The Journal of Microbiology, February 2007, p. 21-28
Vol. 45, No. 1

Comparative Genomics Profiling of Clinical Isolates of *Helicobacter pylori* in Chinese Populations Using DNA Microarray

Yue-Hua Han[1], Wen-Zhong Liu[2,*], Yao-Zhou Shi[3], Li-Qiong Lu[3], Shudong Xiao[2], Qing-Hua Zhang[3], and Guo-Ping Zhao[3]

[1]*Second affiliated hospital, School of medicine, Zhejiang University, Hangzhou, P. R. China*
[2]*Shanghai Institute of Digestive Disease, Renji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, P. R. China*
[3]*National Engineering Center for Biochip at Shanghai, Zhangjiang Hi-Tech Park, Pudong, Shanghai, P. R. China*

(Received September 25, 2006 / Accepted December 22, 2006)

In order to search for specific genotypes related to this unique phenotype, we used whole genomic DNA microarray to characterize the genomic diversity of *Helicobacter pylori* (*H. pylori*) strains isolated from clinical patients in China. The open reading frame (ORF) fragments on our microarray were generated by PCR using gene-specific primers. Genomic DNA of *H. pylori* 26695 and J99 were used as templates. Thirty-four *H. pylori* isolates were obtained from patients in Shanghai. Results were judged based on ln(x) transformed and normalized Cy3/Cy5 ratios. Our microarray included 1882 DNA fragments corresponding to 1636 ORFs of both sequenced *H. pylori* strains. Cluster analysis, revealed two diverse regions in the *H. pylori* genome that were not present in other isolates. Among the 1636 genes, 1091 (66.7%) were common to all *H. pylori* strains, representing the functional core of the genome. Most of the genes found in the *H. pylori* functional core were responsible for metabolism, cellular processes, transcription and biosynthesis of amino acids, functions that are essential to *H. pylori's* growth and colonization in its host. In contrast, 522 (31.9%) genes were strain-specific genes that were missing from at least one strain of *H. pylori*. Strain-specific genes primarily included restriction modification system components, transposase genes, hypothetical proteins and outer membrane proteins. These strain-specific genes may aid the bacteria under specific circumstances during their long-term infection in genetically diverse hosts. Our results suggest 34 *H. pylori* clinical strains have extensive genomic diversity. Core genes and strain-specific genes both play essential roles in *H. pylori* propagation and pathogenesis. Our microarray experiment may help select relatively significant genes for further research on the pathogenicity of *H. pylori* and development of a vaccine for *H. pylori*.

Keywords: *Helicobacter pylori*, genetic diversity, microarray, clinical disease

Helicobacter pylori (*H. pylori*) infection is now recognized as the most important environmental factor in asymptomatic gastritis, peptic ulcer and noncardia gastric cancer. Genetic polymorphisms influencing the virulence of the organism may correlate with the results of infection by *H. pylori*. It has been postulated that this association is accompanied by selection of well adapted, host-specific variants that have particular patterns of expression of various virulence factors (Kuipers *et al.*, 2000; Aspholm-Hurtig *et al.*, 2004; Bäckström *et al.*, 2004). Recombination is frequent during transient colonization with multiple strains of *H. pylori* due to DNA transformation, resulting in variants within individual hosts that differ both in sequence content (Falush *et al.*, 2001) and genomic composition (Israel *et al.*, 2001). As a result of frequent recombination, strains of *H. pylori* differ between individual hosts, as well as between isolates from different continents (Achtman *et al.*, 1999; Salama *et al.*, 2001; Falush *et al.*, 2003; Urwin *et al.*, 2003; Olfat *et al.*, 2005). This genetic variation may be due to either genetic drift as a result of geographic isolation (Falush *et al.*, 2003) or adaptation to genetic differences between different ethnic groups of humans (Aspholm-Hurtig *et al.*, 2004). Although sequencing of bacterial genomes is straightforward, and generally the best method for revealing pathoadaptive mutations, it is still an expensive and logistically demanding process (Welch *et al.*, 2002). Recently, comparative genomic hybridization (CGH) has been used to facilitate comparisons of unsequenced bacterial genomes to reveal characteristic genes or chromosomal regions related to unique phenotypes, and to characterize the extensive intraspecies genetic diversity found in bacteria at the whole-genome level (Israel *et al.*, 2001; Chan *et al.*, 2003; Fukiya *et al.*, 2004). Using this technology, Salama *et al.* (2001) revealed that 22% of *H. pylori* genes are absent in at least one of 15 *H. pylori* strains, and the core genes may aid the bacteria under specific circumstances during long-term infection of genetically diverse hosts. Based on previous research results, it is anticipated that whole genome comparisons based on microarrays would not only provide inferences about phenotypic differences within a species but also reveal the general population structure of the bacteria being studied. However, the population structure of *H. pylori* strains iso-

* To whom correspondence should be addressed.
(Tel) 86-21-6320-0874; (Fax) 86-21-6320-0874
(E-mail) liuwzmd@163.com

lated from Chinese patients has not yet been investigated using microarrays.

Studies have reported that the *cag*A, *vac*A, *ice*A1, and *bab*A2 genes are associated with development of peptic ulcer or gastric carcinoma in western countries (Gatti *et al.*, 2005; Olfat *et al.*, 2005), however this trend has not been observed in Asian countries (Han *et al.*, 2004; Zhou *et al.*, 2004; Yamazaki *et al.*, 2005). Therefore, the specific genotypes of clinical strains of *H. pylori* found in China should be analyzed. In this study, we prepared a DNA microarray of the *H. pylori* genome, so the genomic composition of *H. pylori* clinical isolates could be analyzed to characterize genetic diversity between strains and search for new candidate virulence-associated genes. We specifically intended to identify virulence-associated genes conserved across these strains as vaccine candidates.

Materials and Methods

Bacterial strains and growth

H. pylori strains 26695 and J99, both of which have genomes already sequenced, were donated by Prof. David Y. Graham (Baylor College of medicine, Texas Medical Center, Houston, Texas). Thirty-four *H. pylori* clinical strains were isolated from patients undergoing endoscopy at our hospital in Shanghai between 2002 and 2005, including chronic gastritis (CG, n=10), duodenal ulcer (DU, n=11) and gastric carcinoma (GC, n=13). Those cases were diagnosed using endoscopy and histology examination. *H. pylori* were grown on selective Columbia agar (Oxoid, USA) base plates containing 7% defibrinated horse blood, 5 mg/ml trimethoprim (Sigma, USA), 10 mg/ml vancomycin (Sigma, USA) and 2500 units/ml polymyxin B (Sigma, USA) under microaerobic conditions at 37°C.

PCR primer design

The DNA fragments on our microarray corresponded to unique segments of individual open reading frames (ORFs) in the *H. pylori* genome. We attempted to include the superset of ORFs from both published *H. pylori* genomes in our array, including the 1590 ORFs present in strain 26695 and the 91 ORFs found only instrain J99. The sequences used to design PCR primers were found at http://www.tigr.org. Fragments were then generated by PCR using gene-specific primers. To ensure that the elements of our array would specifically detect specifically their corresponding genes alone, the ORF sequences fed into the primer program were circumscribed such that they would exclude regions with high cross homology to other regions of the genome as well as not overlapping an adjacent ORF. Genes were considered homologous if they had greater than 85% homology to the other genome for more than 60% of their length based on continuous homology. We were unable to define unique regions for the genes identified in Table 1, therefore they were excluded from the array. The PCR primers were designed using Primer Premier 5, which generated primer pairs with an optimum rating score and melting temperatures (52°C-56°C, as well as a minimal possibility of hairpin and secondary structure development. Primer pairs were synthesized in 96-well plates (Hua Da Ding An Biological Ltd., China).

PCR products purification

After PCR amplification, PCR products were purified using a Millipore Multiscreen PCR plate, or isopropanol precipitation for small fragment PCR products (≤300 bp). The products were then resuspended in ddH_20 and DMSO to a final concentration of 300 ng/μl. PCR products were spotted onto polylysine coated glass microarray slides using a Genemachine (USA) at a humidity of 50%-60%, then crosslinked using a CL-1000 Ultraviolet Crosslinker. Probes were printed in quadruplicate on the slides.

Preparation and hybridization of genomic DNA probes

Genomic DNA was prepared from plate-grown bacteria using a Qiagen tissue DNA extraction kit (Qiagen, Germany). Two μg of genomic DNA of each clinical strain (test DNA) was labeled with Cy3, and 1 μg each of strain 26695 and J99 genomic DNA (reference DNA) was labeled with Cy5, by reverse transcription using Superscript (Invitrogen). Unincorporated dyes were removed using a QiaQuick Nucleotide Removal kit (Qiagen, Germany) according to the manufacturer's instructions. Thirty pmol of Cy5 and Cy3 probes were mixed and dried in a Speed Vac, then resuspended in 9 μl ddH_2O. After the mixed probe was denatured by incubation for 5 min at 95°C, 11 μg salmon sperm DNA was then added to the probe mixture which was incubated for 45 min at 75°C, 10 μl 4× Buffer and 20 μl formamide was then added and the mixture applied to the slide and incubated for 6 h at 65°C in the dark. Slides were then washed at 55°C with buffer Ⅰ (1× SSC/0.2% SDS) for 10 min, then washed again at 50°C with buffers Ⅱ (0.1× SSC/0.2% SDS) for 10 min and finally at 50°C with buffers Ⅲ (0.1× SSC) for 5 min. The slides were immediately dried by centrifugation at 2,500 rpm for 2 min. Competitive hybridization was done for each strain.

Table 1. Genes excluded from the microarray

H. pylori strain	Absent Gene code							
26695	HP0008	HP0041	HP0118	HP0119	HP0120	HP0140	HP0161	HP0317
	HP0461	HP0533	HP0560	HP0698	HP0725	HP0789	HP0882	HP0904
	HP0923	HP0988	HP0989	HP0997	HP1008	HP1051	HP1095	HP1096
	HP1097	HP1115	HP1116	HP1188	HP1288	HP1289	HP1297	HP1342
	HP1408	HP1409	HP1410	HP1412	HP1534	HP1535	HP1536	HP1562
J99	jhp0318	jhp0934	jhp0958	jhp1306	jhp1422			

Table 2. Strain-specific genes of 34 *H. pylori* clinical strains

Gene function classification	Gene code
Amino acid synthesis	HP0652 HP0283 HP1277
Transport and binding proteins	HP0600 HP0250 HP0298 HP0475 HP0807 HP1400 HP1561
Biosynthesis of cofactors, prosthetic groups, and carriers	HP0407 HP0293 HP0172 HP0768 HP0769 HP0799 HP0800 HP0006 HP0034 HP0841 HP0002 HP0814 HP0p843 HP0845 HP0329
Cellular processes	HP0332 HP0391 HP1119 HP1192 HP1557 HP0243 HP0333 HP0441 HP0887 HP0315 HP0459 HP0521 HP0523 HP0530 HP0531 HP0535 HP0536 HP0538 HP0545 HP0546 HP0547 HP0967 HP0930 jhp0829 jhp0921
Energy metabolism	HP0056 HP0666 HP1135 HP1133 HP0634 HP1458 HP0903 HP0905 HP1385 HP0027 jhp1429 jhp0834 jhp0870
DNA metabolism	HP0387 HP0440 HP0548 HP0602 HP0661 HP0675 HP0790 HP0883 HP0995 HP1009 HP1231 HP1323 HP1347 HP1553 HP0259 HP0050 HP0054 HP0091 HP0260 HP0462 HP0464 HP0478 HP0481 HP0483 HP0592 HP0593 HP0846 HP0848 HP0850 HP1209 HP1352 HP1354 HP1366 HP1367 HP1368 HP1369 HP1370 HP1371 HP1383 HP1403 HP1404 HP1471 HP1472 HP1522 jhp0755 jhp0919 jhp0931 jhp0045 jhp0046 jhp0164 jhp0414 jhp0756 jhp1284 jhp1296 jhp1297 jhp1409
Transcription	HP0640
Regulatory functions	HP0048 HP0088 HP1287 HP1365 jhp0928
Hypothetical Proteins-Conserved	HP0032 HP0035 HP0066 HP0102 HP0312 HP0318 HP0347 HP0373 HP0395 HP0447 HP0465 HP0466 HP0507 HP0518 HP0519 HP0575 HP0639 HP0644 HP0710 HP0713 HP0718 HP0737 HP0892 HP0894 HP0926 HP0934 HP0944 HP0956 HP0966 HP0980 HP1117 HP1286 HP1335 HP1337 HP1343 HP1417 HP1426 HP1438 HP1449 HP1510 HP1551 HP1589 HP0007 HP0018 HP0023 HP0024 HP0030 HP0040 HP0046 HP0052 HP0053 HP0057 HP0058 HP0059 HP0060 HP0061 HP0062 HP0063 HP0064 HP0065 HP0078 HP0081 HP0101 HP0113 HP0114 HP0128 HP0129 HP0131 HP0135 HP0167 HP0168 HP0174 HP0186 HP0187 HP0188 HP0203 HP0204 HP0205 HP0206 HP0217 HP0225 HP0236 HP0253 HP0256 HP0261 HP0262 HP0292 HP0311 HP0314 HP0316 HP0335 HP0336 HP0337 HP0338 HP0339 HP0340 HP0341 HP0342 HP0343 HP0344 HP0345 HP0356 HP0359 HP0386 HP0398 HP0412 HP0423 HP0424 HP0425 HP0426 HP0427 HP0429 HP0430 HP0434 HP0436 HP0439 HP0442 HP0443 HP0444 HP0445 HP0446 HP0449 HP0450 HP0453 HP0455 HP0456 HP0457 HP0458 HP0460 HP0479 HP0482 HP0484 HP0488 HP0492 HP0495 HP0502 HP0503 HP0504 HP0505 HP0513 HP0556 HP0568 HP0578 HP0579 HP0594 HP0605 HP0609 HP0611 HP0629 HP0636 HP0641 HP0647 HP0664 HP0667 HP0668 HP0669 HP0670 HP0673 HP0674 HP0681 HP0682 HP0689 HP0699 HP0704 HP0712 HP0719 HP0720 HP0721 HP0730 HP0731 HP0732 HP0733 HP0744 HP0762 HP0767 HP0784 HP0788 HP0806 HP0842 HP0849 HP0852 HP0856 HP0880 HP0881 HP0893 HP0895 HP0897 HP0901 HP0906 HP0917 HP0935 HP0937 HP0938 HP0945 HP0947 HP0948 HP0953 HP0963 HP0964 HP0965 HP0982 HP0984 HP0985 HP0986 HP0987 HP0990 HP0991 HP0992 HP0993 HP0994 HP0996 HP0999 HP1001 HP1002 HP1003 HP1004 HP1005 HP1074 HP1078 HP1079 HP1089 HP1093 HP1106 HP1142 HP1144 HP1145 HP1146 HP1187 HP1194 HP1250 HP1265 HP1276 HP1283 HP1322 HP1324 HP1334 HP1351 HP1353 HP1382 HP1388 HP1389 HP1390 HP1396 HP1397 HP1405 HP1411 HP1425 HP1433 HP1437 HP1439 HP1499 HP1500 HP1515 HP1516 HP1518 HP1519 HP1520 HP1524 HP1528 HP1586 HP1590 jhp0165 jhp0331 jhp0332 jhp0616 jhp0813 jhp0814 jhp0825 jhp0828 jhp0830 jhp0916 jhp0920 jhp0922 jhp0923 jhp0924 jhp0925 jhp0926 jhp0927 jhp0929 jhp0930 jhp0932 jhp0933 jhp0936 jhp0940 jhp0943 jhp0944 jhp0945 jhp0946 jhp0947 jhp0948 jhp0950 jhp0953 jhp0954 jhp0956 jhp0957 jhp0959 jhp0961 jhp1043 jhp1049 jhp1160 jhp1307 jhp1408 jhp1437 jhp1462 jhp1463 jhp1495 jhp1132

Gene function classification	Gene code
Mobile and extrachromosomal element functions	HP1000 HP1006 HP0413 HP0414 HP0437 HP1007 jhp0935 jhp0826 jhp0827 jhp0951
Protein fate	HP0011 HP0110 HP0033
Protein synthesis	HP0402 HP0617 HP0643 HP0972 HP0125 HP0514 HP1047 HP1415 HP0077 HP0124
Central intermediary metabolism	HP0067 HP0899
Fatty acid and phospholipid metabolism	HP0090 HP0700 HP0808 HP0962
Cell envelope	HP0160 HP0645 HP0772 HP1418 HP0003 HP0208 HP0326 HP0379 HP0619 HP0651 HP0826 HP0855 HP0957 HP1105 HP1578 HP0009 HP0025 HP0079 HP0227 HP0229 HP0252 HP0289 HP0477 HP0610 HP0671 HP0796 HP0896 HP0922 HP1125 HP1157 HP1177 HP1243 HP0876 jhp0937 jhp0949 jhp0820 jhp0917 jhp0918 jhp1032
Purines, pyrimidines, nucleosides, and nucleotides	HP0618 HP0854 HP1530 HP0005 HP0266 HP1011 HP0104 HP0043
Unknown function	HP0322 HP0381 HP0390 HP0405 HP0872 HP1193 HP0653 jhp0540 jhp0585 jhp0955 jhp0726 jhp0960

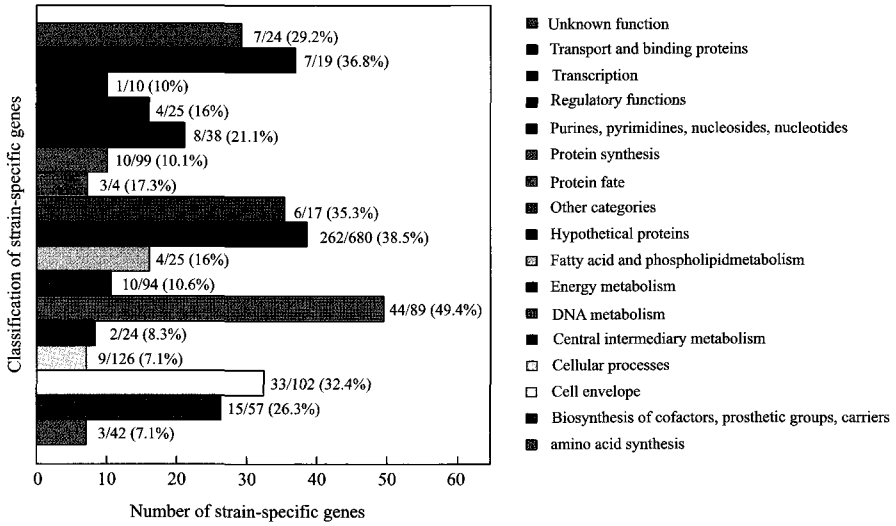

Fig. 1. Homology classes of strain-specific genes according to *H. pylori* 26695. (Left) The number of strain-specific genes. (Right) Different colors represented functional categories of *H. pylori* genome according to http://www.tigr.org/tigr-scripts/CMR2/CMRGenomes.spl. For each bar, numerators represented number of strain-specific gene of all clinical strains, denominator represented the total number of same classification genes in *H. pylori* 26695.

Data analysis and confirmation by PCR

Arrays were scanned using an Agilent scanner and further processed using Imagene 5.1 software. For each test strain, we computed the average signal value of every fragment by using the data from two arrays (each array contains four spots for each fragment). Empty spots and spots with high backgrounds or unusually high or low signal values were excluded from the analysis. Spots with signal/noise <2 were also excluded. The results from empirical studies validate the use of Log ratio thresholds for establishment of gene divergence /absence (Taboada *et al.*, 2005), therefore the average signal values were ln(x) transformed and normalized using the Stanford Microarray Database (http://genome-www5.stanford. edu /resources/restech.shtml). We chose a normalized Cy3/Cy5 (R/G) ratio=0.5 as the cutoff value, which was optimized to control hybridization between strains 26695 and J99 (sensitivity:c98%-100%; specificity: 82%-88%, see results). In other words, if the gene ratio was <0.5 it was considered absent and labeled as 0, however if the ratio was ≥0.5 it was considered present and labeled as

1. Data were further analyzed using the program Cluster and displayed using Treeview. Functional annotations of genes were made according to http://www.tigr.org/tigr-scripts/CMR2/CMRGenomes.spl.

The results revealed by microarray analysis were further confirmed using gene-specific PCR. We randomly selected 96 genes and six *H. pylori* strains to evaluate the hybridization results. PCR reactions were performed as described above.

Results

Quality of prepared micorarray

The total PCR success rate was 99.3%. The final microarray included 1882 DNA fragments, corresponding to 97.3% (1636/1681) of the ORFs of both sequenced *H. pylori* strains, 1549 of which belonged to *H. pylori* 26695 and 87 that belonged to *H. pylori* J99. 10% of spots had signal /noise ratios <2. The microarray had 3.4% (14/412) false positives and 0.27% (19/7116) false negatives for *H. pylori* 26695, and 7.21% (28/388) false positives and 0.24% (1/412) false negatives for *H. pylori* J99, indicating high sensitivity (*H. pylori* J99 99%; *H. pylori* 26695 99%) and specificity (*H. pylori* J99, 82%; *H. pylori* 26695, 86%). The repetitive rate between different dots within the same microarray was 98%, and between genechips was 97%. Gene-specific PCR revealed that among the 96 genes, the average consistency rate between microarray and PCR was 89%-93.1%.

Genetic diversity among H. pylori strains

Our microarray revealed that of the 1636 genes analyzed, 1091 (66.7%) were common to all *H. pylori* strains, representing the functional core of the transcriptom. The *H. pylori* functional core contained most of the genes involved in metabolism (including energetic, fatty acid and phosphorlipid, protein, nucleotide and central intermediary metabolism), cellular processes, transcription and biosynthesis of amino acids, cofactors and carriers. In contrast, 522 (31.9%) genes were missing from at least one of 34 *H. pylori* strains, and therefore designated as strain -specific genes (Table 2). None of these 34 *H. pylori* strains showed a lack of all 522 genes. Strain-specific genes primarily included DNA metabolism components, transposase genes, hypothetical proteins and outer membrane protein (Fig. 1). 23 (1.4%) genes were missing from all 34 *H. pylori* strains being analyzed. It is likely that other strain-specific genes exist but were not represented in the array, resulting in an underestimated number of strain-specific genes.

Cluster analysis of genetic polymorphism of H. pylori clinical strains

Through cluster analysis, we found two diverse regions in the *H. pylori* genome, which were lacking in numerous isolates. Diverse region 1 consisted of HP0424 to HP0462, encoding hypothetical proteins of unknown function, or selfish DNA, such as restriction or modification enzymes. Diverse region 2 was from HP0984 to HP1009 (Fig. 2). Diverse region 1 and 2 respectively corresponded to plasticity zones (PZ) 1 and 2 (Alm *et al.*, 1999), which accounted for only 12.5% of the 522 variably present genes. The remaining 87.5% (457/522) were located in multiple regions scattered around the virtual genome. Thus, hundreds of genes were variably present within *H. pylori* with no obvious genomic clustering. As expected, *cag* pathogenicity island (*cag*PAI) genes were present in nearly all 34 *H. pylori* strains. On the other hand, we found that certain genotypes had higher prevalence in DU or GC groups than in CG groups, such as HP0447 (GC: 23.1%; DU: 0%; CG: 0%), HP0704(GC: 46.2% DU: 27.3%; CG: 10%), and jhp0918 (GC: 38.5%; DU: 18.2%; CG: 0%).

Discussion

It is unlikely that whole-genome sequencing can be used for genotyping or large-scale comparative genomics. Microarray-based CGH provides rich data sets that are useful for both whole-genome genotyping and comparative genomics when whole-genome sequence data is not available. High throughput sequencing of microbial genomes has resulted in relatively rapid accumulation of an enormous amount of genomic sequence data. Comparison of the two *H. pylori* genome sequences reveals that, whereas most of the genes are highly conserved between the two strains (*H. pylori* 26695 and J99), approximately 6% of the genes are unique to each genome (Alm *et al.*, 1999). Genes important to basic metabolism and growth of bacteria are relatively conserved. Similarly, our microarray results revealed that in the 1636 genes analyzed, most were present in all 34 clinical strains, revealing a core set of the *H. pylori* genome, containing most of the genes participating in metabolism, cellular processes and transcription and biosynthesis of amino acids, functions essential to *H. pylori* growth, colonization and pathogenicity in the host. The virulence genes of *H. pylori*, such as most flagellum associated proteins (HP0840 and HP0601) and partial *cag*PAI genes (HP0524, HP0525, HP0526, HP0533, HP0536, and HP0540) also were found in the core DNA, indicating that these virulence genes are the basic pathogenic factors necessary for *H. pylori* to induce clinical diseases. *H. pylori* that possess the *cag*PAI gene are more virulent than strains that do not and previous research has revealed over 80% of Chinese strains are *cag*PAI positive (Liu *et al.*, 2000). Our microarray results also revealed that the *cag*PAI genes were present in nearly all 34 *H. pylori* strains. Due to the size of the *cag*PAI gene (38 kb) and the fact that the median size of DNA fragments exchanged by recombination only being 450 bp, we can infer that selection pressures are not very high (Falush *et al.*, 2001). Selection for the type four secretion system encoded by the *cag*PAI gene may have resulted from descent from an ancestor that had already imported the *cag*PAI genes. The presence of the *cag*PAI genes in all of these populations would then reflect its spread via transformation from the cells that first acquired it, coupled with selection for its expression (Backert *et al.*, 2000; Odenbreit *et al.*, 2000; Stein *et al.*, 2000).

Analysis at the gene level reveals *H. pylori* strains have significant genetic diversity in different infected individuals. Certain genotypes may affect colonization, multiplication of *H. pylori* and the development of its associated clinical

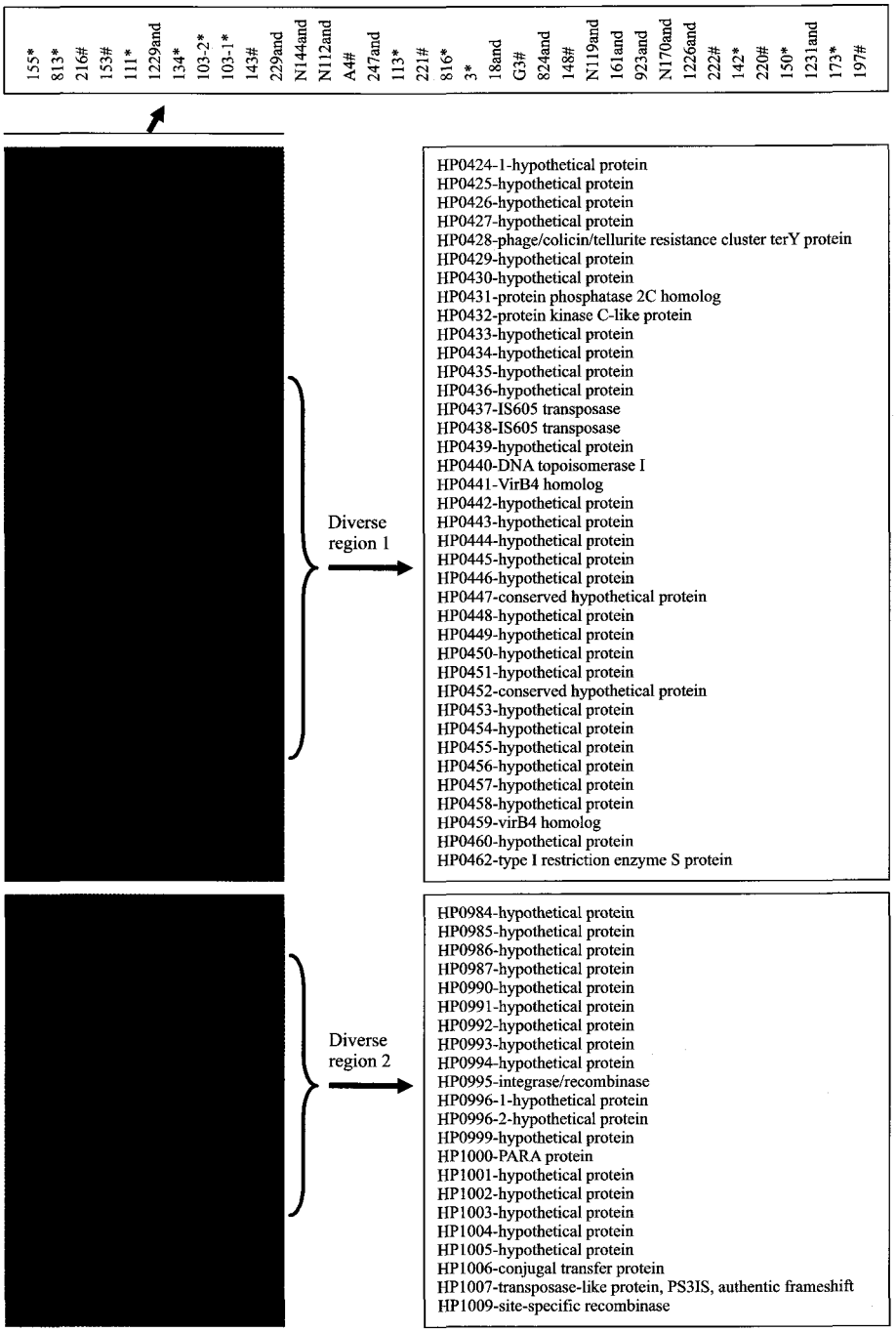

Fig. 2. Cluster analysis of diverse region 1 and 2 in genome composition of the 34 *H. pylori* strains. (Left) The columns marked the name of different *H. pylori* strains(*: DU strains; #: CG strains; and: GC strains). Each row corresponds to a specific spot on the array. (Right) Specific gene description was made for diverse regions 1 and 2. The presence (red) or absence (black) of genes was displayed according to their position on the chromosome for each strain. Gene codes were made according to http://www.tigr.org/tigr-scripts/CMR2/CMRGenomes.spl.

outcome (Lu *et al.*, 2005a; Momynalie*et al.*, 2005). In our experiment, most strain-specific genes encoded hypothetical proteins of unknown function. The two largest classes of strain-specific genes with known function were the restriction modification system components and transposase genes, which regulate DNA exchange among bacteria and may promote genetic diversity of *H. pylori* (Lin *et al.*, 2001; Takata *et al.*, 2002). The results of genomic sequencing and comparative genomics analyses suggest that evolution of pathogenic bacteria is mainly due to large chromosomal alterations, such as horizontal transfers or deletions (Jin *et al.*, 2002; Welch *et al.*, 2002). At present, it is accepted that there are three mechanisms of gene acquisition and loss that may contribute to *H. pylori*'s phenotypic diversity: mutation, recombination mediated by transposon and repetitive sequences, and DNA horizontal transfer. We found

that among strain-specific genes, 12.5% of genes, including many transposase and endonuclease genes, were located in the PZ, which is consistent with the conclusion that PZ is a site of extensive insertion, excision, and recombination (Salama *et al.*, 2001). 87.5% of strain-specific genes were located in multiple regions scattered around the virtual genome. The mosaic distribution of absent regions indicated that the genomes of pathogenic strains were highly diversified due to insertions and deletions. The microenvironment in which *H. pylori* colonization occurs could promote development of DNA absorbance and gene exchange, which in return could result in extensive genomic diversity. These characteristics contribute to long term survival and evolution of *H. pylori* in hosts (Fitzgerald *et al.*, 2001). Most of the strain-specific genes encoded proteins of unknown function or selfish DNA, such as restriction or modification enzymes, and may not be targets for positive selection. Therefore, it seemed likely that repeated loss rather than recent acquisition accounts for the variability in the *H. pylori* genome (Gressmann *et al.*, 2005). The remaining strain-specific genes included those involved in biosynthesis and degradation of surface polysaccharides and lipopolysaccharides, DNA replication, recombination and repair, LPS biosynthesis proteins and glucose metabolism components. These strain-specific genes may represent those with redundant functions required for specific niches or whose DNA sequences have high enough variation between strains that they cannot be detected under our hybridization conditions. We also used hierarchical clustering to explore the relationship of different *H. pylori* strains based on their genomic components. We found certain genotypes were more prevalent in the DU or GC group than in CG groups, such as jhp0918 (GC: 38.5%; DU: 18.2%; CG: 0%). Recently, research has revealed that jhp0918 is present in 42% of *H. pylori* strains associated with duodenal ulcers vs. 21% of strains associated with gastritis. Its presence is also associated with more intense antral neutrophil infiltration and higher IL-8 levels, and is a marker for protection against gastric atrophy, intestinal metaplasia, and gastric cancer (Lu *et al.*, 2005b). Our microarray results indicated which genotypes are most likely related to the clinical outcome of *H. pylori* infection by analyzing its positive rate in different disease groups, enabling research on clinically significant genes without the need for whole genome sequencing. However, other methods such as polymerase chain reaction (PCR), gene mutation and animal models, are required to enlarge sample sizes and test and verify the clinical disease-associated genotypes suggested by our experiment.

Among strain specific genes, those associated with *H. pylori* virulence primarily included a few flagellar motility associated proteins (HP1192 and HP1557), *vac*A (HP0887) partial *cag*PAI genes (HP0521, HP0523, HP0530, HP0531, HP0546 and HP0547), *nap*A (HP0243), HP1177, and HP1243. The discrepancy in virulence genes can also account for differences in virulence and pathogenicity among different *H. pylori* strains. These strain-specific virulence genes may contribute to bacterial adaptations and pathogenesis in genetically diverse hosts.

Our results suggest that CGH microarray analysis could be a rapid and powerful method for extracting candidate regions for pathoadaptive mutations because un-sequenced strains are easily subjected to genomic comparison analysis. Strain-specific genes were candidates for pathoadaptive mutations that contribute to pathogenicity based on their absence. Conserved sequences flanking missing ORFs may serve as good primers for amplifying strain-specific regions. Therefore, our CGH information for pathogenic strains could be useful for rapid identification and isolation of characteristic regions of pathogenic strains. Virulence-associated genes conserved across strains could also be considered as vaccine candidates. An increasing body of *H. pylori* CGH data will enable us to begin formulating hypotheses about *H. pylori* genome evolution and development of the wide variation in virulence.

Acknowledgements

This study supported by Shanghai leading academic discipline project (No. Y0205).

References

Achtman, M., T. Azuma, D.E. Berg, Y. Ito, G. Morelli, Z.J. Pan, S. Suerbaum, S.A. Thompson, A. van der Ende, and L.J. van Doorn. 1999. Recombination and clonal groupings within Helicobacter pylori from different geographical regions. *Mol. Microl. biol.* 32, 459-470.

Alm, R.A., L.S.L. Ling, D.T. Moir, B.L. King, E.D. Brown, P.C. Doig, D.R. Smith, B. Noonan, B.C. Guild, B.L. Dejonge, G. Carmel, P.J. Tummino, A. Caruso, M. Uria-Nickelsen, D.M. Mills, C. Ives, R. Gibson, D. Merberg, S.D. Mills, Q. Jiang, D.E. Taylor, G.F. Vovis, and T.J. Trust. 1999. Genomic sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* (London) 397, 176-180.

Aspholm-Hurtig, M., G. Dailide, M. Lahmann, A. Kalia, D. Ilver, N. Roche, S. Vikström, R.Sjöström, S. Lindén, A.R. Bäckström, C. Lundberg, A. Arnqvist, J. Mahdavi, U.J. Nilsson, B. Velapatiño, R.H. Gilman, M. Gerhard, T. Alarcon, M. López-Brea, T. Nakazawa, J.G. Fox, P. Correa, M.G. Dominguez-Bello, G.I. Perez-Perez, M.J. Blaser, S. Normark, I. Carlstedt, S. Oscarson, S. Teneberg, D.E. Berg, and T. Borén. 2004. Functional adaptation of BabA, the *H. pylori* ABO blood group antigen binding adhesin. *Science* 305, 519 -522.

Backert, S., E. Ziska, V. Brinkmann, U. Zimny-Arndt, A. Fauconnier, P.R. Jungblut, M. Naumann, and T.F. Meyer. 2000. Translocation of the *Helicobacter pylori* CagA protein in gastric epithelial cells by a type IV secretion apparatus. *Cell. Microbiol.* 2, 155-164.

Bäckström, A., C. Lundberg, D. Kersulyte, D.E. Berg, T. Borén, and A. Arnqvist. 2004. Metastability of *Helicobacter pylori* bab adhesin genes and dynamics in Lewis b antigen binding. *Proc. Natl. Acad. Sci. USA* 101, 16923-16928.

Chan, K., S. Baker, C.C. Kim, C.S. Detweiler, G. Dougan, and S. Falkow. 2003. Genomic comparison of *Salmonella enterica serovars* and *Salmonella bongori* by use of an *S. enterica serovar Typhimurium* DNA mircroarray. *J. Bacteriol.* 185, 553-563.

Falush, D., C. Kraft, N.S. Taylor, P. Correa, J.G. Fox, M. Achtman, and S. Suerbaum. 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. USA* 98, 15056-15061.

Falush, D., T. Wirth, B. Linz, J.K. Pritchard, M. Stephens, M. Kidd, M.J. Blaser, D.Y. Graham, S. Vacher, G.I. Perez-Perez, Y. Yamaoka, F. Mégraud, K. Otto, U. Reichard, E. Katzowitsch, X.Y. Wang, M. Achtman, and S. Suerbaum. 2003. Traces of human migrations in *Helicobacter pylori* populations. *Science*

299, 1582-1585.

Fitzgerald, J.R. and J.M. Musser. 2001. Evolutionary genomics of pathogenic bacteria. *Trends Microbiol.* 9, 547-553.

Fukiya, S., H. Mizoguchi, T. Tobe, and H. Mori. 2004. Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* Strains revealed by comparative genomic hybridization microarray. *J. Bacteriol.* 186, 3911-3921.

Gatti, L.L., E. F. Souza, K. Leite, E. de Souza Bastos, L. Vicentini, L. da Silva, M. Smith, and S. Payão. 2005. *cag*A, *vac*A alelles and babA2 genotypes of *Helicobacter pylori* associated with gastric disease in Brazilian adult patients. *Diagn. Microbiol. Infect. Dis.* 51, 231-235.

Gressmann, H., B. Linz, R. Ghai, K.P. Pleissner, R. Schlapbach, Y. Yamaoka, C. Kraft, S. Suerbaum, T.F. Meyer, and M. Achtman. 2005. Gain and loss of multiple genes during the evolution of *Helicobacter pylori*. *PLoS. Genet.* 1, 419-428.

Han, Y.H., W.Z. Liu, H.Y. Zhu, and S.D. Xiao. 2004. Clinical relevance of *ice*A and *bab*A2 genotypes of *Helicobacter pylori* in a Shanghai population. *Chin. J. Dig. Dis.* 5, 181-185.

Israel, D.A., N. Salama, U. Krishna, U.M. Rieger, J.C. Atherton, S. Falkow, and R.M. Peek, Jr. 2001. *Helicobacter pylori* genetic diversity within the gastric niche of a single human host. *Proc. Natl. Acad. Sci. USA* 98, 14625-14630.

Jin, Q. Z.H. Yuan, J.G. Xu, Y. Wang, Y. Shen, W.C. Lu, J.H. Wang, H. Liu, J. Yang, F. Yang, X.B. Zhang, J.Y. Zhang, G.W. Yang, H.T. Wu, D. Qu, J. Dong, L.L. Sun, Y. Xue, A.L. Zhao, Y.S. Gao *et al.* 2002. Genome sequence of *Shigella flexneri* 2a: insights into patho-genicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids. Res.* 30, 4432-4441.

Kuipers, E.J., D.A. Israel, J.G. Kusters, M.M. Gerrits, J. Weel, A. van der Ende, R.W.M. van der Hulst, H.P. Wirth, J. Höök-Nikanne, S.A. Thompson, and M.J. Blaser. 2000. Quasi-species development of *Helicobacter pylori* observed in paired isolates obtained years apart from the same host. *J. Infect. Dis.* 181, 273-282.

Lin, L.F., J. Posfai, R.J. Roberts, and H. Kong. 2001. Comparative genomics of the restriction-modification systems in *Helicobacter pylori*. *Proc. Natl. Acad. Sci. USA* 98, 2740-2745.

Liu, J., G.M. Xu, Z.X. Tu, Z.S. Li, Y.F. Gong, and X.H. Ji. 2000. The distribution and significance of cag pathogenicity island of *Helicobacter pylori* isolated from Chinese patients. *Chin. J. Intern. Med.* 39, 457-460.

Lu, H., P.I. Hsu, D.Y. Graham, and Y. Yamaoka. 2005a. Duodenal ulcer promoting gene of *Helicobacter pylori*. *Gastroenterology* 28, 833-848.

Lu, H., Y. Yamaoka, and D.Y. Graham. 2005b. *Helicobacter pylori* virulence factors: facts and fantasies. *Curr. Opin. Gastroenterol.* 21, 653-659.

Momynaliev, K.T., S.I. Rogov, O.V. Selezneva, V.V. Chelysheva, T.A. Akopian, and V.M. Govorun. 2005. Comparative analysis of transcription profiles of *Helicobacter pylori* clinical isolates. *Biochemistry* 70, 383-390.

Odenbreit, S., J. Püls, B. Sedlmaier, E. Gerland, W. Fischer, and R. Haas. 2000. Translocation of *Helicobacter pylori* CagA into gastric epithelial cells by type IV secretion. *Science* 287, 1497-1500.

Olfat, F.O., Q. Zheng, M. Oleastro, P. Voland, T. Borén, R. Karttunen, L. Engstrand, R. Rad, C. Prinz, and M. Gerhard. 2005. Correlation of the *Helicobacter pylori* adherence factor BabA with duodenal ulcer disease in four European countries. *FEMS Immunol. Med. Microbiol.* 44, 151-156.

Salama, N., K. Guillemin, T.K. McDaniel, G. herlock, L. Tompkins, and S. Falkow. 2001. A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proc. Natl. Acad. Sci. USA* 97, 14668-14673.

Stein, M., R. Rappuoli, and A. Covacci. 2000. Tyrosine phosphorylation of the *Helicobacter pylori* CagA antigen after cag-driven host cell translocation. *Proc. Natl. Acad. Sci. USA* 97, 1263-1268.

Taboada, E.N., R.R. Acedillo, C.C. Luebbert, W.A. Findlay, and J.H. Nash. 2005. A new approach for the analysis of bacterial microarray-based comparative genomic hybridization: insights from an empirical study. *BMC. Genomics* 6, 78-88.

Takata, T., R. Aras, D. Tavakoli, T. Ando, A.Z. Olivares, and M.J. Blaser. 2002. Phenotypic and genotypic variation in methylases involved in type II restriction-modification systems in *Helicobacter pylori*. *Nucleic Acids. Res.* 30, 2444-2452.

Urwin, R. and M.C.J. Maiden. 2003. Multilocus sequence typing: A tool for global epidemiology. *Trends Microbiol.* 11, 479-487.

Welch, R.A., V. Burland, G. Plunkett, P. Redford, P. Roesch, D. Rasko, E.L. Buckles, S.R. Liou, A. Boutin, J. Hackett, D. Stroud, G.F. Mayhew, D.J. Rose, S. Zhou, D.C. Schwartz, N.T. Perna, H.L.T. Mobley, M.S. Donnenberg, and F.R. Blattner. 2002. Extensive mosaic structure revealed by the complete genome sequence of uro-pathogenic *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* 99, 17020-17024.

Yamazaki, S., A. Yamakawa, T. Okuda, M. Ohtani, H. Suto, Y. Ito, Y. Yamazaki, Y. Keida, H. Higashi, M. Hatakeyama, and T. Azuma. 2005. Distinct diversity of *vac*A, *cag*A, and *cag*E genes of *Helicobacter pylori* associated with peptic ulcer in Japan. *J. Clin. Microbiol.* 43, 3906-3916.

Zhou, W., S. Yamazaki, A. Yamakawa, M. Ohtani, Y. Ito, Y. Keida, H. Higashi, M. Hatakeyama, J. Si, and T. Azuma. 2004. The diversity of vacA and *cag*A genes of *Helicobacter pylori* in East Asia. *FEMS Immunol. Med. Microbiol.* 40, 81-87.