

17. Fülöp, V., Moir, J. W. B., Ferguson, S. J. & Hajdu, J. Crystallisation and preliminary crystallographic study of cytochrome c_1 nitrite reductase from *Thiospaera pantotropha*. *J. Mol. Biol.* **232**, 1211–1212 (1993).
18. Berger, H. & Wharton, D. C. Small angle X-ray scattering studies of oxidised and reduced cytochrome oxidase from *Pseudomonas aeruginosa*. *Biochim. Biophys. Acta* **622**, 355–359 (1980).
19. Moore, G. R. & Pettigrew, G. W. *Cytochromes c: Evolutionary, Structural and Physicochemical Aspects* (Springer, Berlin, 1990).
20. Pettigrew, G. W. & Moore, G. R. *Cytochromes c: Biological Aspects* (Springer, Berlin, 1987).
21. Harutunyan, E. H. *et al.* The binding of carbon monoxide and nitric oxide to leghaemoglobin in comparison with other haemoglobins. *J. Mol. Biol.* **264**, 152–161 (1996).
22. Edwards, S. L., Kraut, J. & Poulos, T. L. Crystal structure of nitric oxide inhibited cytochrome-c peroxidase. *Biochemistry* **27**, 8074–8081 (1988).
23. Adman, E. T., Godden, J. W. & Turley, S. The structure of copper nitrite reductase from *Achromobacter cycloclastes* at five pH values, with NO_2^- bound and with type II copper depleted. *J. Biol. Chem.* **270**, 27458–27474 (1995).
24. Williams, P. A. thesis, Oxford Univ. (1996).
25. Poulos, T. L. Ligands and electrons and haem proteins. *Nature Struct. Biol.* **3**, 401–403 (1996).
26. Wittung, P. & Malmstrom, B. G. Redox-linked conformational changes in cytochrome c oxidase. *FEBS Lett.* **388**, 47–49 (1996).
27. Pascher, T., Cheshire, J. P., Winkler, J. R. & Gray, H. B. Protein folding triggered by electron transfer. *Science* **271**, 1558–1560 (1996).
28. Kraulis, P. J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
29. Merritt, E. A. & Murphy, M. E. P. Raster3D Version 2.0. A program for photorealistic molecular graphics. *Acta Crystallogr. D* **50**, 869–873 (1994).
30. Brünger, A. T. The free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–474 (1992).

Acknowledgements. We thank the ESRF and SRS Daresbury for data collection facilities; the EMBL outstation, Grenoble, for use of an image plate detector; M.L.D. Page for expert advice; R. Bryan and R. Esnouf for computing; K. Harlos for help with in-house data collection; F. Armstrong and J. Hirst for providing electrochemically reduced methyl viologen. This work was supported by MRC, BBSRC and EU-BIOTECH. The Oxford Centre for Molecular Sciences is funded jointly by BBSRC, EPSRC and MRC. N.F.W.S. was supported by a Wellcome Trust prize studentship. V.F. is a Royal Society university research fellow.

Correspondence and requests for materials should be addressed to P.A.W. (e-mail: pamela@scripps.edu), V.F. (e-mail: vilmos@biop.ox.ac.uk) or J.H. (e-mail: janos@xray.bmc.uu.se).

errata

The yeast genome directory

Nature **387** (suppl.) (1997)

In the list of authors given on page 5 of this supplement, the names of some authors were omitted or misspelled (asterisks). These were: R. Altmann; W. Arnold*; M. de Haan*; K. Hamberg; K. Hinni; L. Jones; W. Kramer; H. Küster*; K. C. T. Maurer*; D. Niblett; N. Paricio*; A. G. Parle-McDermott*; C. Rebischung; C. Richards; L. Rifkin*; J. Robben; C. Rodrigues-Pousada*; I. Schaaff-Gerstenschläger*; P. H. M. Smits*; Y. Su*; Q. J. M. van der Aart*; J. C. van Vliet-Reedijk*; A. Wach; M. Yamazaki*. \square

Measurements of elastic anisotropy due to solidification texturing and the implications for the Earth's inner core

Michael I. Bergman

Nature **389**, 60–63 (1997)

Owing to a typographical error, this Letter appeared under the title "Measurements of electric anisotropy due to solidification texturing and the implications for the Earth's inner core". The word 'elastic' in the first line was erroneously replaced with 'electric'. \square

cAMP-induced switching in turning direction of nerve growth cones

Hong-jun Song, Guo-li Ming & Mu-ming Poo

Nature **388**, 275–279 (1997)

The order of panels in Fig. 3 of this Letter is incorrect as published. Figure 3a–e should be labelled as f–j, and Fig. 3f–j should be labelled a–e. \square

corrections

Synthesis and X-ray structure of dumb-bell-shaped C_{120}

Guan-Wu Wang, Koichi Komatsu, Yasujiro Murata & Motoo Shiro

Nature **387**, 583–586 (1997)

In this Letter, we overlooked a citation of G. Oszlanyi *et al.*, *Phys. Rev. B* **54**, 11849 (1996), who reported the observation of covalently bound $(\text{C}_{60})_2^{2-}$ dianions from the X-ray powder diffraction patterns of the metastable phases of KC_{60} and RbC_{60} . \square

The complete genome sequence of the gastric pathogen *Helicobacter pylori*

Jean-F. Tomb, Owen White, Anthony R. Kerlavage, Rebecca A. Clayton, Granger G. Sutton, Robert D. Fleischmann, Karen A. Ketchum, Hans Peter Klenk, Steven Gill, Brian A. Dougherty, Karen Nelson, John Quackenbush, Lixin Zhou, Ewen F. Kirkness, Scott Peterson, Brendan Loftus, Delwood Richardson, Robert Dodson, Hanif G. Khalak, Anna Glodek, Keith McKenney, Lisa M. Fitzgerald, Norman Lee, Mark D. Adams, Erin K. Hickey, Douglas E. Berg, Jeanine D. Gocayne, Teresa R. Utterback, Jeremy D. Peterson, Jenny M. Kelley, Matthew D. Cotton, Janice M. Weidman, Claire Fujii, Cheryl Bowman, Larry Watthey, Erik Wallin, William S. Hayes, Mark Borodovsky, Peter D. Karp, Hamilton O. Smith, Claire M. Fraser & J. Craig Venter

Nature **388**, 539–547 (1997)

In this Article, we incorrectly stated that the amino acids lysine and arginine are twice as abundant in *H. pylori* proteins as they are in those of *Haemophilus influenzae* and *Escherichia coli*. This statement was derived from amino-acid analyses that compared absolute differences in abundance, but these do not reflect the frequencies with which amino acids are found in the organisms in question. The actual abundance of arginine in *H. pylori*, *H. influenzae* and *E. coli* is 3.5, 4.5 and 5.5%, respectively; the abundance of lysine in these organisms is 8.9, 6.3 and 4.4%, respectively. This oversight is particularly unfortunate because Russell H. Doolittle, who wrote an accompanying News and Views on our Article and brought this to our attention, was led to comment on the significance of our inaccurate observation. We regret this and any other misunderstanding that our error may have caused. \square

The complete genome sequence of the gastric pathogen *Helicobacter pylori*

Jean-F. Tomb*, Owen White*, Anthony R. Kerlavage*, Rebecca A. Clayton*, Granger G. Sutton*, Robert D. Fleischmann*, Karen A. Ketchum*, Hans Peter Klenk*, Steven Gill*, Brian A. Dougherty*, Karen Nelson*, John Quackenbush*, Lixin Zhou*, Ewen F. Kirkness*, Scott Peterson*, Brendan Loftus*, Delwood Richardson*, Robert Dodson*, Hanif G. Khalak*, Anna Glodek*, Keith McKenney*, Lisa M. Fitzgerald*, Norman Lee*, Mark D. Adams*, Erin K. Hickey*, Douglas E. Berg†, Jeanine D. Gocayne*, Teresa R. Utterback*, Jeremy D. Peterson*, Jenny M. Kelley*, Matthew D. Cotton*, Janice M. Weidman*, Claire Fujii*, Cheryl Bowman*, Larry Watthey*, Erik Wallin‡, William S. Hayes§, Mark Borodovsky§, Peter D. Karp||, Hamilton O. Smith¶, Claire M. Fraser* & J. Craig Venter*

* The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA

† Department of Molecular Biology, School of Medicine, Washington University St Louis, 660 S. Euclid Avenue, St Louis, Missouri 63110, USA

‡ Department of Biochemistry, Arrhenius Laboratory, Stockholm University, S-106 91 Stockholm, Sweden

§ School of Biology, Georgia Tech, Atlanta, Georgia 30332, USA

|| SRI International, Artificial Intelligence Center, 333 Ravenswood Avenue, Menlo Park, California 94025, USA

¶ Department of Molecular Biology and Genetics, School of Medicine, Johns Hopkins University, 725 N. Wolfe Street, Baltimore, Maryland 21205, USA

***Helicobacter pylori*, strain 26695, has a circular genome of 1,667,867 base pairs and 1,590 predicted coding sequences. Sequence analysis indicates that *H. pylori* has well-developed systems for motility, for scavenging iron, and for DNA restriction and modification. Many putative adhesins, lipoproteins and other outer membrane proteins were identified, underscoring the potential complexity of host-pathogen interaction. Based on the large number of sequence-related genes encoding outer membrane proteins and the presence of homopolymeric tracts and dinucleotide repeats in coding sequences, *H. pylori*, like several other mucosal pathogens, probably uses recombination and slipped-strand mispairing within repeats as mechanisms for antigenic variation and adaptive evolution. Consistent with its restricted niche, *H. pylori* has a few regulatory networks, and a limited metabolic repertoire and biosynthetic capacity. Its survival in acid conditions depends, in part, on its ability to establish a positive inside-membrane potential in low pH.**

For most of this century the cause of peptic ulcer disease was thought to be stress-related and the disease to be prevalent in hyperacid producers. The discovery¹ that *Helicobacter pylori* was associated with gastric inflammation and peptic ulcer disease was initially met with scepticism. However, this discovery and subsequent studies on *H. pylori* have revolutionized our view of the gastric environment, the diseases associated with it, and the appropriate treatment regimens².

Helicobacter pylori is a micro-aerophilic, Gram-negative, slow-growing, spiral-shaped and flagellated organism. Its most characteristic enzyme is a potent multisubunit urease³ that is crucial for its survival at acidic pH and for its successful colonization of the gastric environment, a site that few other microbes can colonize². *H. pylori* is probably the most common chronic bacterial infection of humans, present in almost half of the world population². The presence of the bacterium in the gastric mucosa is associated with chronic active gastritis and is implicated in more severe gastric diseases, including chronic atrophic gastritis (a precursor of gastric carcinomas), peptic ulceration and mucosa-associated lymphoid tissue lymphomas². Disease outcome depends on many factors, including bacterial genotype, and host physiology, genotype and dietary habits^{4,5}. *H. pylori* infection has also been associated with persistent diarrhoea and increased susceptibility to other infectious diseases⁶.

Because of its importance as a human pathogen, our interest in its biology and evolution, and the value of complete genome sequence information for drug discovery and vaccine development, we have

Table 1 Genome features

General	
Coding regions (91.0%)	
Stable RNA (0.7%)	
Non-coding repeats (2.3%)	
Intergenic sequence (6.0%)	
RNA	
Ribosomal RNA	Coordinates
23S-5S	445,306-448,642 bp
23S-5S	1,473,557-1,473,919 bp
16S	1,209,082-1,207,584 bp
16S	1,511,138-1,512,635 bp
5S	448,041-448,618 bp
Transfer RNA	
36 species (7 clusters, 12 single genes)	
Structural RNA	
1 species (ssrD)	629,845-630,124 bp
DNA	
Insertion sequences	
IS605 13 copies (5 full-length, 8 partial)	
IS606 4 copies (2 full-length, 2 partial)	
Distinct G + C regions	Associated genes
region 1 (33% G + C) 452-479 kb	IS605, 5SRNA and repeat 7; virB4
region 2 (35% G + C) 539-579 kb	cag PAI (Fig. 4)
region 3 (33% G + C) 1,049-1,071 kb	IS605, 5SRNA and repeat 7
region 4 (43% G + C) 1,264-1,276 kb	β and β' RNA polymerase, EF-G (fusA)
region 5 (33% G + C) 1,590-1,602 kb	two restriction/modification systems
Coding sequences	
1,590 coding sequences (average 945 bp)	
1,091 identified database match	
499 no database match	

sequenced the genome of a representative *H. pylori* strain by the whole-genome random sequencing method as described for *Haemophilus influenzae*⁷, *Mycoplasma genitalium*⁸ and *Methanococcus jannaschii*⁹.

General features of the genome

Genome analysis. The genome of *H. pylori* strain 26695 consists of a circular chromosome with a size of 1,667,867 base pairs (bp) and average G + C content of 39% (Figs 1 and 2). Five regions within the genome have a significantly different G + C composition (Table 1 and Fig. 1). Two of them contain one or more copies of the insertion sequence IS605 (see below) and are flanked by a 5S ribosomal RNA sequence at one end and a 521 bp repeat (repeat 7) near the other. These two regions are also notable because they contain genes involved in DNA processing and one contains 2 orthologues of the *virB4/ptl* gene, the product of which is required for the transfer of oncogenic T-DNA of *Agrobacterium* and the secretion of the *pertussis* toxin by *Bordetella pertussis*¹⁰. Another region is the *cag* pathogenicity island (PAI), which is flanked by 31-bp direct repeats, and appears to be the product of lateral transfer¹¹.

RNA and repeat elements. Thirty-six tRNA species were identified using tRNAscan-SE¹². These are organized into 7 clusters plus 12 single genes. Two separate sets of 23S–5S and 16S ribosomal RNA (rRNA) genes were identified, along with one orphan 5S gene and one structural RNA gene (Table 1). Associated with each of the two 23S–5S gene clusters is a 6-kilobase (kb) repeat containing a possible operon of 5 ORFs that have no database matches.

Eight repeat families (>97% identity) varying in length from 0.47 to 3.8 kb were found in the chromosome (Figs 1 and 2). Members of repeat 7 are found in intergenic regions, while the others are associated with coding sequences and may represent gene duplications. Repeats 1, 2, 3 and 6 are associated with genes that encode outer-membrane proteins (OMP) (Fig. 3).

Two distinct insertion sequence (IS) elements are present. There are five full-length copies of the previously described IS605^{11,13} and two of a newly discovered element designated IS606. In addition, there are eight partial copies of IS605 and two partial copies of IS606. Both elements encode two divergently transcribed transposases (TnpA and TnpB). IS606 has less than 50% nucleotide identity with IS605 and the IS606 transposases have 29% amino-acid identity with their IS605 counterpart. Both copies of the IS606 TnpB may be non-functional owing to frameshifts.

Origin of replication. As a typical eubacterial origin of replication was not identified¹⁴, we arbitrarily designated basepair one at the start of a 7-mer repeat, (AGTGATT)₂₆, that produces translational stops in all reading frames, as this repeated DNA is unlikely to contain any coding sequence.

Open reading frames. One thousand five hundred and ninety predicted coding sequences were identified. They were searched against a non-redundant protein database resulting in 1,091 putative identifications that were assigned biological roles using a classification system adapted from Riley¹⁵ (Table 2). The 1,590 predicted genes had an average size of 945 bp, similar to that observed in other prokaryotes^{7–9}, and no genome-wide strand bias was observed (Fig. 2). More than 70% of the predicted proteins in *H. pylori* have a calculated isoelectric point (pI) greater than 7.0, compared to ~40% in *H. influenzae* and *E. coli*. The basic amino acids, arginine and lysine, occur twice as frequently in *H. pylori* proteins as in those of *H. influenzae* and *E. coli*, perhaps reflecting an adaptation of *H. pylori* to gastric acidity.

Paralogous families. Ninety-five paralogous gene families comprising 266 gene products (16% of the total) were identified (www.tigr.org/tdb/mdb/hpdb/hpdb.html). Of these, 67 (173 proteins) have an assigned role. Sixty-four have only 2 members, while the porin/adhesin-like outer membrane protein family (Fig. 2) is the largest with 32 members. The largest number of paralogues with assigned roles fall into the functional categories of cell

envelope, transport and binding proteins, and proteins involved in replication. The large number of cell envelope proteins might reflect either a reduced biosynthetic capacity or a need to adapt to the challenging gastric environment.

Cell division and protein secretion

The gene content of *H. pylori* suggests that the basic mechanisms of replication, cell division and secretion are similar to those of *E. coli* and *H. influenzae*. However, important differences are noted. For example, apparently missing from the *H. pylori* genome are orthologues of DnaC, MinC, and the secretory chaperonin, SecB. In oriC-type primosome formation, the DnaB and DnaC proteins form a B–C complex that delivers the DnaB helicase to the developing primosome complex¹⁶. The apparent absence of DnaC in *H. pylori* suggests that either a novel mechanism for recruiting DnaB exists or a DnaC orthologue with no detectable sequence similarity is present. Similar arguments can be made for other seemingly missing important functions.

H. pylori has a classical set of bacterial chaperones (DnaK, DnaJ, CbpA, GrpE, GroEL, GroES, and HtpG). The transcriptional regulation of *H. pylori* chaperone genes is likely to be different from that in *E. coli*, as it seems not to have the sigma factors that upregulate chaperone synthesis in *E. coli* (heat-shock sigma 32 and stationary-phase sigma S).

In addition to the SecA-dependent secretory pathway, *H. pylori* has two specialized export systems. One is associated with the *cag* pathogenicity island¹¹ and the other is the flagellar export pathway which is assembled from orthologues of FliH, FliI, FliP, FlhA, FlhB, FliQ, FliR and FliP¹⁷. Apparently absent from *H. pylori* is a type IV signal peptidase and orthologues of the dsbABC system, which in other species are required for the maturation of pili and pilin-like structures¹⁸ and assembly of surface structures involved in virulence and DNA transformation¹⁹.

Recombination, repair and restriction systems

Systems for homologous recombination and post-replication, mismatch, excision and transcription-coupled repair appear to be present in *H. pylori*. Also present are genes with similarity to DNA glycosylases which have associated AP endonuclease activity. The RecBCD pathway, which mediates homologous recombination and double-strand break repair, and RecT and RecE orthologues, proteins involved in strand exchange during recombination²⁰, seem to be absent. The ability of *H. pylori* to perform mismatch repair is suggested by the presence of methyl transferases, mutS and uvrD. However, orthologues of MutH and MutL were not identified. Components of an SOS system also appear to be absent.

Bacteria commonly use restriction and modification systems to degrade foreign DNA. In *H. pylori*, this defence system is well developed with eleven restriction-modification systems identified on the basis of gene order and similarity to endonucleases, methyltransferases, and specificity subunits. Three type I, one type II, and three type IIS systems were identified, as well as four type III systems, including the recently identified epithelial responsive

Figure 1 Linear representation of the *H. pylori* 26695 chromosome illustrating the location of each predicted protein-coding region, RNA gene, and repeat elements in the genome. Symbols are as follows: ++, Co²⁺, Zn²⁺, Cd²⁺; ?, unknown; A/G/S, α -alanine/glycine/ β -serine; B12, B12/ferric siderophores; E, glutamate; Mo, molybdenum; P, proline; P/G, proline/glycine betaine; Q, glutamine; S, serine; a-k, α -ketoglutarate; a/o, arginine/ornithine; aa, amino acids (specificity unknown); aa2, dipeptides; aaX, oligopeptides; fum, fumarate, succinate; glu, glucose/galactose; h, hemin; lac, L-lactate; mal, malate 2-oxoglutarate; nic, nicotinamide mononucleotides; pyr, pyrimidine nucleosides. Numbers associated with tRNA symbols represent the number of tRNAs at a locus. Numbers associated with GES represent the number of membrane-spanning domains according to the Goldman, Engelman and Steitz scale as calculated by TopPred⁴⁷.

endonuclease, *iceA1*, and its associated DNA adenine methyltransferase (M. HypI) genes^{21,22}. In addition to the complete systems, seven adenine-specific, and four cytosine-specific methyltransferases, and one of unknown specificity were found. Each of these has an adjacent gene with no database match, suggesting that they may function as part of restriction-modification systems.

Transcription and translation

Although analysis of gene content suggests that *H. pylori* has a basic transcriptional and translational machinery similar to that of *E. coli*, interesting differences are observed. For example, no genes for a catalytic activity in tRNA maturation (*rnd*, *rph*, or *rnpB*) were identified and of the three known ribonucleases involved in mRNA degradation, only polyribonucleotide phosphorylase was found. Twenty-one genes coding for 18 of the 20 tRNA synthetases normally required for protein biosynthesis were found.

As in most other completely sequenced bacterial genomes, the gene for glutaminyl-tRNA synthetase, *glnS*, is missing, and the existence of a transamidation process is assumed. It is also possible that the product of the second glutamyl-tRNA synthetase gene, *gltX*, present in *H. pylori*, may have acquired the glutaminyl-tRNA synthetase function. *H. pylori* provides the first example of a bacterial genome apparently lacking an asparaginyl-tRNA synthetase gene, *asnS*. A transamidation process to form *Asn-tRNAAsn* from *Asp-tRNAAsn* has been reported for the archaeon *Haloferax volcanii*²² and may also operate in *H. pylori*. Most intriguing, however, is the finding that in *H. pylori* the genes encoding the β and β' subunits of RNA polymerase are fused. In all studied prokaryotes the two genes are contiguous, but separate, and are part of the same transcriptional unit. Whether this gene fusion in *H. pylori* results in a fused protein, or whether the transcriptional or translational product of the fusion is subject to splicing, is currently not known. It is worth noting that an artificial fusion of the *E. coli*

rpoB and *rpoC* genes is viable and results in a transcriptional complex, which has the same stoichiometry as the native complex (K. Severinov, personal communication).

Adhesion and adaptive antigenic variation

Most pathogens show tropism to specific tissues or cell types and often use several adherence mechanisms for successful attachment. *H. pylori* may use at least five different adhesins to attach to gastric epithelial cells⁵. One of them, HpaA (HP0797), was previously identified as a lipoprotein in the flagellar sheath and outer membrane^{5,23}. In addition to the HpaA orthologue, we have identified 19 other lipoproteins. Few have an identifiable function, but some are likely to contribute to the adherence capacity of the organism.

Two adhesins^{24–26}, one of which mediates attachment to the Lewis^b histo-blood group antigens, belong to the large family of outer membrane proteins (OMP) (Fig. 3) (T. Boren and R. Haas, personal communication). It is conceivable that other members of these closely related proteins also act as adhesins. Given the large number of sequence-related genes encoding putative surface-exposed proteins, the potential exists for recombinational events leading to mosaic organization. This could be the basis for antigenic variation in *H. pylori* and an effective mechanism for host defence evasion, as seen in *M. genitalium*²⁷.

At least one other mechanism for antigenic variation could operate in *H. pylori*. The DNA sequence at the beginning of eight genes, including five members of the OMP family, contain stretches of CT or AG dinucleotide repeats (Table 3a). In addition, poly(C) or poly(G) tracts occur within the coding sequence of nine other genes (Table 3b). Slipped-strand mispairing within such repeats are documented features of one mechanism of genotypic variation^{28,29}. These mechanisms may have evolved in bacterial pathogens to increase the frequency of phenotypic variation in genes involved in

Figure 2 Circular representation of the *H. pylori* 26695 chromosome. Outer concentric circle: predicted coding regions on the plus strand classified as to role according to the colour code in Fig. 1 (except for unknowns and hypotheticals, which are in black). Second concentric circle: predicted coding regions on the minus strand. Third and fourth concentric circles: IS elements (red) and other repeats (green) on the plus and minus strand, respectively. Fifth and sixth concentric circles: tRNAs (blue), rRNAs (red), and sRNAs (green) on the plus and minus strand, respectively.

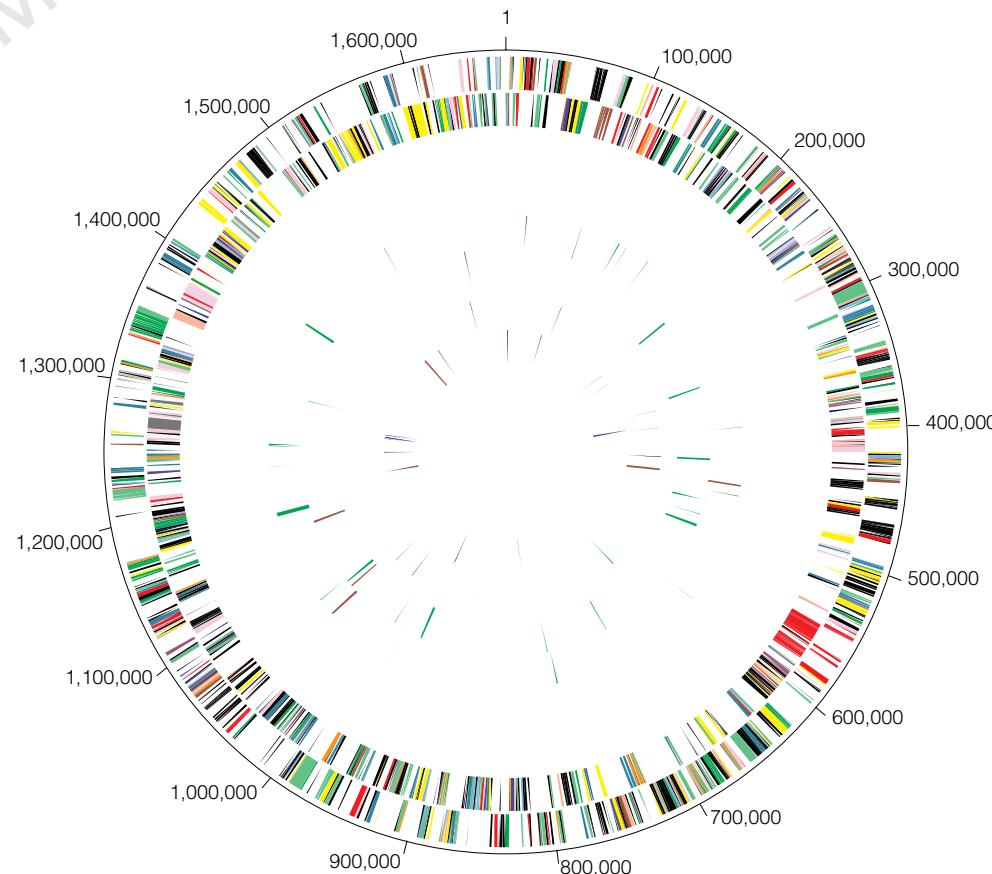




Figure 3 Multiple sequence alignment of members of the outer membrane protein family of *H. pylori*. These proteins were identified as OMPs based on the characteristic alternating hydrophobic residues at their carboxy termini. All members of this family have one domain of similarity at the amino-terminal end and seven domains of similarity at their carboxy-terminal end. Note that the first 11 of these OMPs share extensive similarity over their entire length. Four of the OMPs were identified as porins (Hops) based on identity to published amino-terminal sequences, represented at the top of the alignment⁵⁰. The most likely

candidate for HopD is HP0913, which has 15 matches to the first 20-residue N-terminal peptide sequence⁵⁰. These differences may be due to strain variability. The program Signal-P⁴⁸ was used to identify cleavage sites and signal peptide (underlined). Four of the OMPs have TTG start codons (HP1156, HP0252, HP1131, HP0796). Numbers embedded in the sequences represent amino acids omitted from the alignment. The star symbols indicate that HP722, HP725 and HP726 proteins contain a frameshift in their signal-peptide-coding region. These frame shifts are associated with the presence of dinucleotide repeats (Table 3).

critical interactions with their hosts²⁸. Such 'contingency' genes encode surface structures like pilins, lipoproteins or enzymes that produce lipopolysaccharide molecules²⁸. Our analysis suggests that the seventeen genes reported in Table 3a,b belong to this category and thus may provide an example of adaptive evolution in *H. pylori*.

Phenotypic variation at the transcriptional level may also operate in *H. pylori*. Examples of repetitive DNA mediating transcriptional control have been documented by the presence of oligonucleotide repeats in promoter regions²⁹. Homopolymeric tracts of A or T in potential promoter regions of eighteen genes were found, including eight members of the OMP family (Table 3c).

Virulence

The virulence of individual *H. pylori* isolates has been measured by their ability to produce a cytotoxin-associated protein (CagA) and

an active vacuolating cytotoxin (VacA)⁵. The *cagA* gene, though not a virulence determinant, is positioned at one end of a pathogenicity island containing genes that elicit the production of interleukin (IL)-8 by gastric epithelial cells^{11,30}. Consistent with its more virulent character, *H. pylori* strain 26695 contains a single contiguous PAI region¹¹ (Fig. 4).

VacA induces the formation of acidic vacuoles in host epithelial cells, and its presence is associated epidemiologically with tissue damage and disease³¹. VacA may not be the only ulcer-causing factor as 40% of *H. pylori* strains do not produce detectable amounts of the cytotoxin *in vitro*⁵. Sequence differences at the amino terminus and central sections are noted among VacA proteins derived from Tox⁺ and Tox⁻ strains³¹. This Tox⁺ *H. pylori* strain contains the more toxic S1a/m1 type cytotoxin and three additional large proteins with moderate similarities to the carboxy-terminal end of the active

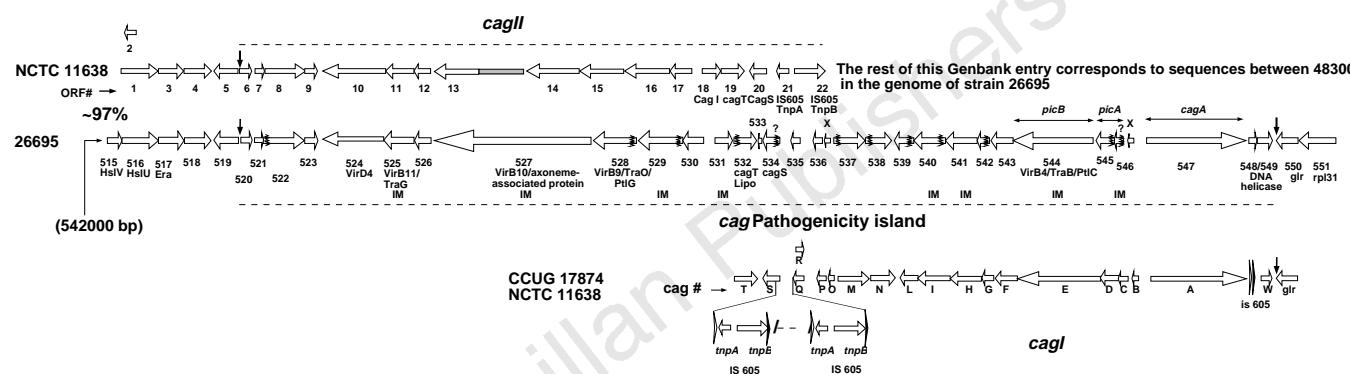


Figure 4 Comparison between the Cag pathogenicity islands of the sequenced strain, 26695 and the NCTC11638 strain. The twenty nine ORFs of the contiguous PAI in strain 26695 are represented together with the corresponding ORFs from the PAI present in NCTC11638 (AC000108 and U60176). The PAI in NCTC11638 is divided by the IS 605 elements into two regions, *cagI* and *cagII*. The PAI in NCTC11638 is flanked by a 31-bp (TTACAATTGAGGCCATTCTTAGCTGTTT) direct repeat (vertical arrows) as described¹¹. Some of the genes encode proteins with similarity to proteins involved either in DNA transfer (Vir and Tra proteins) or in export of a toxin (Ptl protein)¹⁰. However, these genes do not have the conserved contiguous arrangement found in the VirB, Tra and Ptl operons, suggesting that this PAI is not derived from these systems. Most genes of the PAI have no database match, contrary to a previous suggestion¹¹. Thirteen of the proteins have a signal peptide (squiggle line), three of them with a weaker probability (squiggled line+?). The average length of the signal peptides is 25 amino acids, suggesting that this PAI is of Gram-negative origin. Eight proteins are predicted to have at least two membrane-spanning domains and to be integral membrane proteins

(IM)⁴⁷. Although the two PAI are ~97% identical at the nucleotide level, there are several notable and perhaps biologically relevant differences between the two sequences. Four of the genes differ in size. In the PAI of strain 26695, HP 520 and 521 are shorter, whereas HP523 is longer, and HP 527 actually spans both ORF13 and 14. In addition, the N-terminal part of HP527 is 129 amino acids longer than the corresponding region in ORF14. HP548/549 contains a frameshift and is therefore probably inactive in strain 26695. The stippled box preceding ORF13 represents an N-terminal extension not annotated in the Genbank entry for the PAI of NCTC11638. The 'x' indicates ORFs that are neither GeneMark-positive nor GeneSmith-positive, so were not included in our gene list. However, these ORFs may be biologically significant. We do not represent *cagR* as an ORF, because it is completely contained within ORFQ, and is GeneMark-negative.

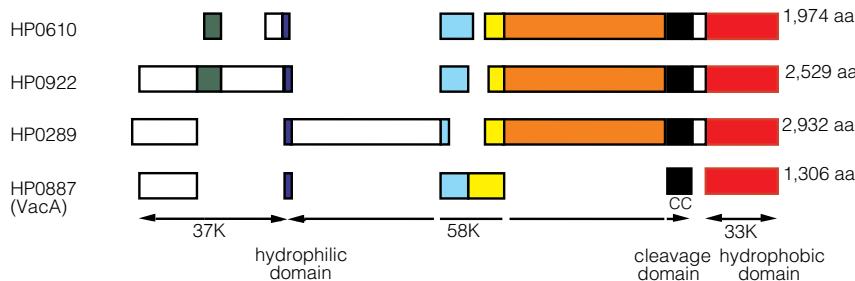


Figure 5 Conserved domains of VacA and related proteins. HP887 is the vacuolating cytotoxin (vacA) gene from *H. pylori* 26695 strain. HP610, HP922 and HP289 are related proteins. Blocks of aligned sequence and the length of each protein are shown. Arrows designate the extents of each VacA domain. The hydrophilic domain (blue boxes) contains the site in VacA at which the N-terminal domain is cleaved into 37K and 58K fragments. The putative cleavage site (ANNNQQNS) differs from that of three cytotoxic strains (CCUG 1784, 60190, G39; AKNDKXES) and is not conserved in the other three VacA-related proteins. The cleavage domain (black boxes) of VacA contains a pair of Cys residues 60 residues upstream from the site at which the C terminus is cleaved. These residues are not conserved in the other three proteins. The 33K C-terminal hydrophobic domain (red boxes) in VacA is thought to form a pore through which the toxin is secreted. The other three proteins show 26–31% sequence similarity to VacA in this region. The other coloured boxes represent regions of similarity.

AKNDKXES) and is not conserved in the other three VacA-related proteins. The cleavage domain (black boxes) of VacA contains a pair of Cys residues 60 residues upstream from the site at which the C terminus is cleaved. These residues are not conserved in the other three proteins. The 33K C-terminal hydrophobic domain (red boxes) in VacA is thought to form a pore through which the toxin is secreted. The other three proteins show 26–31% sequence similarity to VacA in this region. The other coloured boxes represent regions of similarity.

cytotoxin (~26–31%) (Fig. 5). However, they lack the paired-cysteine residues and the cleavage site required for release of the VacA toxin from the bacterial membrane³¹ (Fig. 5). We propose that these proteins may be retained on the outside surface of the cell membrane and contribute to the interaction between *H. pylori* and host cells.

The surface-exposed lipopolysaccharide (LPS) molecule plays an important role in *H. pylori* pathogenesis³². The LPS of *H. pylori* is several orders of magnitude less immunogenic than that of enteric bacteria³³ and the O antigen of many *H. pylori* isolates is known to mimic the human Lewis^x and Lewis^y blood group antigen³². Genes for synthesis of the lipid A molecule, the core region, and the O antigen were identified. Two genes with low similarity to fucosyltransferases (HP379, HP651) were found and may play a role in the LPS-Lewis antigen molecular mimicry. Our analysis also suggests that three genes, two glycosyltransferases (HP208 and HP619) and one fucosyltransferase (HP379), may be subject to phase variation (Table 3a, b).

As with other pathogens, *H. pylori* probably requires an iron-scavenging system for survival in the host⁵. Genome analysis suggests that *H. pylori* has several systems for iron uptake. One is analogous to the siderophore-mediated iron-uptake *fec* system of *E. coli*³⁴, except that it lacks the two regulatory proteins (FecR and FecI) and is not organized in a single operon. Unlike other studied systems, *H. pylori* has three copies of each of *fecA*, *exbB* and *exbD*. A second system, consisting of a *feoB*-like gene without *feoA*, suggests that *H. pylori* can assimilate ferrous iron in a fashion similar to the anaerobic *feo* system of *E. coli*. Other systems for iron uptake present in *H. pylori* consist of the three *frpB* genes which encode proteins similar to either haem- or lactoferrin-binding proteins. Finally, *H. pylori* contains NapA, a bacterioferritin³⁴, and Pfr, a non-haem cytoplasmic iron-containing ferritin used for storage of iron³⁵. The global ferric uptake regulator (Fur) characterized in other bacteria is also present in *H. pylori*. Consensus

sequences for Fur-binding boxes were found upstream of two *fecA* genes, the three *frpB* genes and *fur*.

H. pylori motility is essential for colonization³⁶. It enables the bacterium to spread into the viscous mucus layer covering the gastric epithelium. At least forty proteins in the *H. pylori* genome appear to be involved in the regulation, secretion and assembly of the flagellar architecture. As has been reported for the *flaA* and *flaB* genes, we identified sigma 28 and sigma 54-like promoter elements upstream of many flagellar genes, underscoring the complexity of the transcriptional regulation of the flagellar regulon⁵.

Acidity, pH and acid tolerance

H. pylori is unusual among pathogenic bacteria in its ability to colonize host cells in an environment of high acidity. As it enters the gastric environment by oral ingestion, the organism is transiently subjected to the extreme pH of the lumen side of the gastric mucus layer (pH ~2). The survival of *H. pylori* in acidic environments is probably due to its ability to establish a positive inside-membrane potential³⁷ and subsequently to modify its microenvironment through the action of urease and the release of factors that inhibit acid production by parietal cells⁵. A switch in membrane polarity provides an electrical barrier that prevents the entry of protons (H⁺). A positive cell interior can be created by the active extrusion of anions or by a proton diffusion potential. The latter model appears more likely as no clear mechanism for electrogenic anion efflux is apparent in the genome. A proton diffusion potential would require the anion permeability of the cytoplasmic membrane to be low and, thus far, only three anion transporters have been identified. However, it remains to be determined whether anion conductances are associated with other proteins: the MDR-like transporters (HP600, HP1082 and HP1206) or hypotheticals. Although it has been suggested that proton-translocating P-type ATPases could mediate survival in acid conditions by the extrusion of protons from the cytoplasm³⁸, this idea is not supported by the identified transporter

Table 3 Homopolymeric tracts and dinucleotide repeats in *H. pylori*

HP no.	ID	No. of repeats	Gene status	Poly(A) or Poly(T) tracts in 5' intergenic region				
9	OMP	11 CT	Off	Poly(A)				
208	glycos. transf.	11 AG	Truncated	Poly(A)				
638	OMP	6 CT	On	No				
722	OMP	8 CT	Off	Poly(T)				
725	OMP	6 CT	Off	Poly(T)				
744	Hypo	9 AG	Truncated	No				
896	OMP	11 CT	On	Poly(A)				
1417	Cons. Hypo	9 AG	Truncated	No				
Nucleotide sequence at the beginning of HP0722 showing the CT dinucleotide repeat and the poly T tract. The putative ribosome binding site is shown in green. Translation starting at the designated methionine leads to a truncated product. The addition or deletion of two CT repeats, by 'slipped-strand mispairing', will restore the frame.								
CCAAAAAAATCTTTTTTTTTGAATCCAATAAAATTGGTAAAGT-37bp-TTACAAATAAAAAAATTACTTAAAGGAACATT								
TATGAAAAAGACAATTCTACTCTCTCTCTCGCTCATCGCTTGTGACGCTGAAGACAACCGCTTTTGTGAGCGCCGGCT								
Y E K D N S T L S L S L A S S S L L H A E D N G F F V S A G Y								
M K K T I L L S L S L S L H R S C T L K T T A F L *								
(b) Homopolymeric poly(C) and poly(G) tracts within coding sequence								
HP no.	ID		Tract length	Gene status				
58	Hypo		C15	Off				
217	Hypo		G12	On				
379	fucosyl transf.		C13	On				
464	TypeI R		C15	On				
619	glycos. transf.		C13	Truncated				
651	Hypo		C13	On				
1353	Hypo		C15	Truncated				
1471	TypeII S-R		G14	On				
1522	Methyl ase		G12	Truncated				
Genes possibly regulated by homopolymeric poly(A) or poly(T) tracts in 5' intergenic regions								
HP no.	ID	Tract	HP no.	ID	Tract	HP no.	ID	Tract
9	OMP	A14	25	OMP	T15	208	<i>rfaJ</i>	A11
227	OMP	T14	228	IMP	A14	349	<i>pyrG</i>	T15
350	IMP	A15	547	<i>cagA</i>	A14	629	Hypo	T15
722	OMP	T16	725	OMP	T14	733	Hypo	T13
876	<i>frpB</i>	T16	896	OMP	A14	912	OMP	T13
1342	OMP	A14	1400	<i>fecA</i>	A16			

genes. The P-type ATPase sequences in *H. pylori* (*copAP*, HP791, and HP1503) are more closely related to divalent cation transporters than to ATPases with specificity for protons or monovalent cations. One of them, HP0791, is involved in Ni^{2+} supply, an essential component of urease activity³⁹. The others may be involved in the elimination of toxic metals from the cytoplasm and not in pH regulation.

Additional mechanisms of pH homeostasis may well contribute to *H. pylori* survival. A change in protein content observed in response to a shift of extracellular pH from 7.5 to 3.0 suggests the presence of an acid-inducible response⁴⁰. Although *H. pylori* lacks most orthologues of the genes that are acid-induced in *E. coli* and *Salmonella typhimurium*, including the amino-acid decarboxylases and formate hydrogen lyase, certain virulence factors, outer membrane

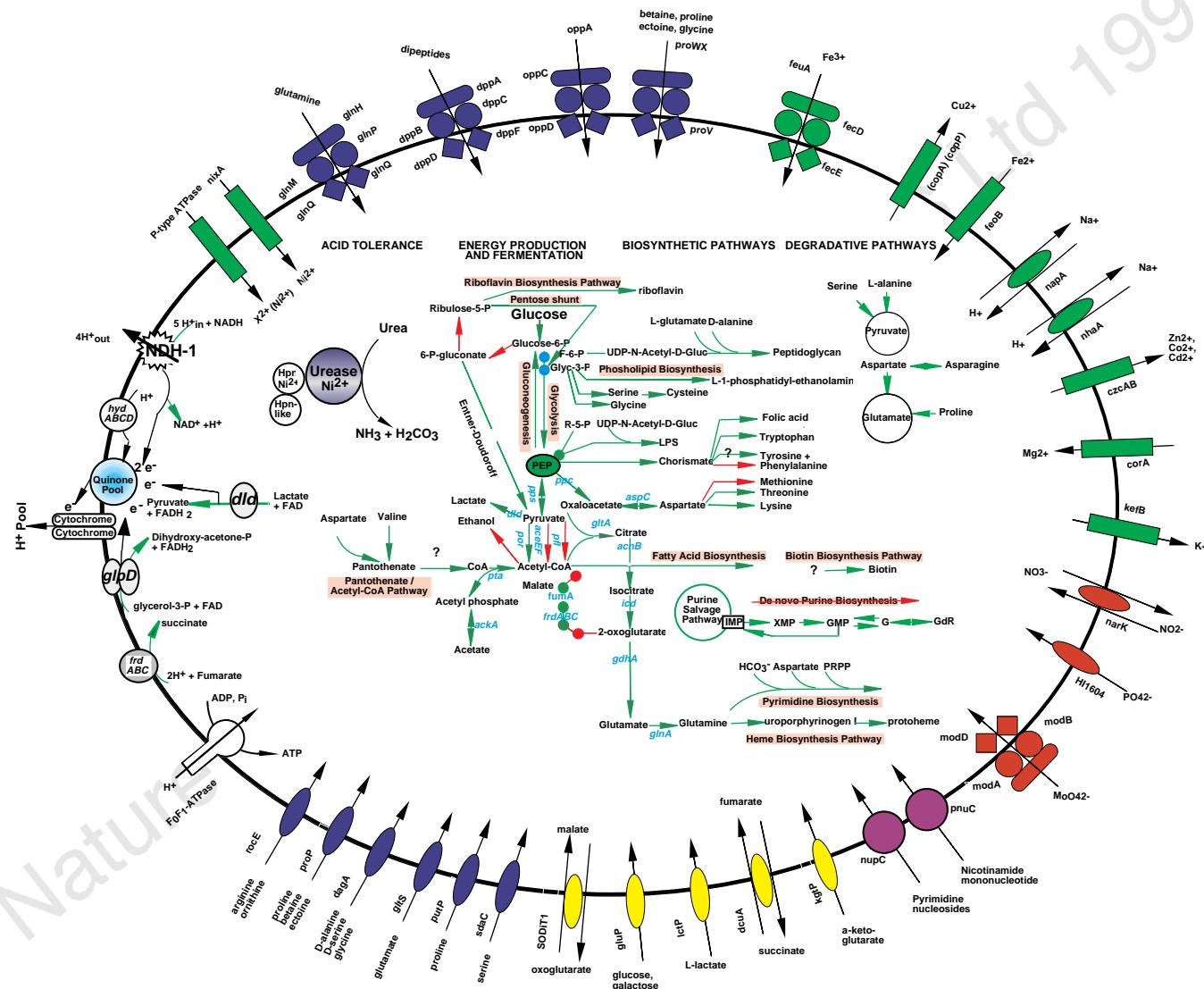


Figure 6 Solute transport and metabolic pathways of *Helicobacter pylori*. Transporters identified by sequence comparisons are characteristic of Gram-negative bacteria. Colours correspond to transport role categories defined by Riley¹⁵: blue, amino acids, peptides and amines; red, anions; yellow, carbohydrates, organic alcohols and acids; green, cations; and purple, nucleosides, purines and pyrimidines. Numerous permeases (ovals) with specificity for amino acids (*recE*, *propP*, *dagA*, *gltS*, *putP* and *sdaC*) or carbohydrates (*SODITI*, *gluP*, *lactP*, *cdtA*, *kgtP*) import organic nutrients. Structurally related permease proteins maintain ionic homeostasis by transporting HPO_4^{2-} (*H1604*), NO_3^- (*narK*), and Na^+ (*nhA*, *napA*). Primary active-transport systems, independent of the proton cycle, are also apparent. Included in this group are ATP-binding protein-cassette (ABC) transporters (composite figures of 2 diamonds, 2 circles, 1 oval) for the uptake of oligopeptides (*oppACD*), dipeptides (*dppABCDF*), proline (*proVWX*), glutamine (*glnHMPQ*), molybdenum (*modABD*), and iron III (*fecED*), P-type ATPases that extrude toxic metals from the cell (*copAP* and *cadA*), and the glutathione-regulated potassium-efflux protein (*kefB*). Transporters for the accumulation of ionic cofactors are encoded by *nixA* (Ni^{2+} for urease activation), *corA* (Mg^{2+} for phosphohydrolases, phosphotransferases, ATPases) and *feoB* (Fe^{2+}

import under anaerobic conditions for cytochromes, catalase). An integrated view of the main components of the central metabolism of *H. pylori* strain 26695 is presented. The use of glucose as the sole carbohydrate source is emphasized. Urease, a multisubunit Ni^{2+} -binding enzyme, is crucial for colonization and for survival of *H. pylori* at acid pH, and is indicated as a complex (purple circle) with HpN, a Ni^{2+} -binding cofactor, and a newly identified HpN-like protein (HP1432). A question mark is attached to pathways that could not be completely elucidated. Pathways or steps for which no enzymes were identified are represented by a red arrow. Pathways for macromolecular biosynthesis (RNA, DNA and fatty acids) have been omitted. *ackA*, acetate kinase; *acnB*, aconitase B; *aspC*, aspartate aminotransferase; *dld*, D -lactate dehydrogenase; *gdhA*, glutamate dehydrogenase; *glnA*, glutamine synthetase; *gltA* citrate synthase; *HydABC*, hydrogenase complex; *icd*, isocitrate dehydrogenase; *pfl*, pyruvate formate lyase; *por*, pyruvate ferredoxin oxidoreductase; *ppc*, phosphoenolpyruvate carboxylase; *pps*, phosphoenolpyruvate synthase; *pta*, phosphate acetyltransferase; *gldD*, glycerol-3-phosphate dehydrogenase; NDH-1, NADH-ubiquinone oxidoreductase complex.

proteins, sensor-regulator pairs and other proteins may be acid-induced.

Regulation of gene expression

Bacteria regulate the transcription of their genes in response to many environmental stimuli, such as nutrient availability, cell density, pH, contact with target tissue, DNA-damaging agents, temperature and osmolarity. In the case of pathogens, the regulated expression of certain key genes is essential for successful evasion of host responses and colonization, adaptation to different body sites, and survival as the pathogen passes to new hosts. In *H. pylori*, global regulatory proteins are less abundant than in *E. coli*. For example, orthologues of many DNA-binding proteins that regulate the expression of certain operons such as OxyR (oxidative stress), Crp (carbon utilization), RpoH (heat shock), and Fnr (fumarate and nitrate regulation) are absent. Only four *H. pylori* proteins have a perfect match to helix-turn-helix (HTH) motifs, a signature of transcription factors; a putative heat-shock protein (HspR), two proteins with no database match (HP1124 and HP1349) and SecA, a component of the general secretory machinery. In contrast, 34 proteins containing an HTH motif were found in *H. influenzae* and 148 in *E. coli*. We identified several other putative regulatory functions, including SpoT and CstA for 'stringent response' to amino-acid starvation and to carbon starvation, respectively.

Environmental response requires sensing changes and transmission of this information to cellular regulatory networks. Two-component regulator systems, consisting of a membrane histidine kinase sensor protein and a cytoplasmic DNA-binding response regulator, provide a well studied mechanism for such signal transduction. Four sensor proteins and seven response regulators were found in *H. pylori*, similar to the number found in *H. influenzae*⁷. This is approximately one third the number found in *E. coli* which, in contrast to *H. pylori* and *H. influenzae*, may be exposed to more environments.

Metabolism

Metabolic pathway analysis of the *H. pylori* genome suggests the following features. *H. pylori* uses glucose as the only source of carbohydrate and the main source for substrate-level phosphorylation. It also derives energy from the degradation of serine, alanine, aspartate and proline. The glycolysis–gluconeogenesis metabolic axis constitutes the backbone of energy production and the start point of many biosynthetic pathways. The biosynthesis of peptidoglycan, phospholipids, aromatic amino acids, fatty acids and cofactors is derived from acetyl-CoA or from intermediates in the glycolytic pathway (Fig. 6). The metabolism of pyruvate reflects the microaerophilic character of this organism. Neither the aerobic pyruvate dehydrogenase (*aceEF*) nor the strictly anaerobic pyruvate formate lyase (*pfl*) associated with mixed-acid fermentation are present. The conversion of pyruvate to acetyl CoA is performed by the pyruvate ferrodoxin oxidoreductase (POR), a four-subunit enzyme thus far only described in hyperthermophilic organisms⁴¹. The tricarboxylic acid cycle (TCA) is incomplete and the glyoxylate shunt is absent. The analysis of degradative pathways, uptake systems and biosynthetic pathways for pyrimidine, purine and haem suggests that *H. pylori* uses several substrates as nitrogen source, including urea, ammonia, alanine, serine and glutamine. The assimilation of ammonia, an abundant product of urease activity, is achieved by the glutamine synthase enzyme and α -ketoglutarate is transformed into glutamate by glutamate dehydrogenase rather than by the glutamate synthase enzyme.

In *H. pylori*, proton translocation is mediated by the NDH-1 dehydrogenase and the different cytochromes, including the primitive-type cytochrome *cbb3* (Table 2). Four respiratory electron-generating dehydrogenases have been identified, glycerol-3-phosphate dehydrogenase (GlpD), α -lactate dehydrogenase, NADH–ubiquinone oxidoreductase complex (NDH-1), and a hydrogenase complex (HydABC). Our analysis also suggests that

H. pylori is not able to use nitrate, nitrite, dimethylsulphoxide, trimethylamine N-oxide or thiosulphate as electron acceptors. Much of our metabolic analysis is supported by experimental evidence^{41,42}.

Evolutionary relationships of *H. pylori*

H. pylori is currently classified in the Proteobacteria, a large, diverse division of Gram-negative bacteria which includes two other completely sequenced species, *H. influenzae* and *E. coli*. Given this taxonomic placement, based primarily on 16S rRNA sequence comparisons, one might expect the proteins of *H. pylori* more closely to resemble their *H. influenzae* and *E. coli* homologues rather than those in other genomes such as *Synechocystis* sp., *M. genitalium*, *M. pneumoniae*, *M. jannaschii*, and *Saccharomyces cerevisiae*. This is indeed the case for many proteins. There are, however, many examples of *H. pylori* proteins in amino-acid biosynthesis, energy metabolism, translation and cellular processes that have greater sequence similarity to those found in non-Proteobacteria. For example, Dhs1, the initial enzyme in the chorismate biosynthesis pathway is 75.5% similar to *Arabidopsis thaliana* chloroplast Dhs1 gene product, and has minimal sequence similarity to the equivalent *E. coli* AroH, AroF or AroG gene products. The remaining enzymes in this pathway have strong sequence similarity to their *E. coli* counterpart. Similarly, the *H. pylori* prephenate dehydrogenase (TyrA), which converts chorismate to tyrosine, and six out of 15 enzymes in the aspartate amino acid biosynthetic pathways, resemble those from *B. subtilis*. A similar pattern can be seen in a different functional category. Nearly all *H. pylori* tRNA synthetases have eubacterial homologues, mostly with best matches to Proteobacteria species. However, histidyl-tRNA synthetase shows several amino-acid sequence signatures in common with eukaryotic and archaeal (*M. jannaschii*) homologues.

Such observations of discordant sequence similarity are often interpreted as evidence of lateral gene transfer in the evolutionary history of an organism. It is also possible that *H. pylori* diverged early from the lineage that led to the gamma Proteobacteria, and retained more ancient forms of enzymes that have been subsequently replaced or have diverged extensively in *H. influenzae* and *E. coli*.

Conclusion

Our whole-genome analysis of *H. pylori* gives new insight into its pathogenesis, acid tolerance, antigenic variation and microaerophilic character. The availability of the complete genome sequence will allow further assessment of *H. pylori* genetic diversity. This is an important aspect of *H. pylori* epidemiology as allelic polymorphism within several loci has already been associated with disease outcome^{5,21,31}. The extent of molecular mimicry between *H. pylori* and its human host, an underappreciated topic, can now be fully explored⁴³. The identification of many new putative virulence determinants should allow critical tests of their roles and thus new insight into mechanisms of initial colonization, persistence of this bacterium during long-term carriage, and the mechanisms by which it promotes various gastroduodenal diseases.

Methods

H. pylori strain 26695 (ref. 44) was originally isolated from a patient in the United Kingdom with gastritis (K. Eaton, personal communication) and was chosen because it colonizes piglets and elicits immune and inflammatory responses. It is also toxicogenic, and transformable, and thus amenable to mutational tests of gene function.

The *H. pylori* genome sequence was obtained by a whole-genome random sequencing method previously applied to genomes of *Haemophilus influenzae*⁷, *Mycoplasma genitalium*⁸, and *Methanococcus jannaschii*⁹. Ninety-two per cent of the genome was covered by at least one λ clone and only 0.56% of the genome had single-fold coverage.

Open reading frames (ORFs) and predicted coding regions were identified using three methods. The predicted protein-coding regions were initially defined by searching for ORFs longer than 80 codons. Coding potential analysis of the entire genome was performed with a version of GeneMark⁴⁵ trained with a set of *H. pylori* ORFs longer than 600 nucleotides. Coding sequences and potential starts of translation were also determined using GeneSmith (H.S., unpublished), a program that evaluates ORF length, separation of ORFs and overlap and quality of ribosome binding site. ORFs with low GeneMark coding potential, no database match, and not retained by GeneSmith were eliminated. GeneSmith identified 25 ORFs that are smaller than 100 codons, had no database match and were GeneMark negative. Frameshifts were detected by inspecting pairwise alignments, families of orthologues (similar proteins derived from different species) and paralogues (similar proteins from within the same organism), and regions containing homopolymer stretches and dinucleotide repeats. Ambiguities were resolved by an alternative sequencing chemistry (terminator reactions), and by sequencing PCR products obtained using the genomic DNA as template. Frameshifts that remain in the genome are considered authentic and not sequencing artefacts.

To determine their identity, ORFs were searched against a non-redundant amino-acid database as previously described⁹. ORFs were also analysed using 175 hidden Markov models constructed for a number of conserved protein families (pfam v1.0) using hmmer⁴³. In addition, all ORFs were searched against the prosite motif database using MacPattern⁴⁶. Families of paralogues were constructed by pairwise searches of proteins using FASTA. Matches that spanned at least 60% of the smaller of the protein pair were retained and visually inspected.

A unix version of the program TopPred⁴⁷ was used to identify membrane-spanning domains (MSD) in proteins. Six hundred and sixty three proteins containing at least one MSD were found; of these, 300 had 2 potential MSDs or more. The presence of signal peptides and the probable position of the cleavage site in secreted proteins were detected using Signal-P, a neural net program that had been trained on a curated set of secreted proteins from Gram-negative bacteria⁴⁸. 367 proteins were predicted to have a signal peptide. Lipoproteins were identified by scanning for the presence of a lipobox in the first 30 amino acids of every protein; 20 lipoproteins were identified, eighteen of which were Signal-P positive. Outer-membrane proteins were found by searching for aromatic amino acids at the end of the proteins.

Homopolymer and dinucleotide repeats were found by using RepScan (H.O.S., unpublished) which finds direct repeats of any length. All features identified using these programs were validated by visual inspection to remove false positives. Metabolic pathways were curated by hand and by reference to EcoCyc⁴⁹.

Received 16 May; accepted 1 July 1997.

- Warren, J. R. & Marshall, B. Unidentified curved bacilli on gastric epithelium in active chronic gastritis. *Lancet* **1**, 1273–1275 (1983).
- Cover, T. L. & Blaser, M. J. *Helicobacter pylori* infection, a paradigm for chronic mucosal inflammation: pathogenesis and implications for eradication and prevention. *Adv. Int. Med.* **41**, 85–117 (1996).
- Mobley, H. L. T., Island, M. D. & Hausinger, R. P. Molecular Biology of Microbial Ureases. *Microbiol. Rev.* **59**, 451–480 (1995).
- Go, M. F. & Graham, D. Y. How does *Helicobacter pylori* cause duodenal ulcer disease: The bug, the host, or both? *J. Gastroenterol. Hepatol.* (suppl.) **9**, 8–12 (1994).
- Labigne, A. & de Reuse, H. Determinants of *Helicobacter pylori* pathogenicity. *Infect. Agents Disease* **5**, 191–202 (1996).
- Clemens, J. et al. Impact of infection by *Helicobacter pylori* on the risk and severity of endemic cholera. *J. Inf. Dis.* **171**, 1653–1656 (1995).
- Fleischmann, R. D. et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
- Fraser, C. M. et al. The *Mycoplasma genitalium* genome sequence reveals a minimal gene complement. *Science* **270**, 397–403 (1995).
- Bult, C. J. et al. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**, 1058–1073 (1996).
- Winans, S. C., Burns, D. L. & Christie, P. J. Adaptation of a conjugal transfer system for the export of pathogenic macromolecules. *Trends Microbiol.* **4**, 64–68 (1996).
- Censini, S. et al. Cag, a pathogenicity island of *Helicobacter pylori*, encodes type-specific and disease-associated virulence factors. *Proc. Natl. Acad. Sci. USA* **93**, 14648–14653 (1996).
- <http://genome.wustl.edu/eddy/low/tRNAscan-SE-Manual/Manual.html>
- Akopyants, N. S., Kersulyte, D. & Berg, D. E. DNA rearrangement in the 40 kb cag (virulence) region in the *Helicobacter pylori* genome. *Gut* **39** (suppl. 2), A67 (1996).
- Marczynski, G. T. & Shapiro, L. Bacterial chromosome origins of replication. *Curr. Opin. Gen. Dev.* **3**, 775–782 (1993).
- Riley, M. Functions of gene products of *Escherichia coli*. *Microbiol. Rev.* **57**, 862–952 (1993).
- Kornberg, A. & Baker, T. A. Replication mechanisms and operations in DNA replication. (ed. Kornberg, A. & Baker, T.) 471–510 (Freeman, New York, 1992).

- Macnab, R. M. in *Escherichia coli and Salmonella Cellular and Molecular Biology* (eds Neidhardt, F. C. et al.) 123–145 (ASM, Washington DC, 1996).
- Strom, M. S., Nunn, D. N. & Lory, S. Posttranslational processing of type IV prepilin and homologs by PilD of *Pseudomonas aeruginosa*. *Meth. Enzymol.* **235**, 527–540 (1994).
- Bardwell, J. C. Building bridges: disulphide bond formation in the cell. *Mol. Microbiol.* **14**, 199–205 (1994).
- Linn, S. in *Escherichia coli and Salmonella Cellular and Molecular Biology* (eds Neidhardt, F. C. et al.) 764–772 (ASM, Washington DC, 1996).
- Peek, R. M., Thompson, S. A., Atherton, J. C., Blaser, M. J. & Miller, G. G. Expression of iceA, a novel ulcer-associated *Helicobacter pylori* gene, is induced by contact with gastric epithelial cells and is associated with enhanced mucosal IL-8. *Gut* **39** (suppl. 2), A71 (1996).
- Curnow, A. W., Ibba, M. & Soll, D. tRNA-dependent asparagine formation. *Nature* **382**, 589–590 (1996).
- Jones, A. C., Foynes, S., Cockayne, A. & Penn, C. W. Gene cloning of a flagellar sheath protein of *Helicobacter pylori* shows its identity with the putative adhesin, HpaA. *Gut* **39** (suppl. 2), A62 (1996).
- Boren, T., Falk, P., Roth, K. A., Larson, G. & Normark, S. Attachment of *Helicobacter pylori* to human gastric epithelium mediated by blood group antigens. *Science* **262**, 1892–1895 (1993).
- Ilver, D. et al. The *Helicobacter pylori* blood group antigen binding adhesin. *Gut* **39** (suppl. 2), A55 (1996).
- Odenbreit, S., Till, M. & Haas, R. Optimized blaM-transposon shuttle mutagenesis of *Helicobacter pylori* allows identification of novel genetic loci involved in bacterial virulence. *Mol. Microbiol.* **20**, 361–373 (1996).
- Peterson, S. N. et al. Characterization of repetitive DNA in the *Mycoplasma genitalium* genome: possible role in the generation of antigenic variation. *Proc. Natl. Acad. Sci. USA* **92**, 11829–11833 (1995).
- Moxon, E. R., Rainey, P. B., Nowak, M. A. & Lenski, R. E. Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr. Biol.* **4**, 24–33 (1994).
- Jonsson, A. B., Nyberg, G. & Normark, S. Phase variation of gonococcal pili by frameshift mutation in pilC, a novel gene for pilus assembly. *EMBO J.* **10**, 477–488 (1991).
- Tummaru, M. K. R., Sharma, S. A. & Blaser, M. J. *Helicobacter pylori* picB, a homologue of the *Bordetella pertussis* toxin secreted protein, is required for induction of IL-8 in gastric epithelial cells. *Mol. Microbiol.* **18**, 867–876 (1995).
- Atherton, J. C. et al. Mosaicism in vacuolating cytotoxin alleles of *Helicobacter pylori*. Association of specific vacA types with cytotoxin production and peptic ulceration. *J. Biol. Chem.* **270**, 17771–17777 (1995).
- Moran, A. P. The role of lipopolysaccharide in *Helicobacter pylori* pathogenesis. *Aliment. Pharmacol. Ther.* **10** (suppl. 1), 39–50 (1996).
- Baker, P. J. et al. Molecular structures that influence the immunomodulatory properties of the lipid A and inner core region oligosaccharides of bacterial lipopolysaccharides. *Infect. Immun.* **62**, 2257–2269 (1994).
- Earhart, C. F. in *Escherichia coli and Salmonella Cellular and Molecular Biology* (eds Neidhardt, F. C. et al.) 1075–1090 (ASM, Washington DC, 1996).
- Evans, D. J. Jr, Evans, D. G., Lampert, H. C. & Nakano, H. Identification of four new prokaryotic bacterioferritins, from *Helicobacter pylori*, *Anabaena variabilis*, *Bacillus subtilis* and *Treponema pallidum*, by analysis of gene sequences. *Gene* **153**, 123–127 (1995); Frazier, B. A. et al. Paracrystalline inclusions of a novel ferritin containing nonheme iron, produced by the human gastric pathogen *Helicobacter pylori*: evidence for a third class of ferritins. *J. Bacteriol.* **175**, 966–972 (1993).
- Suerbaum, S. The complex flagella of gastric *Helicobacter* species. *Trends Microbiol.* **3**, 168–170 (1995).
- Matin, A., Zychlinsky, E., Keyhan, M. & Sachs, G. Capacity of *Helicobacter pylori* to generate ionic gradients at low pH is similar to that of bacteria which grow under strongly acidic conditions. *Infect. Immun.* **64**, 1434–1436 (1996).
- Melchers, K. et al. Cloning and membrane topology of a P type ATPase from *Helicobacter pylori*. *J. Biol. Chem.* **271**, 446–457 (1996).
- Melchers, K. et al. Cloning and analysis of two P type ion pumps of *Helicobacter pylori*, a cation resistance ATPase and a membrane pump necessary for urease activity. *Gut* **39** (suppl. 2), A67 (1996).
- McGowan, C. C., Cover, T. L. & Blaser, M. J. *Helicobacter pylori* and gastric acid: biological and therapeutic implications. *Gastroenterology* **110**, 926–938 (1996).
- Hughes, N. J., Chalk, T. L., Clayton, C. L. & Kelly, D. J. Identification of carboxylation enzymes and characterization of a novel four-subunit pyruvate:flavodoxin oxidoreductase from *Helicobacter pylori*. *J. Bacteriol.* **177**, 3953–3959 (1995).
- Mendz, G. L. & Hazell, S. L. Aminocid utilization by *Helicobacter pylori*. *Int. J. Biochem. Cell. Biol.* **27**, 1085–1093 (1995).
- Sonnhammer, E. L. L., Eddy, S. R. & Durbin, R. Pfam: A comprehensive database of protein families based on seed alignments. *Proteins* (in the press).
- Akopyants, N. S., Eaton, K. A. & Berg, D. E. Adaptive mutation and co-colonization during *Helicobacter pylori* infection of gnotobiotic piglets. *Infect. Immun.* **63**, 116–121 (1995).
- Borodovsky, M., Rudd, K. E. & Koonin, E. V. Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucleic Acids Res.* **22**, 4756–4767 (1994).
- Fuchs, R. MacPattern: protein pattern searching on the Apple Macintosh. *Comput. Appl. Biosci.* **7**, 105–106 (1991).
- Claros, M. G. & von Heijne, G. TopPred II: an improved software for membrane protein structure predictions. *Comput. Appl. Biosci.* **10**, 685–686 (1994).
- Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**, 1–6 (1997).
- Karp, P. D., Riley, M., Paley, S. M., Pellegrini-Toole, A. & Krummenacker, M. EcoCyc: Encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.* **25**, 43–51 (1997).
- Doig, P., Exner, M. M., Hancock, R. E. & Trust, T. J. Isolation and characterization of a conserved porin protein from *Helicobacter pylori*. *J. Bacteriol.* **177**, 5447–5452 (1995).

Acknowledgements. D.E.B., M.B. and W.H. are supported by grants from the NIH; P.K. is supported by a grant from the National Center for Research Resources. We thank N. S. Akopyants for preparing high quality chromosomal DNA from *H. pylori* strain 26695; M. Heaney, J. Scott, A. Saeed and R. Shirley for software and database support; and V. Sapiro, B. Vincent, J. Meehan and D. Mass for computer system support.

Correspondence and requests for materials should be addressed to J.-F.T. (e-mail: ghp@tigr.org). The annotated genome sequence and gene family alignments are available on the World-Wide Web site at <http://www.tigr.org/tdb/mdb/hpdb/hpdb.html>. The sequence has been deposited with GenBank under accession number AE000511.

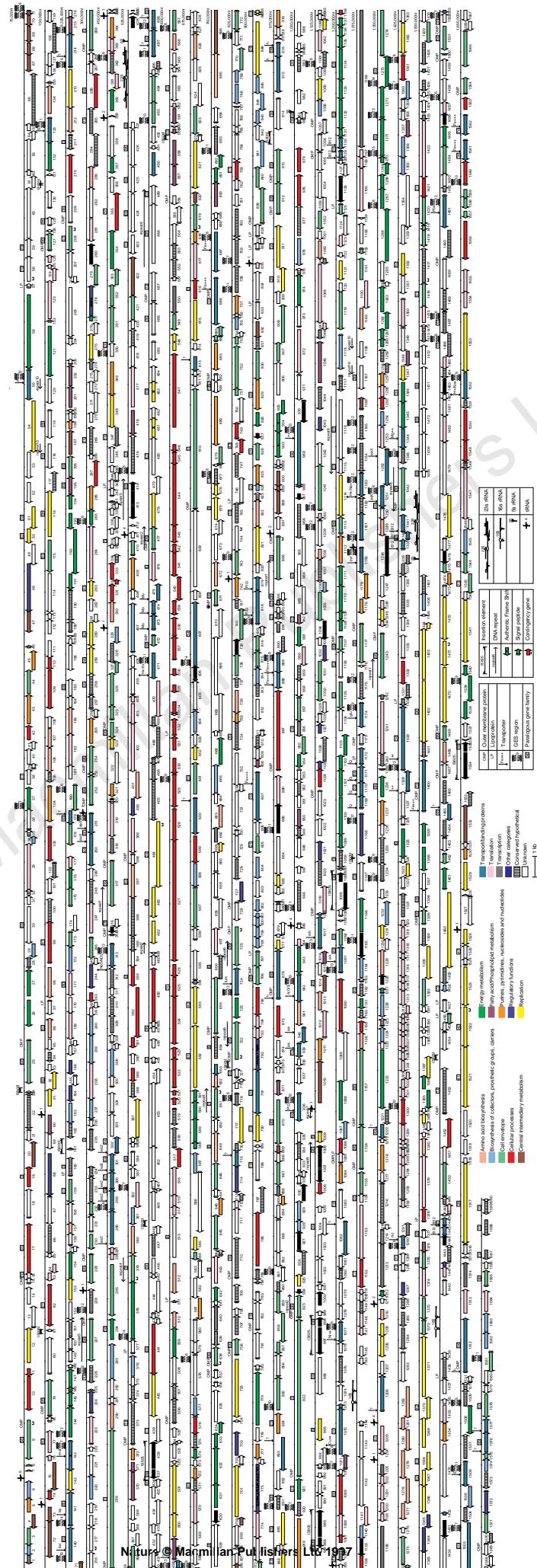


Table 2. List of *H. pylori* genes with putative identifications. Gene numbers correspond to those in Fig. 1. Each identified gene has been assigned a putative role category adapted from ref. 15. Percentages represent per cent identities.

AMINO-ACID BIOSYNTHESIS								
<i>General</i>								
HP0695	hydantoin utilization protein A (hyuA)	28.6%	HP0841	pantothenate metabolism flavoprotein (dfp)	31.3%	HP0855	elginate O-acetylation protein (elgJ)	41.8%
<i>Aromatic amino-acid family</i>			Pyridoxine	pyridoxal phosphate biosynthetic protein A (pdxA)	34.2%	HP0326	CMP-N-acetylneurameric acid synthetase (neuA)	31.9%
HP1038	3-dehydroquinate type II (aroQ)	99.4%	HP1582	pyridoxal phosphate biosynthetic protein J (pdxJ)	42.6%	HP0230	CTP:CM-3-deoxy-D-manno-octulosonate-cyclidytransferase (kdsB)	36.2%
HP0283	3-dehydroquinate synthase (aroB)	38.1%	HP1580	riboflavin biosynthesis protein (ribG)	33.1%	HP1392	fibronectin/fibrinogen-binding protein	25.7%
HP0134	3-deoxy-D-arabinohexitulosonate 7-phosphate synthase (dhs1)	54.6%	HP1581	riboflavin biosynthesis regulatory protein (ribC)	28.9%	HP0379	fucosyltransferase	39.2%
HP0401	3-phosphoshikimate		HP1574	riboflavin synthase alpha subunit (ribC)	32.8%	HP0651	fucosyltransferase	39.2%
	1-carboxyvinyltransferase (aroA)	53.6%	HP0802	GTP cyclohydrolase II (ribA)	47.2%	HP0044	CDP-D-mannose 4,6-dehydratase (ribD)	62.1%
HP1279	anthranilate isomerase (trpC)	47.0%	HP0804	GTP cyclohydrolase II/3,4-dihydroxy-2-butaneone 4-phosphate synthase (ribA, ribB)	44.0%	HP0867	lipid A disaccharide synthetase (lpxB)	32.0%
HP1282	anthranilate synthase component I (trpE)	47.9%	HP1505	riboflavin biosynthesis protein (ribG)	33.1%	HP0193	lipopolysaccharide 1,2-glucosyltransferase (lpxA)	28.9%
HP1280	anthranilate synthase component II (trpD)	42.5%	HP1087	riboflavin biosynthesis regulatory protein (ribC)	28.9%	HP0379	lipopolysaccharide 1,2-glucosyltransferase (lpxB)	26.7%
HP1281	anthranilate synthase component II (trpD)	40.2%	HP1574	riboflavin synthase alpha subunit (ribC)	32.8%	HP0805	lipopolysaccharide 5Gb epitope biosynthesis-associated protein (lexB)	36.9%
HP0663	chorismate synthase (aroC)	47.2%	HP0002	riboflavin synthase beta chain (ribE)	52.4%	HP0262	lipopolysaccharide 5Gb epitope biosynthesis-associated protein (lexB)	39.2%
HP1380	prephenate dehydrogenase (tyrA)	30.2%	HP1164	thioredoxin reductase (trxB)	28.5%	HP1416	lipopolysaccharide 1,2-glucosyltransferase (lpxA)	29.2%
HP1249	shikimate 5-dehydrogenase (aroE)	36.6%	Thioredoxin, glutaredoxin and glutathione		HP0679	lipopolysaccharide biosynthesis protein (wtpB)	42.8%	
HP0157	shikimic acid kinase I (aroK)	36.1%	HP1118	gamma-glutamyltranspeptidase (ggt)	53.2%	HP1475	lipopolysaccharide core biosynthesis protein (katB)	49.0%
HP1277	tryptophan synthase, alpha subunit (trpA)	46.5%	HP1458	thioredoxin	38.3%	HP0279	lipopolysaccharide heptosyltransferase-1 (lpxC)	31.7%
HP1278	tryptophan synthase, beta subunit (trpB)	66.1%	HP0824	thioredoxin (trxA)	51.5%	HP0619	lipopolysaccharide biosynthesis glycosyltransferase (ficB)	37.2%
<i>Aspartate family</i>			HP1164	thioredoxin reductase (trxB)	28.5%	HP1105	LPS biosynthesis protein	28.7%
HP0649	aspartate ammonia-lyase (aspA)	55.5%	Thiamine	thiamin biosynthesis protein (thiF)	34.6%	HP1578	LPS biosynthesis protein	28.1%
HP1189	aspartate-semialdehyde dehydrogenase (asd)	45.7%	HP0841	thiamin phosphosphate pyrophosphorylase/hydroxyethylthiazole kinase (thiB)	35.7%	HP1581	methionine resistance protein (lrrM)	29.2%
HP1229	aspartokinase (lysC)	48.0%	HP0843	thiamin phosphosphate pyrophosphorylase/hydroxyethylthiazole kinase (thiM)	37.7%	HP0857	phosphohexitose isomerase (gmnA)	44.5%
HP0106	cystathione gamma-synthase (metB)	47.7%	HP0845	thiamin biosynthesis protein (thiI)	41.0%	HP1275	phosphomannomutase (algC)	39.6%
HP0290	diaminopimelate decarboxylase (decA)		Pyridine nucleotides	NH3-dependent NAD+ synthetase (nadE)	37.5%	(<i>Pseudomonas aeruginosa</i>)		
	(decA) (lysA)	42.7%	HP0329	NH3-dependent NAD+ synthetase (nadE)	37.5%	HP1429	polysialic acid capsule expression protein (kpsF)	46.0%
HP0566	diaminopimelate epimerase (dapF)	30.0%	HP1355	nicotinate-nucleotide pyrophosphorylase (nadC)	36.3%	HP0366	spore coat polysaccharide biosynthesis protein C	35.3%
HP0510	dihydrodipicolinate reductase (dapB)	95.3%	HP1356	nicotinate-nucleotide pyrophosphorylase (nadA)	34.2%	HP0178	spore coat polysaccharide biosynthesis protein E	36.2%
HP1013	dihydrodipicolinate synthetase (dapA)	39.5%	HP1356	quinolinate synthetase A (nadA)	34.2%	HP0421	type I capsular polysaccharide biosynthesis protein J (capJ)	29.0%
HP0822	homoserine dehydrogenase (metL)	37.7%	CELL ENVELOPE		HP0196	UDP-3-(3-hydroxymyristoyl) glucosamine N-acetyltransferase (lpxD)	39.5%	
HP1050	homoserine kinase (thrB)	27.7%	Membranes, lipoproteins and porins		HP1052	UDP-3-O-acyl N-acetyl glucosamine deacetylase (envA)	44.6%	
HP0672	solute-binding signature and mitochondrial		HP1450	60 kDa inner-membrane protein	40.0%	HP1375	UDP-N-acetylglucosamine acyltransferase (lpxA)	41.8%
	signature protein (aspB)		HP0180	apolipoprotein N-succinyltransferase (cute)	28.0%	Surface structures		
HP0212	succinyl-diaminopimelate desuccinylase (dapE)	42.3%	HP0175	cell binding factor 2	34.9%	HP0840	flaA1 protein	60.2%
HP0626	tetrahydrodipicolinate N-succinyltransferase (dapD)	36.1%	HP0078	Hypothetical protein	28.4%	HP0325	flagellar basal-body L-ring protein (flgH)	32.7%
HP0098	threonine synthase (thrC)	32.9%	HP0567	membrane protein	26.4%	HP0351	flagellar basal-body M-ring protein (flfF)	34.4%
<i>Glutamate family</i>			HP1456	membrane-associated lipoprotein (lpp20)	98.9%	HP2046	flagellar basal-body P-ring protein (flgI)	37.9%
HP0380	glutamate dehydrogenase (gdhA)	59.0%	HP1564	outer membrane protein	39.9%	HP1557	flagellar basal-body protein (flfE)	37.0%
HP0512	glutamine synthetase (glnA)	48.6%	HP0009	outer membrane protein (omp1)	0.0%	HP1559	flagellar basal-body rod protein (flgB)	31.0%
HP1158	pyrroline-5-carboxylate reductase (proC)	28.9%	HP0324	outer membrane protein (omp10)	0.0%	HP1558	flagellar basal-body rod protein (flgC)	46.0%
<i>Pyruvate family</i>			HP0472	outer membrane protein (omp11)	99.5%	HP0192	flagellar basal-body rod protein (flgD)	35.5%
HP0941	alanine racemase, biosynthetic (alr)	32.4%	HP0477	outer membrane protein (omp12)	0.0%	HP0351	flagellar basal-body rod protein (flgG)	35.5%
HP1468	branched-chain-amino-acid aminotransferase (lIvE)	63.5%	HP0638	outer membrane protein (omp13)	0.0%	HP0351	flagellar basal-body rod protein (flgH)	35.5%
HP0330	ketol-acid reductoisomerase (lIvC)	48.1%	HP0671	outer membrane protein (omp14)	36.0%	HP0157	flagellar basal-body rod protein (flgI)	35.5%
<i>Serine family</i>			HP0706	outer membrane protein (omp15)	33.5%	HP1092	flagellar basal-body rod protein (flgJ)	35.5%
HP0107	cysteine synthetase (cysK)	45.7%	HP0722	outer membrane protein (omp16)	43.3%	HP1585	flagellar basal-body rod protein (flgK)	47.7%
HP0096	phosphoglycerate dehydrogenase	31.0%	HP0725	outer membrane protein (omp17)	43.3%	HP1041	flagellar biosynthesis protein (flhA)	43.1%
HP0397	phosphoglycerate dehydrogenase (serA)	32.5%	HP0796	outer membrane protein (omp18)	0.0%	HP1038	flagellar biosynthesis protein (flhF)	35.5%
HP0736	phosphoserine aminotransferase (serC)	30.7%	HP0896	outer membrane protein (omp19)	36.6%	HP0684	flagellar biosynthesis protein (flhP)	43.4%
HP0662	phosphoserine phosphatase (serB)	36.5%	HP0205	outer membrane protein (omp2)	0.0%	HP0770	flagellar biosynthetic protein (flhB)	38.7%
HP1210	serine acetyltransferase (cysE)	98.2%	HP0912	outer membrane protein (omp20)	0.0%	HP0685	flagellar biosynthetic protein (flhP)	55.6%
HP0183	serine hydroxymethyltransferase (glyA)	54.0%	HP0913	outer membrane protein (omp21)	38.2%	HP1419	flagellar biosynthetic protein (flhQ)	52.3%
<i>BIOSYNTHESIS OF COFACTORS, PROSTHETIC GROUPS, AND CARRIERS</i>			HP1107	outer membrane protein (omp23)	0.0%	HP0173	flagellar biosynthetic protein (flhR)	26.4%
<i>General</i>			HP1113	outer membrane protein (omp24)	36.0%	HP0353	flagellar export protein (flhI)	29.1%
HP0220	synthesis of [Fe-S] cluster (nifS)	48.0%	HP1156	outer membrane protein (omp25)	0.0%	HP1420	flagellar export protein ATP synthase (flhII)	47.6%
<i>Biotin</i>			HP1157	outer membrane protein (omp26)	23.0%	HP0670	flagellar hook (flgE)	98.9%
HP0598	8-amino-7-oxononanoate synthase (bioF)	34.9%	HP0227	outer membrane protein (omp5)	36.8%	HP0908	flagellar hook (flgE)	30.5%
HP0976	adenosylmethionine-8-amino-7-oxononanoate		HP1243	outer membrane protein (omp28)	0.0%	HP1119	flagellar hook-associated protein 1 (HAP1) (flgK)	27.6%
	aminotransferase (bioA)	49.2%	HP1342	outer membrane protein (omp29)	0.0%	HP0752	flagellar hook-associated protein 2 (flfD)	28.9%
HP1140	biotin operon repressor/biotin acetyl coenzyme A carboxylase synthetase (birA)	36.9%	HP0079	outer membrane protein (omp3)	0.0%	HP0815	flagellar motor rotation protein (motA)	32.9%
HP0407	biotin sulfoxide reductase (bisC)	42.7%	HP1395	outer membrane protein (omp30)	0.0%	HP0816	flagellar motor rotation protein (motB)	29.7%
HP1254	biotin synthetase (bioC)	32.1%	HP1469	outer membrane protein (omp31)	0.0%	HP0352	flagellar motor switch protein (flgG)	37.0%
HP0029	dethiobiotin synthetase (bioD)	36.0%	HP1501	outer membrane protein (omp32)	0.0%	HP1031	flagellar motor switch protein (flhI)	34.4%
<i>Folic acid</i>			HP0127	outer membrane protein (omp4)	0.0%	HP0753	flagellar protein (flfI)	32.3%
HP1036	7,8-dihydro-6-hydroxymethylpterin- pyrophosphokinase (folK)	34.6%	HP0227	outer membrane protein (omp5)	36.8%	HP0327	flagellar protein G (flaG)	23.3%
HP0587	aminoimidoxylcarboximide lyase (pabC)	32.4%	HP0317	outer membrane protein (omp6)	38.4%	HP0797	flagellar sheath adhesin (hpaA)	98.5%
HP1232	dihydrofolteroate synthase (folP)	34.5%	HP0254	outer membrane protein (omp7)	30.6%	HP0584	flagellar switch protein (flhN)	39.7%
HP1545	folylpolyglutamate synthase (folC)	35.2%	HP0317	outer membrane protein (omp9)	36.3%	HP0601	flagellin A (flaA)	99.8%
HP0928	GTP cyclohydrolase I (folE)	50.9%	HP0639	outer membrane protein P1 (ompP1)	23.3%	HP0295	flagellin B homologue (flaB)	99.0%
HP0577	methylene-tetrahydrofolate dehydrogenase (folD)	48.4%	HP0955	propylpropionate diacylglycerol transferase (lgt)	34.4%	HP1575	flhB protein (flhB)	40.5%
HP0293	para-aminobenzoate synthetase (pabB)	35.1%	HP0655	protective surface antigen D15	27.5%	HP1030	flhB protein (flhY)	29.3%
<i>Haem and porphyrin</i>			HP0610	toxin-like outer membrane protein	26.3%	HP0907	hook assembly protein, flagella (flgD)	25.5%
HP0163	delta-aminolevulinic acid dehydratase (hemB)	50.5%	HP0922	toxin-like outer membrane protein	29.5%	HP1274	paralysed flagella protein (pflA)	23.9%
HP0376	ferrochelatase (hemC)	33.4%	HP0289	toxin-like outer membrane protein	30.6%	HP0751	polar flagellin (flaG)	21.9%
HP0308	glutamate-1-semialdehyde 2,1-aminomutase (hemL)	51.3%	HP0493	transpeptidase (mraY)	45.2%	HP0410	putative neuraminyllactose-binding haemagglutinin homolog (hpaA)	24.2%
HP0239	glutamyl-tRNA reductase (hemA)	32.7%	HP0743	rod shape-determining protein (mreB)	37.9%	HP1192	secreted protein involved in flagellar motility	72.5%
HP0665	oxygen-independent coproporphyrinogen III oxidase (hemN)	42.4%	HP1372	rod shape-determining protein (mreC)	33.6%	HP1462	secreted protein involved in flagellar motility	96.2%
HP1226	oxygen-independent coproporphyrinogen III oxidase (hemN)	37.9%	HP0645	soluble lytic murein transglycosylase (slt)	32.2%	HP0232	secreted protein involved in flagellar motility	99.2%
HP0207	porphobilinogen deaminase (hemC)	45.7%	HP1543	toxR-activated gene (tagE)	37.2%	<i>CELLULAR PROCESSES</i>		
HP0381	protoporphyrinogen oxidase (hemK)	35.9%	HP1544	toxR-activated gene (tagE)	31.2%	<i>General</i>		
HP0604	uroporphyrinogen decarboxylase (hemE)	46.3%	HP1165	transferase, peptidoglycan synthesis (murG)	28.2%	HP0019	chemotaxis protein (cheV)	26.8%
HP1224	uroporphyrinogen III cosynthase (hemD)	27.6%	HP0740	UDP-MurNac-pentapeptide presynthetase (murF)	25.7%	HP0393	chemotaxis protein (cheV)	31.7%
<i>Menadione and ubiquinone</i>			HP1494	UDP-MurNac-tripeptide synthetase (murE)	36.0%	HP0616	chemotaxis protein (cheV)	27.9%
HP1360	4-hydroxybenzoate octaprenyltransferase (ubiA)	26.6%	HP1418	UDP-N-acetylglucosamine-4-epoxy-4-uridyl pyruvyl glucosamine reductase (murB)	32.7%	HP1067	chemotaxis protein (cheV)	99.2%
HP0929	geranyltransferase (ispA)	39.8%	HP0648	UDP-N-acetylglucosamine-4-epoxy-4-uridyl pyruvyl transferase (murZ)	46.7%	HP0617	GTP-binding protein (era)	95.6%
HP0240	octaprenyl-diphosphate synthase (ispB)	31.6%	HP0623	UDP-N-acetylglucosamine-4-epoxy-4-uridyl pyruvyl transferase (kdtA)	35.9%	HP1490	haemolysin	39.2%
<i>Molybdopterin</i>			HP0494	UDP-N-acetylglucosamine-4-epoxy-4-uridyl pyruvyl transferase (murD)	31.1%	HP1068	Haemolysin (lty)	40.2%
HP0768	molbdenum cofactor biosynthesis protein A (moeA)	31.4%	HP0695	3-deoxy-D-manno-octulosonic acid 8-phosphate synthetase (kdsA)	53.4%	HP0593	haemolysin secretion protein precursor (ltxA)	45.4%
HP0798	molbdenum cofactor biosynthesis protein C (moeC)	97.9%	HP1057	3-deoxy-D-manno-octulosonic-acid transferase (kdtA)	46.7%	HP0392	histidine kinase (cheA)	41.4%
HP0712	molbdenopterin biosynthesis protein (moeA)	36.3%	HP0651	ADP-uridylate-4-epoxy-4-uridyl pyruvyl transferase (kdtB)	47.7%	HP0099	methyl-accepting chemotaxis protein (flpA)	32.8%
HP0755	molbdenopterin biosynthesis protein (moeB)	32.2%	HP1030	ADP-uridylate-4-epoxy-4-uridyl pyruvyl transferase (kdtB)	30.7%	HP0103	methyl-accepting chemotaxis protein (flpB)	30.7%
HP0799	molbdenopterin biosynthesis protein (moeG)	50.8%	HP0494	ADP-uridylate-4-epoxy-4-uridyl pyruvyl transferase (kdtD)	46.7%	HP0082	methyl-accepting chemotaxis transducer (ltpC)	28.2%
HP0801	molbdenopterin converting factor, subunit 1 (moeA)	31.1%	HP0868	ADP-heptose synthase (rfaE)	40.6%	HP0391	purine-binding chemotaxis protein (cheW)	34.3%
HP0800	molbdenopterin converting factor, subunit 2 (moeA)	31.1%	HP1191	ADP-heptose-4-epoxy-4-uridyl pyruvyl transferase II (rfaF)	33.2%	<i>Cell division</i>		
HP0769	molbdenopterin-guanine dinucleotide biosynthesis protein A (moeA)	28.3%	HP0869	ADP-L-glycero-D-mannoheptose-6-epimerase (rfaD)	32.7%	HP0331	cell division inhibitor (minD)	50.2%
<i>Pantothenate</i>						HP0749	cell division membrane protein (ftsX)	25.7%
HP1068	3-methyl-2-oxobutanoate hydroxymethyltransferase (panB)	43.7%				HP0978	cell division protein (ftsA)	31.9%
HP0034	aspartate-1-decarboxylase (panD)	50.0%				HP0748	cell division protein (ftsE)	37.6%
HP0006	pantoate-beta-alanine ligase (panC)	44.2%				HP0288	cell division protein (ftsH)	41.2%
						HP1069	cell division protein (ftsI)	98.6%
						HP1566	cell division protein (ftsL)	30.6%
						HP1090	cell division protein (ftsK)	39.8%
						HP1560	cell division protein (ftsW) <i>Escherichia coli</i>	32.7%
						HP0763	cell division protein (ftsY)	46.6%

HP0332	cell division topological specificity factor (minE)	33.8%	HP1270	subunit (NQO10)	-1.0%	HP1101	(devB)	29.2%
HP0979	cell division protein (ftsZ)	43.3%	HP1271	NADH-ubiquinone oxidoreductase, NQO11 subunit (NQO11) [(Paracoccus denitrificans] 42.6%	HP1495	glucose-6-phosphate dehydrogenase (g6PD)	36.7%	
HP1519	cell filamentation protein (flic)	63.2%	HP1271	NADH-ubiquinone oxidoreductase, NQO12 subunit (NQO12)	HP1088	transaldolase (tal)	33.5%	
<i>Cell killing</i>			HP1272	NADH-ubiquinone oxidoreductase, NQO13 subunit (NQO13)	HP1088	transketolase A (tktA)	46.7%	
HP0887	vacuolating cytotoxin	94.7%	HP1273	NADH-ubiquinone oxidoreductase, NQO14 subunit (NQO14)	HP0354	transketolase B (tktB)	39.7%	
<i>Chaperones</i>			HP1266	NADH-ubiquinone oxidoreductase, NQO3 subunit (NQO3)	<i>Sugars</i>			
HP0010	chaperone and heat shock protein (groEL)	99.6%	HP1263	NADH-ubiquinone oxidoreductase, NQO4 subunit (NQO4)[Triticum aestivum]	HP0574	galactosidase acetyltransferase (IaCa)	41.0%	
HP0109	chaperone and heat shock protein 70 (dnkA)	63.4%	HP1262	NADH-ubiquinone oxidoreductase, NQO5 subunit (NQO5)	HP0360	UDP-glucose 4-epimerase	43.1%	
HP0210	chaperone and heat shock protein C62.5 (hspG)	46.5%	HP1261	NADH-ubiquinone oxidoreductase, NQO6 subunit (NQO6)	<i>TCA cycle</i>			
HP0011	co-chaperone (groES)	99.2%	HP1260	NADH-ubiquinone oxidoreductase, NQO7 subunit (NQO7)	HP0779	aconitase B (aconB)	64.0%	
HP1332	co-chaperone and heat-shock protein (dnal)	42.7%	HP1267	NADH-ubiquinone oxidoreductase, NQO8 subunit (NQO8)	HP0026	citrate synthase (cita)	47.8%	
HP0110	co-chaperone and heat-shock protein (grpE)	33.0%	HP1268	NADH-ubiquinone oxidoreductase, NQO9 subunit (NQO9)	HP1325	fumarate (fumC)	63.7%	
HP1024	co-chaperone-curved DNA-binding protein A (CbpA)	37.7%			HP0509	glycolate oxidase subunit (gicD)	98.0%	
<i>Chromosome-associated protein</i>					HP0027	isocitrate dehydrogenase (icd)	70.7%	
HP1138	plasmid replication-partition related protein	40.4%						
<i>Detoxification</i>								
HP1563	alkyl hydroperoxide reductase (tsaA)	98.5%						
HP0875	catalase	99.4%						
HP0267	chlorohydrolase	42.6%						
HP0243	neutrophil activating protein (napA) (bacteriorhodopsin)	95.8%						
HP0389	superoxide dismutase (sodB)	98.6%						
HP1452	thiophene and furan oxidizer (tdhF)	37.6%						
<i>Protein and peptide secretion</i>								
HP0355	GTP-binding membrane protein (lepA)	57.3%						
HP0074	lipoprotein signal peptidase (lspA)	97.0%						
HP0786	preprotein translocase subunit (secA)	54.0%						
HP1300	preprotein translocase subunit (secY)	41.2%						
HP1255	protein translocation protein, low temperature (secG)	30.6%						
HP1550	protein-export membrane protein (secD)	38.9%						
HP1549	protein-export membrane protein (secF)	35.1%						
HP0570	signal peptidase I (lepB)	40.3%						
HP1152	signal recognition particle protein (fifh)	41.4%						
HP0795	trigger factor (tfg)	27.6%						
<i>Transformation</i>								
HP0520	cag pathogenicity island protein (cagI)	96.5%						
HP0530	cag pathogenicity island protein (cagI0)	98.4%						
HP0531	cag pathogenicity island protein (cagI1)	97.2%						
HP0532	cag pathogenicity island protein (cagI2)	98.9%						
HP0534	cag pathogenicity island protein (cagI3)	98.0%						
HP0535	cag pathogenicity island protein (cagI4)	97.6%						
HP0536	cag pathogenicity island protein (cagI5)	96.4%						
HP0537	cag pathogenicity island protein (cagI6)	98.9%						
HP0538	cag pathogenicity island protein (cagI7)	95.3%						
HP0539	cag pathogenicity island protein (cagI8)	98.7%						
HP0540	cag pathogenicity island protein (cagI9)	99.5%						
HP0521	cag pathogenicity island protein (cagI2)	92.5%						
HP0541	cag pathogenicity island protein (cagI20)	97.8%						
HP0542	cag pathogenicity island protein (cagI21)	97.9%						
HP0543	cag pathogenicity island protein (cagI22)	95.5%						
HP0544	cag pathogenicity island protein (cagI23)	99.0%						
HP0545	cag pathogenicity island protein (cagI24)	98.5%						
HP0546	cag pathogenicity island protein (cagI25)	95.7%						
HP0547	cag pathogenicity island protein (cagI26)	92.9%						
HP0522	cag pathogenicity island protein (cagI3)	98.1%						
HP0523	cag pathogenicity island protein (cagI4)	95.7%						
HP0524	cag pathogenicity island protein (cagI5)	99.1%						
HP0526	cag pathogenicity island protein (cagI6)	97.5%						
HP0527	cag pathogenicity island protein (cagI7)	94.6%						
HP0528	cag pathogenicity island protein (cagI8)	99.0%						
HP0529	cag pathogenicity island protein (cagI9)	98.9%						
HP1378	competence lipoprotein (comL)	25.8%						
HP1361	competence locus E (comE)	26.7%						
HP1006	conjugal transfer protein (traG)	27.3%						
HP0333	DNA processing chain A (dprA)	30.7%						
HP0334	trbI protein	31.4%						
HP0525	virB11 homologue	100.0%						
HP0441	VirB4 homologue	23.8%						
HP0017	virB4 homologue (virB4)	25.2%						
HP0459	virB4 homologue (virB4)	25.3%						
CENTRAL INTERMEDIARY METABOLISM								
<i>General</i>								
HP1014	7- α -hydroxysteroid dehydrogenase (hdaA)	33.2%						
HP1186	carbonic anhydrase	37.0%						
HP0004	carbonic anhydrase (icfA)	33.3%						
HP0869	hydrogenase expression/formation protein (hypA)	28.1%						
HP0900	hydrogenase expression/formation protein (hypB)	41.4%						
HP0899	hydrogenase expression/formation protein (hypC)	38.5%						
HP0898	hydrogenase expression/formation protein (hypD)	47.8%						
HP0047	hydrogenase expression/formation protein (hypE)	39.7%						
HP0197	S-adenosylmethionine synthetase 2 (metX)	62.1%						
<i>Amino sugars</i>								
HP1532	glucosamine fructose-6-phosphate aminotransferase (isomerase) (glmS)	41.7%						
<i>Phosphorus compounds</i>								
HP0620	inorganic pyrophosphatase (ppa)	50.0%						
HP0636	N-methylhydantoinase	26.9%						
HP1010	polyphosphate kinase (ppk)	38.5%						
<i>Polyamine biosynthesis</i>								
HP0422	arginine decarboxylase (speA) [Escherichia coli]	33.3%						
HP0020	carboxyornithospermidine decarboxylase (nspC)	45.6%						
HP0832	spermidine synthase (speE)	26.5%						
<i>Other</i>								
HP0070	urease accessory protein (ureE)	97.1%						
HP0069	urease accessory protein (ureF)	94.5%						
HP0068	urease accessory protein (ureG)	95.0%						
HP0067	urease accessory protein (ureH)	96.2%						
HP0071	urease accessory protein (ureI)	98.5%						
HP0073	urease alpha subunit (ureA) (ureamidohydrolase)	100.0%						
HP0072	urease beta subunit (ureA amidohydrolase) (ureB)	100.0%						
HP0075	urease protein (ureC)	98.0%						
ENERGY METABOLISM								
<i>Aerobic</i>								
HP1222	D-lactate dehydrogenase (ldh)	27.0%						
HP0961	glycerol-3-phosphate dehydrogenase (NAD(P)H)	36.8%						
HP0037	NADH-ubiquinone oxidoreductase subunit (NQO10)	19.4%						
<i>Cell killing</i>								
HP0887	vacuolating cytotoxin	94.7%						
<i>Chaperones</i>								
HP0010	chaperone and heat shock protein (groEL)	99.6%						
HP0109	chaperone and heat shock protein 70 (dnkA)	63.4%						
HP0210	chaperone and heat shock protein C62.5 (hspG)	46.5%						
HP0011	co-chaperone (groES)	99.2%						
HP1332	co-chaperone and heat-shock protein (dnal)	42.7%						
HP0110	co-chaperone and heat-shock protein (grpE)	33.0%						
HP1024	co-chaperone-curved DNA-binding protein A (CbpA)	37.7%						
<i>Chromosome-associated protein</i>								
HP1138	plasmid replication-partition related protein	40.4%						
<i>Detoxification</i>								
HP1563	alkyl hydroperoxide reductase (tsaA)	98.5%						
HP0875	catalase	99.4%						
HP0267	chlorohydrolase	42.6%						
HP0243	neutrophil activating protein (napA) (bacteriorhodopsin)	95.8%						
HP0389	superoxide dismutase (sodB)	98.6%						
HP1452	thiophene and furan oxidizer (tdhF)	37.6%						
<i>Protein and peptide secretion</i>								
HP0355	GTP-binding membrane protein (lepA)	57.3%						
HP0074	lipoprotein signal peptidase (lspA)	97.0%						
HP0786	preprotein translocase subunit (secA)	54.0%						
HP1300	preprotein translocase subunit (secY)	41.2%						
HP1255	protein translocation protein, low temperature (secG)	30.6%						
HP1560	protein-export membrane protein (secD)	38.9%						
HP0531	cag pathogenicity island protein (cagI1)	97.2%						
HP0532	cag pathogenicity island protein (cagI2)	98.9%						
HP0533	cag pathogenicity island protein (cagI3)	98.0%						
HP0534	cag pathogenicity island protein (cagI4)	97.6%						
HP0535	cag pathogenicity island protein (cagI5)	95.5%						
HP0536	cag pathogenicity island protein (cagI6)	99.1%						
HP0537	cag pathogenicity island protein (cagI7)	97.5%						
HP0538	cag pathogenicity island protein (cagI8)	99.0%						
HP0539	cag pathogenicity island protein (cagI9)	98.9%						
HP0540	cag pathogenicity island protein (cagI10)	25.8%						
HP0541	cag pathogenicity island protein (cagI11)	27.3%						
HP0542	cag pathogenicity island protein (cagI12)	32.9%						
HP0543	cag pathogenicity island protein (cagI13)	32.9%						
HP0544	cag pathogenicity island protein (cagI14)	31.4%						
HP0545	cag pathogenicity island protein (cagI15)	96.3%						
HP0546	cag pathogenicity island protein (cagI16)	94.6%						
HP0547	cag pathogenicity island protein (cagI17)	95.7%						
HP0548	cag pathogenicity island protein (cagI18)	92.9%						
HP0549	cag pathogenicity island protein (cagI19)	98.7%						
HP0550	cag pathogenicity island protein (cagI20)	27.6%						
HP0551	cag pathogenicity island protein (cagI21)	97.9%						
HP0552	cag pathogenicity island protein (cagI22)	95.5%						
HP0553	cag pathogenicity island protein (cagI23)	99.0%						
HP0554	cag pathogenicity island protein (cagI24)	98.5%						
HP0555	cag pathogenicity island protein (cagI25)	97.6%						
HP0556	cag pathogenicity island protein (cagI26)	95.1%						
HP0557	cag pathogenicity island protein (cagI27)	98.9%						
HP0558	cag pathogenicity island protein (cagI28)	95.3%						
HP0559	cag pathogenicity island protein (cagI29)	97.4%						
HP0560	cag pathogenicity island protein (cagI30)	98.0%						
HP0561	cag pathogenicity island protein (cagI31)	98.1%						
HP0562	cag pathogenicity island protein (cagI32)	97.5%						
HP0563	cag pathogenicity island protein (cagI33)	98.1%						
HP0564	cag pathogenicity island protein (cagI34)	97.9%						
HP0565	cag pathogenicity island protein (cagI35)	98.5%						
HP0566	cag pathogenicity island protein (cagI36)	98.9%						
HP0567	cag pathogenicity island protein (cagI37)	98.4%						
HP0568	cag pathogenicity island protein (cagI38)	98.0%						
HP0569	cag pathogenicity island protein (cagI39)	98.6%						
HP0570	cag pathogenicity island protein (cagI40)	98.2%						
HP0571	cag pathogenicity island protein (cagI41)	98.7%						
HP0572	cag pathogenicity island protein (cagI42)	98.3%						
HP0573	cag pathogenicity island protein (cagI43)	98.8%						
HP0574	cag pathogenicity island protein (cagI44)	98.5%						
HP0575	cag pathogenicity island protein (cagI45)	98.0%						
HP0576	cag pathogenicity island protein (cagI46)	98.6%						
HP0577	cag pathogenicity island protein (cagI47)	98.1%						
HP0578	cag pathogenicity island protein (cagI48)	98.7%						
HP0579	cag pathogenicity island protein (cagI49)	98.3%						
HP0580	cag pathogenicity island protein (cagI50)	98.9%						
HP0581	cag pathogenicity island protein (cagI51)	98.5%						
HP0582	cag pathogenicity island protein (cagI52)	98.2%						
HP0583	cag pathogenicity island protein (cagI53)	98.8%						
HP0584	cag pathogenicity island protein (cagI54)	98.4%						
HP0585	cag pathogenicity island protein (cagI55)	98.0%						
HP0586	cag pathogenicity island protein (cagI56)	98.6%						
HP0587	cag pathogenicity island protein (cagI57)	98.2%						
HP0588	cag pathogenicity island protein (cagI58)	98.7%						
HP0589	cag pathogenicity island protein (cagI59)	98.3%						
HP0590	cag pathogenicity island protein (cagI60)	98.9%						
HP0591	cag pathogenicity island protein (cagI61)	98.5%						
HP0592	cag pathogenicity island protein (cagI62)	98.1%						
HP0593	cag pathogenicity island protein (cagI63)	98.7%						
HP0594	cag pathogenicity island protein (cagI64)	98.3%						
HP0595	cag pathogenicity island protein (cagI65)	98.9%						
HP0596	cag pathogenicity island protein (cagI66)	98.5%						
HP0597	cag pathogenicity island protein (cagI67)	98.1%						
HP0598	cag pathogenicity island protein (cagI68)	98.7%						
HP0599	cag pathogenicity island protein (cagI69)	98						

