Genome-Wide Survey of Mutual Homologous Recombination in a Highly Sexual Bacterial Species

Koji Yahara[1,2,3], Mikihiko Kawai[2,3,4], Yoshikazu Furuta[2,3], Noriko Takahashi[2,3], Naofumi Handa[2,3], Takeshi Tsuru[2,3], Kenshiro Oshima[5], Masaru Yoshida[6], Takeshi Azuma[6], Masahira Hattori[5], Ikuo Uchiyama[4], and Ichizo Kobayashi[2,3,7,]*

[1]Division of Biostatistics, Graduate School of Medicine, Kurume University, Fukuoka, Japan

[2]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, University of Tokyo, Minato-ku, Japan

[3]Institute of Medical Science, University of Tokyo, Minato-ku, Japan

[4]Laboratory of Genome Informatics, National Institute for Basic Biology, Okazaki, Aichi, Japan

[5]Department of Computational Biology, Graduate School of Frontier Sciences, University of Tokyo, Kashiwa, Chiba, Japan

[6]Department of Gastroenterology, Graduate School of Medicine, Kobe University, Chuo-ku, Hyogo, Japan

[7]Graduate Program in Biophysics and Biochemistry, Graduate School of Science, University of Tokyo, Tokyo, Japan

*Corresponding author: E-mail: ikobaya@ims.u-tokyo.ac.jp.

Abstract

The nature of a species remains a fundamental and controversial question. The era of genome/metagenome sequencing has intensified the debate in prokaryotes because of extensive horizontal gene transfer. In this study, we conducted a genome-wide survey of outcrossing homologous recombination in the highly sexual bacterial species *Helicobacter pylori*. We conducted multiple genome alignment and analyzed the entire data set of one-to-one orthologous genes for its global strains. We detected mosaic structures due to repeated recombination events and discordant phylogenies throughout the genomes of this species. Most of these genes including the "core" set of genes and horizontally transferred genes showed at least one recombination event. Taking into account the relationship between the nucleotide diversity and the minimum number of recombination events per nucleotide, we evaluated the recombination rate in every gene. The rate appears constant across the genome, but genes with a particularly high or low recombination rate were detected. Interestingly, genes with high recombination included those for DNA transformation and for basic cellular functions, such as biosynthesis and metabolism. Several highly divergent genes with a high recombination rate included those for host interaction, such as outer membrane proteins and lipopolysaccharide synthesis. These results provide a global picture of genome-wide distribution of outcrossing homologous recombination in a bacterial species for the first time, to our knowledge, and illustrate how a species can be shaped by mutual homologous recombination.

Key words: homologous recombination, horizontal transfer, population genomics, species, *Helicobacter pylori*.

Introduction

The nature of a species has been a fundamental and controversial question in biology for centuries (Darwin 1859; Mayr 1942). The biological species concept defines a species as a reproductively isolated group of organisms that exchange genetic material by interbreeding, and this definition has been widely accepted for eukaryotes since the mid-20th century. However, the era of genome/metagenome sequencing has intensified the debate in prokaryotes (Achtman and Wagner 2008) because extensive

horizontal gene transfer across species boundaries (Nakamura et al. 2004; Fraser et al. 2009) makes the very existence of separate species debatable (Achtman and Wagner 2008; Doolittle and Zhaxybayeva 2009). More than a dozen attempts have been made to establish a conceptual framework for defining prokaryote species, including the ecotype model (Cohan and Perry 2007). However, none of these models is based on genome-wide sequence data.

Recently, effect of homologous recombination between lineages in maintaining cohesion within a bacterial species

has been pointed out (Fraser et al. 2007, 2009; Didelot et al. 2011; Takuno et al. 2012). From this point of view, it is important to reveal flux of mutual homologous recombination between lineages and how it shapes a bacterial species. The extent of outcrossing homologous recombination throughout the entire genome has not been quantitatively analyzed using the genome-wide sequence data in bacteria (Konstantinidis et al. 2006). It thus remains a challenge to reveal genome-wide distribution of the homologous recombination rate in a bacterial species.

A related issue is the relation between genome diversity and homologous recombination. Homologous recombination rate is known to be correlated with DNA diversity in *Drosophila melanogaster* (Begun and Aquadro 1992), although such a correlation is questionable in humans (Spencer et al. 2006). Therefore, it is also important to reveal a genome-wide relationship between homologous recombination rate and DNA diversity in a bacterial species. It will provide a basis to detect genes with high or low recombination rates that deviate from the relationship, which may be a characteristic of the species.

From this perspective, *Helicobacter pylori* is of great interest. This bacterium is present in the stomach of over half the human population, where it is linked with gastritis (stomach inflammation), ulcers, and gastric (stomach) cancer (Yamaoka 2008). It exhibits a remarkable allelic diversity (an "allele" indicates one of the alternative sequences that is possible at a locus in a genome). It is a highly sexual bacterial species, and the allelic diversity is primarily attributed to high homologous recombination between coinfecting lineages following natural transformation in the stomach (Suerbaum and Josenhans 2007). Homologous recombination is much more frequent than point mutation (Suerbaum et al. 1998). One homologous recombination event imports a cluster of small nucleotide polymorphisms into the genome, which increases the relative effect of recombination compared with mutation (Kennemann et al. 2011). Previous population genetic studies on homologous recombination in *H. pylori*, however, used a relatively small number of loci, in particular the seven genes used for multilocus sequence typing (MLST) (Falush et al. 2001, 2003; Linz et al. 2007; Moodley et al. 2009).

In this study, we performed a genome-wide analysis of outcrossing homologous recombination using entire genome sequences of global *H. pylori* strains.

Materials and Methods

Helicobacter pylori Genome Sequences

Helicobacter pylori strain names and accession numbers from GenBank were as follows (Furuta et al. 2011): J99, NC_000921.1; P12, NC_011498.1 and NC_011499.1; G27, NC_011333.1 and NC_011334.1; HPAG1, NC_008086.1 and NC_008087.1; 26695, NC_000915.1; Shi470,

NC_010698.2; F16, DDBJ:AP011940; F30, DDBJ:AP011941 and AP011942; F32, DDBJ: AP011943 and AP011944; F57, DDBJ:AP011945.

Sequence Alignment of Orthologous Genes

An entire data set of orthologous genes (one-to-one orthologous groups in supplementary table S1, Supplementary Material online, which were almost equivalent to the core genes) was prepared by clustering using DomClust (Uchiyama 2006) and RECOG (http://mbgd.genome.ad.jp/RECOG/). The core genes were then extracted using CoreAligner (Uchiyama 2008). Protein sequences were aligned using ClustalW (Thompson et al. 1994). The aligned sequences were then replaced with the corresponding DNA sequences to ensure that gaps occurred only at codon boundaries. We automatically classified orthologous genes based on the functional categories in MBGD (Uchiyama 2003). We then treated outer membrane protein (OMP) and restriction–modification (RM) system as separate categories.

Phylogenetic Analysis of Core, MLST, and Individual Genes

The phylogenetic analysis of core genes and concatenated MLST genes (*atpA*, *efp*, *mutY*, *ppa*, *trpC*, *ureI*, and *yphC*) was conducted using MOLPHY (Adachi and Hasegawa 1996) and Neighbor-Net (Bryant and Moulton 2004).

Bayesian phylogenetic analyses of individual genes were conducted using MrBayes 3.1.2., with a GTR + G + I nucleotide substitution model in a partitioning scheme with three subsets, which corresponded to the three codon positions (Huelsenbeck and Ronquist 2001). All the parameters were unlinked. In Markov chain Monte Carlo procedure, the number of generations was 500,000, and the first 1,250 generations were discarded as a burn-in while the sampling frequency was 100.

Individual genes that did not fit significantly well to the core tree were examined using the Shimodaira–Hasegawa test (Shimodaira and Hasegawa 1999). We calculated the Robinson and Foulds distance (Robinson and Foulds 1981) of an individual tree relative to the core tree, and the overall picture was obtained by multidimensional scaling (Cox TF and Cox MAA 1994) and clustering (Hartigan 1975).

Genome-Wide Detection of Recombination-Derived Mosaics

To obtain an overall view of recombination-derived mosaic structures throughout the entire genome, we extended the bootscan analysis (Salminen et al. 1995) using Hyphy version 2 (Kosakovsky Pond et al. 2005) with a window size of 800 bp and a step size of 30 bp using multiple genome alignments generated by Mauve (Darling et al. 2004). For each window in the genome, the bootscan values were calculated from

a bootstrapped phylogenetic tree. To eliminate noise during phylogenetic estimation, we did not use bootscan values that were less than 90. The window size and step size were set as 800 and 30 bp, respectively. Column containing gaps were not used in the phylogenetic estimation. We validated the settings by confirming that the turnover of bootscan values, as an indicator of a mosaic boundary, was not found in pseudo-sequence alignments where the columns had been randomly shuffled.

Estimation of Minimum Number of Recombination Events and Recombination Rate

For each orthologous gene, the minimum number of recombination events (r_{min}) was calculated using the four-gamete test (Hudson and Kaplan 1985). This test locates pairs of closest segregating sites with four haplotypes that could not have arisen without recombination or a recurrent mutation. Homologous recombination is much more frequent than mutation *in H. pylori* (Suerbaum et al. 1998; Morelli et al. 2010; Kennemann et al. 2011), which makes the four-gamete test method suitable for the estimation of recombination events. We used the method implemented in the PGEToolbox (Cai 2008), which filters gaps in advance. We used the minimum number of recombination events divided by gene length in nucleotides (r_{min}/nt) as a measure of recombination rate.

Identification of Genes with a Particularly High or Low Recombination Rate

We identified these using two approaches.

The first approach (method A) was based on the regression of the minimum number of recombination events on π (nucleotide diversity) without intercept. The Poisson regression was conducted using the r_{min} as the response variable, whereas the linear regression was conducted using the r_{min}/nt as the response variable. We used the no-intercept models so that the regression line includes the origin because recombination cannot be detected without nucleotide diversity. The π for each orthologous gene was calculated using DnaSP (Librado and Rozas 2009). Using the regression, we detected genes that deviated significantly from the regression line. We did not use highly diverged genes where $\pi > 0.08$ during fitting because most of these did not fit the regression model and because we thought that they require another analysis (see below) and biological explanations (see text). We excluded three exceptional genes (HP0462, HP1438, and HP1439) where $\pi > 0.2$ for the same reason.

The second approach (method B) was used for those highly diverged genes where $\pi > 0.08$. Of these genes, we extracted those with r_{min} or r_{min}/nt in the top or bottom 2.5% of all the genes.

In both approaches, we did not use genes with more than 50% of gaps in the alignment.

Identification of Horizontally Transferred Genes from Distantly Related Organisms

We also determined the minimum number of probable recombination events in horizontally transferred genes. Horizontally transferred genes from distantly related organisms were inferred using a Bayesian inference program with training models based on the nucleotide composition. This method exploits fixed order 5-mer nucleotide "words" that deviate from the background genome, and it has been applied successfully to many bacterial genomes including *H. pylroi* 26695 and J99 (Nakamura et al. 2004). Horizontally transferred "alien" genes such as genomic island were inferred using Alien Hunter (Vernikos and Parkhill 2006), which identifies atypical nucleotide compositions based on variable order motif distributions. The transferred genes in the *H. pylori* 26695 genome, which had a one-to-one orthologous relationship, were used to calculate the minimum number of recombination events during comparisons with other genes.

Visualization of a Genome Map

A genome map of *H. pylori* strain 26695 was constructed to show distribution of the r_{min}/nt (as a measure of the homologous recombination rate), which classified the genes into the three categories according to the r_{min}/nt: >top 25%, from bottom 25% to top 25%, <bottom 25%. Also shown were the genes with a particularly high or low recombination rate and horizontally transferred genes and aliens (see above). The genome map was illustrated by DNAPlotter (Carver et al. 2009).

Identification of Functional Motifs/Domains

We searched for PROSITE motifs (Sigrist et al. 2002) in genes with a particularly high recombination rate using the ps_scan program (de Castro et al. 2006). We also searched for conserved domains in the CDD (Conserved Domain Database) using the NCBI Batch Web CD-Search Tool (Marchler-Bauer et al. 2009).

Results

Phylogenetic Analysis of All Genes Suggested Their Mutual Recombination

Complete genome sequences were obtained for ten global *H. pylori* strains (see Materials and Methods). Figure 1a and b shows the maximum likelihood phylogenetic trees for the concatenated MLST genes (3,406 bp) and the core genes (1,097,937 bp), respectively. The tree for the core genes (fig. 1b) had visibly better resolution with higher bootstrap values. However, different patterns were observed in the phylogenetic network analysis (fig. 1c
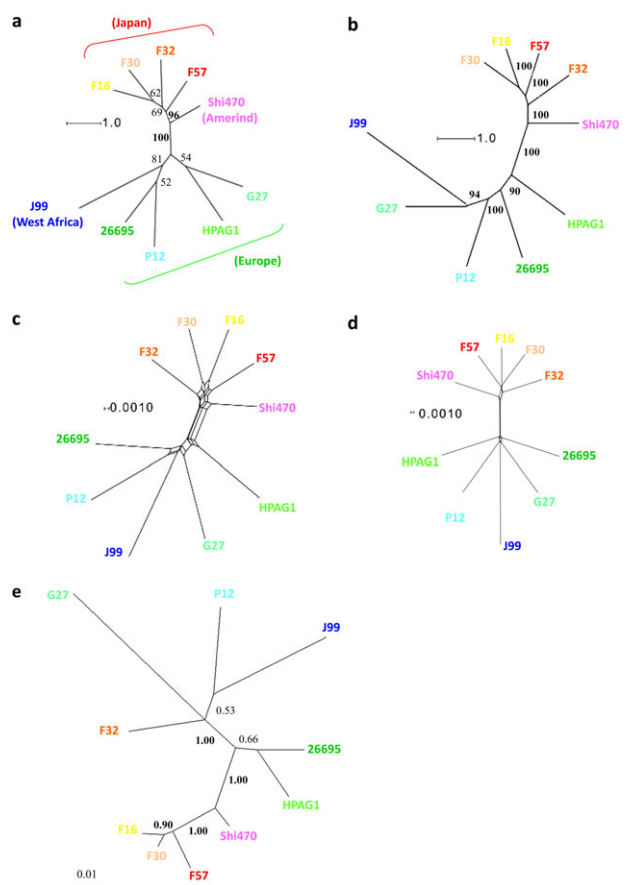
Fig. 1.—Phylogenetic trees of MLST genes, core genes, and individual genes. (*a*) Maximum likelihood trees of MLST genes (3,406 bp) and (*b*) core genes (1,097,937 bp). (*c*) Phylogenetic networks of MLST genes and (*d*) core genes. (*e*) An example of trees with clear topology differences with the core tree for *ubiA*, prenyltransferase (HP1360). Scale bars indicate the number of substitutions per nucleotide site. Numbers indicate the bootstrap values (in *a* and *b*) or posterior probabilities (in *e*).

Fig. 2.—Variable topology of individual gene trees. (*a* and *b*) Multidimensional scaling. (*c* and *d*) Clustering. The colors indicate topological distances to the core tree. The red tree has an identical topology to the core tree. Each branch in the clustering represents one topology. The length of the bar below each branch indicates the number, on a logarithmic scale, of genes with that topology.

and *d*) where the topology was polytomous with no tree-like structures in either of the two major clusters from the east (Japan & Amerind) and west (Europe and West Africa), suggesting substantial recombination between the seven genes within each cluster (fig. 1c). Recombination appeared to have had a similar influence on the core genes (fig. 1*d*).

The probable effect of genetic information transfer among different phylogeographic lineages was also apparent in the individual gene trees, which did not fit significantly well to the core tree (*P* < 0.001, Shimodaira–Hasegawa test, as shown in supplementary table S2, Supplementary Material online). The example shown in figure 1e suggests horizontal transfer from a European/African lineage to a Japanese lineage (F32).

We summarized the topological distances for each of the 1,224 trees relative to the core tree (supplementary table S2, Supplementary Material online) using a multidimensional scaling plot (fig. 2a and b) and a clustering diagram

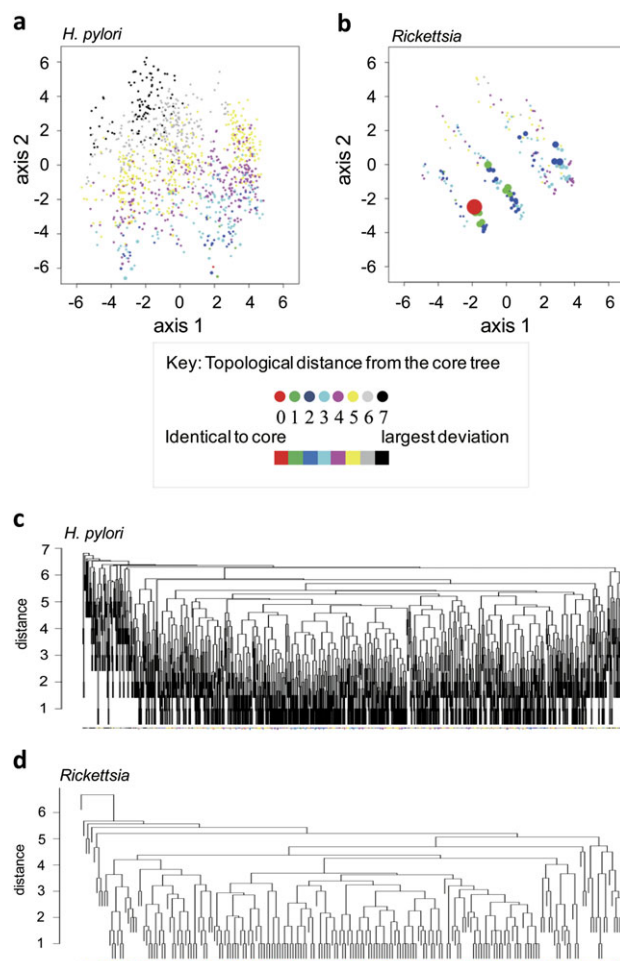(fig. 2c and d). The topological diversity in the gene trees of *H. pylori* was clearly greater than that in *Rickettsia*, which is an endosymbiont bacterium where recombination is expected to be rare. In *H. pylori*, gene trees with an identical topology to the core tree were rare, whereas those with a deviant topology were conspicuous and scattered. Most of the *H. pylori* gene trees had different topologies. Thus, all the phylogenetic lines of evidence suggested frequent recombination between *H. pylori* lineages.

Genome-Wide Mosaics due to Mutual Homologous Recombination

Next, we visualized the consequences of the probable frequent mutual homologous recombinations throughout the entire genome (fig. 3). Using the bootscan analysis
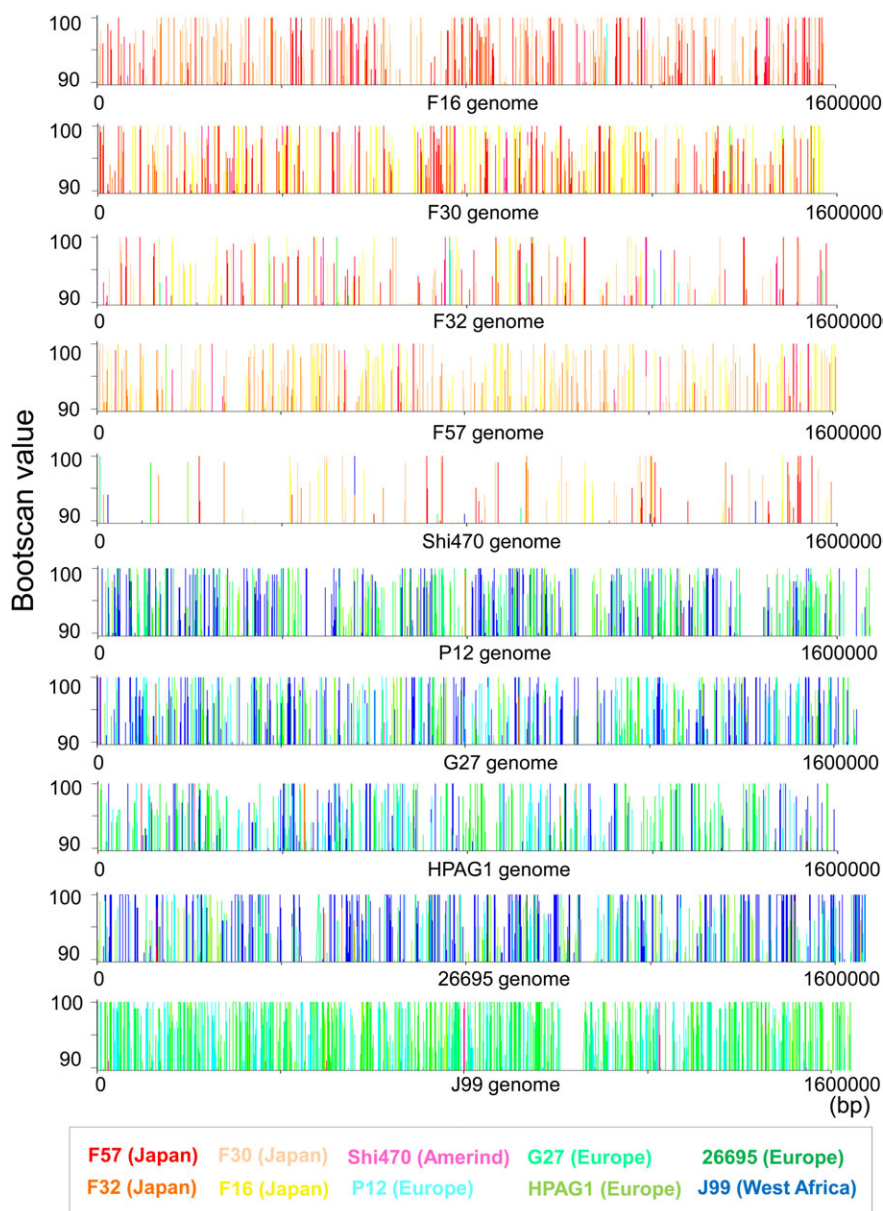
Fig. 3.—Genome-wide mosaic structure indicating recombination. The bootscan values, which are indicative of phylogenetic similarity to the query genome (shown in horizontal axis), were plotted for each of the other nine genome sequences.

(Salminen et al. 1995), a sliding window approach, phylogenetic trees were estimated with bootstrapping throughout the genome. Figure 3 shows the bootscan values (i.e., bootstrap values on a branch grouping the query strain with other strains) for each of the *H. pylori* genomes that were used as queries. Recombination can alter phylogenetic relationships between the query and other genomes, so the turnover of the bootscan values between windows can indicate a recombination-derived mosaic structure (Lole et al. 1999). The turnover of the bootscan values indicative of mosaic boundaries was common throughout the entire genome. This overall view of mosaic structures demon-strated that there had been frequent homologous recombinations between *H. pylori* genomes. Furthermore, this approach revealed more cohesive connections due to recombination within a subgroup, for example, east and west.

Minimum Number of Recombination Events in Each Gene

Next, we calculated the minimum number of recombination events (r_{min}) for every gene (supplementary table S2, Supplementary Material online) using the four-gamete test, a simple method locating a pair of segregating sites with
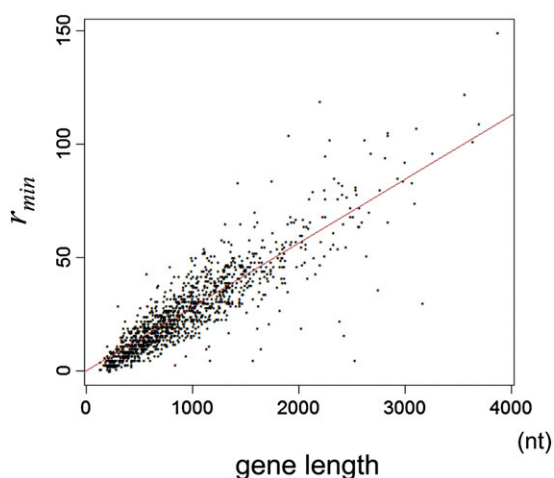
FIG. 4.—Linear relationship between the minimum number of recombination events (r_{min}) and gene length. The line represents linear regression without an intercept.

four haplotypes (Materials and Methods). Almost all genes (>99%) showed at least one indicator of recombination. Only three genes with low nucleotide diversity showed no sign of recombination (supplementary table S2, Supplementary Material online).

Minimum Number of Recombination Events, Recombination Rate, and Nucleotide Diversity

Before examining genes with a particularly high or low frequency of recombination, we examined properties of the minimum number of recombination events.

The minimum number of recombination events (r_{min}) is an indicator of per-gene recombination, which conceivably depends on gene length. A relationship between the minimum number of recombination events (r_{min}) and the gene length among all the genes is shown in figure 4. Clearly, the relationship is linear, indicating r_{min} increases proportionally to gene length. In order to control this effect, we used the minimum number of recombination events divided by gene length in base pairs (r_{min}/nt).

The distribution of r_{min}/nt is summarized into a genome map in which all genes were classified into the three categories: >top 25%, from bottom 25% to top 25%, and <bottom 25% (fig. 5, line 2).

Meanwhile, it is also expected that the minimum number of recombination events depends on nucleotide diversity, even after they are divided by gene length. The relationship between nucleotide diversity and the minimum number of recombination events per nucleotide (r_{min}/nt) is shown in figure 6. The figure indicates linear relationship between π and r_{min}/nt in genes with $\pi \leq 0.08$, which well fitted the regression. Together with figure 4, these genome-wide analyses indicate that the "true" recombination rate is nearly constant across the genome. Meanwhile, another relation-

ship seemed to emerge when $\pi > 0.08$, which we did not explore by regression.

The relationship between nucleotide diversity (π) and the minimum number of recombination events per gene (r_{min}) is shown in supplementary figure S1 (Supplementary Material online). The figures indicate positive relationship for genes with $\pi \leq 0.08$. The relationship appears exponential (supplementary fig. S1, Supplementary Material online).

Detection of Genes with High or Low Recombination Based on the Regression

Using these relationships (fig. 6 and supplementary fig. S1, Supplementary Material online), we identified genes with a particularly high or low recombination rate as those deviated significantly from the regression lines (red and green dots in fig. 6 and supplementary fig. S1, Supplementary Material online) (method A). The genes with particularly high recombination are listed in table 1 (based on "r_{min}/nt") and supplementary table S5 (Supplementary Material online) (based on "r_{min}"), whereas those with particularly low recombination are listed in table 2 (based on "r_{min}/nt") and supplementary table S6 (Supplementary Material online) (based on "r_{min}"). Hereafter, we mainly examined the results using the minimum number of recombination events per nucleotide (r_{min}/nt) as a measure of the recombination rate. These high and low recombination genes are mapped on the genome (fig. 5, line 1).

Genes with High Recombination

Among the genes with a particularly high recombination rate (table 1, red bars in fig. 5, line 1) are several genes responsible for basic cellular functions, such as biosynthesis and metabolism. For example, genes of tryptophan synthase subunit alpha (trpA, HP1277) and L-asparaginase II (ansB, HP0723) showed a high rate of recombination. Recombination breakpoints detected by the four-gamete test were found throughout these genes including functional motifs, as shown in figure 7a and b. L-asparaginase, a putative virulence factor, inhibits host cell function and allows evasion from the immune system (Scotti et al. 2010; Shibayama et al. 2011). Also included is sdaA (HP0132) for L-serine deaminase.

Multiple genes for DNA transformation preceding mutual homologous recombination show a high rate of recombination. comE3 (HP1361) produces a homologue of Bacillus subtilis ComE3, which is essential for DNA transformation (Yeh et al. 2003). HP0033 is a member of the dprA family required for transformation by chromosomal DNA (Ando et al. 1999).

Genes for a transporter and a membrane protein were also included. pnuC (HP1290) produces a membrane-associated protein involved in transport of nicotinamide mononucleotide (Zhu et al. 1991), a key intermediate of NAD biosynthesis. lolA (HP0785) produces an outer membrane lipoprotein carrier protein. glnP (HP1170) produces a glutamine ATP-binding cassette transporter, permease protein.
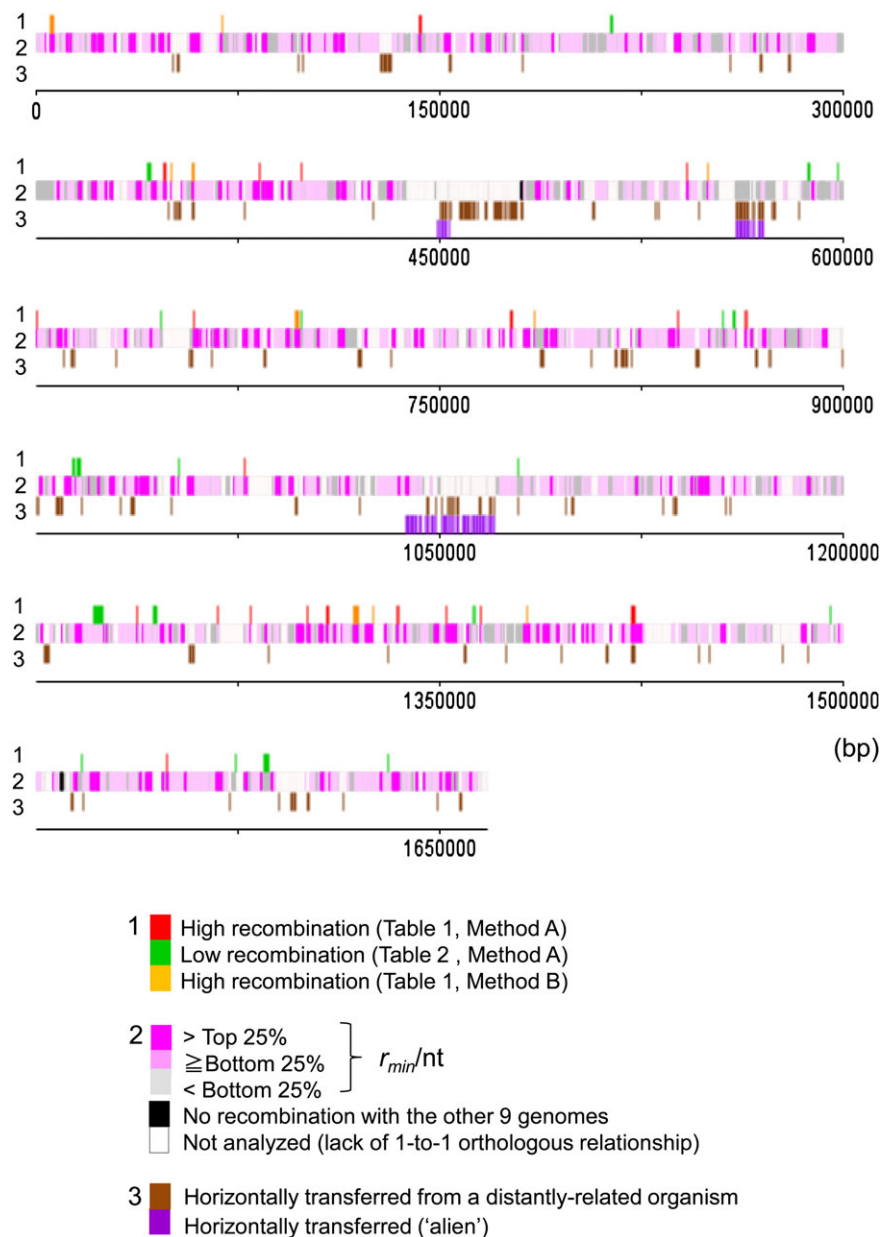
Fig. 5.—Genome map of *Helicobacter pylori* (strain 26695) featuring the level of recombination. 1: Genes with a high or low recombination rate considering gene (nucleotide) diversity. 2: Genes classified according to the recombination rate (minimum number of recombination events/gene length). 3: Horizontally transferred genes from a distance.

Also included were three genes of ribosomal proteins, which represent about 6% of ribosomal protein genes in the genome. They are characterized by short gene length, and their r_{min} is not very large.

Genes with Low Recombination

Of the genes with particularly low recombination rate (table 2, green bars in fig. 5, line 1), there are genes involved in translation and transcription, such as *rpmF* (50S ribosomal protein L32), *tatA* (sec-independent protein translocase

protein), *rnpA* (ribonuclease P, protein component), and HP1287 (putative transcriptional regulator). Also included were genes for lipid metabolism (*pgsA*), protease (*clpS*), and molybdenum cofactor synthesis (*mogA*).

There are also three genes for OMPs (*frpB-4*, *hopQ*, and *hopL*), which represent about 6% of OMP genes in the genome. Genes of OMPs are known to have higher frequency of recombination (Kennemann et al. 2011). Their large gene length makes the r_{min}/nt value smaller.

rpnA, *clpS*, HP1286 (gene of conserved hypothetical secreted protein), and two hypothetical genes (HP0902
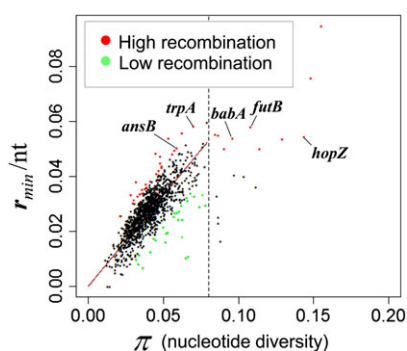
FIG. 6.—Linear relationship between the minimum number of recombination events per gene length in nucleotide (r_{min}/nt) and gene (nucleotide) diversity (π). The broken line indicates nucleotide diversity (π) = 0.08. Red: genes with a particularly high recombination rate in table 1. Green: genes with a particularly low recombination rate in table 2.

and HP0644) are also listed in supplementary table S6 (Supplementary Material online), indicating that both r_{min}/nt and r_{min} are particularly low in these genes.

Highly Divergent Genes with a High Recombination Rate

Of the highly diverged genes where $\pi > 0.08$, we identified genes with a particularly high recombination rate as those in the top 2.5% of all the genes (red dots in fig. 6 and supplementary fig. S1, Supplementary Material online where $\pi > 0.08$, and orange bars in fig. 5, line 1). Sequence divergence inhibits homologous recombination (Fujitani and Kobayashi 1999), but these results show that some are highly divergent and yet with a high rate of recombination. On the other hand, there is no gene with particularly low recombination rates that are in the bottom 2.5% of all genes for $\pi > 0.08$.

These genes with high divergence/recombination are listed in table 1 (method B). Among them, *futB*, *hopZ*, and *babA* are all related to cell surface and expected to be important for host interaction. *futB* is a fucosyltransferase gene responsible for lipopolysaccharide (LPS) synthesis. A previous study reported the high hpEurope–hspEAsia divergence of *futB* (Kawai et al. 2011). Genetic modifications attributable to recombination events within the *futA* and *futB* genes and between the two genes were detected under laboratory conditions (Nilsson et al. 2008). *hopZ* is a phase-variable adhesion gene and plays an important role for colonization (Kennemann et al. 2011). *babA* is responsible for adhesion of *H. pylori* to human gastric epithelium. Recombination in the *babA* locus is unique in that three allele groups are mutually replaced (Hennig et al. 2006). *futB*, *babA*, and *hopZ* were also listed in supplementary table S5 (Supplementary Material online) based on r_{min}, indicating that both r_{min}/nt and r_{min} are

particularly high in these genes. Recombination breakpoints detected by the four-gamete test were found throughout the genes including functional domains and motifs, as shown in figure 7c–e.

Another example is SH3-domain–containing protein (HP1250). It would be interesting if the divergent protein has some interaction with CagA, an SH3-binding oncoprotein, and *cag4* in the list and other cag pathogenicity island proteins.

Homologous Recombination in Horizontally Transferred Genes

We examined the effects of homologous recombination on horizontally transferred genes. The candidates of horizontally transferred genes from distantly related organisms (supplementary table S3, Supplementary Material online) and aliens, such as genomic islands (supplementary table S4, Supplementary Material online), were indicated in the genome map (brown and purple rectangles in fig. 5, line 3). Among the genes with a one-to-one orthologous relationship throughout the ten strains, homologous recombination events were found in all these genes as explained above for supplementary table S2 (Supplementary Material online). There is no significant difference between their average homologous recombination rate (r_{min}/nt) and that of other genes (P = 0.15, Welch's t-test). Thus, even horizontally transferred genes from distantly related organisms appear to have been shared among *H. pylori* strains via active homologous recombination.

Discussion

Examination of a few genes in *H. pylori* indicated that homologous recombination is much more frequent than point mutation, with an estimated rate as high as 6.9×10^{-5}/nt/year (95% credibility region = 3.5×10^{-5}–1.2×10^{-4}) (Falush et al. 2001) or 5.5 (range = 0.5–16.5) $\times 10^{-5}$/initiation sites/year (Kennemann et al. 2011). It was also shown that clusters of polymorphisms were effectively imported into the genome via recombination, which increased the ratio of effect of recombination-derived imports and mutations to 4.3–26.7 (Kennemann et al. 2011). We thus used the four-gamete test to estimate the minimum number of recombination events and the recombination rate. This method is suitable when the recombination rate is significantly higher than the mutation rate and the pairs of segregating sites with four haplotypes arise mainly from recombination. It has been suggested that this test has a low statistical power in detecting recombination events. For example, even if the sample size is 1,000 and mutations are dense, ≤69% of all recombination events may be picked up using this test (Hein et al. 2005). However, we successfully detected recombination events in almost all (>99%) of

Table 1

Genes with High Recombination

Locus Tag	Gene	Description		π	r_{min}/nt	r_{min}	Length (nt)	Method[a]
HP1277	trpA	Tryptophan synthase subunit alpha	Biosynthesis	0.070	0.058	46	789	A
HP0723	ansB	L-asparaginase II	Virulence	0.059	0.050	50	993	A
mHP1361	comE	Competence locus E	Transformation	0.057	0.049	65	1,314	A
HP0808	acpS	4′-phosphopantetheinyl transferase	Fatty acid biosynthesis	0.048	0.047	17	360	A
mHP0333	dprA	Hypothetical protein involved in transformation	Transformation	0.048	0.045	36	801	A
HP1290	pnuC	Nicotinamide mononucleotide transporter	Transport	0.049	0.044	29	663	A
HP0785	lolA	Outer membrane lipoprotein carrier protein	Membrane	0.049	0.043	24	555	A
HP0132	sdaA	L-serine deaminase	Metabolism	0.046	0.041	56	1,368	A
HP0809	fliL	Flagellar basal body–associated protein FliL	Cellular processes	0.032	0.038	21	552	A
HP1170	glnP	Glutamine ABC transporter, permease protein	Transport	0.038	0.037	25	672	A
mHP0514	rplI	50S ribosomal protein L9	Translation	0.038	0.036	16	450	A
HP1261	nuoB	NADH dehydrogenase subunit B	Metabolism	0.036	0.035	17	480	A
mHP1262	nuoC	NADH dehydrogenase subunit C	Metabolism	0.036	0.034	27	798	A
HP1476	ubiD	3-octaprenyl-4-hydroxybenzoate carboxy-lyase	Biosynthesis of cofactors	0.035	0.034	19	564	A
HP0389	sodF	Iron-dependent superoxide dismutase	Cellular processes	0.033	0.033	21	642	A
HP0125	rpmI	50S ribosomal protein L35	Translation	0.022	0.026	5	195	A
HP1196	rpsG	30S ribosomal protein S7	Translation	0.022	0.026	12	468	A
HP0651	futB	Alpha-(1,3)-fucosyltransferase	Cell envelope	0.108	0.058	83	1,431	B
HP0523	cag4	Peptidoglycan hydrolase, Cag island protein (caggamma)	Cellular processes	0.086	0.055	28	510	B
HP0009	hopZ	OMP	OMP	0.144	0.055	104	1,905	B
HP1243	babA	OMP	OMP	0.096	0.054	119	2,202	B
HP1250		Bacterial SH3 domain	Hypothetical	0.091	0.050	29	579	B
HP0374		Hypothetical protein	Hypothetical (other categories)[b]	0.063	0.056	38	681	A
mHP1384		Hypothetical protein	Hypothetical	0.053	0.054	11	204	A
HP1225	crcB	Hypothetical protein	Hypothetical	0.045	0.048	19	393	A
mHP0568		Hypothetical protein	Hypothetical (translation)[b]	0.056	0.048	42	873	A
mHP0614		Hypothetical protein	Hypothetical	0.047	0.042	14	333	A
HP1548		Hypothetical protein	Hypothetical	0.042	0.038	13	339	A
HP0920		Hypothetical protein	Hypothetical (other categories)[b]	0.039	0.036	25	693	A
HP1234		Hypothetical protein	Hypothetical (cell envelope)	0.038	0.036	32	897	A
HP1203a	secE	Preprotein translocase subunit SecE	Hypothetical (no functional assignment)[b]	0.028	0.033	6	180	A
HP1423		Hypothetical protein	Hypothetical (other categories)[b]	0.031	0.031	8	255	A
HP1391		Hypothetical protein	Hypothetical	0.030	0.030	9	297	A
HP0730		Hypothetical protein	Hypothetical	0.155	0.095	29	306	B
HP0338		Hypothetical protein	Hypothetical	0.148	0.076	43	567	B
HP0350		Hypothetical protein	Hypothetical	0.084	0.055	37	669	B
HP0065		Hypothetical protein	Hypothetical	0.129	0.054	19	354	B
mHP1322		Hypothetical protein	Hypothetical	0.114	0.050	29	579	B

NOTE.—ABC, ATP-binding cassette.

[a] A: Top 2.5% of the distribution of deviation from the regression line (fig. 6); B: Top 2.5% of the distribution of r_{min}/nt.

[b] Category in MBGD.

the orthologous genes including the "core" set of genes and horizontally transferred (from distantly related organisms) genes.

To the best of our knowledge, this is the first genome-wide quantitative analysis of homologous recombination in a prokaryote species using population genomic sequence data. Previous genome-wide surveys with other bacteria (Mau et al. 2006; Lefebure and Stanhope 2007; Orsi et al. 2008; Xu et al. 2011) focused on the presence or absence of recombination. In contrast, we quantitatively analyzed intragenic recombination events in each gene. We recognized the dependence of the minimum number of recombination events on nucleotide sequence length and diversity and utilized the linear relationship between nucleotide diversity and recombination rate. Such a linear relationship was reported in several genes of *D. melanogaster* (Begun and Aquadro 1992), but we, for the first time, revealed the relationship using all the orthologous genes

Table 2

Genes with Low Recombination

Locus Tag	Gene	Description		π	r_{min}/nt	r_{min}	Length (nt)	Method[a]
HP0200	rpmF	50S ribosomal protein L32	Translation	0.037	0.007	1	147	A
HP1016	pgsA	Phosphatidylglycerophosphate synthase	Lipid metabolism	0.034	0.008	5	603	A
HP0653	pfr	Nonheme iron-containing ferritin	Transport	0.032	0.010	5	504	A
HP1448	rnpA	Ribonuclease P, protein component	Transcription	0.065	0.010	5	486	A
HP0032	clpS	Hypothetical protein	Other categories	0.066	0.011	3	276	A
HP0320	tatA	Sec-independent protein translocase protein	Translocation	0.041	0.013	3	240	A
HP0799	mogA	Molybdenum cofactor biosynthesis protein	Biosynthesis of cofactors	0.045	0.017	9	531	A
HP1512	frpB-4	Putative IRON-regulated OMP	OMP	0.052	0.019	51	2,634	A
HP0326(2)	neuA	CMP-N-acetylneuraminic acid synthetase	Cell envelope	0.060	0.024	38	1,554	A
HP1287		Putative transcriptional regulator	Transcription	0.063	0.024	16	654	A
HP0566	dapF	Diaminopimelate epimerase	Biosynthesis	0.057	0.026	21	822	A
HP0805		Putative lipopolysaccharide biosynthesis protein	Cell envelope	0.058	0.027	23	855	A
HP1177	hopQ	OMP	OMP	0.074	0.029	56	1,926	A
HP1157(1)	hopL	OMP	OMP	0.078	0.030	109	3,693	A
HP1551	yajC	Preprotein translocase subunit YajC	Cellular processes	0.066	0.031	12	384	A
HP1286		Conserved hypothetical secreted protein	Cell envelope	0.070	0.033	18	549	A
HP1502		Hypothetical protein	Hypothetical	0.034	0.011	5	438	A
HP0552		Hypothetical protein	Hypothetical (other categories)[b]	0.052	0.016	14	864	A
mHP0608		Hypothetical protein	Hypothetical	0.044	0.018	10	570	A
HP0203		Hypothetical protein	Hypothetical	0.063	0.018	5	276	A
mHP0836		Hypothetical protein	Hypothetical	0.053	0.020	7	354	A
HP0863		Hypothetical protein	Hypothetical	0.050	0.021	35	1,629	A
HP0495		Hypothetical protein	Hypothetical (other categories)[b]	0.059	0.027	7	261	A
HP1424		Hypothetical protein	Hypothetical	0.061	0.027	17	621	A
aHP26695_005	rfaJ-2	Putative lipopolysaccharide biosynthesis protein	Cell envelope	0.062	0.029	34	1,155	A
HP0861		Putative thiol:disulfide interchange protein	Hypothetical	0.066	0.032	24	741	A
HP0902		Hypothetical protein	Hypothetical	0.076	0.033	10	300	A
HP0644		Hypothetical protein	Hypothetical (cell envelope)[b]	0.068	0.034	10	294	A

[a] A: Top 2.5% of the distribution of deviation from the regression line (fig. 6); B: Top 2.5% of the distribution of r_{min}/nt.

[b] Category in MBGD.

in a genome. Our results clearly indicated that the "true" recombination rate is nearly constant across the genome. We then identified several genes with a particularly high or low recombination rate based on the relationship and the regression.

The highly divergent genes ($\pi > 0.08$) with a particularly high recombination rate were those of OMPs and related to LPS synthesis, which are important for host interaction. It was interesting genes responsible for DNA transformation, a step preceding the mutual homologous recombination showed high recombination. High recombination genes also included basic cellular functions, such as biosynthesis and metabolism. The active homologous recombination may generate diversity to promote the adaptive evolution of these genes. An examination of this hypothesis is the subject of another publication (Yahara K, Furuta Y, Kawai M, Matelska D, Dunin-Horkawicz S, Bujnicki J, Uchiyama I, Kobayashi I, unpublished data).

From genomic locations of the genes with a particularly high or low recombination rate (fig. 5, line 1), we cannot detect any obvious recombination hot spot or cold spot regions. Two types of hot spots of homologous recombination have been well characterized in bacterial genomes. One is a site for a DNA double-strand breakage (Takahashi and Kobayashi 1990), which initiates homologous recombination just as hot spots in eukaryote meiotic recombination. The other is chi (5'-GCTGGTGG) in *Escherichia coli* and analogous sequences in other bacterial groups (Dillingham and Kowalczykowski 2008). The chi sequence on DNA triggers switching of RecBCD enzyme from DNA degradation to recombination repair and thus serves as an ID sequence of a genome. A homolog of RecBCD enzyme (AddAB) has been characterized in *H. pylori* (Amundsen et al. 2008), but a cognate chi-equivalent sequence has not been identified. *Helicobacter pylori* carries many RM systems. Their recognition sites may serve as
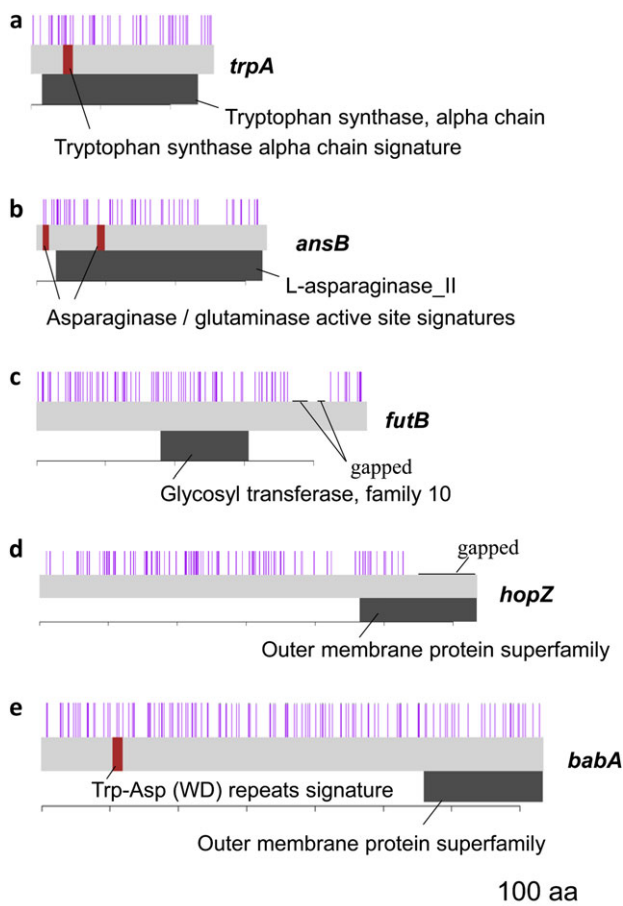
FIG. 7.—Recombination breakpoints and functional motifs/domains of genes with a high recombination rate. A purple bar indicates a recombination breakpoint. A red bar indicates a functional motif. A black belt indicates a functional domain. The locus tags are as follows. (a) HP1277, (b) HP0723, (c) HP0651, (d) HP0009, and (e) HP1243.

a recombination hot spot by DNA double-strand breakage or activate a hot spot elsewhere by providing an entry site for a RecBCD-like recombinase (Stahl et al. 1983) or a restriction enzyme (Ishikawa et al. 2009). Because repertoire of RM systems and, therefore, their recognition sites along the genome are highly variable among *H. pylori* strains (Furuta et al. 2011), hot spot activities related to them may not have been detected by the present method of genome comparison between the ten global strains. Recombination hot/cold spot in *H. pylori* genomes should be analyzed with a higher resolution in the future.

A recent analysis of sequentially sampling *H. pylori* from the same individual (Kennemann et al. 2011) detected a wide distribution of recombination events in several parts of the genome. They found a high frequency of recombination imports in genes in the Hop family of OMPs, such as *babA* and *hopZ* (table 1 and supplementary table S5, Supplementary Material online). This study and our current study are complementary with respect to the time scale,

that is, tens of years of evolution in *Homo sapiens* versus tens of thousands of years of evolution in *H. sapiens*. Moreover, using the entire data set of one-to-one orthologous genes, we identified genes with a particularly high or low recombination rate, which have not been reported previously.

In conclusion, this study provides a genome-wide gene-by-gene view of homologous recombination in this highly sexual bacterial species. From this viewpoint, a species can be considered as a cohesive group of genomes that are closely connected by homologous recombination. We expect that this survey will have implications for evolutionary and population genomic studies of bacteria, which may lead to a reexamination of the species concept.

Supplementary Material

Supplementary figure S1 and tables S1–S7 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

Literature Cited

Achtman M, Wagner M. 2008. Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol. 6:431–440.

Adachi J, Hasegawa M. 1996. MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. Comput Sci Monogr Inst Stat Math. 28:1–150.

Amundsen SK, et al. 2008. Helicobacter pylori AddAB helicase-nuclease and RecA promote recombination-related DNA repair and survival during stomach colonization. Mol Microbiol. 69:994–1007.

Ando T, Israel DA, Kusugami K, Blaser MJ. 1999. HP0333, a member of the dprA family, is involved in natural transformation in Helicobacter pylori. J Bacteriol. 181:5572–5580.

Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature 356:519–520.

Bryant D, Moulton V. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol. 21:255–265.

Cai JJ. 2008. PGEToolbox: a Matlab toolbox for population genetics and evolution. J Hered. 99:438–440.

Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J. 2009. DNAPlotter: circular and linear interactive genome visualization. Bioinformatics 25:119–120.

Cohan FM, Perry EB. 2007. A systematics for discovering the fundamental units of bacterial diversity. Curr Biol. 17:R373–R386.

Cox TF, Cox MAA. 1994. Multidimensional scaling. Boca Raton (FL): Chapman and Hall.

Darling AC, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14:1394–1403.

Darwin C. 1859. On the origins of species by means of natural selection, or the preservation of favoured races in the struggle for life. London: Murray.

de Castro E, et al. 2006. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res. 34:W362–W365.

Didelot X, et al. 2011. Recombination and population structure in Salmonella enterica. PLoS Genet. 7:e1002191.

Dillingham MS, Kowalczykowski SC. 2008. RecBCD enzyme and the repair of double-stranded DNA breaks. Microbiol Mol Biol Rev. 72:642–671, Table of Contents.

Doolittle WF, Zhaxybayeva O. 2009. On the origin of prokaryotic species. Genome Res. 19:744–756.

Falush D, et al. 2001. Recombination and mutation during long-term gastric colonization by Helicobacter pylori: estimates of clock rates, recombination size, and minimal age. Proc Natl Acad Sci U S A. 98:15056–15061.

Falush D, et al. 2003. Traces of human migrations in Helicobacter pylori populations. Science 299:1582–1585.

Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. 2009. The bacterial species challenge: making sense of genetic and ecological diversity. Science 323:741–746.

Fraser C, Hanage WP, Spratt BG. 2007. Recombination and the nature of bacterial speciation. Science 315:476–480.

Fujitani Y, Kobayashi I. 1999. Effect of DNA sequence divergence on homologous recombination as analyzed by a random-walk model. Genetics 153:1973–1988.

Furuta Y, Kawai M, Uchiyama I, Kobayashi I. 2011. Domain movement within a gene: a novel evolutionary mechanism for protein diversification. PLoS One 6:e18819.

Hartigan JA. 1975. Clustering algorithms. New York: Wiley.

Hein J, Schierup MH, Wiuf C. 2005. Gene genealogies, variation and evolution: a primer in coalescent theory. New York: Oxford University Press.

Hennig EE, Allen JM, Cover TL. 2006. Multiple chromosomal loci for the babA gene in Helicobacter pylori. Infect Immun. 74:3046–3051.

Hudson RR, Kaplan NL. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147–164.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Ishikawa K, Handa N, Kobayashi I. 2009. Cleavage of a model DNA replication fork by a Type I restriction endonuclease. Nucleic Acids Res. 37:3531–3544.

Kawai M, et al. 2011. Evolution in an oncogenic bacterial species with extreme genome plasticity: helicobacter pylori East Asian genomes. BMC Microbiol. 11:104.

Kennemann L, et al. 2011. Helicobacter pylori genome evolution during human infection. Proc Natl Acad Sci U S A. 108:5033–5038.

Konstantinidis KT, Ramette A, Tiedje JM. 2006. The bacterial species definition in the genomic era. Philos Trans R Soc Lond B Biol Sci. 361:1929–1940.

Kosakovsky Pond SL, Frost SD, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676–679.

Lefebure T, Stanhope MJ. 2007. Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. Genome Biol. 8:R71.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451–1452.

Linz B, et al. 2007. An African origin for the intimate association between humans and Helicobacter pylori. Nature 445:915–918.

Lole KS, et al. 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J Virol. 73: 152–160.

Marchler-Bauer A, et al. 2009. CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res. 37: D205–D210.

Mau B, Glasner JD, Darling AE, Perna NT. 2006. Genome-wide detection and analysis of homologous recombination among sequenced strains of Escherichia coli. Genome Biol. 7:R44.

Mayr E. 1942. Systematics and the origin of species. New York: Dover Publications.

Moodley Y, et al. 2009. The peopling of the Pacific from a bacterial perspective. Science 323:527–530.

Morelli G, et al. 2010. Microevolution of Helicobacter pylori during prolonged infection of single hosts and within families. PLoS Genet. 6:e1001036.

Nakamura Y, Itoh T, Matsuda H, Gojobori T. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet. 36:760–766.

Nilsson C, et al. 2008. Lipopolysaccharide diversity evolving in Helicobacter pylori communities through genetic modifications in fucosyltransferases. PLoS One 3:e3811.

Orsi RH, Sun Q, Wiedmann M. 2008. Genome-wide analyses reveal lineage specific contributions of positive selection and recombination to the evolution of Listeria monocytogenes. BMC Evol Biol. 8:233.

Robinson DR, Foulds LR. 1981. Comparison of phylogenetic trees. Math Biosci. 53:131–147.

Salminen M, Carr JK, Burke DS, McCutchan FE. 1995. Identification of recombination breakpoints in HIV-1 by bootscanning. AIDS Res Hum Retroviruses. 11:1423–1425.

Scotti C, et al. 2010. Cell-cycle inhibition by Helicobacter pylori L-asparaginase. PLoS One 5:e13892.

Shibayama K, Takeuchi H, Wachino J, Mori S, Arakawa Y. 2011. Biochemical and pathophysiological characterization of Helicobacter pylori asparaginase. Microbiol Immunol. 55:408–417.

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol. 16:1114–1116.

Sigrist CJ, et al. 2002. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform. 3:265–274.

Spencer CC, et al. 2006. The influence of recombination on human genetic diversity. PLoS Genet. 2:e148.

Stahl MM, Kobayashi I, Stahl FW, Huntington SK. 1983. Activation of Chi, a recombinator, by the action of an endonuclease at a distant site. Proc Natl Acad Sci U S A. 80:2310–2313.

Suerbaum S, Josenhans C. 2007. Helicobacter pylori evolution and phenotypic diversification in a changing host. Nat Rev Microbiol. 5:441–452.

Suerbaum S, et al. 1998. Free recombination within Helicobacter pylori. Proc Natl Acad Sci U S A. 95:12619–12624.

Takahashi N, Kobayashi I. 1990. Evidence for the double-strand break repair model of bacteriophage lambda recombination. Proc Natl Acad Sci U S A. 87:2790–2794.

Takuno S, Kado T, Sugino RP, Nakhleh L, Innan H. 2012. Population genomics in bacteria: a case study of Staphylococcus aureus. Mol Biol Evol. 29:797–809.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Uchiyama I. 2003. MBGD: microbial genome database for comparative analysis. Nucleic Acids Res. 31:58–62.

Uchiyama I. 2006. Hierarchical clustering algorithm for comprehensive orthologous-domain classification in multiple genomes. Nucleic Acids Res. 34:647–658.

Uchiyama I. 2008. Multiple genome alignment for identifying the core structure among moderately related microbial genomes. BMC Genomics 9:515.

Vernikos GS, Parkhill J. 2006. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. Bioinformatics 22:2196–2203.

Xu Z, Chen H, Zhou R. 2011. Genome-wide evidence for positive selection and recombination in Actinobacillus pleuropneumoniae. BMC Evol Biol. 11:203.

Yamaoka Y. 2008. Helicobacter pylori: molecular genetics and cellular biology. Norwich (UK): Caister Academic Press.

Yeh YC, Lin TL, Chang KC, Wang JT. 2003. Characterization of a ComE3 homologue essential for DNA transformation in Helicobacter pylori. Infect Immun. 71:5427–5431.

Zhu N, Olivera BM, Roth JR. 1991. Activity of the nicotinamide mononucleotide transport system is regulated in Salmonella typhimurium. J Bacteriol. 173:1311–1320.

Associate editor: Takashi Gojobori