SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
PREDICTIVE MARKERS
GASTRIC CANCER

Received
3 July 2014

Accepted
21 November 2014

Published
11 December 2014

Correspondence and requests for materials should be addressed to J.V. (jamuna@um.edu.my)

Helicobacter pylori Genetic Diversity and Gastro-duodenal Diseases in Malaysia

Selva Perumal Gunaletchumy[1], Indran Seevasant[1], Mun Hua Tan[2], Laurence J. Croft[2], Hazel M. Mitchell[3], Khean Lee Goh[4], Mun Fai Loke[1] & Jamuna Vadivelu[1]

[1]Department of Medical Microbiology, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur Malaysia, [2]Malaysian Genomics Resource Centre Berhad, 59200 Kuala Lumpur Malaysia, [3]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney NSW 2052 Australia, [4]Department of Medicine, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur Malaysia.

Helicobacter pylori **infection results in diverse clinical conditions ranging from chronic gastritis and ulceration to gastric adenocarcinoma. Among the multiethnic population of Malaysia, Indians consistently have a higher *H. pylori* prevalence as compared with Chinese and Malays. Despite the high prevalence of *H. pylori*, Indians have a relatively low incidence of peptic ulcer disease and gastric cancer. In contrast, gastric cancer and peptic ulcer disease incidence is high in Chinese. *H. pylori* strains from Chinese strains predominantly belong to the hspEAsia subpopulation while Indian/Malay strains mainly belong to the hspIndia subpopulation. By comparing the genome of 27 Asian strains from different subpopulations, we identified six genes associated with risk of *H. pylori*-induced peptic ulcer disease and gastric cancer. This study serves as an important foundation for future studies aiming to understand the role of bacterial factors in *H. pylori*-induced gastro-duodenal diseases.**

Helicobacter pylori contributes world-wide to various gastro-duodenal diseases ranging from chronic gastritis to the development of peptic ulcer disease, gastric cancer (GC) and mucosa-associated lymphoid tissue (MALT) lymphoma[1,2]. Significant variation exist in the prevalence and incidence of peptic ulcer disease and gastric cancer due to infection with *H. pylori* among the different multiethnic populations in Malaysia (Malays, Chinese and Indians)[3,4]. The *H. pylori* prevalence amongst Indians is 35.6%, but the incidence of peptic ulcer disease and gastric cancer is relatively low. Despite lower *H. pylori* prevalence of 28.6% among the Chinese, incidence of peptic ulcer disease and gastric cancer are relatively high[3,5].

Phylogenetic analysis of the *H. pylori* genomes was carried out to determine the ancestry relationship among the overall international isolates. Analysis of seven housekeeping genes of *H. pylori* by multi-locus sequence typing (MLST) suggest that the bacteria originated in Africa and later split into seven distinct population groups (hpEurope, hpNEAfrica, hpAfrica1, hpAfrica2, hpAsia2, hpSahul and hpEastAsia) and subpopulations that are strongly associated with geographical localization. As for the subpopulations, AE1 and AE2 are within hpEurope, hspWafrica and hspSAfrica within hpAfrica1, hspIndia within hpAsia2 and hspEAsia, hspAmerind and hspMaori within the population of hpEastAsia[6–8]. In Malaysia, *H. pylori* hpAsia2/hspIndia mainly colonises Indian and Malay subjects while hpEastAsia/hspEAsia predominantly colonises Chinese subjects. These two groups (hpEastAsia/hspEAsia and hpAsia2/hspIndia) accounted for 41.5% and 39.0% of all *H. pylori* isolates respectively[6].

H. pylori genetic factors may play a role in influencing the disease outcome as Chinese strains from Malaysia that are similar to strains from other areas of high gastric cancer incidence (Japan, Korea and China) predominantly belongs to hspEAsia subpopulation. Meanwhile, local Malay/Indian strains together with their counterparts from India (high *H. pylori* prevalence but low incidence of gastric cancer) belong to the hspIndia subpopulation[6]. Besides *H. pylori* genetic factors, host and environmental factors may also influence disease outcome.

In Western countries, the *H. pylori* protein encoded by cytotoxin-associated gene A, CagA, has been shown to be strongly associated with peptic ulcer disease, atrophic gastritis and gastric cancer[1,9]. However, *cagA*, carried by almost all Asian *H. pylori* strains, does not predict disease outcome in this region of the world[1,10,11]. Previous studies demonstrated that *cagA* containing the EPIYA-D tend to be more virulent than those carrying the EPIYA-C motif due to the higher level of IL-8 secretion produced by EPIYA-D strains as compared with EPIYA-C strains[12,13]. Studies have proven that Cag-A containing the EPIYA-D motif possessed among the East Asian

isolates has a stronger affinity for SHP-2 binding activity in contrast with Western Cag-A. Thus, East Asian isolates are found to be greater in countries with a high prevalence of gastric cancer[12,14–19]. However, Schmidt *et al.* (2009) and other studies have demonstrated that no association exists between EPIYA motifs and gastro-duodenal disease progression[11,20,21]. Thus, there is a need to discover new virulence factors by studying the genomic makeup of Asian strains.

H. pylori genetic diversity found in Malaysia, as demonstrated by MLST typing, provides ideal conditions for studying the interaction of co-existing *H. pylori* populations at the genomic level in a multi-ethnic society. In this study, we analyzed *H. pylori* strains from Asians (Malaysian Chinese, Malay, Indian), Japanese and mainland Chinese subjects presenting with different disease status. The aim of the study was to undertake whole genome comparative analysis of these Asian strains to define a subset of *H. pylori* disease-associated genes that may contribute to gastro-duodenal diseases.

Results

In this study, we carried out comparative genome analysis of 27 *H. pylori* strains (4 gastric cancer, 10 peptic ulcer and 13 non-ulcer dyspepsia strains). Strains were selected based on the phylogenetic analysis construction using the full genome sequences (Figure 1). Strains isolated from local Chinese subjects clustered closely with hspEAsia strains from Japan (F30, F32 and F57), China (XZ274 and HLJ271) and Korea (HP51). Most strains from the local Malay and Indian subjects (except UM037) were found clustered together to form the hspIndia branch. Out of 27 genomes selected for comparative analysis, 21 were isolated from Malaysia, three from Japan, two from China and one from Korea (Table 1). The genomes of non-Malaysian isolates were obtained from GenBank. Comparative genomic analysis revealed six genes that showed significant association with peptic ulcer disease and/or gastric cancer (Table 2). Three genes was associated with gastric cancer and the remaining 3 were associated with both peptic ulcer disease and gastric cancer. All 6 genes were statistically significantly correlated with disease with P-value <0.05 when the percentage of identity were compared using Student's t-test (Table 3). Gene ontology and prediction of protein families, domains and functional sites is presented in Table 4. Only genes with predicted structure and functions were selected for further analysis. The Pearson's correlation coefficient test was performed to evaluate the correlation between these genes and disease status, as well as the correlation between different genes (Table 5).

Gastric Cancer. A 456 bp gene encoding for a *H. pylori* membrane protein GC26_77 was absent in all (0/10) peptic ulcer disease (PUD) strains analyzed. This gene was present in all gastric cancer (GC) (4/4) and non-ulcer dyspepsia (NUD) (13/13) strains (Table 2). While no significant statistical significance (P-value ≥ 0.05) was observed between presence of this gene in strains isolated from GC and NUD patients, the gene was absent in PUD patients (P-value < 0.001) (Table 3). Pearson's correlation also revealed a strong negative correlation with PUD (Table 5).

Further, the presence of hypothetical ATPase protein GC26_69 was shown to correlate with strains isolated from GC patients. The ATPase gene was detected in 100% (4/4) GC strains as compared with 30% (3/10) of PUD strains and 0% (0/13) of isolates from NUD. Interestingly, the 3 isolates in which the hypothetical ATPase protein GC26_69 gene was detected originated from PUD patients from China, Japan and Korea. None of the PUD strains isolated from Malaysia carried this gene (Table 2). Highly significant associations were observed in the prevalence of this gene in isolates from GC

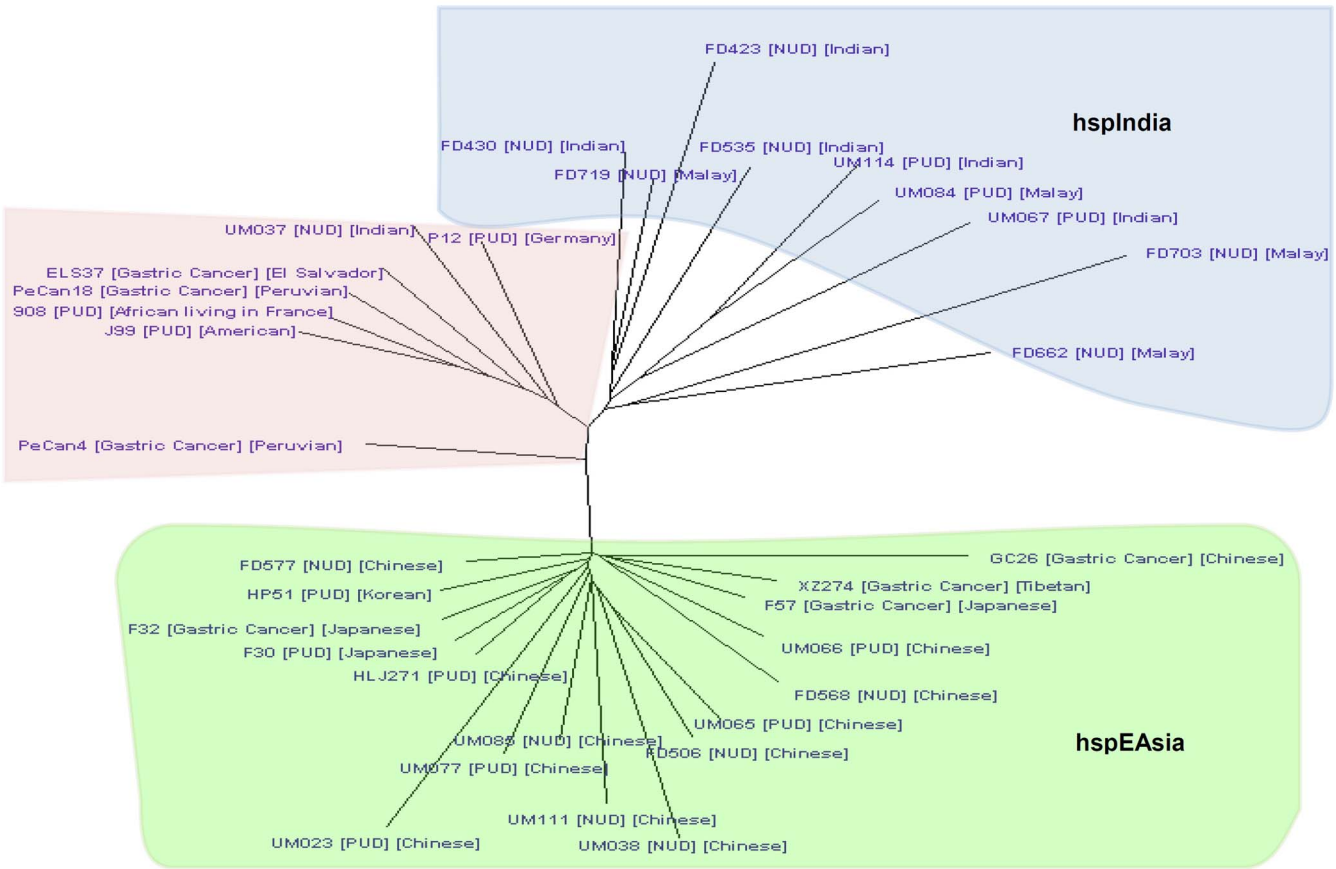

Figure 1 | Phylogenetic tree constructed on a whole genome of 33 *H. pylori* strains. Asian strains used in this study mostly clustered together with hspIndia and hspEAsia strains according to host ethnicity (except UM037).

Table 1 | Details of strains used in this study

Strain	Ethnicity	Age	Gender	Year of Isolation	Geographical	Disease	Reference
GC26	Chinese	70	Female	2005	Malaysia	Gastric cancer	30
F57	Japanese	-	-	-	Japan	Gastric cancer	32
F32	Japanese	-	-	-	Japan	Gastric cancer	32
XZ274	Tibetan	-	-	-	China	Gastric cancer	33
UM023	Chinese	71	Female	2011	Malaysia	Peptic ulcer disease	30
UM065	Chinese	57	Female	2011	Malaysia	Peptic ulcer disease	30
UM066	Chinese	30	Female	2011	Malaysia	Peptic ulcer disease	30
UM077	Chinese	59	Female	2011	Malaysia	Peptic ulcer disease	30
UM067	Indian	64	Male	2011	Malaysia	Peptic ulcer disease	30
UM084	Malay	60	Male	2011	Malaysia	Peptic ulcer disease	30
UM114	Indian	42	Female	2012	Malaysia	Peptic ulcer disease	30
F30	Japanese	-	-	-	Japan	Peptic ulcer disease	32
HP51	Korean	-	-	-	Korea	Peptic ulcer disease	-
HLJ271	-	-	-	-	China	Peptic ulcer disease	34
FD506	Chinese	56	Female	2005	Malaysia	Non-ulcer dyspepsia	30
FD568	Chinese	50	Female	2005	Malaysia	Non-ulcer dyspepsia	30
FD577	Chinese	67	Female	2005	Malaysia	Non-ulcer dyspepsia	30
UM038	Chinese	73	Male	2011	Malaysia	Non-ulcer dyspepsia	30
UM085	Chinese	68	Female	2011	Malaysia	Non-ulcer dyspepsia	30
UM111	Chinese	76	Male	2012	Malaysia	Non-ulcer dyspepsia	30
FD662	Malay	47	Female	2006	Malaysia	Non-ulcer dyspepsia	30
FD719	Malay	60	Female	2006	Malaysia	Non-ulcer dyspepsia	30
FD703	Malay	51	Male	2006	Malaysia	Non-ulcer dyspepsia	30
FD423	Indian	39	Male	2005	Malaysia	Non-ulcer dyspepsia	30
FD430	Indian	34	Male	2005	Malaysia	Non-ulcer dyspepsia	30
FD535	Indian	44	Male	2005	Malaysia	Non-ulcer dyspepsia	30
UM037	Indian	67	Female	2011	Malaysia	Non-ulcer dyspepsia	30

patients as compared with isolates from PUD and NUD (P-value <0.001) (Table 3). The association of hypothetical ATPase protein GC26_69 with GC was confirmed by a strong positive correlation by Pearson's analysis (Table 5).

In addition a hypothetical protein GC26_73 was detected in all (4/4) GC strains but only 20% (2/10) and 7.7% (1/13) of PUD and NUD strains respectively (Table 2). Comparison of the prevalence of this gene in strains isolated from GC patients with that in both PUD and NUD patients showed a highly significant difference to exist (P-value <0.001) (Table 3). The association of hypothetical protein GC26_73 with GC was confirmed by a strong Pearson's positive correlation (Table 5).

Peptic Ulcer Disease and Gastric Cancer. Genes encoding for the *H. pylori* outer membrane protein GC26_66, phospho-2-dehydro-3-deoxyheptonate aldolase, and hypothetical protein GC26_33 were all highly associated with both GC and PUD strains. The gene encoding for outer membrane protein GC26_66 being present in all GC strains (4/4) and 90% (9/10) of the PUD strains. The outer membrane protein gene was not detected in any of the six Chinese NUD strains (Table 2). A highly statistically significant difference was observed between the prevalence of outer membrane protein in GC strains and Chinese NUD strains (P-value <0.001) (Table 3). While outer membrane protein GC26_66 was also present in NUD strains from local Malay and Indians, alignment of the translated protein sequence revealed that these local Malay and Indian strains from NUD patients differed from GC/PUD strains in the 2nd to 5th amino acid positions (Figure 2). For the purpose of discussion, we will differentiate the former as type 1 and the later as type 2. Outer membrane protein GC26_66 (type 1) was found to have strong negative Pearson's correlation with NUD (Table 5).

The phospho-2-dehydro-3-deoxyheptonate aldolase gene was present in all of the GC strains (4/4) and 90% (9/10) of the PUD strains. In contrast, this gene was only found in 15.4% (2/13) of the NUD strains (Table 2). A statistically increased prevalence of this gene was observed in isolates from GC patients as compared with isolates from NUD (P-value <0.001) (Table 3). Phospho-2-dehydro-3-deoxyheptonate aldolase was also found to have strong negative Pearson's correlation with NUD (Table 5).

The gene encoding for hypothetical protein GC26_33 was present in all GC strains (4/4). 60% (6/10) of the PUD strains and only 7.7% (1/13) of the NUD strains (Table 2). Comparison of the prevalence of this gene in *H. pylori* strains isolated from GC and PUD patients showed a significantly increased prevalence (P-value <0.05) in those patients with GC. Comparison of strains isolated from GC and NUD patients showed a significantly increased prevalence in isolates from GC patients (P-value <0.001) (Table 3). Hypothetical protein GC26_33 was also found to have strong negative Pearson's correlation with NUD (Table 5).

Discussion

In this study, we identified six genes that were associated with *H. pylori*-related disease status. In agreement with previous studies, no single *H. pylori* genetic marker on its own was shown to be associated with any specific disease group[22,23]. However, it is possible that these genes, working individually or collaborating with other genetic elements, are potential risk factors influencing the severity of disease or disease progression. Based on Pearson's analysis of the correlation between disease status and genes, as well as the correlation between different genes presented in Table 4, a correlation map was prepared to summarize the relationship between disease and genes (Figure 3). This map illustrates the complexity of *H. pylori* factors that may affect disease outcome. From the map, the number of genes that potentially influence disease status appears to increase with severity of disease. However, it should be noted that the correlation map only illustrates the co-existence of genes but does not imply that these genes actually interacts or work together to have an impact on disease development.

Based on our comparative genomic analysis, the presence of *H. pylori* phospho-2-dehydro-3-deoxyheptonate aldolase in the absence of membrane protein GC-26_77 was demonstrated to be a risk factor for the development of peptic ulcer disease. In contrast, outer membrane protein GC26_66 (type 1), hypothetical ATPase protein

Table 2 | Candidate genes identified as associated with GC, PUD and NUD with percentage of identity. Percentage of identity generated by RAST >80% was considered similar

	GC				PUD				
	Chinese	Japanese		Tibetan	Chinese				
Gene/Function	GC26	F57	F32	XZ274	UM023	UM065	UM066	UM077	HLJ271
Membrane protein GC26_77	100	78.7	86.6	88.67	0	0	0	0	0
Hypothetical ATPase protein GC26_69	100	93.8	98.6	95.65	0	0	0	0	76.88
Hypothetical protein GC26_73	100	92.1	97.4	88.24	0	0	74.1	0	0
Outer membrane protein GC26_66 (type 1)	100	96.2	98.0	97.6	0	98.08	96.63	96.46	98.02
Outer membrane protein GC26_66 (type 2)	0	0	0	0	0	0	0	0	0
Phospho-2-Dehydro-3-Deoxyheptonate Aldolase	100	99.6	99.0	98.66	99.33	98.44	99.33	98.66	99.33
Hypothetical protein GC26_33	100	84.1	100	98.18	100	96.36	90.91	96.36	0

	PUD					**NUD**			
	Japanese	**Korean**	**Malay**	**Indian**		**Chinese**			
	F30	**HP51**	**UM084**	**UM067**	**UM114**	**FD506**	**FD568**	**FD577**	**UM038**
Membrane Protein GC26_77	0	0	0	0	0	89.3	95.2	100	96.0
Hypothetical ATPase Protein GC26_69	82.55	92.22	0	0	0	0	0	0	0
Hypothetical protein GC26_73	89.19	0	0	0	0	0	0	0	0
Outer membrane protein GC26_66 (type 1)	97.12	98.56				0	0	0	0
Outer membrane protein GC26_66 (type 2)	0	0	95.43	93.91	94.42	0	0	0	0
Phospho-2-Dehydro-3-Deoxyheptonate Aldolase	98.66	97.09	97.55	0	98.22	0	0	0	0
Hypothetical protein GC26_33	74.55	0	0	0	81.82	0	0	0	0

	NUD								
	Chinese		**Malay**			**Indian**			
	UM085	**UM111**	**FD662**	**FD703**	**FD719**	**FD423**	**FD430**	**FD535**	**UM037**
Membrane Protein GC26_77	87.3	92.0	93.0	96.7	90.0	92.2	94.0	93.2	89.3
Hypothetical ATPase Protein GC26_69	0	0	0	0	0	0	0	0	0
Hypothetical protein GC26_73	0	94.7	0	0	0	0	0	0	0
Outer membrane protein GC26_66 (type 1)	0	0	0	0	0	0	0	0	0
Outer membrane protein GC26_66 (type 2)	0	0	94.92	95.94	96.45	95.43	93.40	96.45	94.44
Phospho-2-Dehydro-3-Deoxyheptonate Aldolase	99.33	99.0	0	0	0	0	0	0	0
Hypothetical protein GC26_33	100	0	0	0	0	0	0	0	0

Percentage of similarity >80% is considered significa.

Table 3 | Significance of association of candidate genes with disease was examined by Student's t-test

Genes/Function	Student's t-test (P-value) GC vs. PUD	GC vs. NUD
Membrane Protein GC26_77	0.000**	0.390
Hypothetical ATPase Protein GC26_69	0.000**	0.000**
Hypothetical protein GC26_73	0.000**	0.000**
Outer membrane protein GC26_66 (type 1)	0.282	0.005**
Phospho-2-Dehydro-3-Deoxyheptonate Aldolase	0.308	0.000**
Hypothetical protein GC26_33	0.022*	0.000**

*. P-value <0.05 is considered significant.
**. P-value <0.01 is considered highly significant.

GC26_69, hypothetical protein GC26_73 and hypothetical protein GC26_33 were shown to be risk factors for gastric cancer development. Outer membrane GC26_77 may have a protective effect against peptic ulcer disease. However, these data were based on bioinformatic analysis of strains from Malaysia and few strains from other parts of Asia. There is a need to validate these results experimentally in the laboratory. Furthermore, it will be important to be able screen and estimate the carriage rate of these genes in a larger number of strains with different disease presentations from other parts of Asia as well as other parts of the world.

H. pylori is an ancient and permanent resident of the human stomach and has likely been part of the gastric microbiome since the origin of human species. Given this, it is not surprising that more than 80% of those infected with *H. pylori* remain asymptomatic. It has been estimated that *H. pylori*-positive individuals have a 10 to 20% lifetime risk of developing peptic ulcer disease and a 1 to 2% risk of developing gastric cancer[24]. While there is strong evidence to support the role of *H. pylori* in gastro-duodenal disease, there is some evidence that the decline in *H. pylori* infection among human populations has been suggested to have contributed to the increase in other diseases (e.g., esophageal adenocarcinoma [EAC], allergic asthma, rhinitis and atrophy)[25]. Furthermore, it has been reported that in adult males *H. pylori* colonization is associated with reduced circulating leptin levels, a finding that may explain the observation that significant weight gain may occur with *H. pylori* eradication[26]. Furthermore, Blaser, M.J. and Atherton, J.C. have suggested that the fall in *H. pylori* prevalence in developed countries may also contribute to high risk of metabolic syndrome, type II diabetes and metabolic obesity[27]. Current literature would suggest that infection with *H. pylori* is rapidly disappearing in developed countries[28].

These risk factors can be potential biomarkers to identify those strains that present with higher risk of developing severe gastro-duodenal complications. The use of an approach where selective eradication therapy is given to individuals with these risk genes instead of all those infected with *H. pylori* will also help to preserve the usual gut microbiota. The understanding of *H. pylori* genetic factors and its association with severe gastro-duodenal diseases is necessary to decide on the optimal management of *H. pylori* infections.

A similar study has examined the genomic characteristics among 84 *H. pylori* isolates frm China with differing clinical status using microarray[29]. In this study, regions associated with genes involved in bacterial R-M systems and type IV secretion system were identified to be linked to disease status. However, these genes were not identified in this study, instead, a different set of genes were identified in this study. This inconsistency could be due to the different approach adopted and the strains selected for these studies. The microarray

Table 4 | Structural and domain prediction of candidate genes by Blast2Go

Gene/Function	Accession No.	Gene Length (bp)	Hits	Min. e-value	Mean Similarity	Go ID	InterPro
Membrane Protein GC26_77	**KJ716798**	456	20	1.11E-70	92.2%	-	i. Transmembrane protein
Hypothetical ATPase Protein GC26_69	**KJ721220**	1683	20	0	86.35%	Fatty acid biosynthetic process: GTP catabolic process; GTPase activity; GTP binding	i. Dynamin, GTPase domain ii.P-loop containing nucleoside triphosphate hydrolase
Hypothetical protein GC26_73	**KJ721221**	150	6	5.58E-17	93.5%	-	i. Signal peptide
Outer membrane protein GC26_66	**KJ721222**	2394	20	5.99E-140	98.65%	-	i. *Helicobacter pylori* outer membrane protein
Phospho-2-Dehydro-3-Deoxyheptonate Aldolase	**KJ721223**	1353	20	0	99.05%	3-deoxy-7-phosphoheptulonate synthase activity; aromatic amino acid family biosynthetic process	DAHP synthetase, class II
Hypothetical protein GC26_33	**KJ721224**	168	20	4.57E-28	93.55%	-	i. Transmembrane protein

Table 5 | Pearson's correlation coefficient test on relationship between candidate gene and disease and between different genes

		GC	PUD	NUD	Race	Membrane protein GC26_77	Hypothetical ATPase protein GC26_69	Hypothetical protein GC26_73	Outer membrane protein GC26_66 (type 1)	Phospho-2-dehydro-3-deoxyheptonate aldolase	Hypothetical protein GC26_33
GC	Pearson Correlation	1	−0.320	−0.402*	−0.346	0.287	0.760**	0.739**	0.546**	0.377	0.523**
	Sig. (2-tailed)		0.104	0.038	0.077	0.147	0.000	0.000	0.003	0.052	0.005
	N	27	27	27	27	27	27	27	27	27	27
PUD	Pearson Correlation	−0.320	1	−0.739**	−0.168	−0.996**	0.028	−0.138	0.363	0.527**	0.269
	Sig. (2-tailed)	0.104		0.000	0.403	0.000	0.890	0.491	0.063	0.005	0.175
	N	27	27	27	27	27	27	27	27	27	27
NUD	Pearson Correlation	−0.402*	−0.739**	1	0.408*	0.759**	−0.567**	−0.392*	−0.739**	−0.778**	−0.632**
	Sig. (2-tailed)	0.038	0.000		0.035	0.000	.002	0.043	0.000	0.000	0.000
	N	27	27	27	27	27	27	27	27	27	27
Race	Pearson Correlation	−0.346	−0.168	0.408*	1	0.176	−0.488**	−0.310	−0.636**	−0.474*	−0.549**
	Sig. (2-tailed)	0.077	0.403	0.035		0.380	0.010	0.116	0.000	0.012	0.003
	N	27	27	27	27	27	27	27	27	27	27
Membrane protein GC26_77	Pearson Correlation	0.287	−0.996**	0.759**	0.176	1	−0.054	0.113	−0.384*	−0.556**	−0.297
	Sig. (2-tailed)	0.147	0.000	0.000	0.380		0.789	0.575	0.048	0.003	0.132
	N	27	27	27	27	27	27	27	27	27	27
Hypothetical ATPase protein GC26_69	Pearson Correlation	0.760**	0.028	−0.567**	−0.488**	−0.054	1	0.667**	0.771**	0.528**	0.382*
	Sig. (2-tailed)	0.000	0.890	0.002	0.010	0.789		0.000	0.000	0.005	0.049
	N	27	27	27	27	27	27	27	27	27	27
Hypothetical protein GC26_73	Pearson Correlation	0.739**	−0.138	−0.392*	−0.310	0.113	0.667**	1	0.586**	0.531**	0.513**
	Sig. (2-tailed)	0.000	0.491	0.043	0.116	0.575	0.000		0.001	0.004	0.006
	N	27	27	27	27	27	27	27	27	27	27
Outer membrane protein GC26_66	Pearson Correlation	0.546**	0.363	−0.739**	−0.636**	−0.384*	0.771**	0.586**	1	0.687**	0.602**
	Sig. (2-tailed)	0.003	0.063	0.000	0.000	0.048	0.000	0.001		0.000	0.001
	N	27	27	27	27	27	27	27	27	27	27
Phospho-2-dehydro-3-deoxyheptonate aldolase	Pearson Correlation	0.377	0.527**	−0.778**	−0.474*	−0.556**	0.528**	0.531**	0.687**	1	0.740**
	Sig. (2-tailed)	0.052	0.005	0.000	0.012	0.003	0.005	0.004	0.000		0.000
	N	27	27	27	27	27	27	27	27	27	27
Hypothetical protein GC26_33	Pearson Correlation	0.523**	0.269	−0.632**	−0.549**	−0.297	0.382*	0.513**	0.602**	0.740**	1
	Sig. (2-tailed)	0.005	0.175	0.000	0.003	0.132	0.049	0.006	0.001	0.000	
	N	27	27	27	27	27	27	27	27	27	27

*. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).

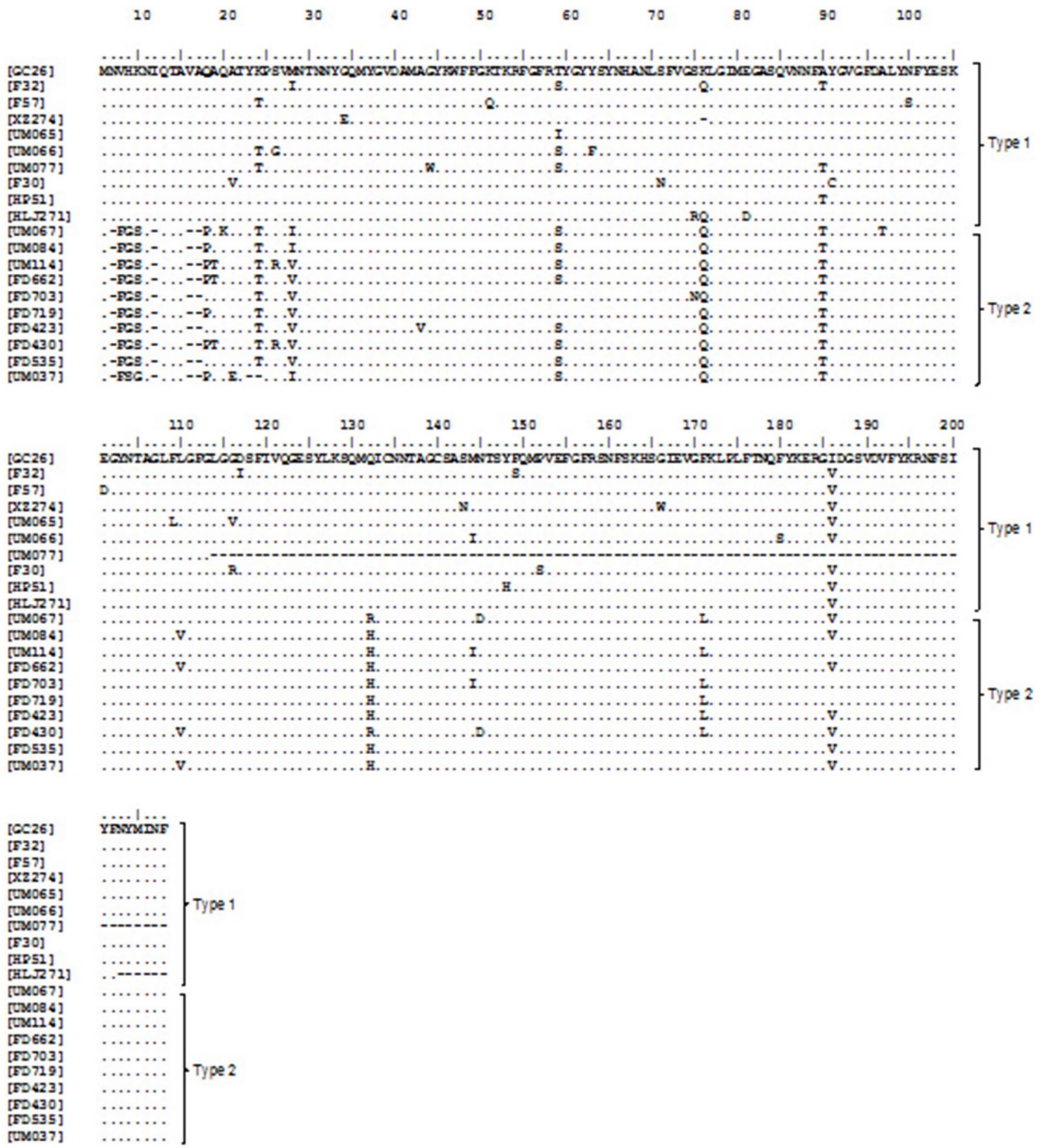

Figure 2 | Sequence alignment of translated outer membrane protein GC26_66 gene. According to the translated peptide, sequences were grouped as type 1 and type 2. Type 1 strains were GC26, F32, F57, XZ274, UM065, UM066, UM077, F30, HP51 and HLJ271. Type 2 strains were UM067, UM084, UM114, FD662, FD703, FD719, FD423, FD430, FD535 and UM037. There was close correlation with host ethnicity.

was designed based on 6 sequenced strains of European and American origins. Thus, genomic variations present only in Asian strains but not in Western strains will not be identified using the microarray approach. Although the next-generation sequencing approach is not limited by probe design, it can be limited by the quality of sequencing data. Also, the higher cost by the next-generation sequencing approach will also complicate mass screening of large number of strains.

In summary, we have identified a number of *H. pylori* genetic factors that may enable the identification of those at risk of peptic ulcer disease strains and gastric cancer using a comparative genomic approach. However, screening of a larger number of *H. pylori* strains from different disease groups is also required. Furthermore, it would be of interest to investigate the functions of these largely hypothetical proteins and how they interact to contribute to *H. pylori*-induced pathogenesis.

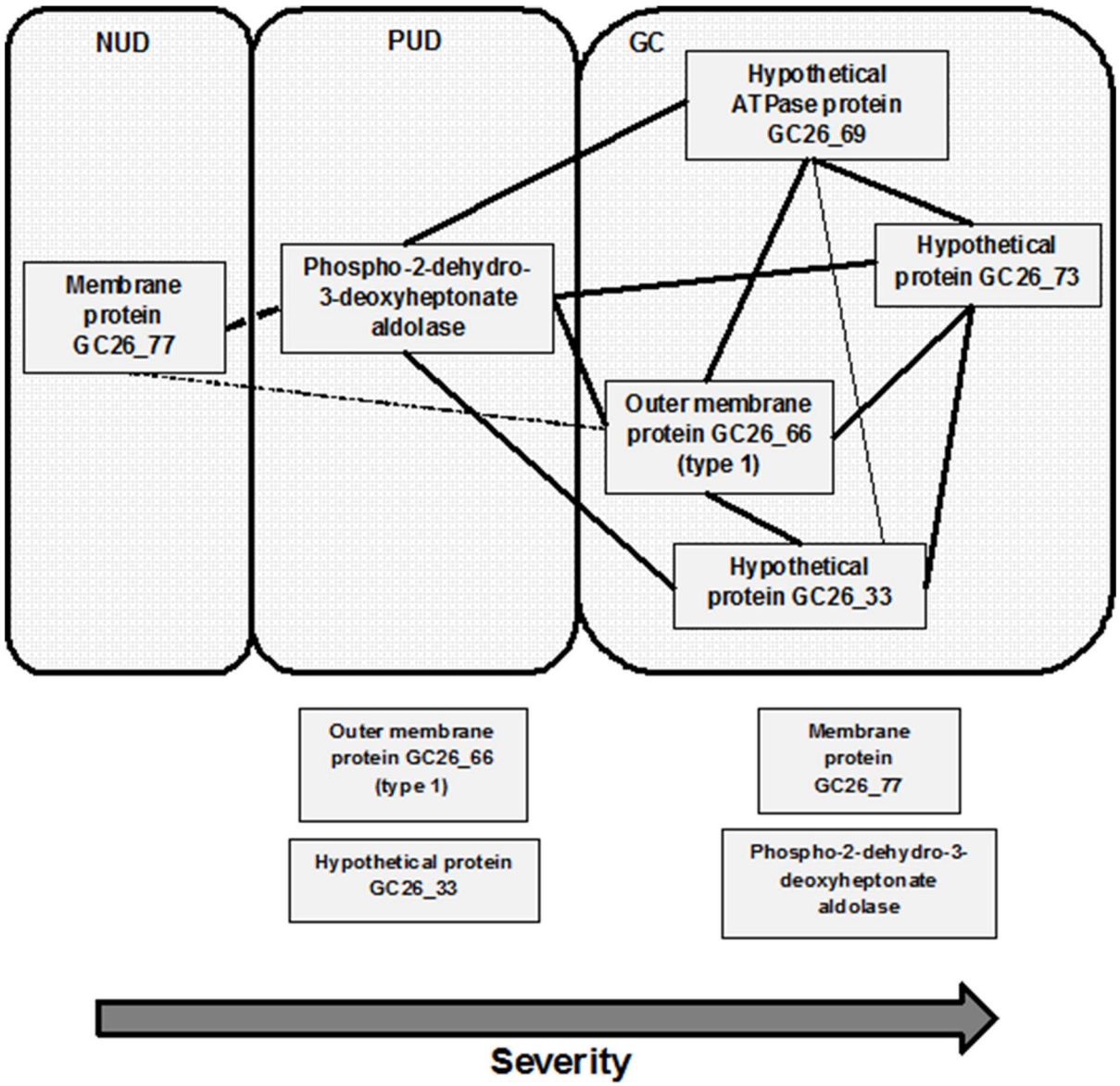

Figure 3 | Correlation map of candidate genes with disease states and between genes. The correlation was determined by Pearson's correlation (Table 5). Dotted line represents negative correlation and full line represents positive correlation. Bold line represents strong correlation with significance <0.01 and fine line represents correlation with significance <0.05. Genes outside the boxes represent genes present in the respective disease strains (Student's t-test) but no significant correlation with respective disease status (Pearson's).

Methods

Sample background. Twenty-one *H. pylori* from Malaysia (GC26, UM023, UM065, UM066, UM077, UM067, UM084, UM114, FD506, FD568, FD577, UM038, UM085, UM111, FD662, FD719, FD703, FD423, FD430, FD535 and UM037) were isolated from symptomatic patients undergoing endoscopy procedure at University of Malaya Medical Centre (UMMC, Kuala Lumpur, Malaysia) between the years 2007 and 2012 (Table 1). Based on endoscopic and histological examinations, patients were diagnosed as having gastric cancer (GC), peptic ulcer disease (PUD) or non-ulcer dyspepsia/functional dyspepsia (NUD/FD). All biopsies were obtained with the informed consent of patients and approval of the Human Ethics Committees of UMMC and UNSW. The genomes of the twenty-one Malaysian *H. pylori* strains were sequenced and assembled *de novo* as previously described[30]. For comparison, we downloaded the genome sequences of another 6 Asian *H. pylori* strains from the National Center for Biotechnology Information (NCBI) GenBank. Further information of these strains are provided in Table 1. Only F30, F32, F57 and HP51 are complete genomes, the rest are available as drafts.

Sequence Alignment and Phylogenetic Analysis. In addition to the 27 genomes analyzed in this study, an additional of six strains regardless of East Asian origin was downloaded from NCBI database resulting in a total of 33 strains that were used to align with MAUVE (version 2.3.1) progressive alignment software and the output data was viewed using the SplitsTree program (version 4.12.8) as Super Network.

Downstream analysis. Annotation of all twenty-seven *H. pylori* strains were performed using the Rapid Annotation using Subsystem Technology (RAST) version 4.0[31]. Comparative genomic analysis was performed using the sequence comparison function available on the SEED viewer version 2.0. A List of core genes present among all strains belonging to the same group was generated by SEED for each disease groups (GC, PUD and NUD). FASTA file containing the sequences these of core genes were generated for the respective disease groups. Further, RAST analysis was performed to this data for annotation and comparing against other *H. pylori* strains in Table 1 to obtain the percentage of similarity. The Student's unpaired two-tailed t-test was performed and the genes with *P-values* of < 0.05 and < 0.01were considered statistically significant and highly significant respectively. Two-tailed Pearsons' correlation analysis was adopted to examine for correlation between candidate genes and disease, as well as between different genes.

Accession Numbers. The accession numbers of the *H. pylori* genome sequences reported in this paper are: GC26 (AKHV00000000), UM023 (AUSK00000000), UM065 (AUSM00000000), UM066 (AUSJ00000000 and CP005493), UM077 (AUSQ00000000), UM067 (AUSN00000000), UM084 (AUSO00000000), UM114 (AUSS00000000), FD506 (AKHO00000000), FD568 (AKHQ00000000), FD577 (AKHR00000000), UM038 (AUSL00000000), UM111 (AUSR00000000), FD662 (AKHT00000000), FD719 (AKHU00000000), FD703 (AKHS00000000), FD423 (AKHM00000000), FD430 (AKHN00000000), FD535 (AKHP00000000), UM037 (AUSI00000000 and CP005492), UM085 (AUSP00000000), F30 (AP011941), F32

(AP011943), F57 (AP011945), HP51 (CP000012), XZ274 (CP003419) and HLJ271 (ALKB00000000).

1. Israel, D. A. *et al. Helicobacter pylori* strain-specific differences in genetic content, identified by microarray, influence host inflammatory responses. *J Clin Invest.* **107**, 611–620 (2001).
2. Chalker, A. F. *et al.* Systematic identification of selective essential genes in *Helicobacter pylori* by genome prioritization and allelic replacement mutagenesis. *J Bacteriol.* **183**, 1259–1268 (2001).
3. Goh, K. L. & Parasakthi, N. The racial cohort phenomenon: seroepidemiology of *Helicobacter pylori* infection in a multiracial South-East Asian country. *Eur J Gastroenterol Hepatol* **13**, 177–183 (2001).
4. Goh, K. L. Epidemiology of *Helicobacter pylori* infection in Malaysia--observations in a multiracial Asian population. *Med J Malaysia.* **64**, 187–192 (2009).
5. Musa, A. F., Yunos, M. N. M., Rahman, S. A. & Nordin, R. B. The seroprevalence and eradication success of *Helicobacter pylori* in indigenous people of Seletar in Southern Malaysia. *Br J Med Med Res.* **4**, 1854–1863 (2014).
6. Tay, C. Y. *et al.* Population structure of Helicobacter pylori among ethnic groups in Malaysia: recent acquisition of the bacterium by the Malay population. *BMC Microbiol* **19**, 126 (2009).
7. Falush, D. *et al.* Traces of human migrations in Helicobacter pylori populations. *Science.* **299**, 1582–1585 (2003).
8. Breurec, S. *et al.* Evolutionary history of Helicobacter pylori sequences reflect past human migrations in Southeast Asia. *PLoS One.* **6**, e22058 (2011).
9. Fock, K. M. & Ang, T. L. Epidemiology of *Helicobacter pylori* infection and gastric cancer in Asia. *J Gastroenterol Hepatol.* **25**, 479–486 (2010).
10. Fock, K. M. *et al.* Asia-Pacific consensus guidelines on gastric cancer prevention. *J Gastroenterol Hepatol* **23**, 351–365 (2008).
11. Conteduca, V. *et al. H. pylori* infection and gastric cancer: state of the art (review). *Int J Oncol.* **42**, 5–18 (2013).
12. Argent, R. H., Hale, J. L., El-Omar, E. M. & Atherton, J. C. Differences in Helicobacter pylori CagA tyrosine phosphorylation motif patterns between western and East Asian strains, and influences on interleukin-8 secretion. *J Med Microbiol.* **57**, 1062–1067 (2008).
13. Schmidt, H. M. *et al.* The cag PAI is intact and functional but HP0521 varies significantly in *Helicobacter pylori* isolates from Malaysia and Singapore. *Eur J Clin Microbiol Infect Dis.* **29**, 439–451 (2010).
14. Kim, S. Y., Lee, Y. C., Kim, H. K. & Blaser, M. J. Helicobacter pylori CagA transfection of gastric epithelial cells induces interleukin-8. *Cell Microbiol.* **8**, 97–106 (2006).
15. Sahara, S. *et al.* Role of Helicobacter pylori cagA EPIYA motif and vacA genotypes for the development of gastrointestinal diseases in Southeast Asian countries: a meta-analysis. *BMC Infect Dis.* **12**, 223 (2012).
16. Jones, K. R. *et al.* Polymorphism in the CagA EPIYA motif impacts development of gastric cancer. *J Clin Microbiol.* **47**, 959–968 (2009).
17. Azuma, T. *et al.* Association between diversity in the Src homology 2 domain--containing tyrosine phosphatase binding siteof *Helicobacter pylori* CagA protein and gastric atrophy and cancer. *J Infect Dis.* **189**, 820–827 (2004).
18. Higashi, H.*et al. Biological activity* of the *Helicobacter pylori* virulence factor CagA is determined by variation in the tyrosine phosphorylation sites. *Proc Natl Acad Sci USA.* **99**, 14428–14433 (2002)
19. Shiota, S., Suzuki, R. & Yamaoka, Y. The significance of virulence factors in *Helicobacter pylori*. *J Dig Dis.* **14**, 341–349 (2013).
20. Choi, K. D. *et al.* Analysis of the 3' variable region of the cagA gene of Helicobacter pylori isolated in Koreans. *Dig Dis Sci.* **52**, 960–966 (2007).
21. Zhu, Y. L., Zheng, S., Du, Q., Qian, K. D. & Fang, P. C. Characterization of CagA variable region of Helicobacter pylori isolates from Chinese patients. *World J Gastroenterol.* **11**, 880–884 (2005).
22. Tan, H. J., Rizal, A. M., Rosmadi, M. Y. & Goh, K. L. Distribution of Helicobacter pylori cagA, cagE and vacA in different ethnic groups in Kuala Lumpur, Malaysia. *J Gastroenterol Hepatol.* **20**, 589–594 (2005).
23. Ahmed, N., Tenguria, S. & Nandanwar, N. *Helicobacter pylori* - a seasoned pathogen by any other name. *Gut Pathog.* **1**, 24 (2009).
24. Kusters, J. G., van Vliet, A. H. M. & Kuipers, E. J. Pathogenesis of *Helicobacter pylori* infection. *Clin Microbiol Rev.* **19**, 449–90 (2006).
25. Blaser, M. J. Who are we? Indigenous microbes and the ecology of human diseases. *EMBO Rep.* **7**, 956–960 (2006).
26. Roper, J. *et al.* Leptin and ghrelin in relation to Helicobacter pylori status in adult males. *J Clin Endocrinol Metab.* **93**, 2350–2357 (2008).
27. Blaser, M. J. & Atherton, J. C. Helicobacter pylori persistence: biology and disease. *J Clin Inves* **113**, 321–333 (2004).
28. Hadley, C. The infection connection. *Helicobacter pylori* is more than just the cause of gastric ulcers--it offers an unprecedented opportunity to study changes in human microecology and the nature of chronic disease. *EMBO Rep.* **7**, 470–473 (2006).
29. You, Y. *et al.* Comparative Genomics of *Helicobacter pylori* Strains of China Associated with Different Clinical Outcome. *PLoS One.* **7**, e38528 (2012).
30. Rehvathy, V. *et al.* Multiple Genome Sequences of *Helicobacter pylori* Strains of Diverse Disease and Antibiotic Resistance Backgrounds from Malaysia. *Genome Announc.* **1** (2013).
31. Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics.* **9**, 75 (2008).
32. Kawai, M. *et al.* Evolution in an oncogenic bacterial species with extreme genome plasticity: *Helicobacter pylori* East Asian genomes. *BMC Microbiol.* **11**, 104 (2011)
33. Guo, Y. *et al.* Genome of *Helicobacter pylori* strain XZ274, an isolate from a tibetan patient with gastric cancer in China. *J Bacteriol* **194**, 4146–4147 (2012).
34. You, Y. *et al.* Genome sequences of three *Helicobacter pylori* strains isolated from atrophic gastritis and gastric ulcer patients in China. *J Bacteriol.* **194**, 6314–6315 (2012).

Acknowledgments

This research was supported by University of Malaya-Ministry of Education (UM-MoE) High Impact Research (HIR) Grant UM.C/625/1/HIR/MoE/CHAN/02 (Account No. H-50001-A000013).

Author contributions

M.F.L., J.V., L.J.C., H.M.M. and K.L.G. designed the study. M.H.T. and I.S. assembled the genomes. S.P.G. analyzed the data and wrote the paper.

Additional information

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

How to cite this article: Gunaletchumy, S.P. *et al. Helicobacter pylori* Genetic Diversity and Gastro-duodenal Diseases in Malaysia. *Sci. Rep.* **4**, 7431; DOI:10.1038/srep07431 (2014).