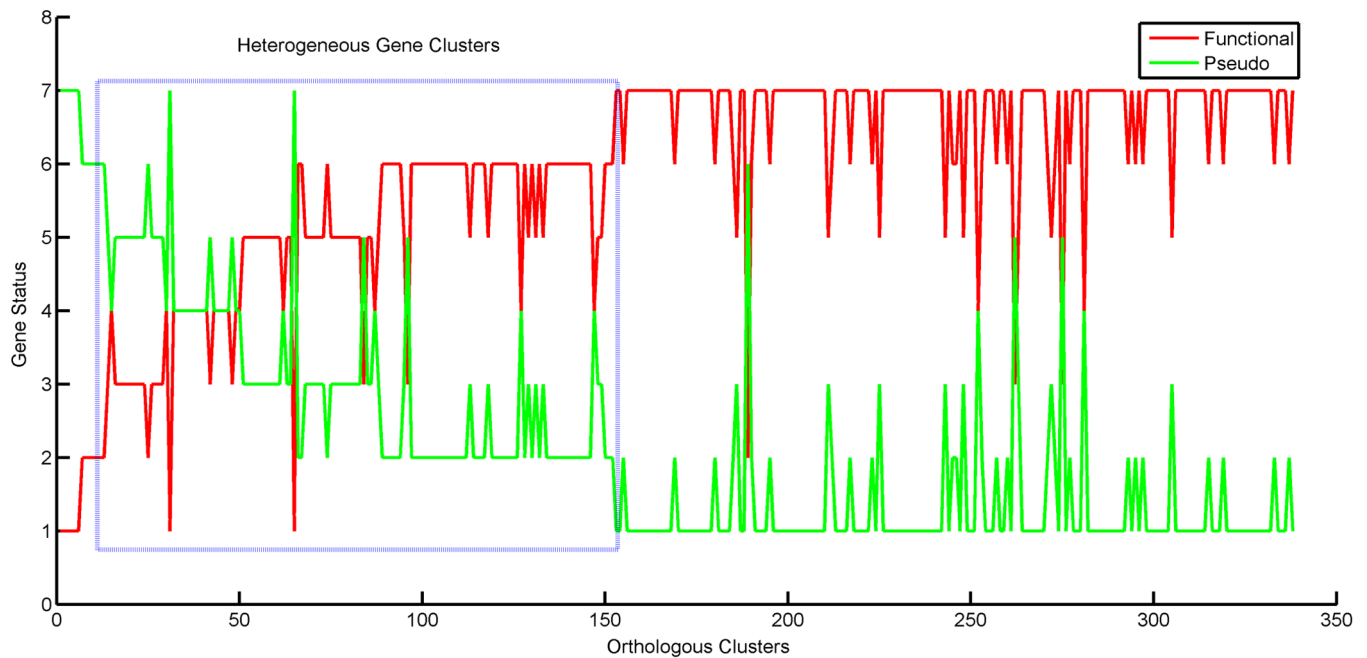


Figure 5 | Accessory pseudogene clusters analysis. The status of genes in each accessory pseudogene cluster was marked as P for pseudogene, F for functional complement and N for absence of gene. The clusters where the orthologs were present in P or F states were considered in plot. This shows the heterogeneous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

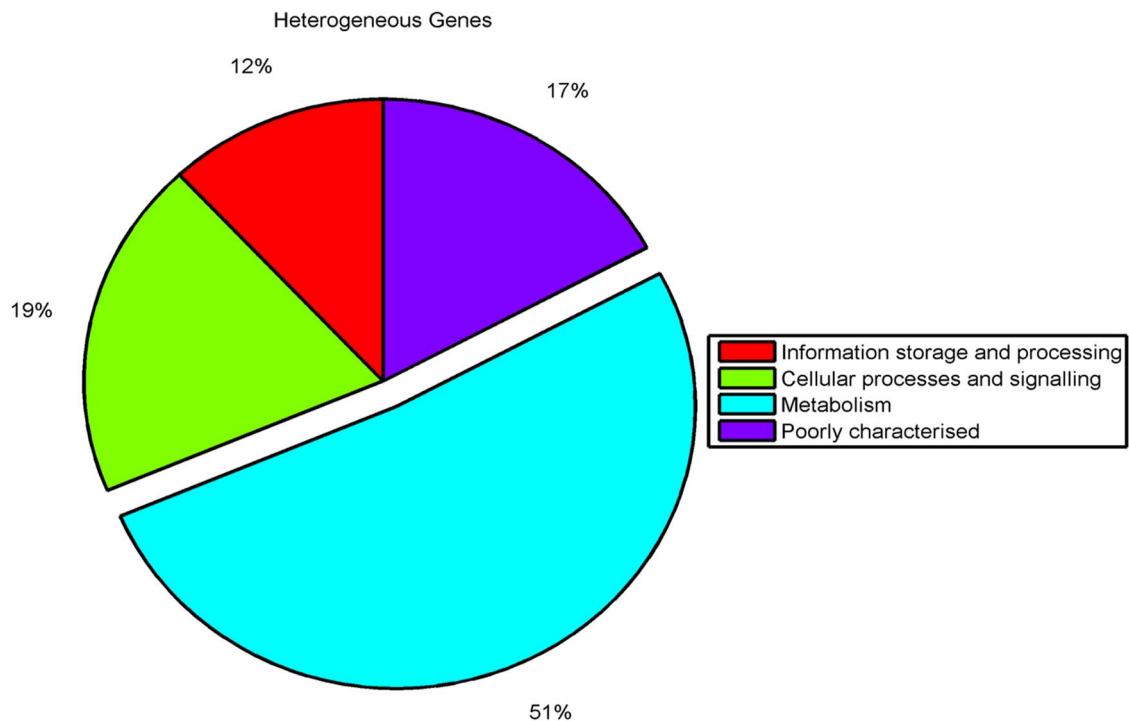


Figure 6 | Proportion of heterogeneous genes classified according to COG functional categories. The figure represents the distribution of accessory pseudogenes (those having variable functional and pseudogene status in at least two strains) among various COG functional categories. The genes related to 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 data¹⁷.

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 donors³⁷. 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 genome³⁸. 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 pseudogenes³⁹ although any gain of function may be rare, or only occurring in a small number of genes through point mutations¹⁷. 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 niche⁴⁰. 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 mutation¹⁷. 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 carriage⁴³. 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 tool⁴⁴. The alignment file was visualized using Tablet alignment viewer⁴⁵ 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 annotation⁴⁶. 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 similarity⁴⁷. The predicted ORFs were scanned to identify protein domains using HMMProfam (<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 Mauve⁵⁰.

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 PanOCT⁵¹. 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 finder⁵⁶. The genomic islands in these strains were identified using IslandViewer⁵⁷.



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 OrthoMCL⁵⁸ 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.*⁵⁴.

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 ClustalW⁶⁰.

- Coburn, B., Grassl, G. A. & Finlay, B. B. *Salmonella*, the host and disease: a brief review. *Immunol Cell Biol* **85**, 112–118 (2007).
- Boyle, E. C., Bishop, J. L., Grassl, G. A. & Finlay, B. B. *Salmonella*: from pathogenesis to therapeutics. *J Bacteriol* **189**, 1489–1495 (2007).
- Crump, J. A., Luby, S. P. & Mintz, E. D. The global burden of typhoid fever. *Bull World Health Organ* **82**, 346–353 (2004).
- Gopinath, S., Carden, S. & Monack, D. Shedding light on *Salmonella* carriers. *Trends Microbiol* **20**, 320–327 (2012).
- 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).
- 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).
- Ploy, M. C. *et al.* Integrin-associated antibiotic resistance in *Salmonella enterica* serovar typhi from Asia. *Antimicrob Agents Chemother* **47**, 1427–1429 (2003).
- Ruby, T., McLaughlin, L., Gopinath, S. & Monack, D. *Salmonella*'s long-term relationship with its host. *FEMS Microbiol Rev* **36**, 600–615 (2012).
- 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).
- 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).
- 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).
- Baumler, A. & Fang, F. C. Host specificity of bacterial pathogens. *Cold Spring Harb Perspect Med* **3**, a010041 (2013).
- McClelland, M. *et al.* Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. *Nature* **413**, 852–856 (2001).
- Holt, K. E. *et al.* Pseudogene accumulation in the evolutionary histories of *Salmonella enterica* serovars Paratyphi A and Typhi. *BMC Genomics* **10**, 36 (2009).
- 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).
- 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).
- Holt, K. E. *et al.* High-throughput sequencing provides insights into genome variation and evolution in *Salmonella Typhi*. *Nat Genet* **40**, 987–993 (2008).
- Rodríguez-Valera, F. U. D. Is the pan-genome also a pan-selectome? *F1000Res* **1**, 16 (2012).
- 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).
- Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Curr Opin Genet Dev* **15**, 589–594 (2005).
- 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).
- Zhou, Z. *et al.* Neutral genomic microevolution of a recently emerged pathogen, *Salmonella enterica* serovar Agona. *PLoS Genet* **9**, e1003471 (2013).
- 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).
- 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).
- 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).
- 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).
- 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).
- Parkhill, J. *et al.* Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* **413**, 848–852 (2001).
- Ong, S. Y. *et al.* Complete Genome Sequence of *Salmonella enterica* subsp. *enterica* Serovar Typhi P-stx-12. *J Bacteriol* **194**, 2115–2116 (2012).
- 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).
- 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).
- 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).
- 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).
- Tettelin, H., Riley, D., Cattuto, C. & Medini, D. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol* **11**, 472–477 (2008).
- 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).
- Nuccio, S. P. & Baumber, A. J. Comparative analysis of *Salmonella* genomes identifies a metabolic network for escalating growth in the inflamed gut. *MBio* **5**, e00929–14 (2014).
- 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).
- Cole, S. T. *et al.* Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007–1011 (2001).
- Olson, M. V. When less is more: gene loss as an engine of evolutionary change. *Am J Hum Genet* **64**, 18–23 (1999).
- Rohmer, L., Hocquet, D. & Miller, S. I. Are pathogenic bacteria just looking for food? Metabolism and microbial pathogenesis. *Trends Microbiol* **19**, 341–348 (2011).
- 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).
- Touchon, M. *et al.* Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS genet* **5**, e1000344 (2009).
- Reis, R. S. & Horn, F. Enteropathogenic *Escherichia coli*, *Salmonella*, *Shigella* and *Yersinia*: cellular aspects of host-bacteria interactions in enteric diseases. *Gut Pathog* **2**, 8 (2010).
- Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
- Milne, I. *et al.* Tablet-next generation sequence assembly visualization. *Bioinformatics* **26**, 401–402 (2010).
- 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).
- 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).
- 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).
- Lagesen, K. *et al.* RNAmmer: consistent annotation of rRNA genes in genomic sequences. *Nucleic Acids Res* **35**, 3100–3108 (2000).
- Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS one* **5**, e11147 (2010).
- Fouts, D. E., Brinkac, L., Beck, E., Inman, J. & Sutton, G. PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic



- analysis of bacterial strains and closely related species. *Nucleic Acids Res* **40**, e172 (2012).
52. Agren, J., Sundstrom, A., Hafstrom, T. & Segerman, B. Gegenees: Fragmented Alignment of Multiple Genomes for Determining Phylogenomic Distances and Genetic Signatures Unique for Specified Target Groups. *PLoS one* **7**, e39107 (2012).
 53. Huson, D. H. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68–73 (1998).
 54. Akhter, S., Aziz, R. K. & Edwards, R. A. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res* **40**, e126 (2012).
 55. Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. & Wishart, D. S. "PHAST: A Fast Phage Search Tool" *Nucleic Acids Res* **39**, W347–352 (2011).
 56. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* **34**, D32–D36 (2006).
 57. Langille, M. G. & Brinkman, F. S. IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* **25**, 664–665 (2009).
 58. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178–2189 (2003).
 59. Liang, W. *et al.* Pan-genomic analysis provides insights into the genomic variation and evolution of *Salmonella* Paratyphi A. *PLoS One* **7**, e45346 (2012).
 60. Larkin, M. A. *et al.* Clustal W and clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).

Acknowledgments

We would like to thankfully acknowledge Prof. Kwai-Lin Thong for information about some of the strains studied herein and to all those authors who contributed to data about

these strains as available in the public domain. We would also like to thank Donepudi RaviTeja for his help with gnuplot and R. We thankfully acknowledge funding received from Department of Biotechnology, Government of India (Ref. No. BT/HRD/NBA/34/01/2011(ix) and BT/PR6921/MED/29/699/2013). RB would like to acknowledge the UGC-RFSMS fellowship.

Author contributions

R.B. and N.A. designed and conducted the study. N.K. helped in analysis of data and preparation of the manuscript. A.K.L. and S.S. provided help in writing scripts.

Additional information

Accession codes CT18 (NC_003198), P-stx-12 (NC_016832), UJ308A (AJTD00000000), UJ816A (AJTE00000000), BL196 (AJGK00000000), CR0063 (AKIC00000000), CR0044 (AKZO00000000) and ST0208 (AJXA00000000).

Supplementary information accompanies this paper at <http://www.nature.com/scientificreports>

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Baddam, R., Kumar, N., Shaik, S., Lankapalli, A.K. & Ahmed, N. Genome dynamics and evolution of *Salmonella* Typhi strains from the typhoid-endemic zones. *Sci. Rep.* **4**, 7457; DOI:10.1038/srep07457 (2014).



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder in order to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>

REVIEW ARTICLE

Helicobacter pylori in 2013: Multiplying Genomes, Emerging Insights

Niyaz Ahmed,^{*,†} Mun Fai Loke,[‡] Narender Kumar^{*} and Jamuna Vadivelu[‡]

^{*}Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, University of Hyderabad, Hyderabad, 500046, India, [†]Institute of Biological Sciences, University of Malaya, Kuala Lumpur, 50603, Malaysia, [‡]Department of Medical Microbiology, Faculty of Medicine, University of Malaya, Kuala Lumpur, 50603, Malaysia

Keywords

Genomes, genomic diversity, evolution, adaptation, virulence.

Reprint requests to: Dr Niyaz Ahmed, Department of Biotechnology and Bioinformatics, University of Hyderabad, Hyderabad 500046, India. E-mail: niyaz.ahmed@uohyd.ac.in

Abstract

We describe features of key additions to the existing pool of publicly accessible *Helicobacter pylori* genome sequences and sequences of *Helicobacter pylori* phages from April 2012 to March 2013. In addition, important studies involving *H. pylori* genomes, especially those pertaining to genomic diversity, disease outcome, *H. pylori* population structure and evolution are reviewed. High degree of homologous recombination contributes to increased diversity of *H. pylori* genomes. New methods of resolving *H. pylori* population structure to an ultrafine level led to the proposal of new subpopulations. As the magnitude of diversity in the *H. pylori* gene pool becomes more and more clear, geographic and demographic factors should be brought to analysis while identifying disease-specific biomarkers and defining new virulence mechanisms.

Background

The vista on *Helicobacter pylori* genomics began in August 1997 with the publication of the complete genome of *Helicobacter pylori* 26695, which was cultured from a gastritis patient in the United Kingdom [1]. The second complete genome of *H. pylori* J99, isolated in the USA from a patient with a duodenal ulcer, was released in January 1999 [2]. It was not until June 2006 that the third *H. pylori* genome, HPAG1, isolated from a Swedish patient with chronic atrophic gastritis, was made known [3]. Recent technological advancement has made next-generation sequencing (NGS) more accessible and less costly resulting in a rapid increase in the number of *H. pylori* patient isolates sequenced, including some of the important laboratory strains [4]. Until March 2013, 43 complete genomes and 198 draft genome sequences were already deposited in GenBank for public access, and the federated genomic databases still growing.

Summarizing on the genomic sequences currently available, based on the country of origin of the source patient: 18 were from North and South America, 14 from Far East Asia (Japan/Korea/China), 11 from Europe, 10 from Malaysia, six from Africa, four from India, and one from Australia. There was insufficient informa-

tion on the origin of the source patients for the remaining 177 strains. When classified by the disease status of the patient: 17 were isolated from gastritis patients, 10 from gastric cancer patients, nine from duodenal ulcer patients, two from gastric cancer patients, and two from patients presented with gastric MALT-lymphoma. The disease status of the remaining 201 strains was unknown. Based on available data of complete genomes in GenBank, the average size of an *H. pylori* genome was estimated as 1.62 Mb (1.51–1.71 Mb) with a GC content of 38.92% (38.40–39.30%). The average *H. pylori* genome was predicted to consist of 1590 (1429–1749) open-reading frames encoding 1532 (1382–1707) proteins.

Notable among many other *H. pylori* genomes that have been sequenced during this period is *H. pylori* (hpEurope) strain N6, which is widely used in research because of its high transformation efficiency [4]. The availability of the genome sequence for N6 will facilitate analysis and evaluation of novel and existing experimental data using this strain.

Newly Sequenced *H. pylori* Genomes

Our survey of the *H. pylori* genomes (Table 1) announced between April 2012 and March 2013 entails

a number of genomes contributed from Asian countries such as South Korea, Japan, and China; these are the countries with the highest incidence of gastric cancer in the world [5]. However, only a few Chinese strains are available. Adding to the collection of strains from geographic regions with high incidence of gastroduodenal diseases, *H. pylori* strain XZ274, isolated from a Tibetan gastric cancer patient, was the first from the high-altitude Tibetan plateau, to be sequenced [6]. The complete genome sequence would likely be helpful in understanding the adaptation of this bacterium to patients living in high-altitude areas and the reason for high infection rate in the Tibetan plateau. Contributing to the diversity of strains from regions with high incidence of gastrointestinal diseases, draft genomes of three strains isolated from two atrophic gastritis (HLJ193 and HLJ256), and one gastric ulcer (HLJ271) disease patients from Heilongjiang province, China, were also made available [7].

In addition, two *H. pylori* strains isolated from duodenal ulcer patients in Bangalore (NAB47) and Delhi (NAD1) in India, where the rate of infection is high but the rate of gastric cancer is low, were also sequenced [8]. Unlike their highly virulent Far East Asian counterparts (that may be linked to the high incidence of gastric cancer), Indian *H. pylori* strains are described as mostly

“innocuous or only mildly pathogenic”. It is certain that these genomes representing Indian patients would rekindle our understanding of the genetic makeup and evolutionary relationships of this pathogen in India. Furthermore, the genome sequence data would allow newer insights into the enigmatic situation, wherein high infection burden predicts the lowest rates of gastric or duodenal ulcers and gastric carcinoma.

Draft genome sequences from nine functional dyspeptic patients and one gastric cancer patient from Kuala Lumpur, Malaysia were also published [9]. In contrast to the situations in China and India, Malaysia presents an interesting canvas to study host–pathogen coevolution in multiple ways. Being an immigrant society located along major shipping routes between China and India, Malaysia has three major ethnic populations, Malay, Chinese, and Indian. Besides its ethnic diversity, the country also presents a coexisting array of *H. pylori* that can be identified to various ancestral populations: hpEastAsia, hpAsia2, and hpEurope [10]. Availability of genomic sequences from this region of South-East Asia will serve as a linking piece on the human migration puzzle, as well as aiding in the study of the role of the transmission and coevolution of *H. pylori* with its host.

The *H. pylori* genome is well known to be “highly sexual” given high frequency of horizontal gene transfers. For the first time, a global survey of the genome-wide distribution of outcrossing homologous recombination in *H. pylori* has been shown and this illustrated how mutual recombination can shape a bacterial species [11]. Mosaic genes, possibly formed following frequent recombination episodes and discordant phylogenies, were distributed throughout *H. pylori* genomes with most genes exhibiting at least one unique recombination event. Genes with high recombination rates comprised those entailing DNA transformation, core cellular functions (biosynthesis and metabolism), and host–pathogen interactions (outer membrane proteins and lipopolysaccharide synthesis).

Helicobacter pylori is highly diverse on a global scale but geographic separation reduces the recombination frequency between strains from a local area relative to those from outside. This leads to the establishment of a clonal population structure for *H. pylori* in a given geographic area. For each population, *H. pylori* has undergone relatively independent evolution processes, resulting in locale specific genomic diversity and differential potential for virulence. Further study to characterize these differences within a local population may help elucidate mechanisms involved in the development of gastroduodenal diseases triggered by *H. pylori* [12]. On this basis, using a whole-genome CombiMatrix custom oligonucleotide tiling microarray, 84 strains

Table 1 *H. pylori* genomes published/deposited during last year

Strains/isolates and origin	Strain/isolate name code	Accession No./Project ID
Strains/isolates from India	NAB47	AJFA00000000
	NAD1	AJGI00000000
Strains/isolates from China	HLJ193	ALJI00000000
	HLJ256	ALKA00000000
	HLJ271	ALKB00000000
	XZ274	CP003419
	pXZ274	CP003420
<i>Helicobacter pylori</i> plasmid		
<i>Helicobacter pylori</i> phages (Japan)	KHP30	AB647160
	KHP40	AB731695
Strains/isolates from Malaysia	FD423	AKHM00000000
	FD430	AKHN00000000
	FD506	AKHO00000000
	FD535	AKHP00000000
	FD568	AKHQ00000000
	FD577	AKHR00000000
	FD703	AKHS00000000
	FD662	AKHT00000000
	FD719	AKHU00000000
	GC26	AKHV00000000
Laboratory strains	N6	CAHX00000000
	17874	PRJNA76569
	P79	PRJNA76567

with different clinical outcomes from various parts of China were screened and several highly variable regions correlating with gastritis, gastric ulceration, and gastric cancer presentations were identified within the *cag* pathogenicity island [13].

In a recent study, *in silico* chromosome painting was applied to *H. pylori* to reveal finer population structure that was not recognized previously using conventional multi-locus sequence typing (MLST) analysis of several housekeeping genes, and the latter failed to account for recombination across the genome [14]. Using this method, novel subpopulations were found in European, Amerindian, and East Asian groups. In addition, some singleton strains were shown to be hybrids of subgroups and demonstrated signs of population admixture in Africa, Europe, and parts of Asia. These results enhance our current understanding of the intraspecific bacterial evolution.

Sequencing *H. pylori* Phages

Bacteriophages make up part of the bacterial genomes. They can contribute to bacterial evolution and may affect host attributes, such as physiological behavior, pathogenesis, or adaptation *via* their possible roles in horizontal gene acquisition and bacteria–phage antagonistic coevolution [15]. The temperate bacteriophage 1961P was isolated from the lysate of a *H. pylori* clinical isolate cultured in Taiwan, characterized and sequenced [16]. The bacteriophage was reported to be typical of the *Podoviridae* family and may be transduced and integrated into the host bacterial chromosome *via* a mechanism similar to that of lambda phage. The complete genome sequences of two *H. pylori* bacteriophages isolated from culture supernatants of East Asian-type strains from Japanese patients were also sequenced [17]. Bacteriophages/prophages are not uncommon as phage-associated genes were identified in nine of ten *H. pylori* strains from Kuala Lumpur [Vadivelu J, Loke MF, et al. (unpublished)]. The functional and evolutionary roles of these *H. pylori* bacteriophages/prophages remain largely unknown.

Epilogue

It appears that *H. pylori* research community has ushered in to its good times while the technology revolution flooding the bacterial genomic landscape [18, 19]. Significant progress has been made in augmenting the understanding of the biology of *H. pylori* through replicative genomics. New genome data from high burden countries would add to the story in a more meaningful manner. The much-needed genome sequences of unex-

plored strains from remote/tribal and mainstream populations will facilitate understanding of the true virulence potentials of *H. pylori*, its transmission, global epidemiology, and adaptive coevolution with its host. Chronological and replicate genomics of serial isolates obtained from single patients are likely to enrich vistas of the host–microbe interactions occurring during colonization by *H. pylori*.

The core genome of *H. pylori* is limited to approximately 1,200 genes although an up to date pan-genome size is difficult to estimate, because it is infinitely open like any other enterobacterial pan-genomes, due to an ever growing pool of novel genes and genetic elements, contributed by each new genome. However, its variable genome or “plastiome”, which can be portrayed as a fraction of the pan-genome, is dominated by large plasticity zones and is of significant attraction because it contains seemingly novel genes and genetic elements predicted to occur at the frequency of about 20–40 genes per newly sequenced genome. Given this, the wealth of novel genes emanating from multiple whole-genome sequencing projects should help us systematically understand and decipher functional attributes and consequences of horizontal gene transfer, especially in terms of virulence potentials, adaptation, and persistence in different host populations. Such descriptive analyses supported by functional level associations as well as patient data are expected to form the baseline observations that define newly emerging areas of functional molecular infection epidemiology and systems epidemiology [18].

Acknowledgements and Disclosures

We would like to acknowledge funding support from the department of Biotechnology, Government of India, as well as support from the University of Malaya, Kuala Lumpur, Malaysia, under the HIR grant (UM.C/625/1/HIR/MOHE/CHAN-02; account no. A000002-50001, “Molecular Genetics”).

Competing interests: the authors have no competing interests.

References

- 1 Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, et al. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* 1997;388: 539–47.
- 2 Alm RA, Ling LS, Moir DT, King BL, Brown ED, Doig PC, et al. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* 1999;397:176–80.
- 3 Oh JD, Kling-Bäckhed H, Giannakis M, Xu J, Fulton RS, Fulton LA, et al. The complete genome sequence of a chronic atrophic gastritis *Helicobacter pylori* strain: evolution during disease progression. *Proc Natl Acad Sci USA* 2006;103:9999–10004.

- 4 Behrens W, Bönig T, Suerbaum S, Josenhans C. Genome sequence of *Helicobacter pylori* hpEurope strain N6. *J Bacteriol* 2012;194:3725–6.
- 5 Yamaoka Y. Mechanisms of disease: *Helicobacter pylori* virulence factors. *Nat Rev Gastroenterol Hepatol* 2010;7:629–41.
- 6 Guo Y, Wang H, Li Y, Song Y, Chen C, Liao Y, et al. Genome of *Helicobacter pylori* strain XZ274, an isolate from a Tibetan patient with gastric cancer in China. *J Bacteriol* 2012;194:4146–7.
- 7 You Y, Liu L, Zhang M, Han X, He L, Zhu Y, et al. Genome sequences of three *Helicobacter pylori* strains isolated from atrophic gastritis and gastric ulcer patients in China. *J Bacteriol* 2012;194:6314–15.
- 8 Kumar N, Mukhopadhyay AK, Patra R, De R, Baddam R, Shaik S, et al. 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* 2012;194:5963–4.
- 9 Gunaletchumy SP, Teh X, Khosravi Y, Ramli NS, Chua EG, Kavitha T, et al. Draft genome sequences of *Helicobacter pylori* isolates from Malaysia, cultured from patients with functional dyspepsia and gastric cancer. *J Bacteriol* 2012;194:5695–6.
- 10 Tay CY, Mitchell H, Dong Q, Goh KL, Dawes IW, Lan R. Population structure of *Helicobacter pylori* among ethnic groups in Malaysia: recent acquisition of the bacterium by the Malay population. *BMC Microbiol* 2009;9:126.
- 11 Yahara K, Kawai M, Furuta Y, Takahashi N, Handa N, Tsuru T, et al. Genome-wide survey of mutual homologous recombination in a highly sexual bacterial species. *Genome Biol Evol* 2012;5:628–40.
- 12 Dong Q-J, Zhan S-H, Wang L-L, Xin Y-N, Jiang M, Xuan S-Y. Relatedness of *Helicobacter pylori* populations to gastric carcinogenesis. *World J Gastroenterol* 2012;18:6571–6.
- 13 You Y, He L, Zhang M, Fu J, Gu Y, Zhang B, et al. Comparative genomics of *Helicobacter pylori* strains of China associated with different clinical outcome. *PLoS ONE* 2012;7:e38528.
- 14 Yahara K, Furuta Y, Oshima K, Yoshida M, Azuma T, Hattori M, et al. Chromosome painting *in silico* in a bacterial species reveals fine population structure. *Mol Biol Evol* 2013;30:1454–64.
- 15 Clokie MR, Millard AD, Letarov AV, Heaphy S. Phages in nature. *Bacteriophage* 2011;1:31–45.
- 16 Luo C-H, Chiou P-Y, Yang C-Y, Lin N-T. Genome, integration, and transduction of a novel temperate phage of *Helicobacter pylori*. *J Virol* 2012;86:8781–92.
- 17 Uchiyama J, Takeuchi H, Kato S, Takemura-Uchiyama I, Ujihara T, Daibata M, et al. Complete genome sequences of two *Helicobacter pylori* bacteriophages isolated from Japanese patients. *J Virol* 2012;86:11400–1.
- 18 Ahmed N. Coevolution and adaptation of *Helicobacter pylori* and the case for 'functional molecular infection epidemiology'. *Med Princ Pract* 2011;20:497–503.
- 19 Ahmed N. A flood of microbial genomes-do we need more? *PLoS ONE* 2009;4:e5831.

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, IS606, and mobilization proteins, together with replication protein A. CagA protein in both of the strains contained EPIYA D-type motifs, which are typical of Indo-European strains. Important plasticity region genes, such as *jhp0940*, *jhp0947*, and *dupA*, were absent, and *hp0986* was detected only in NAB47. Finally, whole-genome phylogeny incorporating all the available genomes reconfirmed an Indo-European ancestry (HpEurope).

We believe that the genomes described herein are likely to rekindle our knowledge of the genetic makeup and evolutionary relationships of *H. pylori* in India. Comparative genomic analyses extending out to other unexplored strains from the tribal and mainstream populations will facilitate understanding of the true pathogenic potential (amid adaptive evolution) of the Indian *H. pylori*. Furthermore, they will be immensely helpful in global epidemiological studies and also for the development of diagnostic tools tailored to a particular host population.

Received 31 July 2012 Accepted 17 August 2012

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

N.K. and A.K.M. contributed equally to this work.

Copyright © 2012, American Society for Microbiology. All Rights Reserved.

doi:10.1128/JB.01371-12

Nucleotide sequence accession numbers. The genome sequences of *H. pylori* NAB47 and NAD1 have been deposited with GenBank and assigned accession numbers [AJFA000000000](#) and [AJGJ000000000](#), 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

- Ahmed N. 2011. Coevolution and adaptation of *Helicobacter pylori* and the case for 'functional molecular infection epidemiology.' *Med. Princ. Pract.* 20:497–503.
- Ahmed N. 2009. A flood of microbial genomes: do we need more? *PLoS One* 4:e5831. doi:10.1371/journal.pone.0005831.
- Atherton JC, Blaser MJ. 2009. Coadaptation of *Helicobacter pylori* and humans: ancient history, modern implications. *J. Clin. Invest.* 119:2475–2487.
- 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.
- Aziz RK, et al. 2008. The RAST server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi:10.1186/1471-2164-9-75.
- Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 5:578–579.
- Delcher AL, et al. 1999. Improved microbial gene identification with Glimmer. *Nucleic Acids Res.* 27:4636–4641.
- Devi SH, et al. 2010. Genome of *Helicobacter pylori* strain 908. *J. Bacteriol.* 192:6488–6489.
- Devi SM, et al. 2007. Ancestral European roots of *Helicobacter pylori* in India. *BMC Genomics* 8:184. doi:10.1186/1471-2164-8-184.
- 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.
- Falush D, et al. 2003. Traces of human migrations in *Helicobacter pylori* populations. *Science* 299:1582–1585.
- Kent WJ. 2002. BLAT: the BLAST-like alignment tool. *Genome Res.* 12:656–664.
- 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.
- 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.
- 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.
- Linz B, et al. 2007. An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* 445:915–918.
- 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.
- 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.
- Yamaoka Y. 2009. *Helicobacter pylori* typing as a tool for tracking human migration. *Clin. Microbiol. Infect.* 15:829–834.
- Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18:821–829.



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 Baddam¹, Narender Kumar¹, Sabiha Shaik¹, 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 aetiological 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

¹Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad, Hyderabad, India

²Institute 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

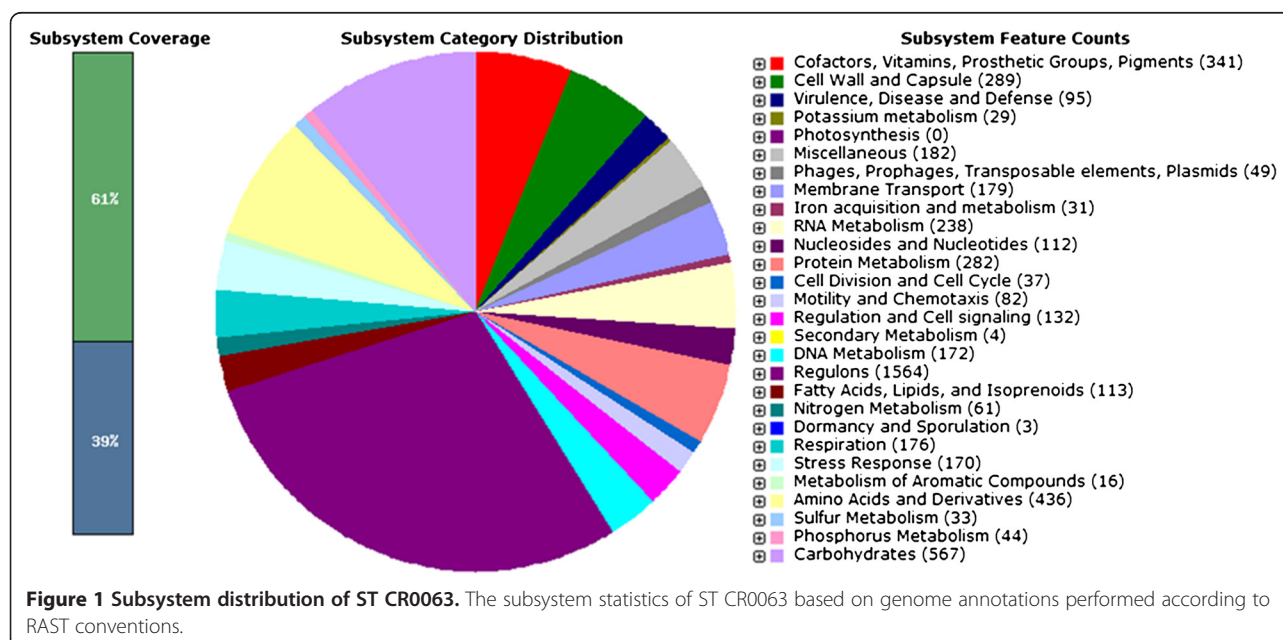
The gene *shdA*, a key factor predicted to be involved in persistence of the bacterium in the intestines [14] by binding to its extracellular matrix, was identified and annotated. This gene, by mimicking the host heparin, is able to bind to the extracellular matrix proteins, fibronectin and collagen, 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

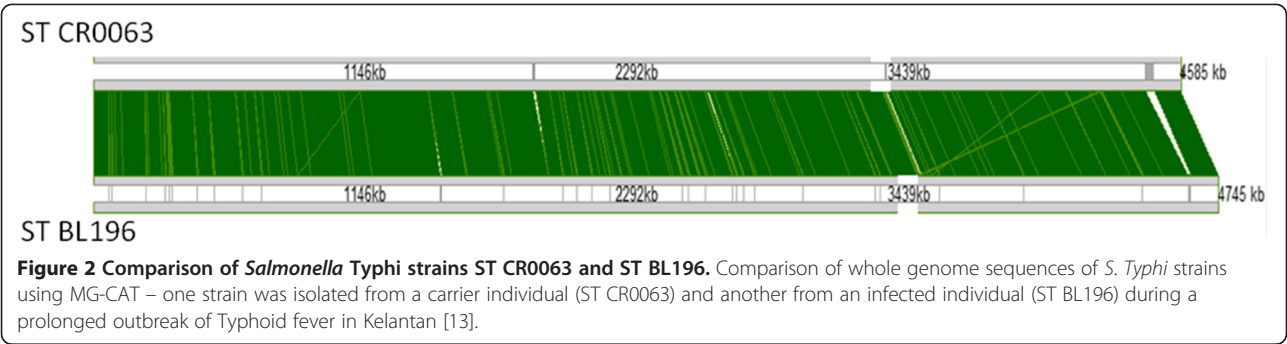
which aid in attachment of the bacterium to intestinal villi and also with each other, were found in the genome. These adherence factors determine the sites of bacterial colonisation and thereby adaptation and pathogenicity of a particular strain [19,20].

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 (*tvxA-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)

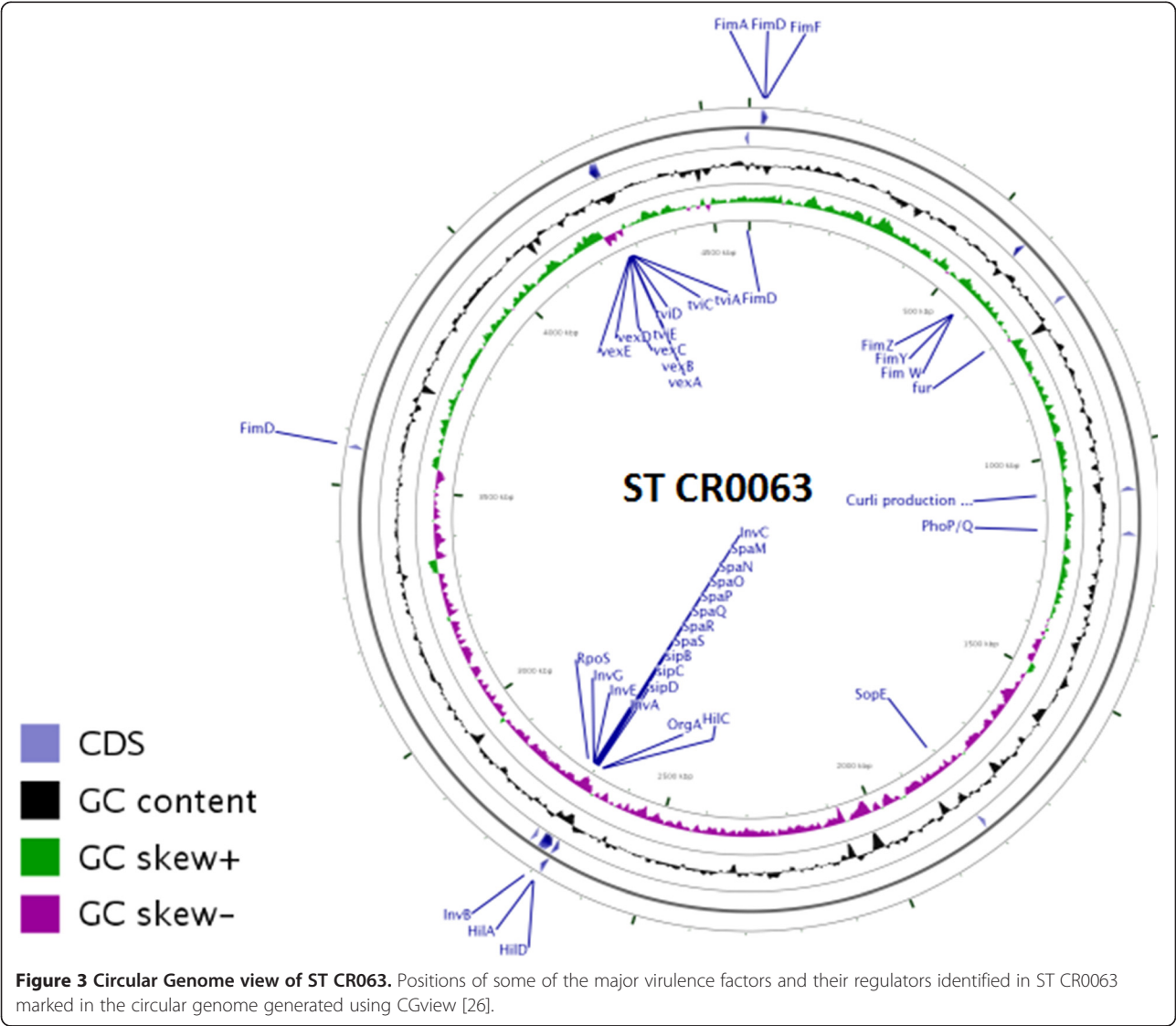
The phages gifsy-1 and fels-2 [27] together with many phage proteins and a few hypothetical proteins were identified in the genome of ST CR0063 by various algorithms (See Methods for details). It is expected that these phages are acquired by horizontal gene transfer (HGT) events as they were embedded 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].





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*,



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 1.6). 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 AKIC000000000. 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

¹Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad, Hyderabad, India. ²Institute of Biological sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia. ³Laboratory 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, Puthucherry S, Yassin RM, Sudarmono P, Padmidevi M, Soewandjo E, Handojo I, Sarasombath S, Pang T: **Analysis of Salmonella typhi isolates from southeast Asia by pulsed-field gel electrophoresis.** *J Clin Microbiol* 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, Messegger 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, Bäuml 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, Bäuml 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, Schlauch 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, Brummell 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 J, Wallis TS, Galvay 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, Sebahia 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, Barrall 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, Mégraud F, Ahmed N: Genome of *Helicobacter pylori* strain 908. *J Bacteriol* 2010, **192**:6488–6489.
 43. Siddavattam D, Karegoudar TB, Mudde SK, Kumar N, Baddam R, Avasthi TS, Ahmed N: Genome of a novel isolate of *Paracoccus denitrificans* capable of degrading N, N-dimethylformamide. *J Bacteriol* 2011, **193**:5598–5599.
 44. Yap KP, Teh CS, Baddam R, Chai LC, Kumar N, Avasthi TS, Ahmed N, Thong KL: Insights from the genome sequence of a *Salmonella enterica* serovar Typhi strain associated with a sporadic case of typhoid fever in Malaysia. *J Bacteriol* 2012, **194**:5124–5125.
 45. Yap KP, Gan HM, Teh CS, Baddam R, Chai LC, Kumar N, Tiruvayipati SA, Ahmed N, Thong KL: Genome sequence and comparative pathogenomics analysis of a *Salmonella enterica* serovar Typhi strain associated with a typhoid carrier in Malaysia. *J Bacteriol* 2012, **194**:5970–5971.
 46. Zerbino DR, Birney E: Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 2008, **18**:821–829.
 47. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W: Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 2011, **27**:578–579.
 48. Nadalin F, Vezzi F, Policriti A: GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics* 2012, **13**(14):8. doi:10.1186/1471-2105-13-S14-S8.
 49. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* 2009, **25**:1754–1760.
 50. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D: Tablet—next generation sequence assembly visualization. *Bioinformatics* 2010, **26**:401–402.
 51. Aziz RK, Devoid S, Disz T, Edwards RA, Henry CS, et al: SEED Servers: high-performance access to the SEED genomes, annotations, and metabolic models. *PLoS ONE* 2012, **7**(10):48053. doi:10.1371/journal.pone.0048053.
 52. Hemmerich C, Buechlein A, Podicheti R, Revanna KV, Dong Q: An Ergatis-based prokaryotic genome annotation web server. *Bioinformatics* 2010, **26**:1122–1124.
 53. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: Artemis: sequence visualization and annotation. *Bioinformatics* 2000, **16**:944–945.
 54. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL: Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 1999, **27**:4636–4641.
 55. Larsen TS, Krogh A: EasyGene—a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics* 2003, **4**:21.
 56. Lagesen K, Hallin PF, Rødland E, Stærfeldt HH, Ussery DW RT: RNAMmer: consistent annotation of rRNA genes in genomic sequences. *Nucleic Acids Res* 2007, **35**:3100–3108.
 57. Schattner P, Brooks AN, Lowe TM: tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 2005, **33**:W686–W689.
 58. Akhter S, Aziz RK, Edwards RA: PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res* 2012, **40**:e126.
 59. Bose M, Barber RD: Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. *In Silico Biol* 2006, **6**:223–227.
 60. Vernikos GS, Parkhill J: Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics* 2006, **22**:2196–2203.
 61. Langille MG, Brinkman FS: IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* 2009, **25**:664–665.

doi:10.1186/1757-4749-4-20

Cite this article as: Baddam et al.: Genome sequencing and analysis of *Salmonella enterica* serovar Typhi strain CR0063 representing a carrier individual during an outbreak of typhoid fever in Kelantan, Malaysia. *Gut Pathogens* 2012 **4**:20.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



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, Seksyen U13, 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 Illumina 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](https://www.ncbi.nlm.nih.gov/nuclink/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.

Copyright © 2012, American Society for Microbiology. All Rights Reserved.

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, Casades 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-ulose 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 gamma-proteobacteria. *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.

Genomes of Two Clinical Isolates of *Mycobacterium tuberculosis* from Odisha, India

Mohammad Majid,^{a,d} Narender Kumar,^a Asifa Qureshi,^d Priyadarshini Yerra,^a Ashutosh Kumar,^a Mandala Kiran Kumar,^a Suma Tiruvayipati,^{a,c} Ramani Baddam,^a Sabiha Shaik,^a Aparna Srikantam,^b Niyaz Ahmed^{a,c}

Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, University of Hyderabad, Hyderabad, India^a; Blue Peter Research Centre (Lepra Society), Cherlapally, Hyderabad, India^b; Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia^c; Environmental Genomics Division, CSIR–National Environmental Engineering Research Institute (NEERI), Nagpur, India^d

We report whole-genome sequences of two clinical isolates of *Mycobacterium tuberculosis* isolated from patients in Odisha, India. The sequence analysis revealed that these isolates are of an ancestral type and might represent some of the “pristine” isolates in India that have not admixed with other lineages.

Received 20 February 2014 Accepted 27 February 2014 Published 20 March 2014

Citation Majid M, Kumar N, Qureshi A, Yerra P, Kumar A, Kumar MK, Tiruvayipati S, Baddam R, Shaik S, Srikantam A, Ahmed N. 2014. Genomes of two clinical isolates of *Mycobacterium tuberculosis* from Odisha, India. *Genome Announc.* 2(2):e00199-14. doi:10.1128/genomeA.00199-14.

Copyright © 2014 Majid et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 3.0 Unported license](http://creativecommons.org/licenses/by/3.0/).

Address correspondence to Narender Kumar, kumarnaren13@gmail.com, or Niyaz Ahmed, niyazahmed@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* CDC 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 [ALYG000000000](https://www.ncbi.nlm.nih.gov/nuccore/ALYG000000000) and [ALYH000000000](https://www.ncbi.nlm.nih.gov/nuccore/ALYH000000000), 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](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, Tzolaki 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>.

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 resis-

tance 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 [AJXA000000000](https://www.ncbi.nlm.nih.gov/nuccore/AJXA000000000). The version described in this paper is the first version, [AJXA010000000](https://www.ncbi.nlm.nih.gov/nuccore/AJXA010000000). The BioProject designation for this project is [PRJNA160181](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA160181).

ACKNOWLEDGMENTS

This research is supported by the University of Malaya High Impact Research Grant (reference UM.C/625/1HIR/MOHE/-02 [A000002-5000 1]), 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.

Copyright © 2012, American Society for Microbiology. All Rights Reserved.

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 pan-genome. *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,^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 to 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 [AJTD000000000](#) (UJ308A) and [AJTE000000000](#) (UJ816A).

Received 12 June 2012 Accepted 29 June 2012

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

Copyright © 2012, American Society for Microbiology. All Rights Reserved.

doi:10.1128/JB.01051-12

ACKNOWLEDGMENTS

We thankfully acknowledge support received from the University of Malaya High Impact Research Grant (UM.C/625/1HIR/MOHE/02 [A000002-5000 1]) Molecular Genetics. This genome program was completed under the wider umbrella of the Indo-German International Research Training Group, Internationales Graduiertenkolleg (GRK1673), Functional Molecular Infection Epidemiology, an initiative of the German Research Foundation (DFG) and the University of Hyderabad (India), of which N.A. is a speaker. N.A. is an Adjunct Professor of Molecular Biosciences at the University of Malaya, Kuala Lumpur, Malaysia, and an Adjunct Professor of Chemical Biology at the Institute of Life Sciences, Hyderabad, India. We are also grateful to M/s Genotypic Technology Pvt. Ltd., Bengaluru, India, for their untiring efforts with Illumina sequencing. We acknowledge the Bioinformatics Facility (BIF) at the Department of Biotechnology, University of Hyderabad, for the use of its computational infrastructure. We are thankful to Akash Ranjan for enabling access to the high-speed computing infrastructure at CDFD, Hyderabad.

All of the members of the Ahmed and Thong labs are gratefully acknowledged for their help and support.

REFERENCES

1. Avasthi TS, et al. 2011. Genomes of two chronological isolates (*Helicobacter pylori* 2017 and 2018) of the West African *Helicobacter pylori* strain 908 obtained from a single patient. *J. Bacteriol.* **193**:3385–3386.
2. Avasthi TS, et al. 2011. Genome of multidrug-resistant uropathogenic *Escherichia coli* strain NA114 from India. *J. Bacteriol.* **193**:4272–4273.
3. Aziz RK, et al. 2008. The RAST server: rapid annotations using subsystems technology. *BMC Genomics* **9**:75. doi:10.1186/1471-2164-9-75.
4. Baddam R, et al. 2012. Genetic fine structure of a *Salmonella enterica* serovar Typhi strain associated with the 2005 outbreak of typhoid fever in Kelantan, Malaysia. *J. Bacteriol.* **194**:3565–3566.
5. Boetzer M, et al. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**:578–579.
6. Combs BG, et al. 2005. Ribotyping of *Salmonella enterica* serovar Typhi isolates from Papua New Guinea over the period 1977 to 1996. *P. N. G. Med. J.* **48**:158–167.
7. Delcher AL, et al. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**:4636–4641.
8. Devi SH, et al. 2010. Genome of *Helicobacter pylori* strain 908. *J. Bacteriol.* **192**:6488–6489.
9. Haghjoo E, Galan JE. 2004. *Salmonella typhi* encodes a functional cytolethal distending toxin that is delivered into host cells by a bacterial internalization pathway. *Proc. Natl. Acad. Sci. U. S. A.* **101**:4614–4619.
10. Hirose K, et al. 1997. Survival of Vi-capsulated and Vi-deleted *Salmonella typhi* strains in cultured macrophage expressing different levels of CD14 antigen. *FEMS Microbiol. Lett.* **147**:259–265.
11. Lagesen K, et al. 2007. RNAmmer: consistent annotation of rRNA genes in genomic sequences. *Nucleic Acids Res.* **35**:3100–3108.
12. Larsen TS, Krogh A. 2003. EasyGene—a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics* **4**:21. doi:10.1186/1471-2105-4-21.
13. Liu SL, Sanderson KE. 1995. Rearrangements in the genome of the bacterium *Salmonella typhi*. *Proc. Natl. Acad. Sci. U. S. A.* **92**:1018–1022.
14. Nair S, Schreiber E, Thong KL, Pang T, Altwegg M. 2000. Genotypic characterization of *Salmonella typhi* by amplified fragment length polymorphism fingerprinting provides increased discrimination as compared to pulsed-field gel electrophoresis and ribotyping. *J. Microbiol. Methods* **41**:35–43.
15. Pang T, Bhutta ZA, Finlay BB, Altwegg M. 1995. Typhoid fever and other salmonellosis: 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.

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. AJGK0000000. The version described herein is the first version, AJGK0100000. 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.

Copyright © 2012, American Society for Microbiology. All Rights Reserved.

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 Chakravorty for critical reading of the manuscript. All other members of the Ahmed and Thong laboratories are gratefully acknowledged for their help and support.

REFERENCES

- Ahmed N. 2011. Coevolution and adaptation of *Helicobacter pylori* and the case for 'functional molecular infection epidemiology'. *Med. Princ. Pract.* 20:497–503.
- Alekshun MN, Levy SB. 1997. Regulation of chromosomally mediated multiple antibiotic resistance: the mar regulon. *Antimicrob. Agents Chemother.* 41:2067–2075.
- Alekshun MN, Levy SB. 1999. The mar regulon: multiple resistance to antibiotics and other toxic chemicals. *Trends Microbiol.* 7:410–413.
- 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.
- Avasthi TS, et al. 2011. Genome of multidrug-resistant uropathogenic *Escherichia coli* strain NA114 from India. *J. Bacteriol.* 193:4272–4273.
- Aziz RK, et al. 2008. The RAST server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi:10.1186/1471-2164-9-75.
- 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.
- Crump JA, Mintz ED. 2010. Global trends in typhoid and paratyphoid fever. *Clin. Infect. Dis.* 50:241–246.
- Crump JA, Luby SP, Mintz ED. 2004. The global burden of typhoid fever. *Bull. World Health Organ.* 82:346–353.
- Deng W, et al. 2003. Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18. *J. Bacteriol.* 185:2330–2337.
- Devi SH, et al. 2010. Genome of *Helicobacter pylori* strain 908. *J. Bacteriol.* 192:6488–6489.
- Everest P, Wain J, Roberts M, Rook G, Dougan G. 2001. The molecular mechanisms of severe typhoid fever. *Trends Microbiol.* 9:316–320.
- Holt KE, et al. 2008. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nat. Genet.* 40:987–993.
- Hongo E, et al. 1994. The methyl viologen-resistance-encoding gene *smvA* of *Salmonella typhimurium*. *Gene* 148:173–174.
- House D, Bishop A, Parry C, Dougan G, Wain J. 2001. Typhoid fever: pathogenesis and disease. *Curr. Opin. Infect. Dis.* 14:573–578.
- 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.
- Jones BD, Falkow S. 1996. Salmonellosis: host immune responses and bacterial virulence determinants. *Annu. Rev. Immunol.* 14:533–561.
- 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.
- Monack DM. 2012. *Salmonella* persistence and transmission strategies. *Curr. Opin. Microbiol.* 15:100–107.
- Parkhill J, et al. 2001. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* 413:848–852.
- Siddavattam D, et al. 2011. Genome of a novel isolate of *Paracoccus denitrificans* capable of degrading N,N-dimethylformamide. *J. Bacteriol.* 193:5598–5599.
- Thong KL, Cheong YM, Puthucherry 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.
- Thong KL, et al. 1995. Analysis of *Salmonella* Typhi from Southeast Asia by pulsed-field gel electrophoresis. *J. Clin. Microbiol.* 33:1938–1941.
- Thong KL, Puthucherry SD, Pang T. 1996. Genome size variation among recent human isolates of *Salmonella typhi*. *Res. Microbiol.* 148:229–235.
- 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.
- Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18:821–829.

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, India¹; Department of Biochemistry, Gulbarga University, Gulbarga, Karnataka, India²; Pathogen Biology Laboratory, Department of Biotechnology, School of Life Sciences, University of Hyderabad, Hyderabad, India³; Institute of Life Sciences, University of Hyderabad, Hyderabad, India⁴; and Institute of Biological Sciences, University of Malaya, Kuala Lumpur, Malaysia⁵

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 1.6) 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 selenocysteine) 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, 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.* **33**: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.

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, India¹;
Department of Microbiology, Dr. D. Y. Patil Medical College, Dr. D. Y. Patil University, Pimpri, Pune, India²; Institute of
Life Sciences, University of Hyderabad Campus, Hyderabad, India³; and Institute of Biological Sciences,
University of Malaya, Kuala Lumpur, Malaysia⁴

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. 1.6) 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 Nava-mallika.

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 β -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.



RESEARCH

Open Access

Genome anatomy of the gastrointestinal pathogen, *Vibrio parahaemolyticus* of crustacean origin

Suma Tiruvayipati^{1,2}, Subha Bhassu^{1,3*}, Narender Kumar², Ramani Baddam², Sabiha Shaik², Anil Kumar Gurindapalli², 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

¹Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia

²Pathogen 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 sero-variants 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 sero-variants 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.

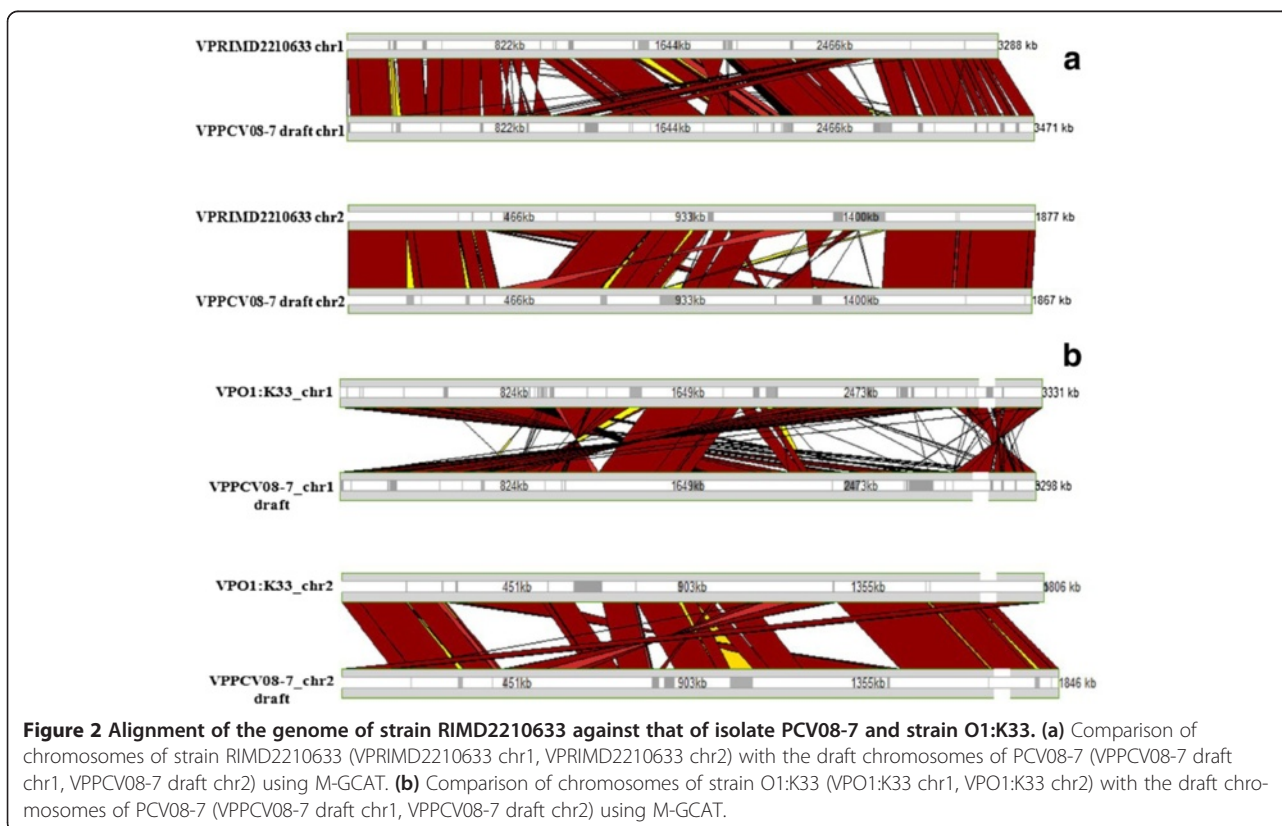
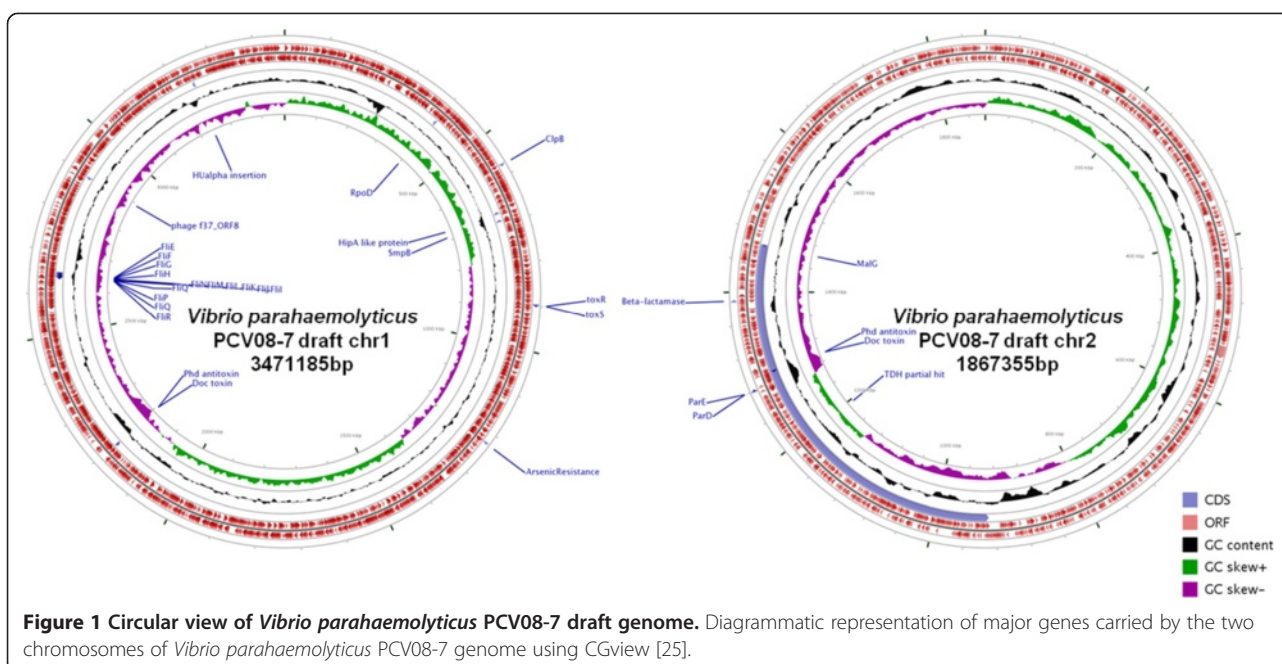
The sequencing reads obtained by us were primarily passed through a quality control step using FASTX toolkit [23] to obtain high quality reads free from adaptor and primer contamination which was further standardized to an optimal parameter p value of 70. High quality reads thus obtained were assembled 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

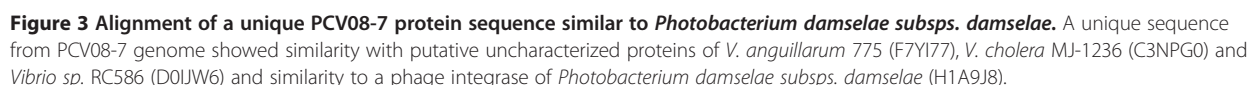
Genome assembly

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



RNAmmmer [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.



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 damsela* subsp. *damsela* 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 VPAl1-VPAl7 [37] to develop into a new pandemic clone of which VPAl4-VPAl6 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, VPAl2, was detected completely, whereas VPAl3 and VPAl7 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 (VPAl1-VPAl7) 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) <i>Vibrio parahaemolyticus</i> RIMD2210633	<i>V. parahaemolyticus</i> PCV08-7
VPAl1 (VP0380-0403)	Absent
Vpal2 (VP0635-0643)	Present
Vpal3 (VP1071-1094)	Partially present
Vpal4 (VP2131-2144)	Absent
Vpal5 (VP2900-2910)	Absent
Vpal6 (VPA1254-1270)	Absent
Vpal7 (VP1312-1398)	Present
Type VI secretion system (VP1386-1420)	Absent
(2) Other VP clusters	<i>V. parahaemolyticus</i> PCV08-7
VP1355-1368	Partially present
VPA0074-0089	Present
VPA0713-0732	Present
VPA1194-1210	Present

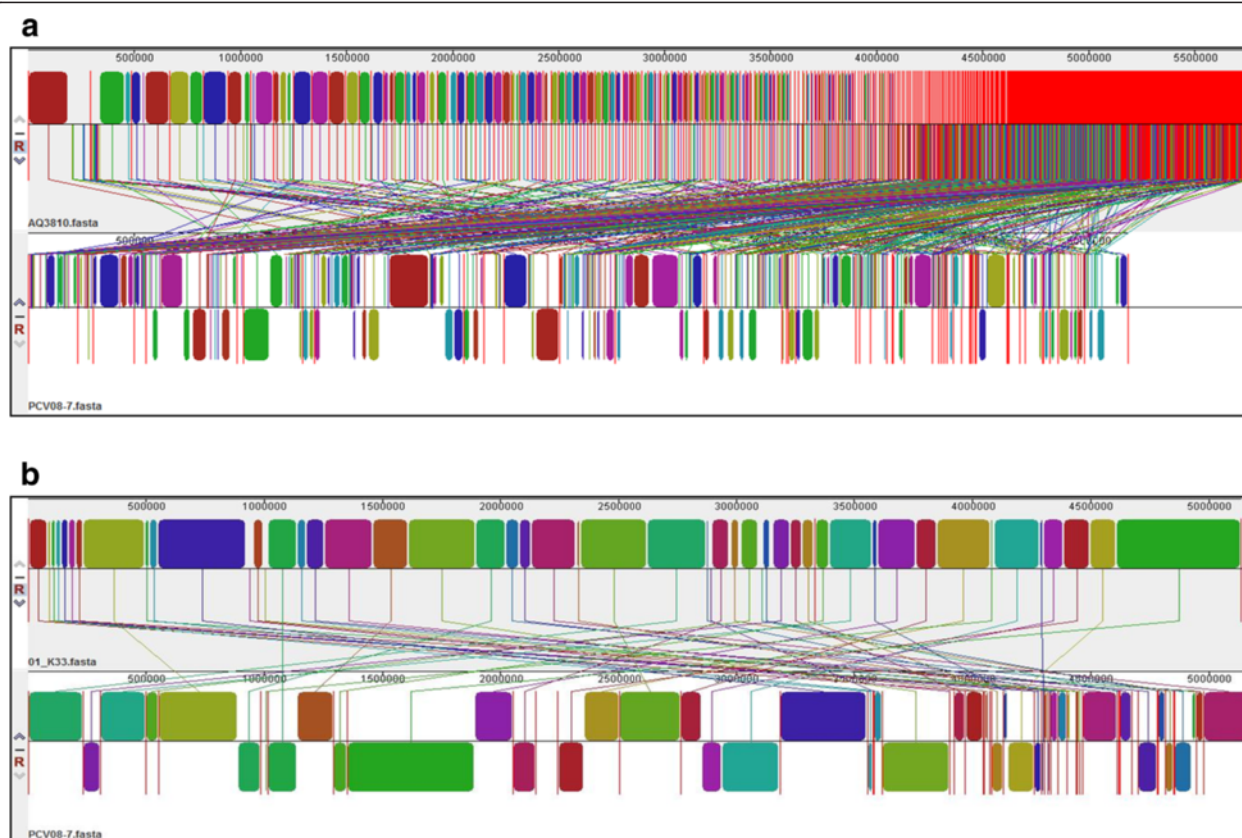


Figure 4 (a) *V. parahaemolyticus* AQ3810 alignment against PCV08-7 genome: Concatenated chromosome 1 and chromosome 2 of *V. parahaemolyticus* AQ3810 (AQ3810.fasta) against *V. parahaemolyticus* PCV08-7 (PCV08-7.fasta) (b) *V. parahaemolyticus* O1:K33 alignment against PCV08-7. Concatenated chromosome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

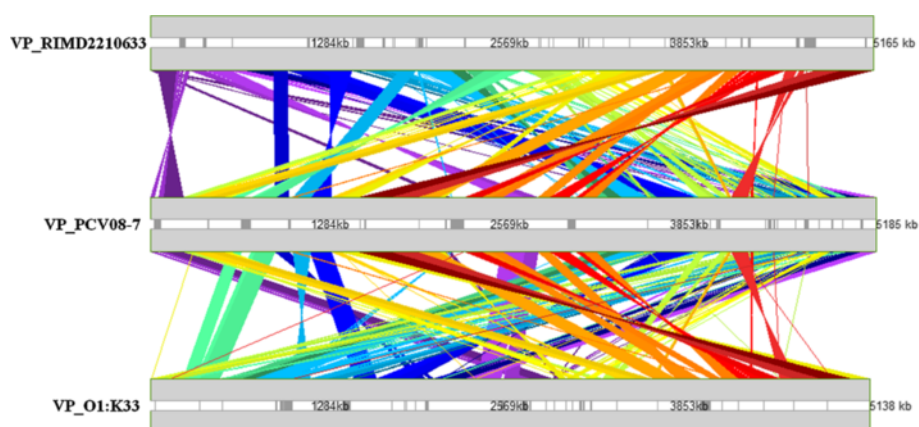


Figure 5 Comparison of whole genome sequences of strains RIMD2210633, PCV08-7 and O1:K33. Alignment of complete genomes of *V. parahaemolyticus* RIMD2210633, *V. parahaemolyticus* PCV08-7 and *V. parahaemolyticus* O1:K33, showing PCV08-7 being more similar to RIMD2210633.

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

¹Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia. ²Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad, Hyderabad, India. ³Center of Biotechnology for Agriculture (CEBAR), University of Malaya, Kuala Lumpur, Malaysia. ⁴Laboratory 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. Sujeeva 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* Superintegrans.** *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: **RNAmer: 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 persist cell formation and the general stress response.** *Appl Environ Microbiol* 2011, **77**(16):5577–5583.
36. Park KS, Ono T, Rokuda M, Jang MH, Okada K, Iida T, Honda T: **Functional characterization of two type III secretion systems of *Vibrio parahaemolyticus*.** *Infect Immun* 2004, **72**(11):6659–6665.
37. Hurley CC, Quirke A, Reen FJ, Boyd EF: **Four genomic islands that mark post-1995 pandemic *Vibrio parahaemolyticus* isolates.** *BMC Genomics* 2006, **7**:104.

doi:10.1186/1757-4749-5-37

Cite this article as: Tiruvayipati et al.: Genome anatomy of the gastrointestinal pathogen, *Vibrio parahaemolyticus* of crustacean origin. *Gut Pathogens* 2013 **5**:37.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit



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, India¹; Quantitative Biology Center, RIKEN, 1-7-22 Suehiro, Tsurumi, Yokohama, Kanagawa 230-0045, Japan²; Graduate School of Frontier Sciences, University of Tokyo, Tokyo, Japan³; CHU de Bordeaux, Hospital Saint André, 33077 Bordeaux, France⁴; Université de Bordeaux, Laboratoire de Bactériologie, 33000 Bordeaux, France⁵; INSERM U853, 33076 Bordeaux Cedex, France⁶; Institute of Life Sciences, University of Hyderabad Campus, Hyderabad, India⁷; and Institute of Biological Sciences, University of Malaya, Kuala Lumpur, Malaysia⁸

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 recrudescence 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 1.6) 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 ref-

erence 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 *cagPAI*, 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 Niranjana of Genotypic for assistance with resequencing of the genome of strain 908.

REFERENCES

- Ahmed, N. 2010. Replicative genomics can help *Helicobacter* fraternity usher in good times. *Gut Pathog.* **2**:25.
- Ahmed, N. 2009. A flood of microbial genomes—do we need more? *PLoS One* **4**:e5831.
- Ahmed, N., S. Tenguria, and N. Nandanwar. 2009. *Helicobacter pylori*—a seasoned pathogen by any other name. *Gut Pathog.* **1**:24.
- 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.
- 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.
- 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.
- Alm, R. A., et al. 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397**:176–180.
- 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.
- Atherton, J. C., and M. J. Blaser. 2009. Coadaptation of *Helicobacter pylori* and humans: ancient history, modern implications. *J. Clin. Invest.* **119**:2475–2487.
- Aziz, R. K., et al. 2008. The RAST server: Rapid Annotations using Subsystems Technology. *BMC Genomics* **9**:75.
- Baltrus, D. A., et al. 2009. The complete genome sequence of *Helicobacter pylori* strain G27. *J. Bacteriol.* **191**:447–448.
- 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.
- Besemer, J., and M. Borodovsky. 2005. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.* **33**:W451–W454.
- 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.
- Devi, S. H., et al. 2010. Genome of *Helicobacter pylori* strain 908. *J. Bacteriol.* **192**:6488–6489.
- 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.
- Gressmann, H., et al. 2005. Gain and loss of multiple genes during the evolution of *Helicobacter pylori*. *PLoS Genet.* **1**:e43.
- 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.
- 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.
- Kent, W. 2002. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**:656–664.
- 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.
- Kersulyte, D., H. Chalkauskas, and D. E. Berg. 1999. Emergence of recombinant strains of *Helicobacter pylori* during human infection. *Mol. Microbiol.* **31**:31–43.
- Larsen, T. S., and A. Krogh. 2003. EasyGene—a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics* **4**:21.
- 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.
- 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.
- Rutherford, K., et al. 2000. Artemis: sequence visualization and annotation. *Bioinformatics* **16**:944–945.
- Tomb, J. F., et al. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**:539–547.
- 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.
- Zerbino, D. R., and E. Birney. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**:821–829.