COMPARATIVE ANALYSIS OF FOUR CAMPYLOBACTERALES

Mark Eppinger[*§], *Claudia Baar*[*§], *Guenter Raddatz*[*], *Daniel H. Huson*[‡] *and Stephan C. Schuster*[*]

Abstract | Comparative genome analysis can be used to identify species-specific genes and gene clusters, and analysis of these genes can give an insight into the mechanisms involved in a specific bacteria–host interaction. Comparative analysis can also provide important information on the genome dynamics and degree of recombination in a particular species. This article describes the comparative genomic analysis of representatives of four different Campylobacterales species — two pathogens of humans, *Helicobacter pylori* and *Campylobacter jejuni*, as well as *Helicobacter hepaticus*, which is associated with liver cancer in rodents and the non-pathogenic commensal species, *Wolinella succinogenes*.

CHEMOLITHOTROPHIC
An organism that is capable of using CO, CO_2 or carbonates as the sole source of carbon for cell biosynthesis, and that derives energy from the oxidation of reduced inorganic or organic compounds.

CHEMOORGANOTROPHIC
An organism that derives energy from organic sources in a light-independent manner.

*Max-Planck-Institute for Developmental Biology, Genome Centre, Spemannstr. 35, 72076 Tübingen, Germany. ‡ZBIT Zentrum für Bioinformatik, Tübingen University, 72076 Tübingen, Germany. §These authors contributed equally to this work. Correspondence to S.C.S. e-mail: stephan.schuster@tuebingen.mpg.de
doi:10.1038/nrmicro1024

The ε-subdivision of the Proteobacteria is a large group of CHEMOLITHOTROPHIC and CHEMOORGANOTROPHIC microorganisms with diverse metabolic capabilities that colonize a broad spectrum of ecological habitats. Representatives of the ε-subgroup can be found in extreme marine and terrestrial environments ranging from oceanic hydrothermal vents to sulphidic cave springs. Although some members are free-living, others can only persist in strict association with a host organism. The host-associated members of this group include both commensal and pathogenic species. As most of the environmental ε-proteobacteria are not cultivable, their physiology remains obscure[1–3].

Within the ε-proteobacteria, the order Campylobacterales comprises two families, the Helicobacteraceae and the Campylobacteraceae. There are presently five complete genome sequences available for this order: four from the Helicobacteraceae — the genomes of two distinct strains of *Helicobacter pylori*, 26695 and J99, *Wolinella succinogenes* DSM 1740 and *Helicobacter hepaticus* ATCC 51449 — and one from the Campylobacteraceae, *Campylobacter jejuni* NCTC 11168 (REFS 4–8). In this article, we focus on these ε-proteobacterial representatives.

Both of these *Helicobacter* species, the gastric colonizer *H. pylori* and the enterohepatic species *H. hepaticus*, are classified as cancer-causing microorganisms as infection can lead to gastric cancer in humans and liver cancer in rodents, respectively[9–11]. The *Campylobacter* representative *C. jejuni* is one of the main causes of bacterial food-borne illness worldwide, causing acute gastroenteritis, and is also the most common microbial antecedent of Guillain–Barré syndrome[12–15]. Besides their pathogenic potential in humans, *C. jejuni* persists as a commensal in its avian host reservoir and *H. pylori* can persist in the human stomach asymptomatically. The genus *Wolinella* is represented by *W. succinogenes*, which persists as a commensal in the gastrointestinal tract of cattle, where it utilizes metabolites produced in the rumen[16].

The availability of complete genome sequences for these microorganisms allows a comparative analysis of the genome content and architecture[4–8]. Our current understanding of bacterial pathogenicity mechanisms has been greatly advanced by our increased knowledge of the genome inventory that is common to bacterial pathogens. As many pathogenic microorganisms are strictly host-adapted, their genomes have undergone a process known as reductive evolution, leading to reduced genome sizes[17,18]. This process of deleting genetic information has resulted in 'orphaned' cellular processes that can only be understood in their ancestral context. To truly understand

the origins and emergence of a pathogen, it is essential to analyse the genomes of any living relatives that have largely maintained the gene pool of the last common ancestor. For the ε-proteobacteria group, *W. succinogenes* is such an organism as it encodes complete metabolic pathways and environmental-sensing systems that are absent from the genomes of its pathogenic relatives[8]. On the basis of phylogenetic comparison of 16S rRNA sequences, *W. succinogenes* seems to be closely related to the three other species considered here[19,20]. As all four species belong to the same subclass of the Proteobacteria, they share many analogous cell structures and biological features. By comparing the individual gene content and identifying both shared and species-specific genes, it is possible to describe specific similarities and differences in the lifestyle of these organisms, as well as common and divergent mechanisms for host adaptation. The genes that are present in all species are likely to be essential for maintenance of the microorganism in a mammalian or vertebrate host, thereby providing a general strategy for survival and growth as well as providing common mechanisms for spread and transmission into larger host populations[21]. However, although identification of the genes that are unique to a species provides an excellent starting point for functional analysis, a definitive role for these gene products with respect to their physiological function in a symbiotic, commensal or pathogenic interaction with the host cannot be deduced by analysis alone.

The availability of these ε-proteobacterial genomes also allows us to study any genome rearrangements that have taken place since these microorganisms diverged from their last common ancestor. These ongoing dynamic processes and the level of genome plasticity can be investigated on an intraspecies level — between the two *H. pylori* strains[22,23] — as well as on an interspecies level. In this article, we use comparative genomic analysis to identify factors that define the host specificity of each microorganism, as well as discussing the limitations of this comparative approach.

General genome features of the ε-proteobacteria

The *W. succinogenes* genome consists of one 2.11-Mb circular chromosome and is larger than the genomes of *H. hepaticus* (1.80 Mb), *H. pylori* 26695 (1.67 Mb), *H. pylori* J99 (1.64 Mb)and *C. jejuni* (1.64 Mb) (TABLE 1). This increased genome size corresponds with a greater number of predicted genes (2,046) compared with the revised annotations of the *H. pylori* 26695 genome (1,552), and the *H. pylori* J99 (1495), *H. hepaticus* (1,875) and *C. jejuni* (1,654) genomes[24]. The genomes of *W. succinogenes*, *C. jejuni* and *H. hepaticus* have a high gene density, with a coding area of 94.3% for *C. jejuni*, 94.0% for *W. succinogenes* and 93.0% for *H. hepaticus*, whereas only 91.0% of the *H. pylori* 26695 and 90.8% of the *H. pylori* J99 genomes are predicted to be transcribed. The overall GC content of the *W. succinogenes* genome (48.5%) is greater than that of *C. jejuni* (30.6%), *H. hepaticus* (35.9%) and *H. pylori* 26695 and J99 (39.0%).

Table 1 | General features of sequenced Helicobacteraceae and Campylobacteraceae genomes

Species	*Wolinella succinogenes*	*Helicobacter pylori*		*Helicobacter hepaticus*	*Campylobacter jejuni*
Strain	DSM 1740	26695	J99	ATCC 51449	NCTC 11168
Host(s)	Bovine	Human	Human	Rodent	Human & avian
Genome size (bp)	2,110,355	1,667,867	1,643,831	1,799,146	1,641,481
G+C content (%)	48.5	39.0	39.0	35.9	30.6
Open-reading frames					
Predicted number	2,046	1,590 (1,552*)	1,495	1,875	1,654
Coding area (%)	93.0	91.0	90.8	93.0	94.3
Average length (bp)	964	945	998	1,082	948
Flexible genome pool					
Reported plasmids	None	None‡	None‡	None	None§
Phages and phage-like elements	Prophage-tRNA^{met}	None	None	3 phage genes	None
IS elements	ISWsu*1302*, ISWsu*1203*	IS*605*, IS*606*	IS*605* (partial), IS*606*	None	Cj0752 (partial)
Genomic islands	WsuGI I and II	*cag* PAI	*cag* PAI	HHGI1	None
Regions of deviating GC content	Genomic islets & islands, DNA-restriction/ modification system, translation machinery	Genomic islets & islands, DNA-restriction/ modification system, translation machinery	Genomic islets & islands, DNA-restriction/ modification system, translation machinery	Genomic islets & islands, DNA-restriction/ modification system, translation machinery	EPS/LOS synthesis & flagella modification
References	8	4	5	7	6

*Revised number of predicted *Helicobacter pylori* 26695 ORFs[24]. ¶In *W. succinogenes*. ‡Eight plasmids (pHP51, pHP489, pHPM8, pHP180, pHP186, pHel4, pHe5 and pHPO100) have been detected in a variety of other *H. pylori* strains. §Three plasmids (pVir, pCJ419 and pTet) have been sequenced in other *C. jejuni* strains. *cag* PAI, cytotoxin-associated pathogenicity-associated island; EPS, extracellular polysaccharide biosynthesis; GI, genomic island; IS, insertion sequence; LOS, lipooligosaccharide biosynthesis.

Although all members of the ε-proteobacteria generally have small genomes, a feature that has been observed as an adaptive trait for various other host-adapted bacteria, the commensal *W. succinogenes* has a markedly larger genome compared with its close pathogenic relatives. Together with the observation of an asymmetric GC-skew in *W. succinogenes* (BOX 1), this indicates that all four genomes are derived from a larger ancestral genome.

Phylogeny and gene-content analysis

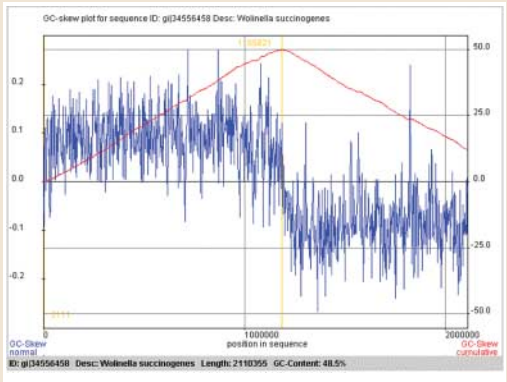
The close relatedness of these four species has been demonstrated using morphological, physiological and molecular classification methods, and this is reflected in their gene content as they share ~50% of their genes[19,20,25] (FIG. 1). A high level of functional conservation has been maintained, as many ORTHOLOGUES with a high degree of identity were detected in the predicted protein sequences of these organisms. For genes shared between *W. succinogenes* and *C. jejuni*, the identities range from 22% to 91%, with an average of 48.4%. Orthologue identities between *W. succinogenes* and *H. pylori* 26695 range from 21% to 91% (average 50.8%) and from 19% to 90% (average 45.9%) for *H. pylori* 26695 and *C. jejuni*. The orthologue identities for the two *Helicobacter* species range from 21% to 90% (average 51.2%), for *W. succinogenes* and *H. hepaticus* from 19% to 90% (average 47%) and 22% to 96% (average 47%) for *C. jejuni* and *H. hepaticus*. These interspecies comparisons show identity ranges that are similar to those reported for *Chlamydia pneumoniae* and *Chlamydia trachomatis* (22–95% with an average of 62%, based on 859 open reading frames, ORFs)[26]. This indicates that the four compared organisms, despite being separate species, share a pool of homologous gene products. As expected, the proteins with the greatest sequence identity are associated with translation and important metabolic functions, whereas those with the least identity are associated with uncharacterized or hypothetical proteins.

The flexible genome pool

Of these four ε-proteobacterial species, only *H. pylori* and *W. succinogenes* share multiple sets of flexible genomic elements — such as insertion sequence (IS) elements — that enhance intragenomic recombination and therefore contribute to the dynamics and diversity of a genome and have a role in speciation[27]. Genome analysis of the

ORTHOLOGUES
Homologous genes that originated through speciation.

Box 1 | GC-skew in microbial genomes



In most bacterial genomes, a difference in base composition between the leading and lagging strand is observed — usually, the leading strand is enriched in G and T, whereas the lagging strand is enriched in A and C. Deviations from the base frequencies of A=T and G=C are called AT- and GC-skews. The GC-skew is usually stronger than the AT-skew, so frequently only the GC-skew is considered. A method of measuring GC-skew is to plot (G–C)/(G+C) in a sliding window along the sequence. This gives the percentage of excess of G over C — a value that is positive in the leading strand and negative in the lagging strand.

What causes GC-skew? The reasons are still poorly understood. The leading and the lagging strands are probably subject to different mutational pressure owing to differences in the time spent as single-stranded DNA during replication and thereby different exposure to DNA-damaging conditions. As T–G and G–T mispairs are dominant over complementary C–A and A–C pairs, the more error-prone strand would be relatively richer in G and T. Another theory relies on hydrolytic deamination of cytosine, which occurs predominantly in single-stranded DNA. The asymmetric structure of the replication fork leaves the lagging-strand template temporarily single-stranded, making it more susceptible to cytosine deamination. Deamination of cytosine leads to the formation of uracil, which pairs with adenine during replication causing a C to T mutation. Consequently, C to T deamination would increase the percentage of G and T in that strand and the percentage of C and A in the complementary strand.

Why is a GC-skew analysis so important? As GC-skew is positive in the leading strand and negative in the lagging strand, the GC-skew changes sign at the origin and terminus where the leading strand becomes the lagging strand and vice versa. This makes GC-skew analysis a useful tool to identify the origin and the terminus in circular chromosomes. Local changes, which are visible as diagram distortions, can mark recent rearrangements such as sequence inversions or integration of foreign DNA. The loss of DNA would not change the basic shape of the GC-skew curve, whereas recent incorporation of external DNA would probably result in a local deviation.

In practice, the visualization of the GC-skew can suffer from local fluctuations. So it is better to work with the cumulative GC-skew, which sums up the values of neighbouring sliding windows from an arbitrary start to a given point in the sequence. The figure shows the GC-skew and cumulative GC-skew of the *Wolinella succinogenes* DSM1740 genome and illustrates how the GC-skew changes sign at the origin and the terminus of replication. The cumulative GC-skew shows a minimum and maximum at these positions, respectively.

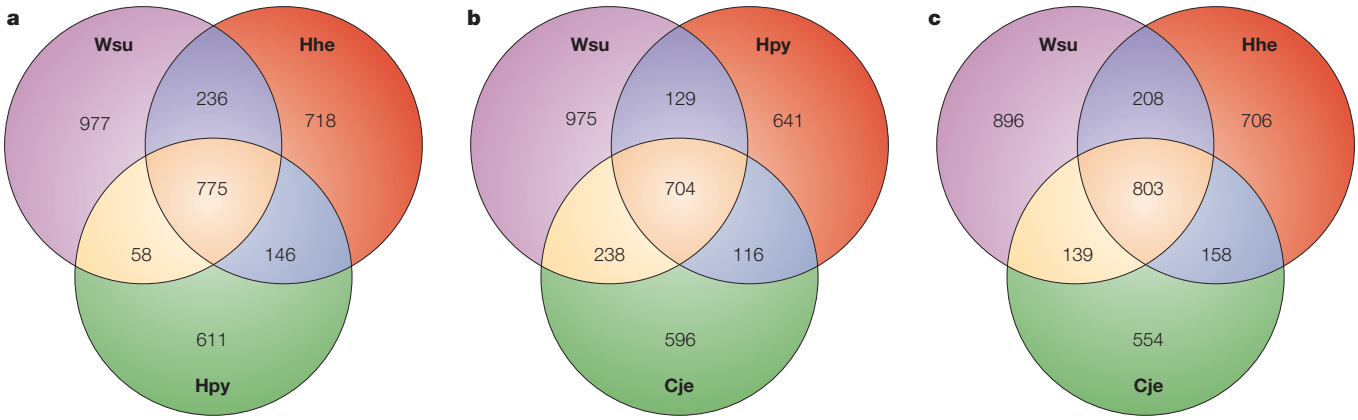


Figure 1 | **Comparative gene content analysis of sequenced ε-proteobacteria.** Comparing the gene content of *Wolinella succinogenes* DSM 1740 (Wsu) against *Helicobacter pylori* strain 26695 (Hpy), *Helicobacter hepaticus* ATCC 51449 (Hhe) and *Campylobacter jejuni* NCTC 11168 (Cje). The Venn diagrams show the number of shared and species-specific open reading frames (ORFs) among the four genomes.

two *H. pylori* strains revealed the presence of two types of IS element, IS*605* and IS*606*. Although *H. pylori* 26695 harbours five complete copies of IS*605* and two complete copies of IS*606*, the genome of the second sequenced *H. pylori* strain, J99, only encodes one complete copy of IS*606* and no complete IS*605* element[4,5]. These *H. pylori*-specific IS elements are not only found integrated into the core chromosome, but are also encoded on plasmids as part of the FLEXIBLE GENOME POOL in some *H. pylori* strains[28]. The *W. succinogenes* chromosome harbours two types of IS elements, multiple copies of IS*1302* and a single complete copy of ISWsu*1203*, as well as disrupted partial copies[8,29]. By contrast, the genomes of *H. hepaticus* and *C. jejuni* do not contain any detectable full-length IS elements, although *C. jejuni* encodes one ORF with partial similarity to an IS*605* transposase of *H. pylori*[5], which might be indicative of the presence of IS elements in other *Campylobacter* isolates. All IS elements found in these ε-proteobacterial species belong to the IS*3* family of IS elements, which are characterized by the highly conserved DDE motif, which is also found in retroviral integrases[30].

The lack of IS elements and the absence of repetitive DNA sequences, except for ribosomal RNA genes and three other examples of duplicated genes, are characteristic features of the *C. jejuni* genome, even though this is not reflected in a reduced amount of genetic recombination[6]. The lack of IS elements might be compensated for by the abundance of LOW-COMPLEXITY ZONES and simple sequence repeats that function as potential target sites for recombination[31,32]. In the genomes of *W. succinogenes* and *H. pylori*, the IS elements flank regions that are characterized by a GC content that is different to the rest of the genome, indicating that these regions were acquired by the incorporation of foreign DNA into the core genome by lateral gene transfer[33,34]. This observation is supported by the co-localization of mobility genes, such as transposases and integrases with homology to bacteriophage genes and transfer RNA loci[35]. The three regions of deviating GC content in the *C. jejuni* core genome encode gene products that are involved in extracellular polysaccharide (EPS) and lipooligosaccharide (LOS) biosynthesis and flagellar modification. No genomic islands can be observed in the *C. jejuni* genome; this is in contrast to the genomic islets and islands that have been reported for the three related Helicobacteraceae genomes. So far, no plasmids have been found in *W. succinogenes* and *H. hepaticus*, although several plasmids have been reported for *H. pylori* and *C. jejuni*[28,36,37]. The presence of the WsuGI genomic island in the *W. succinogenes* genome might indicate that this genomic region once existed as an autonomous plasmid, as has been observed for the SYNTENIC virulence plasmid pVir of *C. jejuni*[36]. These findings indicate the present status of the flexible gene pool of these organisms and show that the ε-proteobacteria, in principle, have the ability to harbour plasmids; their occurrence and distribution could be dependent on the particular isolate examined.

Functional classification of the genomic inventory

To compare the gene content of all four genomes by a standardized method, the predicted ORFs were classified on the basis of the cluster of orthologous groups (COG) database[38]. When the distribution of the COG categories is compared between the four species, areas in which each of the microorganisms specializes become apparent (FIG. 2).

The *W. succinogenes* genome contains fewer genes in the categories of translation and cell-envelope components, which have important roles in bacteria–host interactions, than the other three compared organisms. This finding parallels the observation that the genomes of the two *Helicobacter* species and *C. jejuni* contain many genes that have been annotated as outer membrane proteins, albeit belonging to different protein families.

Another important feature of the *H. pylori* 22695 genome is the high percentage of genes that are predicted to be involved in DNA replication and modification[39].

FLEXIBLE GENOME
A bacterial chromosome represents a mosaic structure composed of ancestral DNA, the core genome, and the horizontally acquired flexible genome pool.

LOW-COMPLEXITY ZONES
Genomic regions with a high level of repetitiveness of distinct nucleotides.

SYNTENIC
A genomic region found in two organisms with a colinear order of genes (or nucleotides). Syntenic regions are found in chromosomal or plasmid-encoded replicons.

PARALOGOUS
Homologous genes that originated by gene duplication

TWO-COMPONENT SIGNAL TRANSDUCTION SYSTEMS
A signal-transduction system using two components — a histidine protein kinase (HPK) and a response regulator (RR) — to sense and respond to external stimuli. The HPK autophosphorylates at a histidyl residue following stimulation and transfers that phosphoryl group to a cognate RR at its aspartyl residue to induce a conformational change in the regulatory domain, which in turn activates an associated domain.

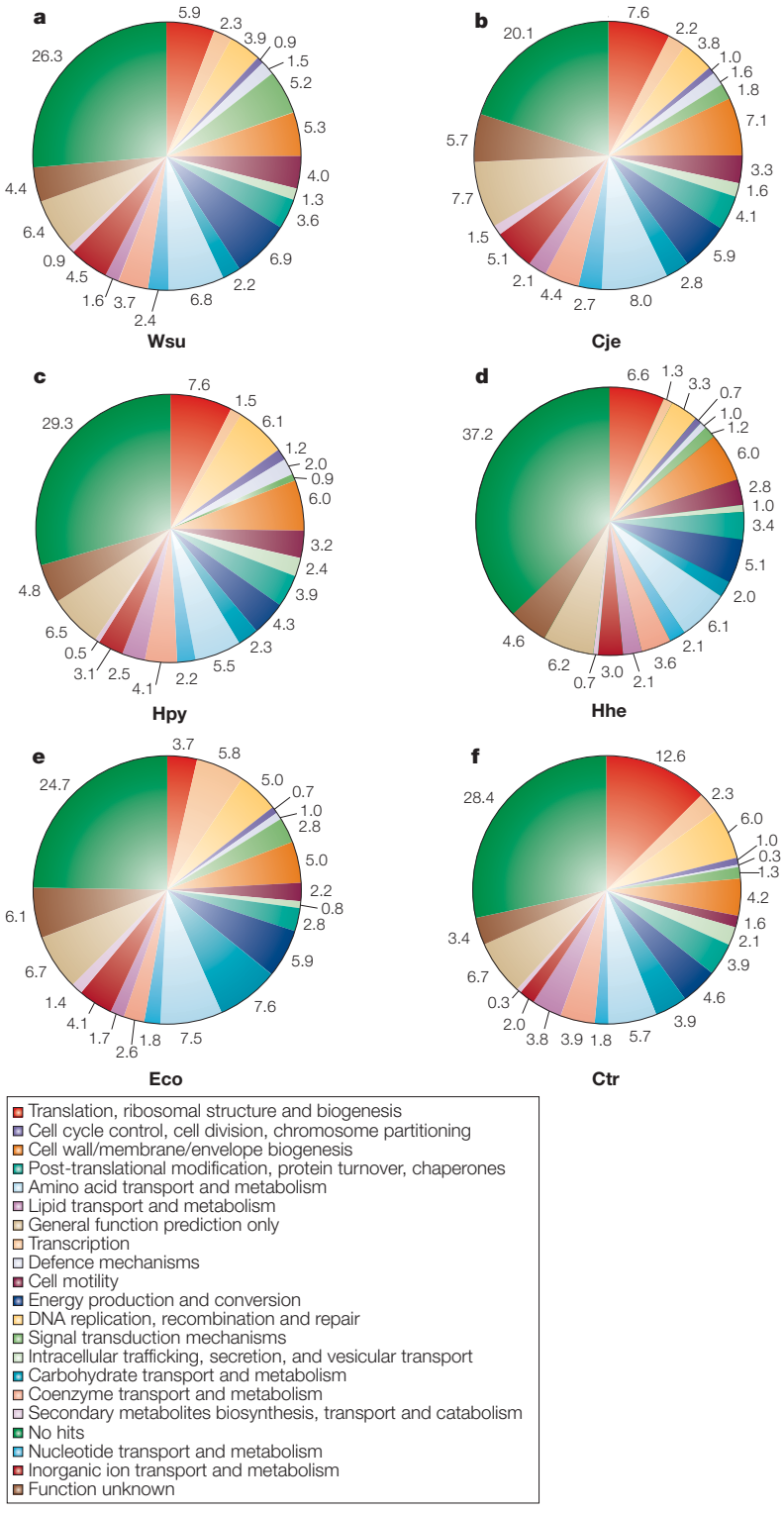


Figure 2 | **Functional classification of the genome inventory of sequenced ε-proteobacteria.** All predicted protein-coding genes of the four organisms (*Wolinella succinogenes* DSM 1740 (Wsu) (**a**); *Campylobacter jejuni* NCTC 11168 (Cje) (**b**); *Helicobacter pylori* strain 26695 (Hpy) (**c**); and *Helicobacter hepaticus* ATCC 51449 (Hhe) (**d**)) are classified and coloured according to the 26 categories of the cluster of orthologous groups (COG) database. Values given represent percentage of genome content. Proteins with no hit <e-15 were assigned to the 'No hits' category. Proteins that match COG database entries that are not assigned to a defined COG category are also grouped into the 'No hits' category. About 20–35% of all genes in each genome could not be classified in any functional COG category. Parts **e** and **f** show *Escherichia coli* K12 MG-1655 and *Chlamydia trachomatis* D/UW-3/CX for comparison.

As *H. pylori* is naturally competent, the genome encodes many DNA restriction and methylation systems — consisting of many pseudogenes as well as active genes — to degrade foreign DNA[4,5]. The large number of pseudogenes present in the *H. pylori* genome compared with those of the other species might be indicative of a genome in the process of decay, as has been observed for many bacterial pathogens and symbionts[40]. By contrast, both *W. succinogenes* and *C. jejuni* have few pseudogenes. Also, only these two organisms have a large number of genes that have been categorized as important for energy metabolism and ion transport, which is indicative of an organism that is not restricted to a distinct ecological niche in a host.

The number of signal-transduction proteins in *C. jejuni*, *H. pylori* and *H. hepaticus* is surprisingly small compared with the number in *W. succinogenes*. This is due to the presence of a large number of PARALOGOUS signalling genes, histidine kinases and their corresponding regulators of TWO-COMPONENT SIGNAL-TRANSDUCTION SYSTEMS, as well as numerous methyl-accepting chemotaxis proteins in the *W. succinogenes* genome[41,42] (TABLE 2). Limited environmental-sensing capabilities and reduced sets of metabolic pathways have been suggested to be characteristic of host-adapted pathogens because specialized adaptation to a particular host can result in a small genome[17,18,43]. In accordance with this, the non-pathogenic *W. succinogenes* has the largest set of transcriptional regulators of the four compared microorganisms, with an additional copy of the sigma factor σ^{54} and two copies of σ^{24} that are not present in the three other genomes. Interestingly, no orthologues of *cheB*/*cheR* can be detected in the genomes of *W. succinogenes* and *H. pylori* (TABLE 2). As the function of these methylases and demethylases has not been proven experimentally in *H. hepaticus* and *C. jejuni*, the omission of these genes in *H. pylori* and *W. succinogenes* could be suggestive of a gene function that has also been targeted for deletion in other ε-proteobacteria. It therefore remains to be shown how this class of bacteria confers an adaptive response to a given stimulus[44]. Although the COG classification is extremely useful, it is expected that the largest advance in our understanding of the host specificity of these microorganisms will come from analysing the 20–35% of predicted genes that are specific to each species and show no homology to sequences already deposited in the databases.

Common and unique features

Despite their different lifestyles, all four microorganisms share a large number of genes that have been identified as virulence factors in the two *Helicobacter* species and *C. jejuni*. These include genes that encode haemolysins and related proteins, adhesion factors, proteins involved in type IV secretion, invasins, antigenicity factors and toxins[10,45,46] (TABLE 2).

One of the most important differences between the four species is the urease activity of *H. pylori* and *H. hepaticus*. The physiological function of urease activity is the conversion of urea into ammonia and carbon dioxide. In *H. pylori*, the ammonia that is produced

Table 2 | **Common and unique features of the sequenced Campylobacterales genomes**

Species	***W. succinogenes***	***H. pylori***		***H. hepaticus***	***C. jejuni***
Strain	DSM 1740	26695	J99	ATCC 51449	NCTC 11168
Virulence factors & toxins					
Urease	*ure**	Urease operon	Urease operon	Urease operon	None
Cytolethal-distending toxin	None	None	None	*cdtABC*	*cdtABC*
C. jejuni invasion antigen B	*ciaB*	None	None	*ciaB*	*ciaB*
Vacuolating cytotoxins	*vacB*	*vacA*‡, 3 *vacA* paralogues, *vacB*	*vacA*‡, 3 vacA paralogues, *vacB*	*vacB*	*vacB*
Cytotoxin-associated genes	None	*cagA*, *cagE*§	*cagA*, *cagE*§	None	None
Neutrophil-activation protein	*nap*	*nap*	*nap*	*dps*	Bacterioferritin
Adhesion & antigenicity factors					
Hop and hor proteins	None	33 genes	32 genes	10 genes	None
Flagellar sheath adhesin	None	*hpaA*	*hpaA*	None	None
Major outer membrane protein	2 orthologues	None	None	None	*momP*
Fibronectin-binding protein	*cadF*	None	None	None	*cadF*
Fibronectin- & fibrinogen-binding protein	Yes	Yes	Yes	Yes	Yes
Major cell binding factor	None	None	None	HH1481	*peb1*, *peb2*/*peb3*
Adhesive lipoprotein	None	None	None	None	*jlpA*
Immunogenic protein	Yes	None	None	None	None
Motility					
Flagellum	Yes	Yes	Yes	Yes	Yes
Contingency genes	Yes	Yes	Yes	Yes	Yes
Carbohydrate-active enzymes**					
(Trans-)glycosydases	4 genes	2 genes	2 genes	4 genes	3 genes
Glycosyltransferases	20 genes	20 genes	22 genes	23 genes	25 genes
Carbohydrate esterases	2 genes	2 genes	2 genes	3 genes	1 gene
Glycosylation system					
General *N*-linked glycosylation	15 genes‖	None	None	None	13 genes‖
LOS locus	None	None	None	None	Yes¶
Lipoproteins##	18 genes	14 genes	16 genes	19 genes	22 genes
Transport and metabolism					
Nickel transport#	*nikA*,*A2*,*B*,*D*,*E*	*nixA*, *hpn*	*nixA*	*nikABCDE*$	None
Iron metabolism (Fe^{2+}, Fe^{3+})	*fur*, *tonB/exbBD*, *feoAB*, *fec* type, *frpB*	*fur*, *tonB/exbBD*, *feoB*, *fec* type, *frpB*, *ceuE*	*fur*, *tonB/exbBD*, *feoB*, *fec* type, *frpB*, *ceuE*	*fur*, *tonB/exbBD*, *feoB*, *frpB*, *cfrA*	*fur*, *tonB/exbBD*, *feoAB*, *chuABCD*, *ceuBCDE*, *cfrA*
Multidrug efflux transporter	*cmeABC*	*cme(A)B*	*cmeB*	*cmeAB*	*cmeABC*+
N_2-fixation locus	Yes	None	None	None	None
Sensing and regulation					
Two-component signal transduction system	39 HK, 52 RR	4 HK, 6 RR	4 HK, 6 RR	9 HK, 6 RR	7 HK, 12 RR
Transcription factors and regulators	σ^{70}, σ^{54} (2), σ^{28} σ^{24}(2), *flgMR*	σ^{70}, σ^{54}, σ^{28}, *flgMR*	σ^{70}, σ^{54}, σ^{28}, *flgMR*	σ^{70}, σ^{54}, σ^{28}, *flgMR*	σ^{70}, σ^{54}, σ^{28}, *flgMR*
Methyl-accepting chemotaxis	31 genes	3 genes	4 genes	8 genes + *cheB/cheR*	10 genes + *cheB/cheR*
Restriction and modification~					
R/M system(s)	Type I (1), type II (3)	Type I (5), type II (18), type III (4)	Type I (4), type II (18), type III (4)	Type I (1), type II (7)	Type I (1), type II (4)
Secretion machinery					
Type IV secretion system	WsuGI	*cag* PAI	*cagPAI*	HHGI1	pVir°
Flagellar export system	Yes	Yes	Yes	Yes	Yes

**W. succinogenes* harbours a single urease accessory gene. ‡Reported virulent *H. pylori* strains lacking *vacA*. §Part of *H. pylori* cag PAI. **Numbers according to CAZy (see Online links.) ‖*N*-linked glycosylation cluster arranged syntenically between *C. jejuni* and *W. succinogenes*. ¶Genetic basis for molecular mimicry in *C. jejuni*. ##Numbers according to DOLOP (see Online links). #Used for urease production. $Next to urease operon. +Mediates antibiotic resistance of *C. jejuni*. ~Numbers according to the REBASE (see Online links). °Type IV secretion system arranged syntenically between *W. succinogenes* and *C. jejuni* pVir. DPS starvation-inducible DNA-binding protein, HK, histidine kinase, LOS, lipooligosaccharide; RR, response regulator.

buffers the acidic environment of the human stomach, while increasing the pH level. Both *C. jejuni* and *W. succinogenes* lack this activity, although *W. succinogenes* has retained a single orthologue of a urease accessory protein. *H. pylori* and *H. hepaticus* use two different systems for nickel uptake and transport, which are required for optimum urease activity. *C. jejuni* has no nickel transport system. Although *W. succinogenes* has no urease activity, it has a nickel transport system that is orthologous to the system that is found in *H. hepaticus*. The location of the nickel operon next to the urease cluster on the *H. hepaticus* chromosome reflects the composition of the urease enzyme as a metalloprotein with two bound nickel ions. In *W. succinogenes*, nickel uptake is a prerequisite for [NiFe] hydrogenase maturation, whereas in the related pathogenic *Helicobacter* species, nickel is required to form the catalytic centre of the urease complex.

In all four microorganisms, iron metabolism is an elaborate process, and involves up to five iron-uptake systems, which demonstrates the importance of an extended iron metabolism for this class of bacteria. Due to their strict association with mammalian hosts, iron acquisition is a major concern as most of the iron in a mammal is complexed with various proteins[47,48]. The ability of pathogens to obtain iron from transferrins, ferritin, haemoglobin and other iron-containing host proteins is essential for persistence in the mammalian host environment[49].

The *cmeABC* locus encodes a multidrug efflux pump, which contributes to the resistance of *C. jejuni* to a broad range of antimicrobials as well as to the colonization of host tissue[50]. Orthologous genes can be detected in the related Helicobacteraceae, in particular in the *W. succinogenes* genome, which carries a *cmeABC* cluster that resembles the *C. jejuni* locus (TABLE 2). As in the related *Helicobacter* species *H. pylori* and *H. hepaticus* this gene cluster is incomplete, the reported resistance of *H. pylori* strains to diverse antibiotics might not be mediated via this multidrug efflux transport system and instead might be mediated by unrelated ABC-type multidrug transport systems.

A key pathogenicity factor of *C. jejuni*, the invasion antigen CiaB, is essential for host-cell invasion[51]. An orthologue has been described in *W. succinogenes*, and a gene with high homology is found in the genome of *H. hepaticus*[8]. However, a putative physiological role for these two gene products in the host interaction process has still to be determined. In *H. pylori*, this gene seems to have been selectively lost from a shared ancestral gene pool, and it is therefore not required for successful interaction with the host. A similar pattern is found for the cytolethal-distending toxins (CdtABC), which are found in *C. jejuni* and *H. hepaticus*, but not in *W. succinogenes* and *H. pylori*. These cytotoxins cause host cells to arrest during their cell cycle by inflicting DNA damage[52]. This suggests that the *cdt* genes might be part of the flexible gene pool as they have been identified in several Gram-negative pathogenic bacteria and, like the CiaB invasion antigen, they are not generally required for host interaction[53]. Virulence factors that are exclusively found in *H. pylori* include the cytotoxin-associated protein CagA, which is secreted via the cytotoxin-associated gene (*cag*) pathogenicity-associated island (PAI)-encoded type IV secretion machinery, and disrupts multiple signalling pathways in the host cell. Another example is the vacuolating cytotoxin autotransporter VacA and three further paralogues[54,55]. Recently an alternative contact-dependent mechanism for the transfer of VacA into host cells was discovered[56].

Secretion systems. Secretion systems such as the type III and type IV secretion systems have a crucial role in bacterial virulence[35,57]. No type III system other than the flagellar export apparatus is observed in the compared microorganisms. The two *Helicobacter* species and *W. succinogenes* all carry numerous orthologues of virulence (*vir*) genes in their core genome, which are likely to assemble the type IV secretion machinery. These type IV systems are encoded on genomic regions categorized as genomic islands, namely *W. succinogenes* WsuGI I, *H. hepaticus* HHGI1 and *H. pylori cag* PAI. The deviating GC content of these regions, as well as their co-localization with mobility loci such as IS elements and transposases, is indicative of acquisition via horizontal gene transfer. In *C. jejuni*, the type IV secretion system is part of the flexible genome pool and is encoded on the *C. jejuni* virulence plasmid pVir[36,58]. The genetic variability and diversity that has been reported for many *Campylobacter* strains is therefore supported by the presence of several plasmids, the repertoire of contingency genes and the exchange of genetic material due to natural transformation[59,60].

Remarkably, the comparative analysis reveals an almost undisturbed syntenic arrangement of the type IV system in the *W. succinogenes* genomic island WsuGI I and the *C. jejuni* virulence plasmid pVir[8,36]. By contrast, the genomic islands of the two *Helicobacter* species do not have a similar gene order, indicating a different origin for the secretion systems of these species. The type IV secretion machinery that is found in these ε-proteobacteria species therefore seems to have been acquired independently from different sources[57].

Although bacterial secretion pathways can be expressed in non-pathogenic and pathogenic organisms, their usage might differ according to their ecological habitat. For example, in pathogenic *H. pylori* strains these virulence factors contribute to the pathogenicity of the organism, whereas in *W. succinogenes* the same factors might be essential for persistence in its ecological niche as a rumen symbiont or commensal. In general, the use of the term virulence factor is based on the initial identification of these virulence determinants in the context of pathogenic organisms, and does not take into account the presence and different physiological functions of these factors in non-pathogenic organisms. Therefore, these factors might be better described as 'host-interaction factors'. However, it could take just a few recombination or mutational events to cross the border between commensalism and pathogenicity[40,61].

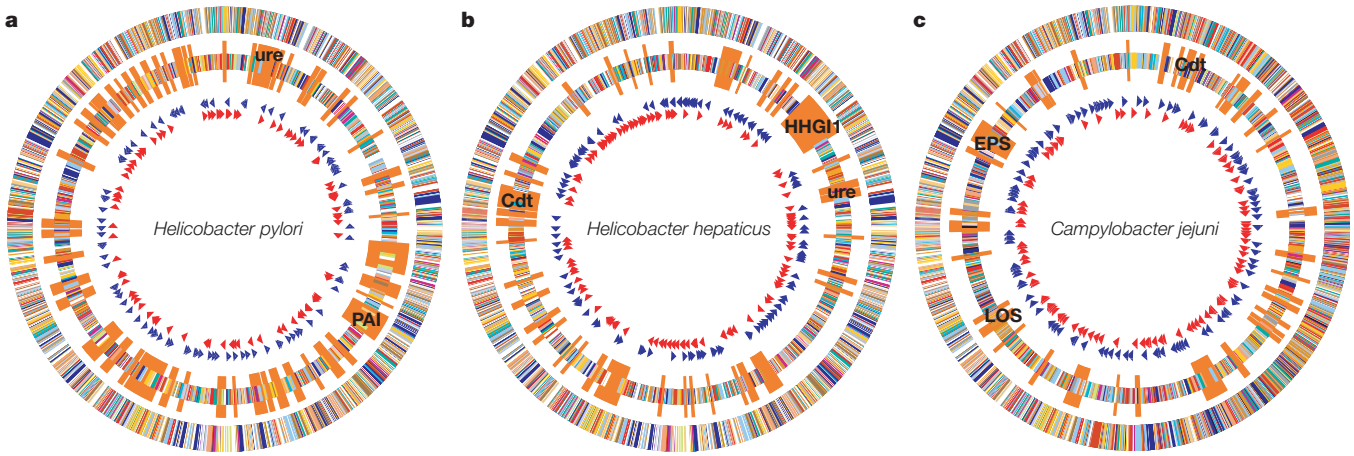


Figure 3 | **Species-specific and syntenic regions of sequenced ε-proteobacteria.** Comparing the gene content of *Wolinella succinogenes* DSM 1740 with *Helicobacter pylori* strain 26695 (**a**), *Helicobacter hepaticus* ATCC 51449 (**b**) and *Campylobacter jejuni* strain NCTC 11168 (**c**). The corresponding orthologous genes are shown in the middle circle, their order is given with respect to their original location in the genome. Orange blocks highlight species-specific clusters (SSC) spanning more than four open reading frames (ORFs) without orthologues in *W. succinogenes*. The genomic islands of *H. pylori* (*cag* PAI) and *H. hepaticus* (HHGI1) are clearly detectable. The two innermost circles reveal syntenic regions between *W. succinogenes* and the three other microorganisms. The order and orientation is given by the compared organism; syntenic regions of the same orientation are labelled with red arrowheads, whereas inverted regions are shown by blue arrowheads. The other colours match those in FIG. 2. Cdt, cytolethal-distending toxins, EPS, exopolysaccharide biosynthesis, HHGI 1, *H. hepaticus* genomic island 1, LOS, lipooligosaccharide biosynthesis, PAI, pathogenicity-associated island, *ure*, urease operon.

Phase-variable contingency genes. Contingency genes are characterized by intragenic repetitive sequences that mediate antigenic variation. These genes are prone to replication slippage and mispairing, which affects the expression level and can lead to the on/off-switching of a gene, known as phase variation[62,63]. The potential phase-variable genes containing these characteristic intragenic motifs code for proteins that are intimately involved in microorganism–host interactions for the compared ε-proteobacterial species (see online supplementary material S1 (table)). Phase-variable genes provide these organisms with the capability to alter their host-recognizable epitopes and adhesins. For *C. jejuni* it has been shown experimentally that the encoded phase-variable genes are involved in the biosynthesis of the capsule and modification of the lipopolysaccharide components[64]. Certain *H. pylori* strains mimic human LEWIS^{X,Y} ANTIGENS using fucosyltransferases, which have a role in adhesion to the gastric host tissue[15,65,66].

Expanded HOMOPOLYMERIC TRACTS are important features of the genomes of the two *Helicobacter* species , particularly *H. hepaticus*, as well as *C. jejuni*[5,67]. The obvious lack of these motifs in *W. succinogenes* is compensated by an abundance of genes with repetitive dinucleotide sequences. The gene products of these putative phase-variable genes in *W. succinogenes* function in identical molecular pathways to the known phase-variable proteins of *C. jejuni* and the two *Helicobacter* species. It therefore seems likely that antigenic variation is a common strategy used by these microorganisms for adaptation, specialization and persistence in hepatic or gastrointestinal niches within their distinct hosts, such as the human stomach, the rodent liver or the bovine rumen.

Species-specific and syntenic regions. The high genomic variability of the individual species has contributed to the distinct physiological differences that were originally used to discriminate between these species. Apart from the known virulence determinants, species-specific gene clusters (SSC) were identified in each genome. These gene clusters might have important roles in the observed host and organ specificity (stomach versus gut versus liver), as well as the pathogenic potential of the four microorganisms.

FIGURE 3 shows all predicted ORFs in the *W. succinogenes* genome and their functional classification; the corresponding orthologous genes in the *H. pylori*, *H. hepaticus* and *C. jejuni* genomes are depicted according to their respective positions. This representation shows that the orthologues are spread throughout the entire chromosome. However, there are clusters of species-specific genes in all three microorganisms. As a common characteristic, many of the uncategorized genes are arranged in clusters and co-localize with reported virulence genes (see online supplementary material S2 (table)). The largest clusters in *H. pylori* contain the cytotoxin-associated genes of the *cag* PAI as well as the urease operon, one of the main virulence factors in *H. pylori*. Additionally, we identified clusters of hypothetical genes of unknown function that do not have any matching homologues in the databases (see online supplementary material S2 (table)); as well as two large sets of consecutive predicted ORFs centred around a homologue of the *virB4* gene (FIG. 3). In the *H. hepaticus* genome, the HHGI1 genomic island — which, like the *H. pylori cag* PAI, encodes a set of type IV virulence genes — and the urease operon are clearly detectable (FIG. 3). In *C. jejuni* a species-specific cluster

LEWIS ANTIGENS
Fucosylated carbohydrate antigens usually found on the surface of eukaryotic cells. They are structurally related to the human ABH blood group system.

HOMOPOLYMERIC TRACTS
The simplest but most frequent low-complexity zones in prokaryotes are simple sequence repeats (SSR), which consist either of homopolymeric tracts or multimeric repeats. These genomic regions are prone to slippage and mispairing.

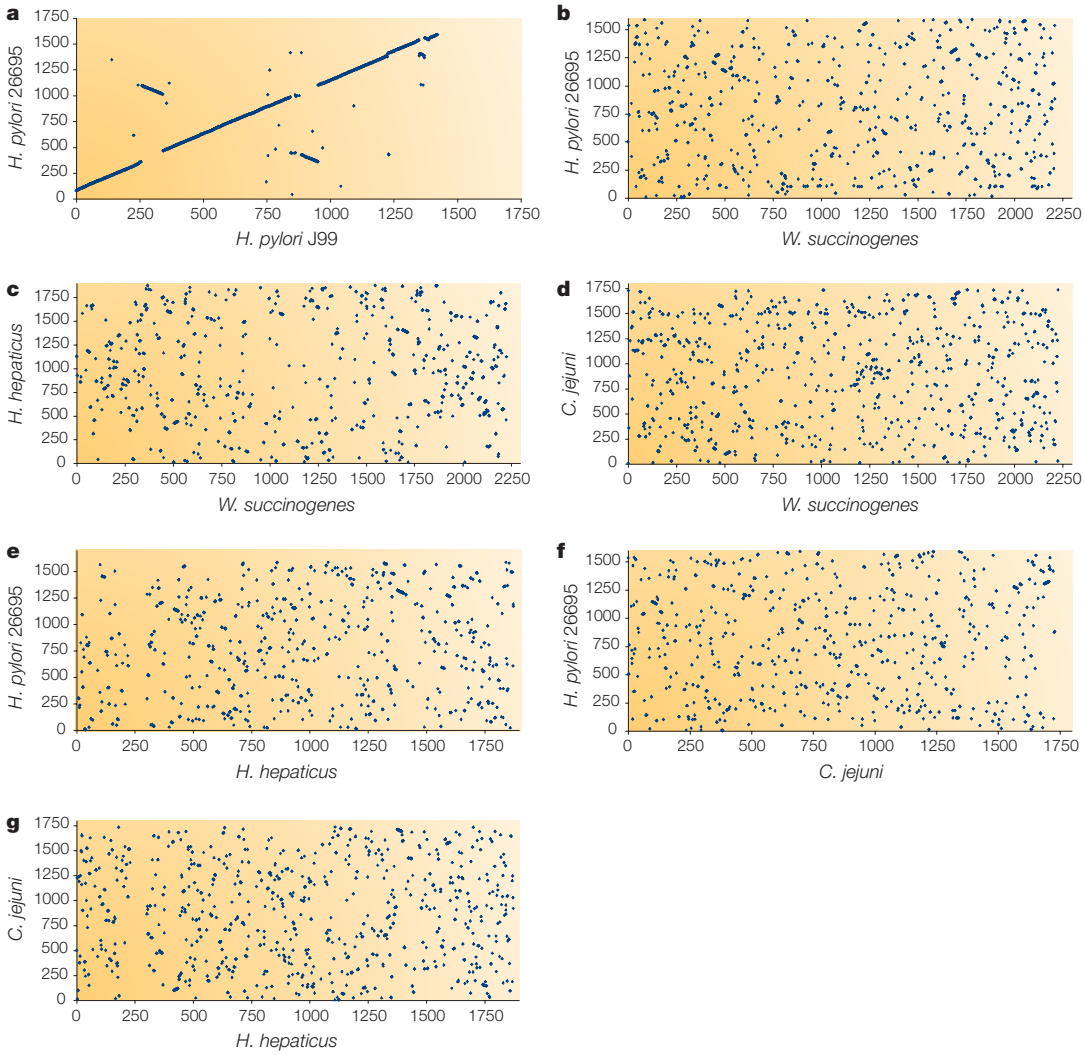


Figure 4 | **Genome-wide co-linearity analysis.** The positions of orthologous genes along the chromosome are plotted as open reading frame (ORF) numbers. The *x*- and *y*-axes represent the two genomes being compared. The origin of replication (*ori*) is clearly detectable in *Wolinella succinogenes* DSM 1740 and *Campylobacter jejuni* NCTC 11168, as predicted by a bias of G towards the leading strand (G+C skew), therefore the positions of each pair of homologous genes relative to their *ori* can be calculated[87,88]. **a** | A co-linearity plot between *Helicobacter pylori* strains 26695 and J99. The presence of a diagonal line clearly illustrates the conserved gene order between the two strains. The genus-level comparisons are shown in **b**–**g**. To derive the co-linearity factor, the *x* and *y* coordinates of the orthologous pairs were used. For each pair of neighbouring ORFs (x_i, x_{i+1}) on the query genome, the position of the orthologues (y_j, y_{j+1}) on the target genomes was found. The difference Min (| yi+1 - yi |, #Orfs - | yi+1 - yi |) between the positions on the target genome was calculated. The difference values for all the ORF pairs were summed and divided by the number of ORFs to obtain the co-linearity factor. The resulting co-linearity factor for the strain-to-strain comparison is 18, whereas the species-to-species comparison gives 213 in the case of *W. succinogenes*/*C. jejuni* (**d**) and 214 in the case of *W. succinogenes*/*H. pylori* 26695 (**b**). Comparison of *H. pylori* 26695 and *C. jejuni* (**f**) results in a factor of 247, *H. hepaticus* and *H. pylori* 26695 (**e**) of 173, *H. hepaticus* and *C. jejuni* (**g**) of 212 and *H. hepaticus* and *W. succinogenes* (**c**) 112.

encoding the *H. hepaticus* shared cytolethal-distending toxins (CdtABC) and multiple hypothetical proteins can be detected as well as the *C. jejuni*-specific LOS and EPS loci (FIG. 3). The *H. pylori* OMPs are often flanked by hypothetical proteins within these species-specific regions. This might indicate that these proteins contribute to the structure of these outer membrane proteins as part of the *H. pylori*-specific adhesion mechanism to a host cell (see online supplementary material S2 (table)).

Genome-wide and local co-linearity analysis

As these four closely related species share a last common ancestor, the genome rearrangements that have contributed to their evolution have been investigated (FIG. 4). The two sequenced strains of *H. pylori* are examples of the co-linear succession that would be expected for organisms that belong to the same species and show only strain-to-strain variation[4,5,68]. For *H. pylori* strains 26695 and J99, the comparison results in an almost undisrupted diagonal (FIG. 4a).

The only two positions at which the order is interrupted are the result of a single recombination event[69]. The random pattern of the other plots (FIG. 4b–g) illustrates the lack of genome-wide co-linearity (FIG. 4b–g). The calculated co-linearity factors reflect the phylogenetic relationship between the organisms and outlines the position of *W. succinogenes* as phylogenetically intermediate between the two *Helicobacter* species and *C. jejuni*, which is in agreement with the published 16S rRNA phylogenetic trees[19,20]. Therefore, *W. succinogenes* might be considered as an out-group to the three pathogenic microorganisms, allowing differential analysis not only of gene content, but also their respective genome structures.

The high degree of disorder has caused the separation of genes that are jointly regulated in other subgroups of the Proteobacteria. The structural genes of the flagellar motor provide an example of this observation. In all four species, the genes for this structure, which encode more than 20 gene products, are dispersed throughout the entire genome[70]. How and whether they are jointly regulated remains unclear from the functional analysis currently available.

The rearrangement described in the co-linearity analysis reflects the extensive recombination that has occurred since these species diverged from their last common ancestor and that presumably had an important role in the speciation process. In this regard, it is tempting to speculate on the allowable degree of differential recombination that is possible within an organismal lineage, while maintaining the status of being the same biological species. The observed lack of operon organization within the genomes of this bacterial subgroup raises the question of how the functionality of the regulatory network is achieved. The study of environmental strains of species that are closely related to human pathogens has identified many 'orphaned' processes in the metabolism and the signalling circuitry of an organism. It will therefore be useful to re-assemble the complete set of physiological capabilities of the last common ancestor of this group of bacteria to draw conclusions on the emergence of their interaction with a distinct host.

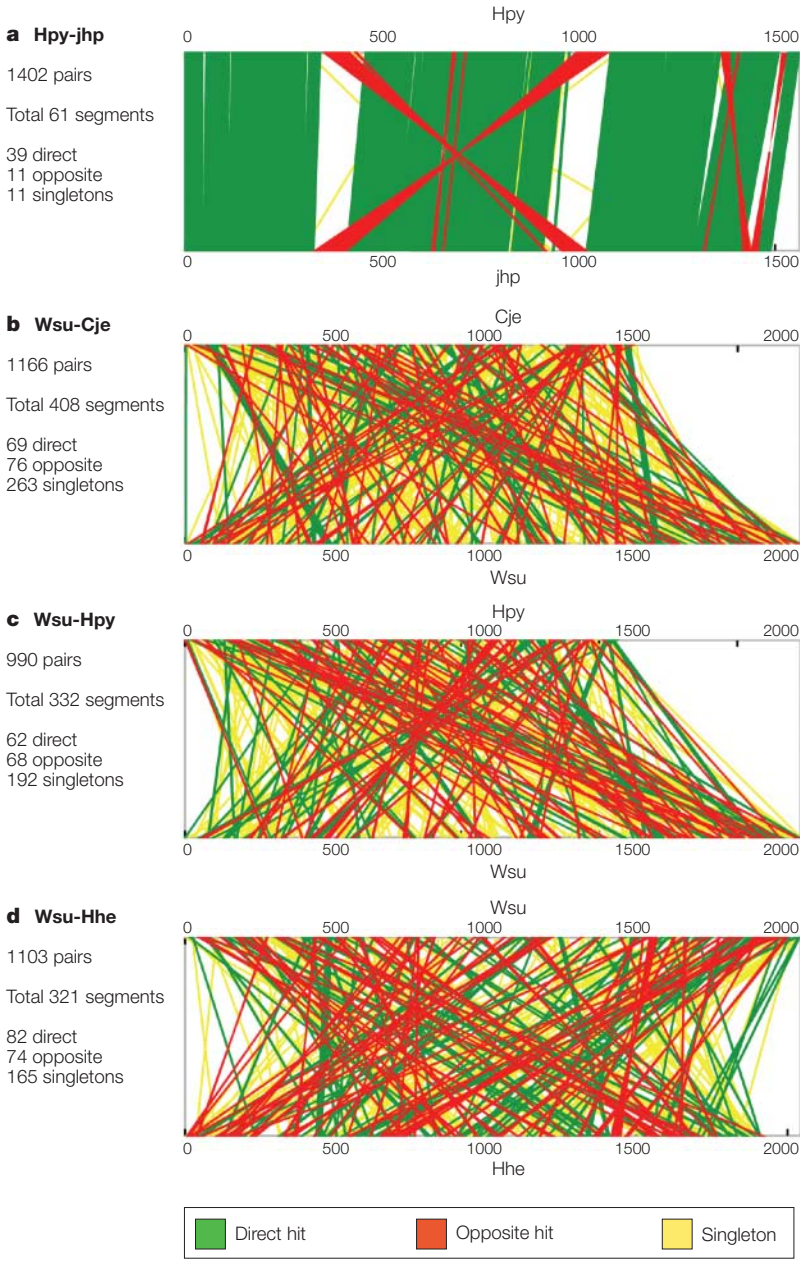


Figure 5 | **Visualization of intra- and interspecies recombinatorial events.** Direct matches are two or more pairs of orthologues that have maintained their order with respect to the origin (*ori*). Opposite matches have reversed their order with respect to the *ori* but maintained their relative order. Singletons are orthologues that are not adjacent to other orthologues. When two or more neighbouring orthologous pairs are considered in the pair-wise comparisons, local co-linearity (**a**–**d**) between *Wolinella succinogenes* DSM 1740 and the related pathogens can be seen. The intraspecies comparison between the *Helicobacter pylori* strains 26695 (Hpy) and J99 (jhp) revealed recombinatorial rearrangements (**a**). Parts **b**–**d** show the calculated interspecies comparisons. Cje, *Campylobacter jejuni* NCTC 11168.

Recombinatorial rearrangements

To visualize all the recombinatorial events that have occurred, direct matches (two or more pairs of orthologues that have maintained their order with respect to the origin (*ori*)), opposite matches (two or more pairs of orthologues that have reversed their order with respect to the *ori* but maintained their relative order) and singletons (orthologues that are not adjacent to other orthologues) are plotted against each other (FIG. 5). Despite an almost random arrangement of many orthologues on a genome-wide scale, one can still observe local co-linearity by a representation that takes into account only two or more pairs of co-linear genes. Examples are those that encode metabolic and multi-subunit enzyme complexes, such as the ribosomal genes or the genes coding for the F-type ATPase. These groups of clustered genes are of particular interest because they have maintained their order in the genomes of all four species, even though in some cases their orientation with regard to the origin of replication has changed.

Gene loss and acquisition

Through a series of deletion and recombination events driven by host adaptation and speciation, the genomes of all four species have lost genes, albeit at different rates. In

this regard, it is likely that *W. succinogenes*, due to its genome size and gene content, has lost the least genetic information so far, as it still has the largest amount of genes coding for extended signalling and metabolic capabilities. Examples of gene loss include signalling cascades in *H. pylori* that are missing 'upstream' components but which are still present in *W. succinogenes*. The drive to a smaller genome might also have been responsible for the particularly high coding frequency, which is highest in *C. jejuni*, at 94.3%. The lower coding frequency of the *Helicobacter* species is indicative of an ongoing effort to degrade these genomes even further. This is reflected by the large number of pseudogenes that are found in the *H. pylori* genome and the general tendency for genome plasticity, which can be observed in the high rates of mutation and recombination and which, as a consequence, lead to the high variability of strains that is observed for *Helicobacter* species.

Although this class of bacteria tend to delete genetic information, there is evidence that gain of genetic information is also important. This is particularly true for the two *Helicobacter* species and *W. succinogenes*, where horizontal gene transfer accounts for 5–6% of the present genetic information[34,71]. In the case of *W. succinogenes*, this percentage does not include the additional 450 kb of genetic information that is present compared with the *Helicobacter* genomes and therefore cannot be explained by horizontal gene transfer alone.

This comparative genomic analysis has identified genes from a shared ancestral gene pool that are essential for the maintenance of these organisms in mammalian hosts, as well as species-specific genes that mediate host and tissue specificity[72]. A large proportion of the species-specific genes co-localize in regions of the individual genomes that have not been identified as genomic islands, which contrasts with the observation that no genome-wide co-linearity is present between these closely related genomes.

Comparative genomics: power and limitations

Genome sequencing and comparative genomics are undoubtedly of immense value for researchers who wish to investigate common and unique traits in microbial adaptation and speciation[73]. The genome data are starting points for subsequent research that considers the many interactions between a microorganism and its ecological niche and surrounding environment. The initial findings of a sequence analysis must be followed by experimental work to elucidate or confirm the physiological function of the identified gene products. Sequence data cannot be used to elucidate all aspects of the biology of a particular species, but the continued generation of genomic and experimental data gives researchers an insight into the complexity of an organism's biology and allows questions to be posed regarding genome dynamics and diversity (FIG. 6).

Comparisons between the genome sequences of commensal bacteria and those of pathogenic bacteria can identify genes that are involved in the observed host specificity and the mechanisms of the host–microorganism interaction. A comparative genome analysis of a bacterial subgroup can identify strain-specific, lineage-specific or even niche-specific genomic regions of the analysed species or phylum, and can help to explain differences in lifestyle and disease manifestation[26,68,74–76].

The comparative approach is valuable from an evolutionary point of view as it is an excellent tool to track the evolutionary processes of loss, acquisition and variation of genetic information[77]. In a more global approach, shared ancestral gene pools can be detected. The reconstruction of 'minimal gene sets' of the last common ancestor can elucidate the origin and adaptive traits of microbial species[72]. For *C. jejuni* and *H. pylori*, a reconstruction of a tentative common ancestor has been presented using a modular protein approach[78].

In addition to the core chromosome, analysis of the flexible genome pool should also be part of a comparative analysis. The almost complete identity of the chromosome in some bacterial species provides no clues to the differences observed in their biology and pathogenic potential. These distinctions are largely anchored in the varied presence and gene content of specific plasmids and other mobile genetic elements; examples include members of the *Bacillus cereus sensu lato* group and isolates of *C. jejuni*[36,79]. Comparative analysis of these elements could help to elucidate their evolutionary origins, as observed for the *W. succinogenes* WsuGI genomic island, which is part of an autonomous plasmid in *C. jejuni*. These analyses allow the identification of unique and common evolutionary traits that contribute to speciation and help to identify the proportion of horizontally acquired genes in a particular organism.

However, it is noteworthy that access to a microbial genome provides just a snapshot of a momentary state in microbial evolution. It does not provide information on the genome plasticity in different isolates. Genotyping and re-sequencing isolates with different physiological capabilities or pathogenic potential can address these questions by analysing stretches of sequence that are dispersed throughout the genome of the compared species[31,80]. For the compared ε-proteobacteria, the intraspecies diversity of different *C. jejuni* and *H. pylori* isolates has been studied in detail[22,23,59,81–84].

The sequenced genome of a type strain can be a reference for further fundamental analysis of the similarities and differences among diverse isolates. Using this approach, highly variable regions in strains or different clinical isolates that are mutational 'hot spots' can be identified. In the ε-proteobacteria subgroup examples include the high-plasticity zones (HPZ) of *H. pylori*, which are composed of highly repetitive DNA, these regions are crucial for this pathogen to persist and evade immune clearance[31,85].

Transcriptional analysis based on sequence data is focused on the analysis of promoter regions, the overall operon organization and identification of transcriptional regulators, as well as the classification of sensing capabilities. However, comparative genome analysis is inadequate to understand those cases in which the observed phenotype or pathogenic potential is based

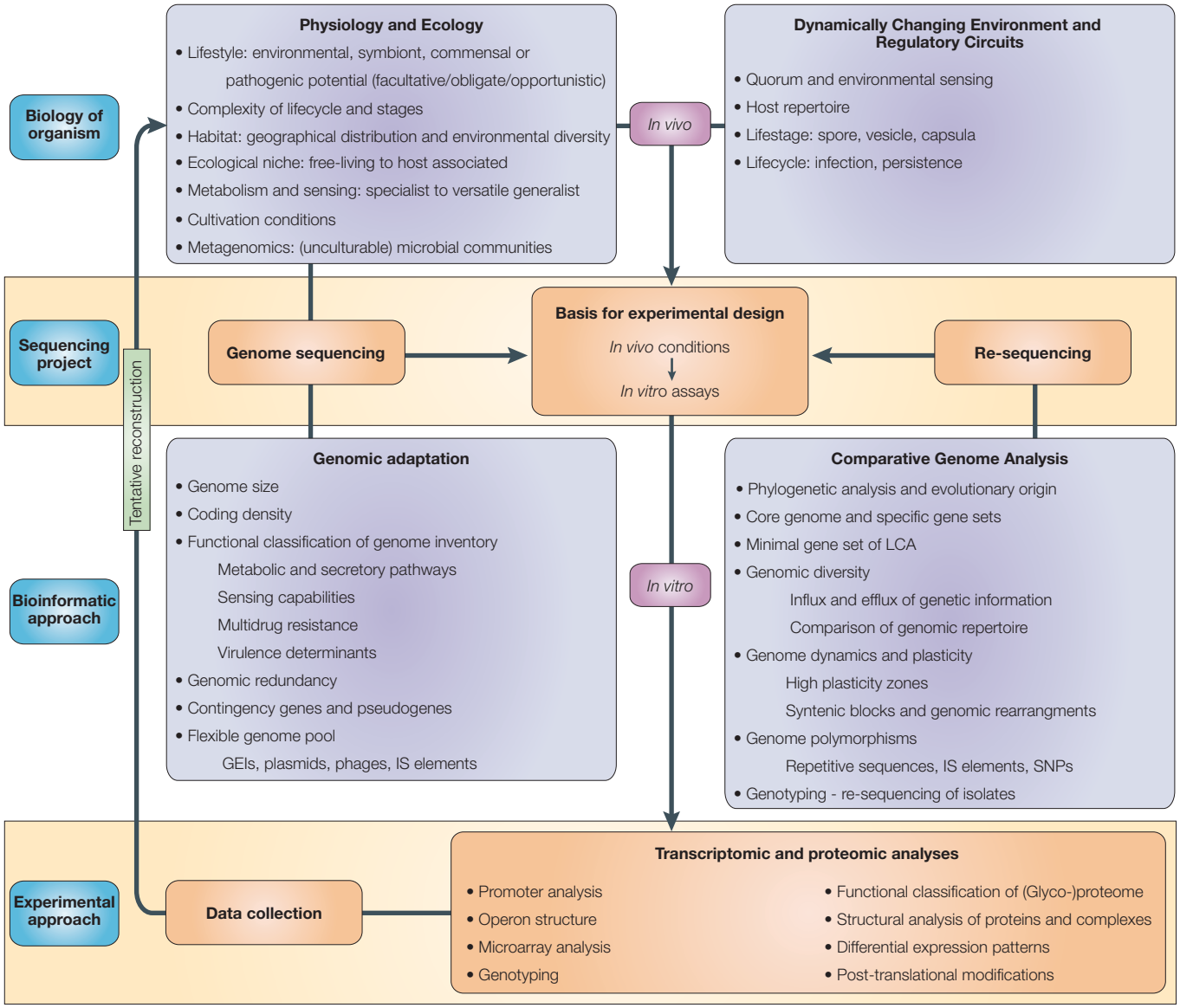


Figure 6 | **Microbial genomics and the study of bacterial pathogenicity.** General information on the *in vivo* biology of a microorganism, including knowledge of the lifestyle and habitat, feeds into *in vitro* experimental design. Information from a genome sequencing project, particularly a functional classification of the genome inventory, can also inform *in vitro* analysis. Comparative genome analysis provides further information, including the composition of the core genome and the flexible genome pool and can be useful in identifying the genetic makeup of the last common ancestor of the microorganism in question. For genes of unknown function, or genetic variations such as single nucleotide polymorphisms, further 'omics' technologies such as transcriptomics and proteomics can provide more information.

on alterations in the regulatory circuitry in response to a dynamically changing environment. The *Pseudomonas aeruginosa* quorum-sensing circuitry is such an example, wherein the expression of virulence determinants depends on the density of the *Pseudomonas* population[86].

Monitoring RNA levels in transcriptomic arrays is a powerful tool but is not sufficient to cover all aspects of an organisms' regulatory network to answer questions of post-transcriptional regulation and post-translational modification. To answer these questions, proteomic and structural analysis of proteins and their native complexes and investigation of their enzymatic and catalytic functions are essential. As more sequencing data are gathered, this will provide the fundamental basis to answer various questions covering all aspects of microbial life and could help to clarify the evolutionary origin, phylogenetic relationship and emergence of an organism.

Conclusions

Studying the evolutionary process of adaptation to a specific host at the genomic level identifies sets of genes that are either species-specific or are part of the species gene pool. The species-specific genes identified in a comparative analysis can then be used by all laboratories studying the molecular mechanisms of the interactions of this bacterial species with its host. It is therefore

important to note that many hypothetical genes form clusters in species-specific regions that are not defined by the traditional hallmarks of genomic islands. In this analysis, this is particularly apparent for the species-specific clusters (SSC) that were noted for the first time (FIG. 3), and which cannot be identified by any other methodology. An in-depth understanding of the mechanisms of specific bacteria–host interactions will not be possible without understanding the function of these genes.

Using genomic comparison, it was also possible to visualize the high degree of genomic rearrangements that seems to be a general characteristic within this bacterial lineage. Understanding the different rates of recombination among the studied strains and species will not only give insight into the microbial speciation process, but also into the common and unique traits of the genome evolution of commensal and pathogenic species.

A limitation of the comparative approach is the detection threshold that can be applied at the genomic level. This is particularly the case for mutations based on single nucleotide polymorphisms (SNPs) and insertions and deletions (indels) that affect either genes or regulatory elements, which cannot be assessed by the methodology used in this study; functional genomics techniques, such as transcriptomics and proteomics, will therefore be required to study the effects of these genomic variations. The same is true for the hypothetical genes with no assigned function, which account for as much as one-third of the gene content of the sequenced ε-proteobacteria. The impact of these genes on the physiology of these microorganisms must be further elucidated by experimental studies.

Last, but not least, it is important to note that, although there is a stringent correlation between the absence of certain genes and a potential physiological function, the opposite does not hold true. For example, the absence of a particular set of genes can rule out certain physiological capabilities but the presence of virulence factors is essential, but not sufficient, for an obligatory pathogenic lifestyle. It will therefore be rewarding to continue to tackle the thin line between gut commensalism and pathogenicity using diverse techniques including comparative genomics.

1. Miroshnichenko, M. L. *et al. Caminibacter profundus* sp. nov., a novel thermophile of Nautiliales ord. nov. within the class 'Epsilonproteobacteria', isolated from a deep-sea hydrothermal vent. *Int. J. Syst. Evol. Microbiol.* **54**, 41–45 (2004).
2. Engel, A. S. *et al.* Filamentous 'Epsilonproteobacteria' dominate microbial mats from sulfidic cave springs. *Appl. Environ. Microbiol.* **69**, 5503–5511 (2003).
3. Moyer, C. L., Dobbs, F. C. & Karl, D. M. Phylogenetic diversity of the bacterial community from a microbial mat at an active, hydrothermal vent system, Loihi Seamount, Hawaii. *Appl. Environ. Microbiol.* **61**, 1555–1562 (1995).
4. Tomb, J. F. *et al.* The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**, 539–547 (1997).
5. Alm, R. A. *et al.* Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397**, 176–180 (1999).
6. Parkhill, J. *et al.* The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403**, 665–668 (2000).
7. Suerbaum, S. *et al.* The complete genome sequence of the carcinogenic bacterium *Helicobacter hepaticus*. *Proc. Natl Acad. Sci. USA* **100**, 7901–7906 (2003).
8. Baar, C. *et al.* Complete genome sequence and analysis of *Wolinella succinogenes*. *Proc. Natl Acad. Sci. USA* **100**, 11690–11695 (2003).
References 4–8 report the genome sequences of the ε-proteobacterial species compared in this review.
9. Cover, T. L. & Blaser, M. J. *Helicobacter pylori* infection, a paradigm for chronic mucosal inflammation: pathogenesis and implications for eradication and prevention. *Adv. Intern. Med.* **41**, 85–117 (1996).
10. Solnick, J. V. & Schauer, D. B. Emergence of diverse *Helicobacter* species in the pathogenesis of gastric and enterohepatic diseases. *Clin. Microbiol. Rev.* **14**, 59–97 (2001).
11. Ward, J. M. *et al.* Chronic active hepatitis and associated liver tumors in mice caused by a persistent bacterial infection with a novel *Helicobacter* species. *J. Natl Cancer Inst.* **86**, 1222–1227 (1994).
12. Blaser, M. J. Epidemiologic and clinical features of *Campylobacter jejuni* infections. *J. Infect. Dis.* **176** (Suppl.), S103–S105 (1997).
13. Nachamkin, I., Allos, B. M. & Ho, T. *Campylobacter* species and Guillain–Barré syndrome. *Clin. Microbiol. Rev.* **11**, 555–567 (1998).
14. Ang, C. W., Jacobs, B. C. & Laman, J. D. The Guillain–Barré syndrome: a true case of molecular mimicry. *Trends Immunol.* **25**, 61–66 (2004).
15. Moran, A. P. & Prendergast, M. M. Molecular mimicry in *Campylobacter jejuni* and *Helicobacter pylori* lipopolysaccharides: contribution of gastrointestinal infections to autoimmunity. *J. Autoimmun.* **16**, 241–256 (2001).
16. Wolin, M. J., Wolin, E. A. & Jacobs, N. J. Cytochrome-producing anaerobic vibrio, *Vibrio succinogenes* sp. n. *J. Bacteriol.* **81**, 911–917 (1961).
17. Andersson, J. O. & Andersson, S. G. Insights into the evolutionary process of genome degradation. *Curr. Opin. Genet. Dev.* **9**, 664–671 (1999).
18. Oshima, K. *et al.* Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma. *Nature Genet.* **36**, 27–29 (2004).
19. Simon, J., Gross, R., Klimmek, O. & Kröger, A. in *The Prokaryotes: An Evolving Electronic Resource for the Microbiological Community*. 3rd ed. (Balows. A., Trüper, H. G., Dworkin, M., Harder, W. & Schleifer, K. H. eds) (Springer–Verlag, New York, 2000).
20. Vandamme, P. *et al.* Revision of *Campylobacter*, *Helicobacter*, and *Wolinella* taxonomy: emendation of generic descriptions and proposal of *Arcobacter* gen. nov. *Int. J. Syst. Bacteriol.* **41**, 88–103 (1991).
21. Tamas, I., Klasson, L. M., Sandström, J. P. & Andersson, S. G. Mutualists and parasites: how to paint yourself into a (metabolic) corner. *FEBS Lett.* **498**, 135–139 (2001).
22. Garcia-Vallve, S., Janssen, P. J. & Ouzounis, C. A. Genetic variation between *Helicobacter pylori* strains: gene acquisition or loss? *Trends Microbiol.* **10**, 445–447 (2002).
23. Janssen, P. J., Audit, B. & Ouzounis, C. A. Strain-specific genes of *Helicobacter pylori*: distribution, function and dynamics. *Nucleic Acids Res.* **29**, 4395–4404 (2001).
24. Boneca, I. G. *et al.* A revised annotation and comparative analysis of *Helicobacter pylori* genomes. *Nucleic Acids Res.* **31**, 1704–1714 (2003).
25. Henz, S. R., Auch, A. F., Huson, D. H., Nieselt-Struwe, K. & Schuster, S. C. Whole genome-based prokaryotic phylogeny. *Bioinformatics* 2004 (epub ahead of print).
26. Kalman, S. *et al.* Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nature Genet.* **21**, 385–389 (1999).
27. Dobrindt, U., Hochhut, B., Hentschel, U. & Hacker, J. Genomic islands in pathogenic and environmental microorganisms. *Nature Rev. Microbiol.* **2**, 414–424 (2004).
Adresses the impact of horizontal gene transfer on the evolution of pathogenic as well as non-pathogenic bacteria.
28. Hofreuter, D. & Haas, R. Characterization of two cryptic *Helicobacter pylori* plasmids: a putative source for horizontal gene transfer and gene shuffling. *J. Bacteriol.* **184**, 2755–2766 (2002).
29. Simon, J. & Kröger, A. Identification and characterization of IS*1302*, a novel insertion element from *Wolinella succinogenes* belonging to the IS3 family. *Arch. Microbiol.* **170**, 43–49 (1998).
30. Mahillon, J. & Chandler, M. Insertion sequences. *Microbiol. Mol. Biol. Rev.* **62**, 725–774 (1998).
General overview of the diverse types of insertion sequence elements, their classification, structural composition and versatile transposition mechanisms within a genome.
31. Aras, R. A., Kang, J., Tschumi, A. I., Harasaki, Y. & Blaser, M. J. Extensive repetitive DNA facilitates prokaryotic genome plasticity. *Proc. Natl Acad. Sci. USA* **100**, 13579–13584 (2003).
32. Petersen, L., On, S. L. W. & Ussery, D. W. Visualization and significance of DNA structural motifs in the *Campylobacter jejuni* genome. *Genome Lett.* **1**, 16–25 (2002).
33. Doolittle, R. F. Biodiversity: microbial genomes multiply. *Nature* **416**, 697–700 (2002).
34. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).
35. Hacker, J. & Carniel, E. Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep.* **2**, 376–381 (2001).
36. Bacon, D. J. *et al.* Involvement of a plasmid in virulence of *Campylobacter jejuni* 81-176. *Infect. Immun.* **68**, 4384–4390 (2000).
37. Alfredson, D. A. & Korolik, V. Sequence analysis of a cryptic plasmid pCJ419 from *Campylobacter jejuni* and construction of an *Escherichia coli–Campylobacter* shuttle vector. *Plasmid* **50**, 152–160 (2003).
38. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
39. Takata, T. *et al.* Phenotypic and genotypic variation in methylases involved in type II restriction–modification systems in *Helicobacter pylori*. *Nucleic Acids Res.* **30**, 2444–2452 (2002).
40. Ochman, H. & Moran, N. A. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* **292**, 1096–1099 (2001).
Highlights the influence of gene acquisition and gene loss on genome evolution as major factors promoting the spectrum of interactions between bacteria and their hosts.
41. Adler, J., Hazelbauer, G. L. & Dahl, M. M. Chemotaxis toward sugars in *Escherichia coli*. *J. Bacteriol.* **115**, 824–847 (1973).
42. Hoch, J. A. & Silhavy, T. J. *Two-component Signal Transduction*. (ASM Press, Washington DC, 1995.
43. Cases, I., de Lorenzo, V. & Ouzounis, C. A. Transcription regulation and environmental adaptation in bacteria. *Trends Microbiol.* **11**, 248–253 (2003).
44. Marchant, J., Wren, B. & Ketley, J. Exploiting genome sequence: predictions for mechanisms of *Campylobacter* chemotaxis. *Trends Microbiol.* **10**, 155–159 (2002).
45. Covacci, A., Telford, J. L., Del Giudice, G., Parsonnet, J. & Rappuoli, R. *Helicobacter pylori* virulence and genetic geography. *Science* **284**, 1328–1333 (1999).

46. Martino, M. C. *et al. Helicobacter pylori* pore-forming cytolysin orthologue TlyA possesses *in vitro* hemolytic activity and has a role in colonization of the gastric mucosa. *Infect. Immun.* **69**, 1697–1703 (2001).
47. Ratledge, C. & Dover, L. G. Iron metabolism in pathogenic bacteria. *Annu. Rev. Microbiol.* **4**, 881–941 (2000).
48. Wooldridge, K. G. & Williams, P. H. Iron uptake mechanisms of pathogenic bacteria. *FEMS Microbiol. Rev.* **12**, 325–48 (1993).
49. Braun, V. Iron uptake mechanisms and their regulation in pathogenic bacteria. *Int. J. Med. Microbiol.* **291**, 67–79 (2001).
50. Lin, J., Michel, L. O. & Zhang, Q. CmeABC functions as a multidrug efflux system in *Campylobacter jejuni*. *Antimicrob. Agents Chemother.* **46**, 2124–2131 (2002).
51. Konkel, M. E., Kim, B. J., Rivera–Amill, V. & Garvis, S. G. Bacterial secreted proteins are required for the internaliztion of *Campylobacter jejuni* into cultured mammalian cells. *Mol. Microbiol.* **32**, 691–701 (1999).
52. Lara-Tejero, M. & Galan, J. E. A bacterial toxin that controls cell cycle progression as a deoxyribonuclease I-like protein. *Science* **290**, 354–357 (2002).
53. Thelestam, M. & Frisan, T. Cytolethal distending toxins. *Rev. Physiol. Biochem. Pharmacol.* 2004 Aug 27 (epub ahead of print).
54. Odenbreit, S. *et al.* Translocation of *Helicobacter pylori* CagA into gastric epithelial cells by type IV secretion. *Science* **287**, 1497–1500 (2000).
55. Gebert, B., Fischer, W. & Haas, R. The *Helicobacter pylori* vacuolating cytotoxin: from cellular vacuolation to immunosuppressive activities. *Rev. Physiol. Biochem. Pharmacol.* 2004 (epub ahead of print).
A detailed review that summarizes the manifold interferences of the vacuolating cytotoxin VacA with the targeted host cell, inducing — apart of vacuole formation — apoptosis and modulatory effects on the host's immune system.
56. Ilver, D., Barone, S., Mercati, D., Lupetti, P. & Telford, J. L. *Helicobacter pylori* toxin VacA is transferred to host cells via a novel contact-dependent mechanism. *Cell. Microbiol.* **6**, 167–174 (2004).
A recent paper that reports the discovery of a novel contact-dependent mechanism for transfer of bacterial surface associated vacuolating cytotoxin VacA into the epithelial host cells.
57. Cascales, E. & Christie, P. J. The versatile bacterial type IV secretion systems. *Nature Rev. Microbiol.* **1**, 137–149 (2003).
Describes the phylogentic distribution and structural composition of the versatile bacterial type IV secretion systems and summarizes the effects of the secreted molecules on host-interacting targets.
58. Bacon, D. J. *et al.* DNA sequence and mutational analyses of the pVir plasmid of *Campylobacter jejuni* 81-176. *Infect. Immun.* **70**, 6242–6250 (2002).
Presents the sequence of the *C. jejuni* virulence plasmid pVir that encodes for virulence *(vir)* homologues of the type IV secretion system.
59. Dorrell, N. *et al.* Whole genome comparison of *Campylobacter jejuni* human isolates using a low-cost microarray reveals extensive genetic diversity. *Genome Res.* **11**, 1706–1715 (2001).
60. Wilson, D. L. *et al.* Variation of the natural transformation frequency of *Campylobacter jejuni* in liquid shake culture. *Microbiology* **149**, 3603–3615 (2003).
61. Gilmore, M. S. & Ferretti, J. J. The thin line between gut commensal and pathogen. *Science* **299**, 1999–2002 (2003).
62. 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).
Reports the existence of highly mutable contingency genes, containing simple sequence repeats that are prone to slipping and mispairing.
63. Metzgar, D. & Wills, C. Evidence for the adaptive evolution of mutation rates. *Cell* **101**, 581–584 (2000).
64. Linton, D., Karlyshev, A. V. & Wren, B. W. Deciphering *Campylobacter jejuni* cell surface interactions from the genome sequence. *Curr. Opin. Microbiol.* **4**, 35–40 (2001).
In this study, a β-1,3 galactosyltransferase was shown to be phase variable expressed due to an intragenic homopolymeric tract, leading to alternate ganglioside-mimicking LOS structures.
65. Wang, G., Rasko, D. A., Sherburne, R. & Taylor, D. E. Molecular genetic basis for the variable expression of Lewis Y antigen in *Helicobacter pylori*: analysis of the α(1,2) fucosyltransferase gene. *Mol. Microbiol.* **31**, 1265–1274 (1999).
66. Lozniewski, A. *et al.* Influence of Lewis antigen expression by *Helicobacter pylori* on bacterial internalization by gastric epithelial cells. *Infect. Immun.* **71**, 2902–2906 (2003).
67. Saunders, N. J., Peden, J. F., Hood, D. W. & Moxon, E. R. Simple sequence repeats in the *Helicobacter pylori* genome. *Mol. Microbiol.* **27**, 1091–1098 (1998).
68. da Silva, A. C. *et al.* Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. *Nature* **417**, 459–463 (2002).
69. Tillier, E. R. & Collins, R. A. Genome rearrangement by replication-directed translocation. *Nature Genet.* **26**, 195–197 (2000).
70. O'Toole, P. W., Lane, M. C. & Porwollik, S. *Helicobacter pylori* motility. *Microbes Infect.* **2**, 1207–1214 (2000).
71. Garcia-Vallve, S., Romeu, A. & Palau, J. Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res.* **10**, 1719–1725 (2000).
72. Koonin, E. V. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Rev. Microbiol.* **1**, 127–136 (2003).
73. Nierman, W. C. & Fraser, C. M. The power in comparison. *Trends Microbiol.* **12**, 62–63 (2004).
74. Read, T. D. *et al.* Genome sequence of *Chlamydophila caviae* (*Chlamydia psittaci* GPIC): examining the role of niche-specific genes in the evolution of the Chlamydiaceae. *Nucleic Acids Res.* **31**, 2134–2147 (2003).
75. Parkhill, J. *et al.* Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nature Genet.* **35**, 32–40 (2003).
76. Seshadri, R. *et al.* Comparison of the genome of the oral pathogen *Treponema denticola* with other spirochete genomes. *Proc. Natl Acad. Sci. USA* **101**, 5646–5651 (2004).
77. Wren, B. W. The yersiniae — a model genus to study the rapid evolution of bacterial pathogens. *Nature Rev. Microbiol.* **1**, 55–64 (2003).
78. Le Bouder-Langevin, S., Capron-Montaland, I., De Rosa, R. & Labedan, B. A strategy to retrieve the whole set of protein modules in microbial proteomes. *Genome Res.* **12**, 1961–1973 (2002).
79. Rasko, D. A. *et al.* The genome sequence of *Bacillus cereus* ATCC 10987 reveals metabolic adaptations and a large plasmid related to *Bacillus anthracis* pXO1. *Nucleic Acids Res.* **32**, 977–988 (2004).
80. Snyder, L. A. S., Davies, J. K. & Saunders, N. J. Microarray genomotyping of key experimental strains of *Neisseria gonorrhoeae* reveals gene complement diversity and five new neisserial genes associated with minimal mobile elements. *BMC Genomics* **5**, 23 (2004).
81. Poly, F., Threadgill, D. & Stintzi, A. Identification of *Campylobacter jejuni* ATCC 43431-specific genes by whole microbial genome comparisons. *J. Bacteriol.* **186**, 4781–4795 (2004).
82. Leonard, E. E., Tompkins, L. S., Falkow, S. & Nachamkin, I. Comparison of *Campylobacter jejuni* isolates implicated in Guillain–Barré syndrome and strains that cause enteritis by a DNA microarray. *Infect. Immun.* **72**, 1199–1203 (2004).
83. Pearson, B. M. *et al.* Comparative genome analysis of *Campylobacter jejuni* using whole genome DNA microarrays. *FEBS Lett.* **554**, 224–230 (2003).
84. Gaynor, E. C. *et al.* The genome-sequenced variant of *Campylobacter jejuni* NCTC 11168 and the original clonal clinical isolate differ markedly in colonization, gene expression, and virulence-associated phenotypes. *J. Bacteriol.* **186**, 503–517 (2004).
85. Aras, R. A. *et al.* Plasticity of repetitive DNA sequences within a bacterial (Type IV) secretion system component. *J. Exp. Med.* **198**, 1349–1360 (2003).
86. de Kievit, T. R., Kakai, Y., Register, J. K., Pesci, E. C. & Iglewski, B. H. Role of the *Pseudomonas aeruginosa* las and rhl quorum-sensing systems in rhll regulation. *FEMS Microbiol. Lett.* **212**, 101–106 (2002).
87. Grigoriev, A. Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res.* **26**, 2286–2290 (1998).
88. Frank, A. C. & Lobry, J. R. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* **238**, 65–77 (1999).

Acknowledgements

The authors would like to thank Ramkumar Nandakumar and Daniel Richter for assistance with the computation and graphics. This work was funded by the Max-Planck Society.

Competing interests statement

The authors declare no competing financial interests.

Online links

DATABASES

The following terms in this article are linked online to:

Entrez: http://www.ncbi.nlm.nih.gov/Entrez/
CagA | *Campylobacter jejuni* NCTC 11168 | *Chlamydia pneumoniae* | *Chlamydia trachomatis* | *cmeABC* | *Helicobacter hepaticus* ATCC 51449 | *Helicobacter pylori* 26695 | *Helicobacter pylori* J99 | pVir | *Wolinella succinogenes* DSM 1740

FURTHER INFORMATION

Stephen C. Schuster's laboratory:
http://www.eb.tuebingen.mpg.de/schuster/home.htm
COG database: http://www.ncbi.nlm.nih.gov/Entrez/
PyloriGene database:
http://genolist.pasteur.fr/PyloriGene/
Helicobacter Foundation: http://www.helico.com/
Restriction Enzyme Database:
http://rebase.neb.com/rebase/rebase.html

SUPPLEMENTARY INFORMATION

See online article: S1 (table) | S2 (table)
Access to this links box is available online.

Biographies

Mark Eppinger, Claudia Baar and Günter Raddatz are Research Associates at the Max-Planck Institute for Developmental Biology, Tübingen, Germany. Their research interests focus on predictions made from genomic analysis and comparisons. Using bioinformatics techniques they identify targets for subsequent functional studies.

Daniel H. Huson is a Professor for Bioinformatics at the Eberhard-Karls University in Tübingen, Germany. After working with the team that assembled the first draft of the human genome at Celera, his current research interests include genomic assemblies and synteny, as well as whole-genome phylogeny.

Stephan C. Schuster is head of the genome centre at the Max-Planck Institute for Developmental Biology, Tübingen, Germany, as well as of a research group that studies genome evolution in bacteria. His work concentrates on δ- and ε-proteobacteria, in which he studies the mechanisms of host-adaptation at the genomic level.

Online links

Entrez

CagA
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=Retrieve&dopt=Graphics&list_uids=889201

Campylobacter jejuni NCTC 11168
http://www.ncbi.nlm.nih.gov/genomes/framik.cgi?db=genome&gi=152

Chlamydia pneumoniae
http://www.ncbi.nlm.nih.gov/genomes/framik.cgi?db=genome&gi=311

Chlamydia trachomatis
http://www.ncbi.nlm.nih.gov/genomes/framik.cgi?db=genome&gi=137

cmeABC
http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=45269108

Helicobacter hepaticus ATCC 51449
http://www.ncbi.nlm.nih.gov/genomes/framik.cgi?db=genome&gi=307

Helicobacter pylori 26695
http://www.ncbi.nlm.nih.gov/genomes/framik.cgi?db=genome&gi=128

Helicobacter pylori J99
http://www.ncbi.nlm.nih.gov/genomes/framik.cgi?db=genome&gi=139

pVir
http://www.ncbi.nlm.nih.gov/genomes/framik.cgi?db=genome&gi=17287

Wolinella succinogenes DSM 1740
http://www.ncbi.nlm.nih.gov/genomes/framik.cgi?db=genome&gi=323

COG database
http://www.ncbi.nlm.nih.gov/COG/