324–335 Nucleic Acids Research, 2015, Vol. 43, No. 1 Published online 1 December 2014
doi: 10.1093/nar/gku1271

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

Narender Kumar[1], Vanitha Mariappan[2], Ramani Baddam[1], Aditya K. Lankapalli[1], Sabiha Shaik[1], Khean-Lee Goh[3], Mun Fai Loke[2], Tim Perkins[4], Mohammed Benghezal[4], Seyed E. Hasnain[5], Jamuna Vadivelu[2], Barry J. Marshall[4] and Niyaz Ahmed[1,6,*]

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

Received September 30, 2014; Revised November 12, 2014; Accepted November 19, 2014

ABSTRACT

The discordant prevalence of *Helicobacter pylori* and its related diseases, for a long time, fostered certain enigmatic situations observed in the countries of the southern world. Variation in *H. pylori* infection rates and disease outcomes among different populations in multi-ethnic Malaysia provides a unique opportunity to understand dynamics of host–pathogen interaction and genome evolution. In this study, we extensively analyzed and compared genomes of 27 Malaysian *H. pylori* isolates and identified three major phylogeographic lineages: hspEastAsia, hpEurope and hpSouthIndia. The analysis of the virulence genes within the core genome, however, revealed a comparable pathogenic potential of the strains. In addition, we identified four genes limited to strains of East-Asian lineage. Our analyses identified a few strain-specific genes encoding restriction modification systems and outlined 311 core genes possibly under differential evolutionary constraints, among the strains representing different ethnic groups. The *cagA* and *vacA* genes also showed variations in accordance with the host genetic background of the strains. Moreover, restriction modification genes were found to be significantly enriched in East-Asian strains. An understanding of these variations in the genome content would provide significant insights into various adaptive and host modulation strategies harnessed by *H. pylori* to effectively persist in a host-specific manner.

INTRODUCTION

Helicobacter pylori, the human gastric pathogen, colonizes almost 50% of the world population (~70% of the population of developing countries and ~40% of developed countries) (1,2). It is a major etiological agent for a wide range of gastric diseases such as gastritis, peptic ulcers, gastric carcinoma and mucosa-associated lymphoid tissue lymphoma (3,4). Generally acquired during the childhood (intra-familial transfer), *H. pylori* establishes a lifelong persistent infection unless cleared by antibiotics (5).

The analysis of *H. pylori* strains has revealed existence of populations that are geographically localized (6,7). These populations have been classified based on multi-locus sequence typing (MLST) (7–11) into seven major lineages or genotypes depending on their regional prevalence: hpEurope, hpSahul, hpEastAsia, hpAfrica1, hpAfrica2, hpWestAfrica and hpAsia2. The ability to undergo frequent mutation and recombination serves as one of the major contributors for the observed genetic heterogeneity among various *H. pylori* isolates (12–14). It also allows the bacterium to quickly adapt to the changing gastric niches and establish a persistent infection. Certain countries with people of different ethnicities, cultures, lifestyles and religions present a pertinent model to examine the effects of migration and co-evolution on bacteria–host interaction. Studies entailing such settings would provide a better understanding of the evolutionary and adaptive strategies employed by *H. pylori* which might aid in design of intervention strategies (15).

[*]To whom correspondence should be addressed. Tel: +91 40 23134585; Fax: +91 40 23134585; Email: niyaz.ahmed@uohyd.ac.in; ahmed.nizi@gmail.com

Malaysia is one such multicultural, developing nation with a population comprising four major ethnic groups: Malay/'Bumiputera', Chinese, Indians and others (http://www.statistics.gov.my). In general, the Malays are considered natives (Bumiputera) and are in majority. The Malaysian-Chinese comprise the second largest ethnic group and are documented to have migrated from Southern China while the Malaysian-Indian group is comprised of migrants from Southern India (16). Apart from these major ethnic groups, there are a number of indigenous groups ('Orang Asli') living together, particularly in East Malaysia, Sabah and Sarawak who do not share the same ethnic origin as the Malays (17).

Previous reports have shown high prevalence of *H. pylori* infection among the Malaysian-Indians (69–75%), followed by Malaysian-Chinese (45–60%) and Malays (8–43%), and a minuscule number of inter-racial/inter-community or inter-religion marriages result in a putatively reduced chance of cross-infection occurring between ethnic groups (18,19). A majority of the *H. pylori* isolates from Malays and Malaysian-Indians were suggested to be of a recent common origin, while those from Malaysian-Chinese exhibited East-Asian ancestry (6,19). Generally, the *H. pylori* isolate collections representative of Asian populations are composed of strains from hpEastAsia, hpEurope and hpAsia2 (10,19). A significant proportion of the Malay isolates were found similar to their Indian counterparts, suggesting a possible acquisition of *H. pylori* from Indians (19), although there is not enough genomic evidence to support this interpretation. Further, the reason for low prevalence of *H. pylori* infection among Malay/'Bumiputera' population remains unclear and is likely to involve a number of environmental, genetic and host-related factors (20).

Recent sequencing efforts have reported multiple genome sequences of *H. pylori* isolates from patients of different ethnicities and various disease manifestations from Malaysia (21–23). To date, there are 29 Malaysian *H. pylori* genomes (27 clinical strains and two mice-adapted strains) available in NCBI database, a majority of them sequenced and deposited as a part of this work. In this study, we carried out an in-depth whole genome comparative analysis of 27 clinical isolates. The comparison of their core and accessory gene pools demonstrated close similarity among the strains according to their respective host genetic backgrounds. The status of various virulence genes and outer membrane proteins (OMPs) was also compared among the strains in order to unleash novel co-ordinates of adaptive evolution. The study aimed at understanding the genomic heterogeneity among these isolates and their possible role in observed enigmas related to disease outcomes in the region. Further, the analysis of strain-specific genes would allow us to better understand the disease biology and might open avenues for developing effective control strategies.

MATERIALS AND METHODS

Strain collection and ethics approval

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

Bacterial culture and DNA isolation

The *H. pylori* isolates were cultured from gastric biopsies by inoculating them on chocolate agar fortified with 4% blood base agar No. 2 (Oxoid) containing defibrinated horse blood (Oxoid) and antimicrobials such as trimethoprim, vancomycin and polymyxin B added to it at standard concentrations. Primary cultures were kept for incubation for up to 10 days (with daily observation) at 37°C in an incubator with 10% CO_2. For isolation of pure cultures, a single colony was identified and sub-cultured on chocolate agar for 3–5 days. Morphological identification of *H. pylori* isolates was carried out based on microscopic features and based on characteristic biochemical tests for the detection of enzymes such as urease, oxidase and catalase. A plateful of *H. pylori* culture was suspended into 500 μl of Tris buffer. The suspension was centrifuged at 5000 rpm for 10 min and the resulting pellet was collected. The *H. pylori* DNA was isolated using the QIAamp DNA Mini kit (Qiagen) according to the manufacturer's instruction.

Genome sequencing, assembly and annotation

Whole genome sequencing of the collected strains was carried out on Illumina GAIIx sequencer. The 100-bp paired-end sequencing run generated ~1-GB read data per strain with an average insert size of ~400 bp. The raw reads were then filtered using NGS QC toolkit (threshold quality >20) and were assembled into contigs using Velvet *de novo* assembler (24,25). The contigs were aligned against the NCBI refSeq database to identify a suitable complete genome as a reference. The contigs were sorted and re-oriented according to the chosen reference using in-house written scripts utilizing BLASTn. The sort-order was subjected to manual curation based on the paired-end information which helped us to order most of the unaligned contigs. These ordered contigs were joined together to form a draft genome by inserting a linker sequence (NNNNNCACACACTTAATT AATTAAGTGTGTGNNNNN) encoding start and stop codons in all six frames at the ends. The draft genomes were submitted for gene prediction and annotation to RAST annotation server, and the results were validated using GeneMarkS, Easygene and Glimmer (26–31). Genome statistics of respective strains were obtained through Artemis (32). tRNAs were identified using tRNAscan-SE program while rRNA genes were identified by using RNAmmer program (33,34).

Functional annotation, calculation of core and specific content

The predicted genes were scored using BLASTp against a protein database consisting of genes from 36 complete genomes from the NCBI refseq database. The output was filtered with an identity and query coverage of 90 and 70%, respectively. The proteins were then assigned functional categories based on their best hit obtained in the Basic Local Alignment Search Tool (BLAST) alignment. The other 22 draft genomes reported from Malaysia were downloaded and their gene prediction was performed as mentioned in the previous section. The core genome was calculated by identifying orthologs in every genome by applying Markov cluster (MCL) algorithm included in the OrthoMCL program (35). The parameters for deciding orthologs such as identity and e-value cutoff were set to 80% and 0.00001, respectively. The genes with less than 50 amino acids were excluded from the analysis. The clusters that contained orthologs in all the strains constituted the core, while those that did not have corresponding ortholog in any of the other genomes were considered as strain specific. The identified gene clusters were assigned functional categories after comparison with the COG database using rpsBLAST program followed by manual curation of the results (36,37).

Phylogenetic analysis

A whole genome alignment of 27 genomes of the Malaysian isolates was carried out with 43 other *H. pylori* genomes (draft or complete) from NCBI database using Gegenees tool (38,39). The tool utilizes a fragmented alignment algorithm to calculate average similarity among the compared genomes using BLASTn. The fragment size can be optimized according to the user. The tool was run with the fragment size set to 200 and a step size of 100 using BLASTn. The average similarity was calculated with a BLAST score threshold of 40% generating a heat plot matrix that was further used to deduce phylogenetic relationships exported in the form of a .nexus file. This nexus tree file was supplied as an input to SplitsTree (40) program for building an unrooted phylogenetic tree employing Neighbor-Joining algorithm.

Virulence genes and phage detection

The available whole genomes were screened for the presence of virulence genes enlisted in the Virulence Factor Database (VFDB) using BLAST program (41). The cutoff identity and query coverage was set to 70 and 60%, respectively. Further, comparison of the amino acid sequences of CagA and VacA was carried out using sequence-similarity-based alignments. Phage-related sequences in the genome were identified using PHAST server that integrates the analysis against various phage databases and compares key phage attributes to detect similar phage sequences in the query genome sequence (42).

RESULTS AND DISCUSSION

Genome statistics and phylogenetic analysis

The whole genome sequencing of five Malaysian isolates (Figure 1) revealed their chromosome sizes ranging from 1.56 to 1.62 Mb. The genomes also revealed a low G+C content of 39% which is characteristic of *H. pylori*. The draft genomes were predicted to encode ∼1600 genes with an average coding DNA sequence (CDS) measuring up to 930 bp. All the sequenced genomes harbored three rRNA operons as well as 36 tRNA genes. Two out of five sequenced strains (UM045 and UM054) also harbored phage sequences encoding 136 and 12 putative phage genes, respectively. A detailed genome statistics of these five isolates has been mentioned in Table 1 and a comparison with the remaining 22 genomes under the study is given in Supplementary Table S1. The genomes were also compared using BLASTn against a reference strain G27 as shown in Figure 1.

The genomes of sequenced isolates were pooled together with others from NCBI database to construct a whole genome based phylogenetic tree. The phylogenetic tree demonstrated a similar clustering pattern of various isolates as reported by other MLST-based phylogenetic trees (5,7,9). The strains co-clustered according to their genetic relatedness, exhibited by the formation of distinct clusters, and could be grouped according to their geographical affinities as shown in Figure 2. The strains affiliated to European countries formed hpEurope cluster, whereas those from African continent clustered into hpAfrica1 and hpAfrica2. The East-Asian genotype (hpEastAsia) has been further subdivided into three subpopulations: hspEastAsia (found in Japan and China), hspAmerind (found among Native Americans) and hspMaori (found among Taiwanese aboriginals, Melanesians and Polynesians) (43). Although studies based on MLST have indicated existence of three lineages (6,10,19) in South Asia: hpEurope, hpAsia2 and hpEastAsia, a comprehensive understanding of their phylogeny, evolution and adaptation could not be achieved perhaps because of scarcity of available genome sequences. A rapid increase in the number of genome sequences being available provided a better opportunity to achieve greater resolution in classifying *H. pylori* strains by whole genome comparative studies.

The isolates from South India and those from Malaysian-Indians clustered tightly forming a group that we named as hpSouthIndia. This close phylogenetic association among Indian and Malaysian-Indian strains is in accordance with the findings of Tay *et al.* based on MLST typing (19). Moreover, UM045 and UM037 isolated from Malay and Malaysian-Indian patients, respectively, clustered with hpEurope genotype suggesting their European ancestry (11). Further analysis of only Malaysian *H. pylori* genomes also revealed a bipartite clustering that was supplemented by a similarity score matrix, as shown in Figure 3. All the strains of Malay and Malaysian-Indian (European) origin exhibited more similarity to each other allowing them to cluster away from the Malaysian-Chinese (East-Asian) strains. These findings also appear to support the hypothesis of ancient human migration entailing ancestral Indians (11) and their subsequent migration to South Asia including Malaysia (6,19). Similarly, the affinity of Malaysian-Chinese strains with hspEastAsia reiterated their common ancestry. On the whole, the phylogenetic analysis explained a mixed population genetic structure of *H. pylori* existing in Malaysian population. These differential genotypes might explain the observed discrepancy in the colonization rates and disease outcome among various ethnic groups (44–46).

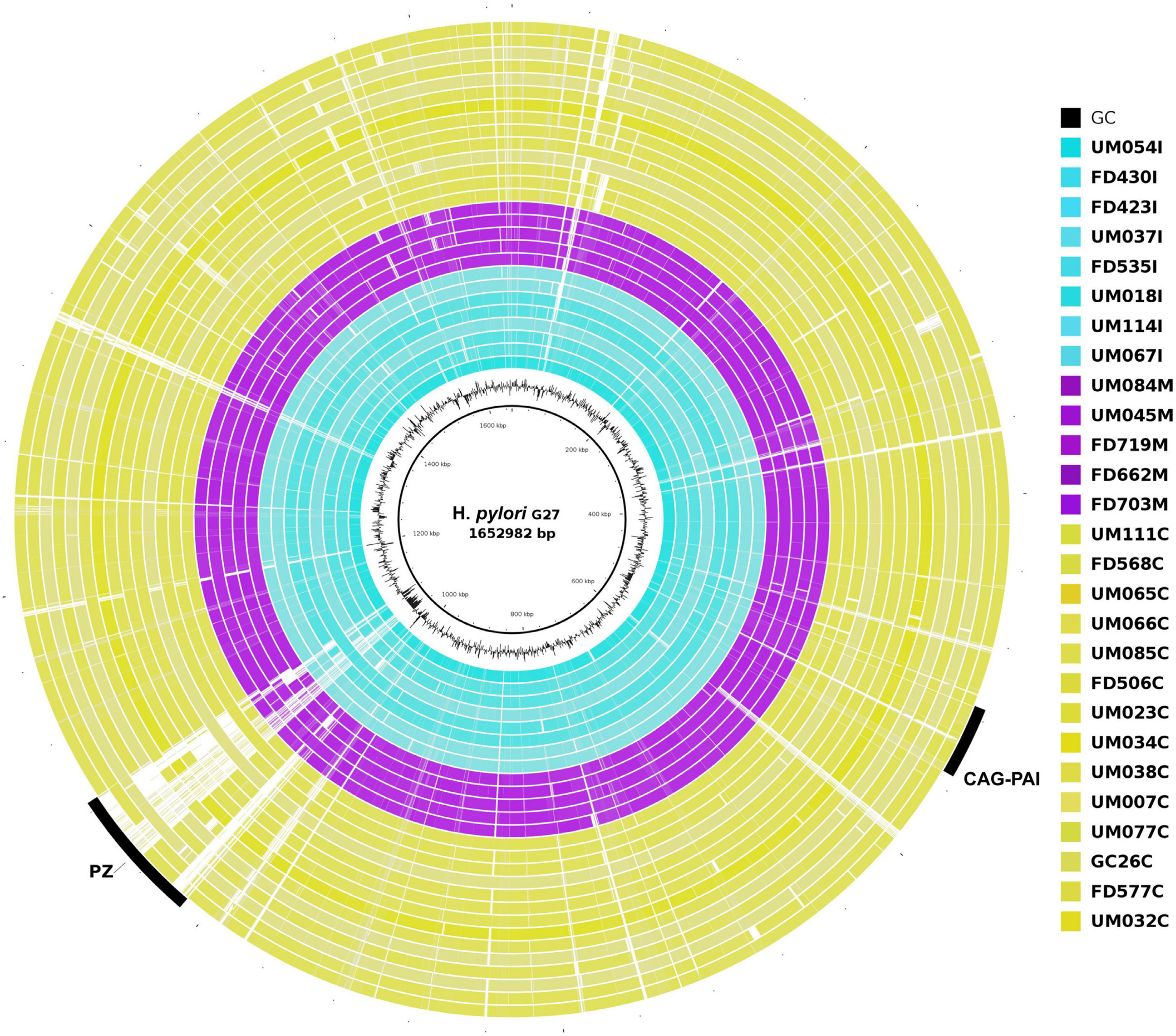


Figure 1. A circular representation of the genomes of Malaysian isolates: the draft genomes of 27 strains were aligned against the genome of reference strain *H. pylori* G27. Each genome is represented by a ring. The yellow rings represent *H. pylori* genomes from Malaysian-Chinese, purple represents those from Malays and light blue represents genomes of Malaysian-Indian strains. The G+C content (%) of the reference genome (strain G27) is represented by a ragged inner circle in black (GC). The variable regions such as plasticity zones (PZ) and *cag*PAI are compared across all the genomes using BRIG image generator (http://brig.sourceforge.net).

Virulence potential

The observed phylogenetic distinction among the Malaysian isolates was further investigated for the presence of differential virulence gene content. Various comparative studies have reported high polymorphism among different *H. pylori* lineages. Therefore, we sought to analyze the status of OMPs among 27 Malaysian *H. pylori* isolates. All the Malaysian genomes revealed a conserved nature for most of the 62 OMPs with minor exceptions. The BLASTn similarity percentage for these genes varied from 84 to 100 indicating their polymorphic nature. Few of the genes such as *hopZ*, *hopMN*, *hopQ* (*sabB*) varied among the strains, but we could not succeed in identifying a lineage/group-specific pattern among the East-Asian and other strains. The genes such as *homA* and *homB* were also found to variably exist among the genomes. The status of genes encoding these OMPs has been shown in Supplementary Table S2. Some of these genes correspond to critical virulence determinants induced upon host cell contact. These OMPs play an important role in adhesion and are reported to be associated with increased pro-inflammatory responses (47). OMPs in *H. pylori* have been categorized into five different families based on their structural composition and are known to carry out various functions ranging from host–surface interactions to non-selective porins for import of ions (48,49).

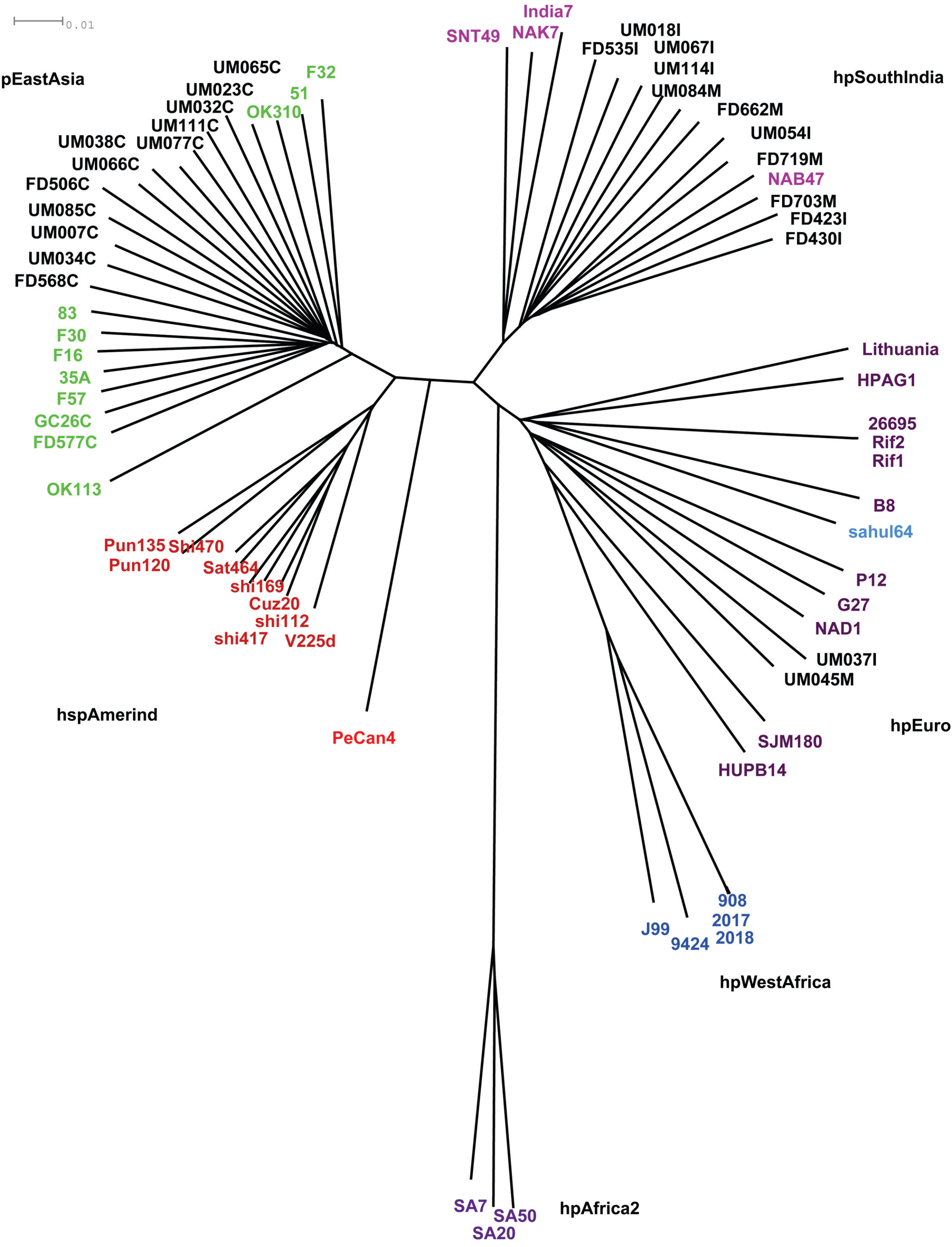
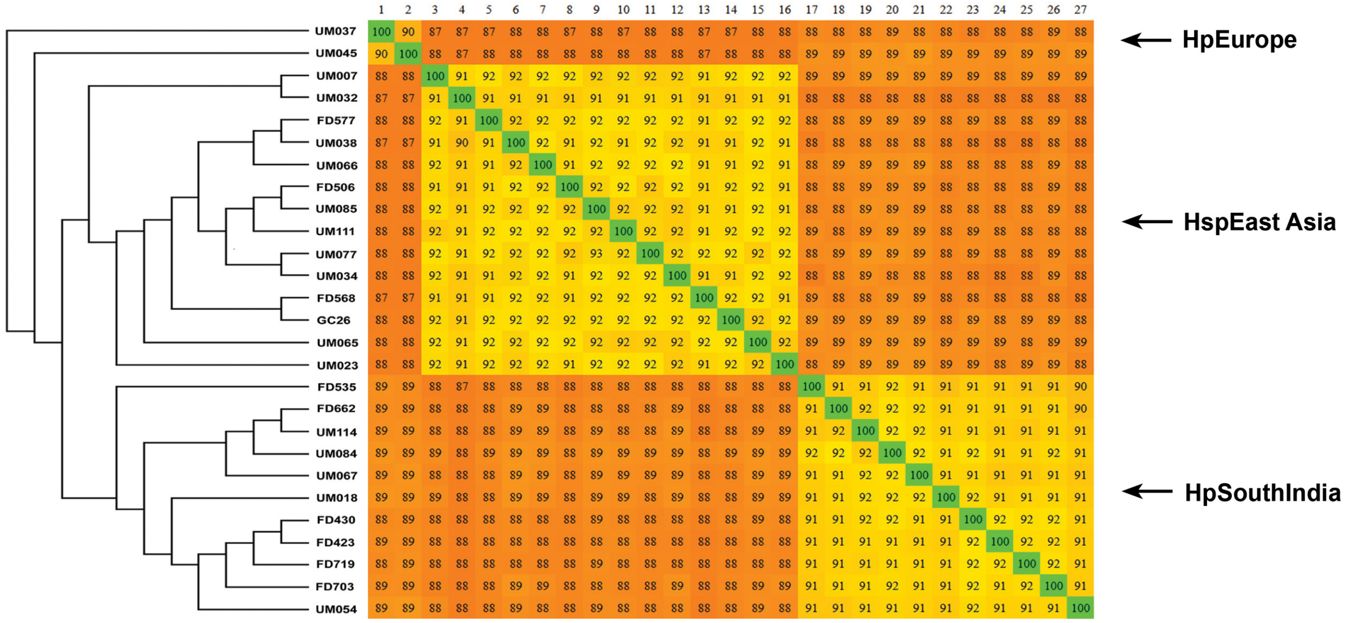


Figure 2. The whole genome phylogenetic analysis: the figure represents a whole genome comparison-based phylogenetic tree of various complete and draft *H. pylori* genomes from different geographical regions. The tree was constructed based on neighbor joining algorithm using SplitsTree. The Malaysian strains used in the analysis are labeled in black whereas the other genomes are colored to represent their genotypes.

Table 1. Genome statistics of the sequenced Malaysian *H. pylori* isolates

	UM018	UM054	UM007	UM034	UM045
Origin	Malaysian-Indian	Malaysian-Indian	Chinese	Chinese	Malay
Avg. genome coverage	170X	170X	150X	180X	200X
No. of contigs	72	89	80	27	28
Genome size	1 617 433	1 603 218	1 568 678	1 714 278	1 623 876
G+C	39.05	39.12	38.84	38.59	38.96
CDS	1579	1585	1557	1669	1595
Avg. length	937	926	924	940	933
Coding%	91.5	91.6	91.7	91.5	91.7
rRNA	3	4	3	4	3
tRNA	36	36	36	36	36
cagA (EPIYA-motif)	AB-C	AB-C	AB-D	AB-D	AB-C
vacA	s1m1	s2m2	s1m2	s1m2	s1m2
Prophage	Absent	Present (incomplete)	Absent	Absent	Present (intact)



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
UM037	100	90	87	87	87	88	88	87	88	87	88	88	87	87	88	88	88	88	88	89	88	88	88	88	88	89	88
UM045	90	100	88	87	88	88	88	88	88	88	88	88	87	88	88	88	89	89	89	89	89	89	89	89	89	89	89
UM007	88	88	100	91	92	92	92	92	92	92	92	92	91	92	92	92	89	89	89	89	89	89	89	88	89	89	89
UM032	87	87	91	100	91	91	91	91	91	91	91	91	91	91	91	91	88	88	88	88	88	88	88	88	88	88	88
FD577	88	88	92	91	100	92	92	92	92	92	92	92	91	92	92	92	88	88	89	89	89	88	89	88	88	89	88
UM038	87	87	91	90	91	100	92	91	92	91	92	92	91	91	92	91	88	88	88	88	88	88	88	88	88	88	88
UM066	88	88	92	91	91	92	100	91	92	92	92	92	91	91	92	91	88	89	89	89	89	88	88	88	88	89	88
FD506	88	88	91	91	91	92	92	100	92	92	92	92	91	92	92	91	88	88	89	89	89	88	88	88	88	89	88
UM085	88	88	92	91	92	92	92	92	100	92	92	92	91	91	92	91	88	88	89	89	89	88	88	88	88	89	88
UM111	88	88	92	91	92	92	92	92	92	100	92	92	91	92	92	92	89	88	89	89	89	88	89	88	88	88	88
UM077	88	88	92	91	92	92	92	92	93	92	100	92	92	92	92	92	88	89	89	89	89	88	89	88	88	89	88
UM034	88	88	92	91	91	92	92	91	92	92	92	100	91	91	92	92	88	88	89	88	88	88	88	88	88	89	88
FD568	87	87	91	91	91	92	92	91	92	92	92	92	100	92	92	91	89	88	88	89	89	88	88	88	88	88	88
GC26	88	88	92	91	92	92	92	92	92	92	92	92	92	100	92	92	89	89	89	89	89	88	89	88	89	89	88
UM065	88	88	92	91	92	92	92	92	92	92	92	92	92	92	100	92	89	89	89	89	89	89	89	89	88	89	89
UM023	88	88	92	91	92	92	92	91	92	92	92	92	91	92	92	100	88	89	89	89	89	89	89	88	89	89	88
FD535	89	89	88	87	88	88	88	88	88	88	88	88	88	88	88	88	100	91	91	92	91	91	91	91	91	91	90
FD662	89	89	88	88	88	89	89	88	88	88	88	89	88	88	88	88	91	100	92	92	92	91	91	91	91	91	90
UM114	89	89	88	88	88	89	89	88	89	88	88	89	88	88	89	89	91	92	100	92	92	91	91	91	91	91	91
UM084	89	89	89	88	89	89	89	88	89	89	89	89	88	89	89	89	92	92	92	100	92	91	92	91	91	92	91
UM067	89	89	88	88	88	89	89	88	89	89	89	89	88	88	89	89	91	91	92	92	100	91	91	91	91	91	91
UM018	89	89	89	88	88	89	89	88	89	89	88	89	88	88	89	89	91	91	92	92	92	100	92	91	91	91	91
FD430	88	89	88	88	88	88	88	88	89	88	88	88	88	88	89	88	91	91	92	92	91	91	100	92	92	92	91
FD423	89	89	88	88	88	88	88	88	88	88	88	88	88	88	88	88	91	91	91	91	91	91	92	100	92	92	91
FD719	88	89	88	88	88	88	88	88	89	88	88	88	88	88	88	88	91	91	91	91	91	91	92	92	100	92	91
FD703	89	89	88	88	88	89	89	88	88	88	88	89	88	88	89	89	91	91	91	92	91	91	92	91	92	100	91
UM054	89	89	88	88	88	89	88	88	89	88	88	88	88	88	89	88	91	91	91	91	91	91	92	91	91	91	100

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

The high recombination capability of *H. pylori* (50) and its natural competence (51) makes it difficult to draw conclusive inferences about its virulence apparatus. Therefore, various computational and functional efforts have revealed a number of genes implicated in pathogenesis of *H. pylori*. The virulence factor database (VFDB) (41) lists all the reported and predicted virulence markers for pathogenic organisms including *H. pylori*. We determined the status of all 57 virulence markers in the Malaysian *H. pylori* genomes using BLASTp (Supplementary Table S3). All the strains harbored intact *cag*PAI including a conserved *cagA* gene. Other virulence markers such as *oipA*, *vacA* and *flgG* which have been associated with severe disease phenotypes were also consistently present. Further, all the genomes possessed components of the urease cluster which allows *H. pylori* to survive under low pH conditions. The analysis thus revealed a high virulence potential encoded by the genomes irrespective of the ethnic groups that they represented. However, analysis of gene polymorphisms in *cag*A revealed lineage-specific patterns. CagA protein encoded by the *cag*PAI is highly correlated with severe gastric outcomes (45,52). The extraordinary virulence potential of CagA has earned its name as a bacterial ‘oncoprotein’ (53). Phylogenetic analysis of CagA could clearly differentiate East-Asian (Malaysian-Chinese) strains from their non-East-Asian (Malay and Malaysian Indian) counterparts (Supplementary Figure S1). The analysis of alignment revealed a lineage-specific variation not only at C-terminal EPIYA motifs but also at N-terminal region. The Malaysian-Indian and Malay strains possessed AB-C-type EPIYA motifs, whereas all the Malaysian-Chinese strains had AB-D-type motifs. These findings are in line with others and suggest a differential evolution of this protein among isolates of different lineages and its probable role in the observed disparity of the disease outcomes (13).

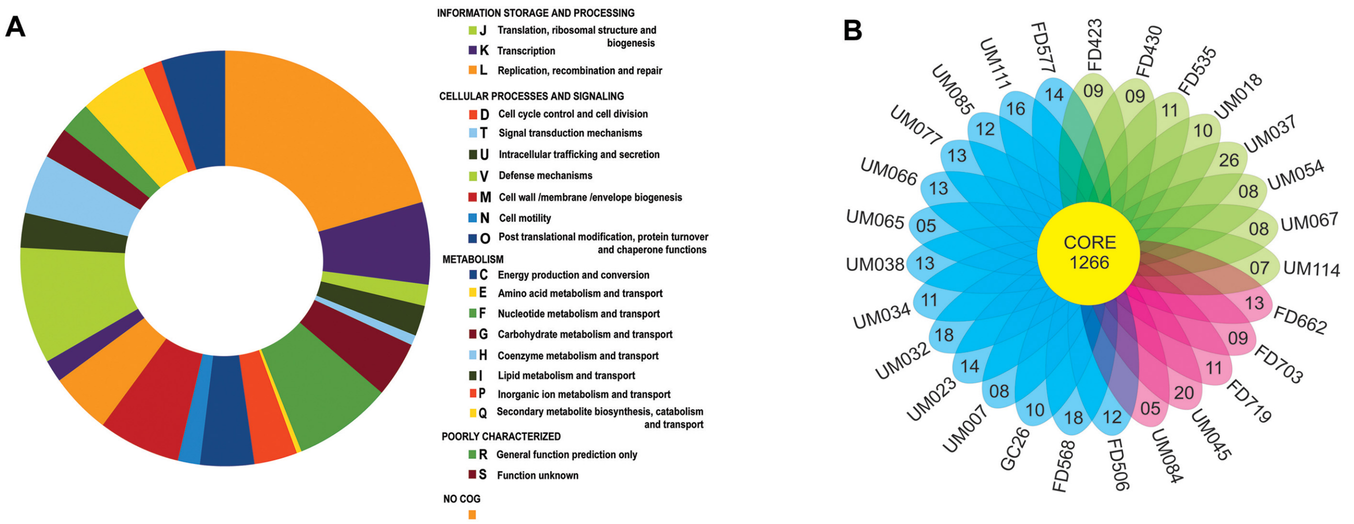


Figure 4. The functional COG classification of genes: (**A**) the COG functional classification representing core genome of the Malaysian *H. pylori* isolates. (**B**) Core and specific gene content observed among various strains compared in the study.

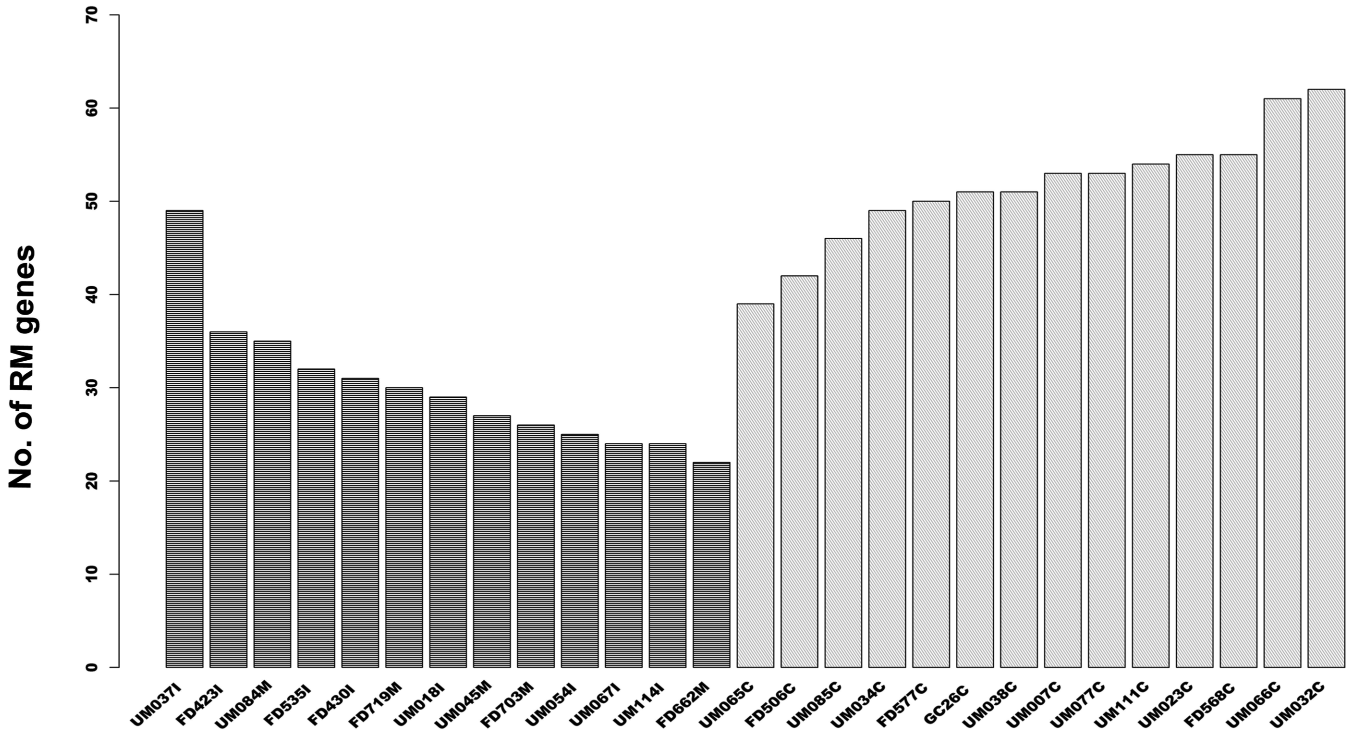


Figure 5. The distribution of RM genes in various strains: the graph shows the number of genes annotated to encode putative RM functions in each strain. The Y-axis represents the number of genes and the X-axis denotes strain names.

The core genome of Malaysian *H. pylori*

The gene content analysis of Malaysian isolates was carried out by calculating the core and accessory genome content. The genes that shared orthologs in all the genomes constituted the core while the accessory gene pool was constituted by those gene clusters which did not have orthologs in all the genomes. All the genes from the 27 strains formed a total of 1993 orthologous gene clusters. Among them, 1266 clusters comprised orthologs in all the genomes representing the core gene pool, whereas the remaining 727 gene clusters formed the variable or accessory gene pool which is also in accordance with the previous reports (54). Out of 1266 core gene clusters, 1005 clusters did find a significant match with the COG database and were assigned functional categories as shown in Figure 4A and the rest 261 gene clusters remained uncategorized. Among these 261 gene clusters, a majority were found to be encoding putative hypothetical proteins based on their comparison with other *H. pylori* genes. Out of 1005 functionally categorized gene clus-

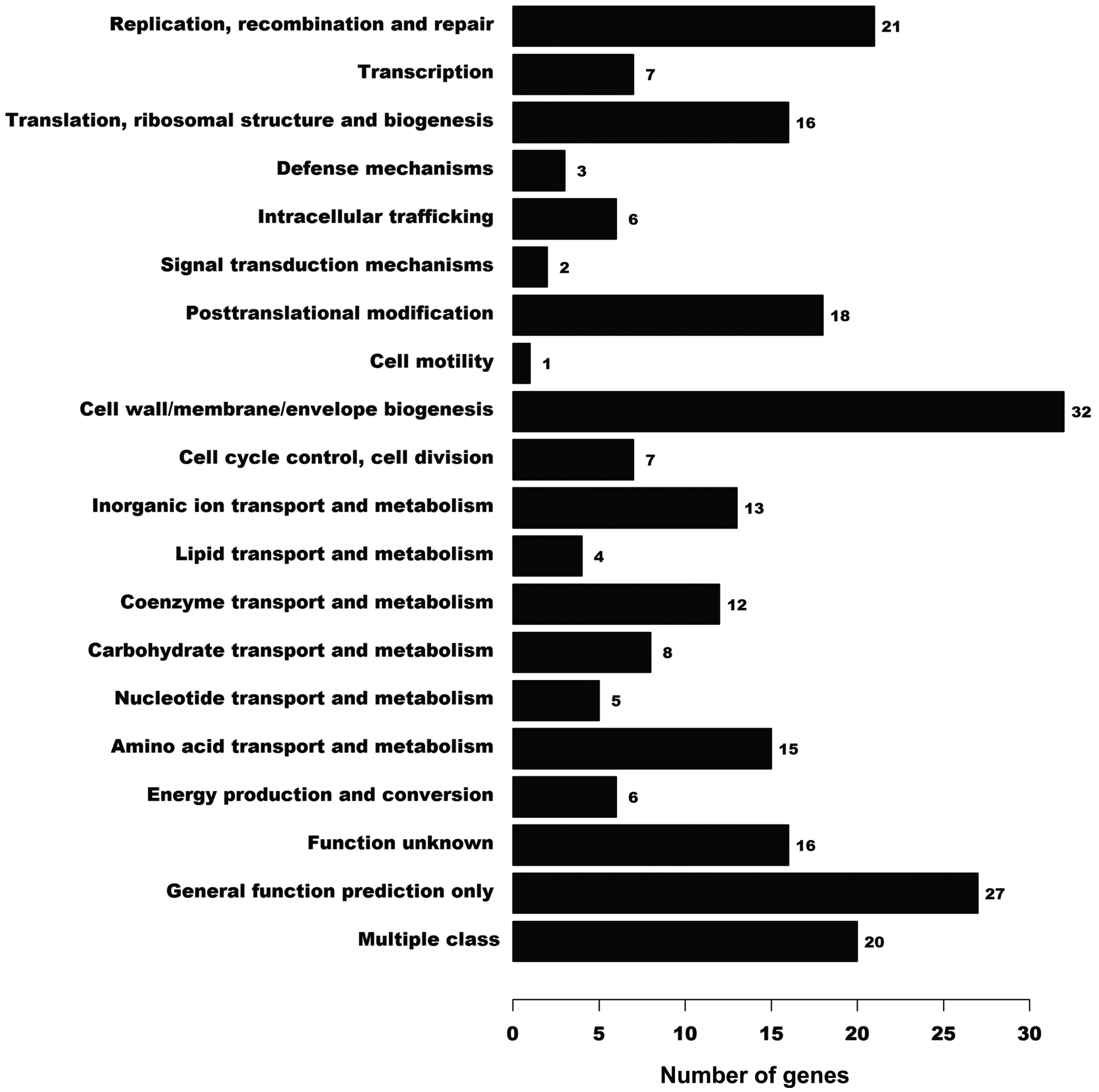


Figure 6. The functional classification of differentially evolving genes: the graph shows the COG functional classification of various core genes with signs of differential evolution among East-Asian and non-East-Asian strains. The Y-axis denotes the functional category and the X-axis represents the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

ters, 115 clusters encoded proteins involved in translation, ribosomal structure and biogenesis. Other than performing housekeeping functions, studies have shown that some proteins like Pol I also aid in generating genome plasticity (55). We found the core genome to be enriched with the genes related to cell wall biosynthesis and amino acid/ion transport. The existence of a significant proportion of these transport related genes may be suggestive of an increased dependence of *H. pylori* on the host metabolites which could possibly be a result of its long association with the host. Interestingly, 81 core gene clusters were identified as belonging to multiple functional classes. These genes could represent the proteins involved in multiple pathways (56). Multifunctional proteins could also be advantageous to *H. pylori* with a small genome and limited coding potential, but this requires further functional validation. Moreover, a high proportion of hypothetical proteins in the core genome also warrants their functional characterization to ascertain their role in the biology and pathogenesis of this gastric pathogen.

H. pylori has been reported to possess a high strain-specific gene content majorly localized in hypervariable regions known as plasticity zones (57). Few genes from plasticity regions have been reported to be associated with increased pro-inflammatory secretion in cell culture studies. Recent studies on *jhp0940* and *hp0986* have provided strong evidence for their role in induction of pathogenic phenotypes (58–60). Of the 27 Malaysian strains, 11 harbored *hp0986* and belonged to non-East-Asian genotype. Further, *jhp0940* was found to be present in seven of the East-Asian strains and one non-East-Asian strain. The presence of these genes also reflects their importance in the pathogenesis of this pathogen. The strain-specific content of Malaysian genomes varied from 10 to 12 genes per genome (Figure 4B). A high proportion of these strain specific genes was predicted to encode hypothetical proteins while few of them encoded putative restriction-modification (RM) systems in some strains. A total of 749 strain-specific genes were identified among the Malaysian strains of which 15 genes were found to encode putative type II restriction or modification related functions. Interestingly, these were mostly prevalent among the Malaysian-Chinese strains that appear highly virulent and therefore warrant further functional characterization of their possible role in the pathogenesis of *H. pylori.*

Lineage-specific genes

Our phylogenetic analysis classified Malaysian isolates into three distinct genotypes. The strains belonging to hpEurope shared a close similarity to hpSouthIndia compared to hspEastAsia strains. We divided the strains into two groups in accordance with their phylogenetic clustering and genomic identity. In total, 14 Malaysian-Chinese strains represented the East-Asian group, while 13 Malaysian-Indian and Malay strains constituted the non-East-Asian group. The core genome content was calculated for each group separately from the same orthologous cluster file. The East-Asian core genome possessed 1299 orthologous gene clusters, whereas 1301 gene clusters formed the core content among non-East-Asian genomes. The comparison of these two core genomes revealed 33 clusters conserved among East-Asian but varied among non-East-Asian genomes. Out of these 33 gene clusters, four gene clusters did not have orthologs in any of the non-East-Asian strains; one of them was predicted to encode a putative lysozyme-like protein (Table 2). The lysozyme-like proteins have been observed to be upregulated during DNA-damage-induced stress in *H. pylori* (50). The other three encoded hypothetical proteins await further functional characterization. A clear understanding of the proteins encoded by these genes could provide significant insights into the underlying distinction between East-Asian and non-East-Asian genotypes and associated disease outcomes (61).

RM genes

Previous studies on *H. pylori* genomes revealed a proportion of genes encoding RM systems (62). In our collection of 27 Malaysian genomes, a total of 1077 genes were predicted to encode RM-related genes. Further, clustering by UCLUST (63) with an identity of 80% arranged these genes into 149 clusters. We then analyzed the RM gene content of East-Asian (Malaysian-Chinese) and non-East-Asian (Malay and Malaysian-Indian) strains separately. It was observed that East-Asian strains together contained 698 genes, whereas non-East-Asian strains had only 379 genes. Thus, East-Asian strains harbored, on average, 52 RM genes per strain, much higher compared to 29 RM genes per strain for non-East-Asian strains (Figure 5). In addition, the distribution of RM genes among compared genomes revealed a higher proportion of genes in East-Asian strains as shown in Figure 5. This analysis, therefore, clearly outlines the extent of diversity both in terms of numbers and allelic diversity of RM genes present in *H. pylori*. This might explain the observed strain to strain diversity in *H. pylori*. Higher proportion of RM genes in case of East-Asian strains is striking and warrants further functional validation. The role of RM genes in regulating gene expression and virulence of *H. pylori* is being earnestly pursued. Recent studies have proved that inactivation of the RM genes leads to changes in the expression of several genes in *H. pylori* (64). Moreover, these RM genes have also been shown to exhibit phase variation (65). Therefore, a clear understanding of the roles played by these RM genes in a host/lineage-specific manner would be necessary to better understand the mechanisms of differential host adaptation in *H. pylori*.

Differentially evolving genes

It has been proposed that pathogenic bacteria that resort to long-term adaptation to a particular niche modulate their core gene repertoire in synchrony with their virulence complement to gain fitness advantage (66). Therefore, we attempted to identify core gene clusters that show some evidence of differential evolution between East-Asian and non-East-Asian strains. The core gene clusters were analyzed by constructing a gene-based phylogeny to search for those gene clusters which distinguished East-Asian strains from non-East-Asian strains. The analysis identified 311 out of 1266 core gene clusters with possible signs of differential evolution. Out of 311, only 239 genes could be assigned to functional categories while the rest did not find a significant hit with the COG database. Their functional categorization revealed an enrichment of genes with functions related to cell-wall/membrane biogenesis, recombination and repair, plus some others with poorly characterized functions (Figure 6). The latter included various OMPs that have been proven to be differentially evolving among East-Asian and non-East-Asian genomes (67). Even *cagA* and *vacA* that are known to be differentially evolving among East-Asian-type and non-East-Asian-type strains showed up in our analysis. The core genome thus possesses a significant number of differentially evolving genes. This also mirrors the differential adaptive and evolutionary pressures experienced by these isolates.

CONCLUSION AND FUTURE PERSPECTIVES

This study was aimed at understanding the genetic structure of *H. pylori* in Malaysia and cues obtained therefrom to gain insights into observed variation in disease outcomes. The

Table 2. Genes differentially present among the core of East-Asian (EA) and non-East-Asian (Non-EA) *H. pylori*

	Status in *H. pylori* strains			
Cluster ID	EA ($n = 14$)	Non-EA ($n = 13$)	Predicted functions	Orthologs in 26695
2426	14	0	Lysozyme family protein	HP0339
2425	14	0	Hypothetical protein	HP0344
2424	14	0	Hypothetical protein	HP0346
2423	14	0	Hypothetical protein	Absent
2403	14	2	Lipopolysaccharide biosynthesis protein	Absent
2376	14	4	Type II restriction endonuclease	HP1537
2375	14	4	Type II methylase	Absent
2370	14	5	Glycosyltransferase	Absent
2364	14	4	DNA methyltransferase	HP0051
2402	1	13	Hypothetical protein	Absent

whole genome phylogenetic analysis resolved the strains into three lineages representing patients/individuals from various ethnic groups in a multicultural setting such as Malaysia. The conservation of most of the virulence related genes in the core genome revealed a high pathogenic potential of the strains. Few genes were found to be more prevalent in East-Asian strains as compared to others but await further confirmation considering the draft status of the genomes we analyzed. Further investigation of the core gene pool revealed a significant proportion of genes differentially represented/evolving among East-Asian and non-East-Asian strains. These differentially evolving genes included RM genes and OMPs. Given these findings, it is tempting to believe that *H. pylori* could possibly harness various mechanisms like surface antigen variation and virulence gene regulation to effectively evade the inhospitable microenvironment of the host. A careful analysis of these molecular interactions would also open avenues for the development of specific control strategies and drug intervention for *H. pylori*. A functional level understanding of the preponderances and interplay of the virulence and core gene complements among different strains/lineages would allow us to gain better understanding of the pathogen biology and host–pathogen interactions in different endemic settings.

ACCESSION NUMBERS

The whole genome sequences of five Malaysian strains sequenced in this study have been submitted to NCBI genome database with the following accession numbers: UM018 (AONK00000000), UM054 (AONL00000000), UM007 (AONM00000000), UM034 (AONN00000000) and UM045 (AONO00000000).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

We would like to thank members of the Ahmed lab, Vadivelu lab and Marshall labs for constructive suggestions and discussions. N.A. is a Visiting Professor at the University of Malaya, Malaysia, and is an Adjunct Professor of the Academy of Scientific and Innovative Research (AcSIR), India. We are thankful to several researchers for the use of genome sequence data from NCBI.

FUNDING

Department of Biotechnology, Government of India [BT/PR6921/MED/29/699/2013 to N.A.]; Senior Research Fellowship, Council for Scientific and Industrial Research, India [to N.K.]; HIR Project of the University of Malaya [UM.C/625/1/HIR/MOHE/CHAN-02-Molecular Genetics to J.V., K.L.G., N.A., M.B.G., B.J.M.]. Funding for openaccess charge: University of Malaya [UM.C/625/1/HIR/MOHE/CHAN-02-Molecular Genetics to J.V., K.L.G., N.A., M.B.G., B.J.M.].
Conflict of interest statement. None declared.

REFERENCES

1. Perez-Perez,G.I., Rothenbacher,D. and Brenner,H. (2004) Epidemiology of *Helicobacter pylori* infection. *Helicobacter*, **9**(Suppl. 1), 1–6.
2. Khalifa,M.M., Sharaf,R.R. and Aziz,R.K. (2010) *Helicobacter pylori*: a poor man's gut pathogen? *Gut Pathog.*, **2**, 2.
3. Amieva,M.R. and El-Omar,E.M. (2008) Host-bacterial interactions in *Helicobacter pylori* infection. *Gastroenterology*, **134**, 306–323.
4. Cover,T.L. and Blaser,M.J. (2009) *Helicobacter pylori* in health and disease. *Gastroenterology*, **136**, 1863–1873.
5. Kuipers,E.J., Israel,D.A., Kusters,J.G., Gerrits,M.M., Weel,J., van Der Ende,A., van Der Hulst,R.W., Wirth,H.P., Hook-Nikanne,J., Thompson,S.A. *et al.* (2000) Quasispecies development of *Helicobacter pylori* observed in paired isolates obtained years apart from the same host. *J. Infect. Dis.*, **181**, 273–282.
6. Breurec,S., Guillard,B., Hem,S., Brisse,S., Dieye,F.B., Huerre,M., Oung,C., Raymond,J., Tan,T.S., Thiberge,J.M. *et al.* (2011) Evolutionary history of *Helicobacter pylori* sequences reflect past human migrations in Southeast Asia. *PloS one*, **6**, e22058.
7. Linz,B., Balloux,F., Moodley,Y., Manica,A., Liu,H., Roumagnac,P., Falush,D., Stamer,C., Prugnolle,F., van der Merwe,S.W. *et al.* (2007) An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature*, **445**, 915–918.
8. Falush,D., Wirth,T., Linz,B., Pritchard,J.K., Stephens,M., Kidd,M., Blaser,M.J., Graham,D.Y., Vacher,S., Perez-Perez,G.I. *et al.* (2003) Traces of human migrations in *Helicobacter pylori* populations. *Science*, **299**, 1582–1585.
9. Achtman,M., Azuma,T., Berg,D.E., Ito,Y., Morelli,G., Pan,Z.J., Suerbaum,S., Thompson,S.A., van der Ende,A. and van Doorn,L.J. (1999) Recombination and clonal groupings within *Helicobacter pylori* from different geographical regions. *Mol. Microbiol.*, **32**, 459–470.
10. Moodley,Y., Linz,B., Yamaoka,Y., Windsor,H.M., Breurec,S., Wu,J.Y., Maady,A., Bernhoft,S., Thiberge,J.M., Phuanukoonnon,S. *et al.* (2009) The peopling of the Pacific from a bacterial perspective. *Science*, **323**, 527–530.
11. Devi,S.M., Ahmed,I., Francalacci,P., Hussain,M.A., Akhter,Y., Alvi,A., Sechi,L.A., Megraud,F. and Ahmed,N. (2007) Ancestral

European roots of *Helicobacter pylori* in India. *BMC Genomics*, **8**, 184.
12. Blaser,M.J. and Berg,D.E. (2001) *Helicobacter pylori* genetic diversity and risk of human disease. *J. Clin. Invest.*, **107**, 767–773.
13. Suerbaum,S. and Josenhans,C. (2007) *Helicobacter pylori* evolution and phenotypic diversification in a changing host. *Nat. Rev. Microbiol.*, **5**, 441–452.
14. Suzuki,R., Shiota,S. and Yamaoka,Y. (2012) Molecular epidemiology, population genetics, and pathogenic role of *Helicobacter pylori*. *Infect. Genet. Evol.*, **12**, 203–213.
15. Stein,M., Ruggiero,P., Rappuoli,R. and Bagnoli,F. (2013) *Helicobacter pylori* CagA: from pathogenic mechanisms to its use as an anti-cancer vaccine. *Front. immunol.*, **4**, 328.
16. Andaya,L.Y. (2001) The search for the 'origins' of Melayu. *J. Southeast Asian Stud.*, **32**, 315–330.
17. Hill,C., Soares,P., Mormina,M., Macaulay,V., Meehan,W., Blackburn,J., Clarke,D., Raja,J.M., Ismail,P., Bulbeck,D. *et al.* (2006) Phylogeography and ethnogenesis of aboriginal Southeast Asians. *Mol. Biol. Evol.*, **23**, 2480–2491.
18. Tan,H.J., Rizal,A.M., Rosmadi,M.Y. and Goh,K.L. (2005) Distribution of *Helicobacter pylori* cagA, cagE and vacA in different ethnic groups in Kuala Lumpur, Malaysia. *J. Gastroenterol. Hepatol.*, **20**, 589–594.
19. Tay,C.Y., Mitchell,H., Dong,Q., Goh,K.L., Dawes,I.W. and Lan,R. (2009) Population structure of *Helicobacter pylori* among ethnic groups in Malaysia: recent acquisition of the bacterium by the Malay population. *BMC Microbiol.*, **9**, 126.
20. Lee,Y.Y., Mahendra Raj,S. and Graham,D.Y. (2013) *Helicobacter pylori* infection–a boon or a bane: lessons from studies in a low-prevalence population. *Helicobacter*, **18**, 338–346.
21. Rehvathy,V., Tan,M.H., Gunaletchumy,S.P., Teh,X., Wang,S., Baybayan,P., Singh,S., Ashby,M., Kaakoush,N.O., Mitchell,H.M. *et al.* (2013) Multiple genome sequences of *Helicobacter pylori* strains of diverse disease and antibiotic resistance backgrounds from Malaysia. *Genome Announcements*, **1**, doi:10.1128/genomeA.00687-13.
22. Khosravi,Y., Rehvathy,V., Wee,W.Y., Wang,S., Baybayan,P., Singh,S., Ashby,M., Ong,J., Amoyo,A.A., Seow,S.W. *et al.* (2013) Comparing the genomes of *Helicobacter pylori* clinical strain UM032 and Mice-adapted derivatives. *Gut Pathog.*, **5**, 25.
23. Gunaletchumy,S.P., Teh,X., Khosravi,Y., Ramli,N.S., Chua,E.G., Kavitha,T., Mason,J.N., Lee,H.T., Alias,H., Zaidan,N.Z. *et al.* (2012) Draft genome sequences of *Helicobacter pylori* isolates from Malaysia, cultured from patients with functional dyspepsia and gastric cancer. *J. Bacteriol.*, **194**, 5695–5696.
24. Patel,R.K. and Jain,M. (2012) NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PloS one*, **7**, e30619.
25. Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.
26. Overbeek,R., Olson,R., Pusch,G.D., Olsen,G.J., Davis,J.J., Disz,T., Edwards,R.A., Gerdes,S., Parrello,B., Shukla,M. *et al.* (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.*, **42**, D206–D214.
27. Besemer,J., Lomsadze,A. and Borodovsky,M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.*, **29**, 2607–2618.
28. Larsen,T.S. and Krogh,A. (2003) EasyGene—a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics*, **4**, 21.
29. Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.
30. Salzberg,S.L., Delcher,A.L., Kasif,S. and White,O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.*, **26**, 544–548.
31. Aziz,R.K., Devoid,S., Disz,T., Edwards,R.A., Henry,C.S., Olsen,G.J., Olson,R., Overbeek,R., Parrello,B., Pusch,G.D. *et al.* (2012) SEED servers: high-performance access to the SEED genomes, annotations, and metabolic models. *PloS one*, **7**, e48053.
32. Rutherford,K., Parkhill,J., Crook,J., Horsnell,T., Rice,P., Rajandream,M.A. and Barrell,B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.
33. Lagesen,K., Hallin,P., Rodland,E.A., Staerfeldt,H.H., Rognes,T. and Ussery,D.W. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.*, **35**, 3100–3108.
34. Schattner,P., Brooks,A.N. and Lowe,T.M. (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.*, **33**, W686–W689.
35. Li,L., Stoeckert,C.J. Jr and Roos,D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.
36. Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.
37. Natale,D.A., Galperin,M.Y., Tatusov,R.L. and Koonin,E.V. (2000) Using the COG database to improve gene recognition in complete genomes. *Genetica*, **108**, 9–17.
38. Agren,J., Sundstrom,A., Hafstrom,T. and Segerman,B. (2012) Gegenees: fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups. *PloS one*, **7**, e39107.
39. Kumar,N., Mukhopadhyay,A.K., Patra,R., De,R., Baddam,R., Shaik,S., Alam,J., Tiruvayipati,S. and Ahmed,N. (2012) Next-generation sequencing and de novo assembly, genome organization, and comparative genomic analyses of the genomes of two *Helicobacter pylori* isolates from duodenal ulcer patients in India. *J. Bacteriol.*, **194**, 5963–5964.
40. Huson,D.H. (1998) SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, **14**, 68–73.
41. Chen,L., Yang,J., Yu,J., Yao,Z., Sun,L., Shen,Y. and Jin,Q. (2005) VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.*, **33**, D325–D328.
42. Zhou,Y., Liang,Y., Lynch,K.H., Dennis,J.J. and Wishart,D.S. (2011) PHAST: a fast phage search tool. *Nucleic Acids Res.*, **39**, W347–W352.
43. Moodley,Y., Linz,B., Bond,R.P., Nieuwoudt,M., Soodyall,H., Schlebusch,C.M., Bernhoft,S., Hale,J., Suerbaum,S., Mugisha,L. *et al.* (2012) Age of the association between *Helicobacter pylori* and man. *PLoS Pathog.*, **8**, e1002693.
44. Kaur,G. and Naing,N.N. (2003) Prevalence and ethnic distribution of *Helicobacter pylori* infection among endoscoped patients in north eastern peninsular malaysia. *Malays. J. Med. Sci.*, **10**, 66–70.
45. Peek,R.M. Jr and Crabtree,J.E. (2006) Helicobacter infection and gastric neoplasia. *J. Pathol.*, **208**, 233–248.
46. Ahmed,N., Loke,M.F., Kumar,N. and Vadivelu,J. (2013) *Helicobacter pylori* in 2013: multiplying genomes, emerging insights. *Helicobacter*, **18**(Suppl. 1), 1–4.
47. Gerhard,M., Rad,R., Prinz,C. and Naumann,M. (2002) Pathogenesis of *Helicobacter pylori* infection. *Helicobacter*, **7**(Suppl. 1), 17–23.
48. Alm,R.A., Ling,L.S., Moir,D.T., King,B.L., Brown,E.D., Doig,P.C., Smith,D.R., Noonan,B., Guild,B.C., deJonge,B.L. *et al.* (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature*, **397**, 176–180.
49. Alm,R.A., Bina,J., Andrews,B.M., Doig,P., Hancock,R.E. and Trust,T.J. (2000) Comparative genomics of *Helicobacter pylori*: analysis of the outer membrane protein families. *Infect. Immun.*, **68**, 4155–4168.
50. Dorer,M.S., Fero,J. and Salama,N.R. (2010) DNA damage triggers genetic exchange in *Helicobacter pylori*. *PLoS Pathog.*, **6**, e1001026.
51. Dorer,M.S., Cohen,I.E., Sessler,T.H., Fero,J. and Salama,N.R. (2013) Natural competence promotes *Helicobacter pylori* chronic infection. *Infect. Immun.*, **81**, 209–215.
52. Wiedemann,T., Loell,E., Mueller,S., Stoeckelhuber,M., Stolte,M., Haas,R. and Rieder,G. (2009) *Helicobacter pylori* cag-Pathogenicity island-dependent early immunological response triggers later precancerous gastric changes in Mongolian gerbils. *PloS one*, **4**, e4754.
53. Ohnishi,N., Yuasa,H., Tanaka,S., Sawa,H., Miura,M., Matsui,A., Higashi,H., Musashi,M., Iwabuchi,K., Suzuki,M. *et al.* (2008) Transgenic expression of *Helicobacter pylori* CagA induces gastrointestinal and hematopoietic neoplasms in mouse. *Proc. Natl Acad. Sci. U.S.A.*, **105**, 1003–1008.
54. Lu,W., Wise,M.J., Tay,C.Y., Windsor,H.M., Marshall,B.J., Peacock,C. and Perkins,T. (2014) Comparative analysis of the full genome of *Helicobacter pylori* isolate Sahul64 identifies genes of high divergence. *J. Bacteriol.*, **196**, 1073–1083.

55. Garcia-Ortiz,M.V., Marsin,S., Arana,M.E., Gasparutto,D., Guerois,R., Kunkel,T.A. and Radicella,J.P. (2011) Unexpected role for *Helicobacter pylori* DNA polymerase I as a source of genetic variability. *PLoS Genet.*, **7**, e1002152.
56. Boneca,I.G., de Reuse,H., Epinat,J.C., Pupin,M., Labigne,A. and Moszer,I. (2003) A revised annotation and comparative analysis of *Helicobacter pylori* genomes. *Nucleic Acids Res.*, **31**, 1704–1714.
57. Kersulyte,D., Lee,W., Subramaniam,D., Anant,S., Herrera,P., Cabrera,L., Balqui,J., Barabas,O., Kalia,A., Gilman,R.H. *et al.* (2009) *Helicobacter pylori*'s plasticity zones are novel transposable elements. *PloS one*, **4**, e6859.
58. Devi,S., Ansari,S.A., Vadivelu,J., Megraud,F., Tenguria,S. and Ahmed,N. (2014) *Helicobacter pylori* antigen HP0986 (TieA) interacts with cultured gastric epithelial cells and induces IL8 secretion via NF-kappaB mediated pathway. *Helicobacter*, **19**, 26–36.
59. Alvi,A., Ansari,S.A., Ehtesham,N.Z., Rizwan,M., Devi,S., Sechi,L.A., Qureshi,I.A., Hasnain,S.E. and Ahmed,N. (2011) Concurrent proinflammatory and apoptotic activity of a *Helicobacter pylori* protein (HP986) points to its role in chronic persistence. *PloS one*, **6**, e22530.
60. Rizwan,M., Alvi,A. and Ahmed,N. (2008) Novel protein antigen (JHP940) from the genomic plasticity region of *Helicobacter pylori* induces tumor necrosis factor alpha and interleukin-8 secretion by human macrophages. *J. Bacteriol.*, **190**, 1146–1151.
61. Ahmed,N., Dobrindt,U., Hacker,J. and Hasnain,S.E. (2008) Genomic fluidity and pathogenic bacteria: applications in diagnostics, epidemiology and intervention. *Nat. Rev. Microbiol.*, **6**, 387–394.
62. Lin,L.F., Posfai,J., Roberts,R.J. and Kong,H. (2001) Comparative genomics of the restriction-modification systems in *Helicobacter pylori*. *Proc. Natl Acad. Sci. U.S.A.*, **98**, 2740–2745.
63. Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.
64. Furuta,Y., Namba-Fukuyo,H., Shibata,T.F., Nishiyama,T., Shigenobu,S., Suzuki,Y., Sugano,S., Hasebe,M. and Kobayashi,I. (2014) Methylome diversification through changes in DNA methyltransferase sequence specificity. *PLoS Genet.*, **10**, e1004272.
65. Krebes,J., Morgan,R.D., Bunk,B., Sproer,C., Luong,K., Parusel,R., Anton,B.P., Konig,C., Josenhans,C., Overmann,J. *et al.* (2014) The complex methylome of the human gastric pathogen *Helicobacter pylori*. *Nucleic Acids Res.*, **42**, 2415–2432.
66. Rohmer,L., Hocquet,D. and Miller,S.I. (2011) Are pathogenic bacteria just looking for food? Metabolism and microbial pathogenesis. *Trends Microbiol.*, **19**, 341–348.
67. Duncan,S.S., Valk,P.L., McClain,M.S., Shaffer,C.L., Metcalf,J.A., Bordenstein,S.R. and Cover,T.L. (2013) Comparative genomic analysis of East Asian and non-Asian *Helicobacter pylori* strains identifies rapidly evolving genes. *PloS one*, **8**, e55120.