76 Anderson, J.M. *et al.* (1990) Ultrastructure and antigenicity of the unique cell and pimple of the *Candida albicans* opaque phenotype. *J. Bacteriol*. 172, 224–235

77 Morrow, B. *et al.* (1993) Coordinate regulation of two opaque phase specific-genes during white–opaque switching in *Candida albicans*. *Infect. Immun*. 61, 1823–1828

78 Srikantha, T. and Soll, D.R. (1993) A white-specific gene in the white-opaque switching system of *Candida albicans*. *Gene* 131, 53–60

79 Srikantha, T. *et al.* (1995) Functional analysis of the promoter of the phase-specific *WH11* gene of *Candida albicans*. *Mol. Cell. Biol*. 15, 1797–1805

80 Rustchenko-Bulgac, E.P. *et al.* (1990) Chromosomal rearrangements associated with morphological mutants provide a means for genetic variation of *Candida albicans*. *J. Bacteriol*. 173, 7436–7442

81 Perez-Martin, J. *et al.* (1999) Phenotypic switching in *Candida albicans* is controlled by a *SIR2* gene. *EMBO J*. 18, 2580–2592

82 Sonneborn, A.B. *et al.* (1999) Control of white–opaque phenotypic switching in *Candida albicans* by the Efg1p morphogenetic regulator. *Infect. Immun*. 67, 4655–4660

83 Soll, D. (1988) High-frequency switching in *Candida albicans* and its relationship to vaginal candidiasis. *Am. J. Obstet. Gynecol*. 158, 997–1001

84 Jones, S. *et al.* (1994) Increased phenotypic switching in strains of *Candida albicans* associated with invasive infections. *J. Clin. Microbiol*. 132, 2869–2870

85 Kvaal, C. *et al.* (1999) Missexpression of the opaque-phase specific gene *PEP1* (*SAP1*) in the white phase of *Candida albicans* confers increased in a mouse model of cutaneous infection. *Infect. Immun*. 67, 6652–6662

Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes

Samuel Karlin

A gene in a genome is defined as putative alien (pA) if its codon usage difference from the average gene exceeds a high threshold and codon usage differences from ribosomal protein genes, chaperone genes and protein-synthesis-processing factors are also high. pA gene clusters in bacterial genomes are relevant for detecting genomic islands (GIs), including pathogenicity islands (PAIs). Four other analyses appropriate to this task are G+C genome variation (the standard method); genomic signature divergences (dinucleotide bias); extremes of codon bias; and anomalies of amino acid usage. For example, the *cagA* domain of *Helicobacter pylori* is highly deviant in its genome signature and codon bias from the rest of the genome. Using these methods we can detect two potential PAIs in the *Neisseria meningitidis* genome, which contain hemagglutinin and/or hemolysin-related genes. Additionally, G+C variation and genome signature differences of the *Mycobacterium tuberculosis* genome indicate two pA gene clusters.

Samuel Karlin
Dept of Mathematics,
Stanford University,
Stanford, CA 94305-2125,
USA.
e-mail: karlin@
math.stanford.edu

Pathogenicity islands (PAIs) contain genes involved in diseases, such as genes encoding invasins, adhesins and secretion factors, and are often sources of toxins. In many cases, genes found in PAIs are similar to the genes in other pathogens that encode subunits of type IV secretion systems, which deliver bacterial virulence factors to the surface of, and possibly into, host cells. PAIs are a subset of genomic islands (GIs) comprising linked blocks of genes, such as the cobalamin operon of *Salmonella typhimurium*[1] and iron-uptake systems in *Yersinia* spp.[2], which are required for fitness and survival. However, under adverse conditions they can contribute to pathogenicity. PAIs are associated with virulence functions that are present in some closely related species but not in others. GIs have generally been found to differ significantly in G+C frequency from the average genome G+C frequency. Here, we describe new methods (Box 1) for discerning GIs and PAIs: by the genome signature profile (Box 2); by assessing codon biases from the complete genome (Box 3); and by identifying extremes of amino acid usages in the proteome (Box 4).

PAIs and GIs

Genes that deviate substantially in codon usage from the average gene but that are similar in codon usage to ribosomal protein genes (RPs) tend to be ‘highly expressed’. By these criteria, highly expressed genes in most prokaryotic genomes include RPs, translation and transcription processing factors (TF), and chaperone and degradation protein complexes (CH)[3]. By contrast, a gene is deemed putative alien (pA) if its codon usage difference relative to the average gene in the genome (Box 3) exceeds a high threshold and if its codon usage differences from RPs, chaperone genes and protein synthesis genes are also high. pA genes include: many open reading frames (ORFs) of unknown function (many of which are probably recent lateral transfers); genes acquired via conjugative plasmid integration (e.g. restriction/modification enzymes or neutralizers of antibiotics); cryptic prophages; transposases; insertion sequence (IS) elements; pilus and fimbrial genes; lipopolysaccharide (LPS) biosynthesis genes; and some genes expressed under particular environmental conditions.

PAIs have been identified in, *inter alia*, some *Escherichia coli* strains, *S. typhimurium*, *Vibrio cholerae*, *Yersinia* sp., *Pseudomonas aeruginosa* and *Helicobacter pylori*. Recent reviews have discussed

0966-842X/01/$ – see front matter PII: S0966-842X(01)02079-0

Box 1. Genomic characterizations of anomalous gene regions

Take a sliding window *W* (length 10, 20...50 kb) and evaluate the following:

- **Compositional contrasts (standard method):** compare G+C frequency within *W* to the average genomic G+C frequency.
- **Genome signature contrasts:** compare δ* differences of each window segment to the average genomic signature.
- **Codon usage contrasts:** compare codon biases of the gene set of each window to the average gene codon usages.
- **Amino acid contrasts:** compare amino acid biases of proteins in each window relative to the average proteome amino acid frequencies.
- **Putative alien (pA) gene clusters:** compare differences in codon usages from the RP, TP and CH gene classes, and from the average gene.

the potential mechanisms of bacterial pathogenicity and the acquisition of PAIs by lateral gene transfer[4–7]. These authors mainly describe primary examples of PAIs in enterobacteria located at or near tRNA genes. Actually, tRNA-like elements are common sites for the integration of foreign DNA for prophage lysogenization, which might convert to virulence[5]. In enterobacteria, most PAIs are found adjacent to *selC* (selenocysteine-specific tRNA^{sel})[8]. PAIs of 10–200 kb in length are often flanked by small direct repeats or IS elements and generally encode an integrase that is inserted into the bacterial chromosome[5]. Virulence is often abetted by the absence of specific resident genes, for example the gene encoding the surface protease OmpT, which attenuates pathogenicity[7]. Various major virulence genes are often incorporated into bacterial cells via plasmids and bacteriophages[5]; several PAIs are located on plasmids in *Shigella* and *Yersinia*[8]. A variety of toxins are encoded by bacteriophages, including Shiga toxin in enterobacteria, cholera toxin in *V. cholerae* and neurotoxins in some strains of *Clostridium*[9]. Plasmids carry restriction systems, antibiotic resistance genes, heavy metal cofactors, *nif* (nitrogen fixation) genes and other contingency functions[5–7]. Additionally, microbial genomics has also helped elucidate microbial evolution with respect to similarities and differences among PAIs, disease mechanisms, cellular localization, antigenic variation, host–microorganism interaction,

Box 2. Genome signature comparisons

Every genome has its characteristic signature calculated from genomic sequences[a,b]. Explicitly, the genome signature profile consists of the array of dinucleotide relative abundance values:

$$\left\{\rho^*_{XY} = f^*_{XY}/f^*_X f^*_Y\right\}$$

where f^*_X denotes the frequency of the mononucleotide *X* and f^*_{XY} the frequency of the dinucleotide *XY*, both computed from the sequence concatenated with its inverted complement sequence. These $\{\rho^*_{XY}\}$ (actually $\rho^*_{XY}-1$, termed dinucleotide biases) essentially assess differences between the observed dinucleotide frequencies and those expected from random associations of the component mononucleotide frequencies. Biochemical experiments in the 1960s and 1970s measuring nearest-neighbor frequencies[c–e] established that the $\{\rho^*_{XY}\}$ values present a remarkably stable (ten-component) array for the DNA of an organism. Genomic DNA samples (≥50 kb) from different chromosomal locations of the same genome have approximately constant signatures, and closely related species have more similar genome signatures than distantly related species. From this perspective, the $\{\rho^*_{XY}\}$ array constitutes a genomic signature that is diagnostic and can discriminate among sequences from different organisms[a,f,g].

The mechanisms that generate and maintain the signature are not understood. A reasonable explanation postulates that these are the result of differences in the replication and repair machinery of different species, which either preferentially generate or preferentially select specific dinucleotides in the DNA. These effects might operate through local DNA structures (base-step stacking energies and conformational tendencies), context-dependent mutation rates, methylation and/or other DNA modifications. A measure of genomic signature difference between two sequences *f* and *g* (from different organisms or from different regions of the same genome) is the average absolute dinucleotide relative abundance difference calculated as:

$$\delta^*(f,g) = \frac{1}{16}\sum|\rho^*_{XY}(f) - \rho^*_{XY}(g)|$$

(termed δ* difference),
where the sum extends over all dinucleotides.

We propose that evolutionarily successful DNA integration including cell fusions requires that the genome signatures of the two partners be moderately similar, meaning δ* values ≤0.90 or, at least, weakly similar[h] (δ* ≤1.20). Plasmids (specific or of broad host range) and hosts tend to have moderately similar genome signatures[b]. We speculate that laterally transferred genes evolve rapidly toward the signature of their new host. Concomitantly, we expect generally that G+C content, codon bias and amino acid usage adopt the genome-wide tendencies of these measurements.

References

[a] Karlin, S. (1998) Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr. Opin. Microbiol.* 1, 598–610

[b] Campbell, A. *et al.* (1999) Genome signature comparisons among prokaryote, plasmid and mitochondrial DNA. *Proc. Natl. Acad. Sci. U. S. A.* 96, 9184–9189

[c] Josse, J. *et al*. (1961) Enzymatic synthesis of deoxyribonucleic acid VIII. Frequencies of nearest neighbor base sequences in dexoyribonucleic acid. *J. Biol. Chem*. 263, 864–875

[d] Russell, G.J. *et al.* (1976) Doublet frequency analysis of fractionated vertebrate nuclear DNA. *J. Mol. Biol.* 108, 1–23

[e] Russell, G.J. and Subak-Sharpe, J.H. (1977) Similarity of general designs of protochordates and invertebrates. *Nature* 266, 533–536

[f] Karlin, S. *et al*. (1997) Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol*. 179, 3899–3913

[g] Gentles, A. and Karlin, S. (2001) Genome-scale compositional comparisons in eukaryotes. *Genome Res.* 11, 540–546

[h] Karlin, S. *et al*. (1999) A chimeric prokaryotic ancestry of mitochondria and primitive eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 96, 9190–9195

Box 3. Codon usage contrasts between different gene classes

Let *G* be a family of genes with average codon frequencies $g(x,y,z)$ for the codon nucleotide triplets (x,y,z) normalized so that:

$$\sum_{(x,y,z)=a} g(x,y,z) = 1$$

where the sum extends over all codons translated to amino acid *a*.

Let $f(x,y,z)$ indicate the average codon frequencies for the gene family *F* (*F* can be a single gene *g*) again normalized to 1 in each amino acid codon family. The codon usage difference of the gene family *F* relative to the gene family *G*, termed the codon bias of *F* with respect to *G*, is calculated by the formula:

$$\mathrm{B}(F|G) = \sum_a p_a(F)\left[\sum_{(x,y,z)=a} |f(x,y,z) - g(x,y,z)|\right]$$

where $\{p_a(F)\}$ are the average amino acid frequencies of the gene family *F* (Refs a,b). For the collection of all genes of length at least 100 amino acids, which we designate C, B(F|C) is referred to as the codon bias of the gene set *F* relative to the average gene of the genome.

Let RP denote the collection of all ribosomal protein genes of the genome, CH the major chaperone/degradation genes and TF the principal protein synthesis genes[c]. Formally, a gene, *g*, is considered putative alien (pA) if the biases B(g|RP), B(g|CH), B(g|TF) and B(g|C) all exceed $M + 0.15$, where *M* is the median codon bias of B(g|C) over all genes. The statistical significance of pA gene clusters or clusters of predicted highly expressed genes is identified using the scan algorithms set forth in Ref. d; see also Ref. e on the distribution of uptake signal sequences in the *H. influenzae* genome.

References

a Karlin, S. *et al*. (1998) Codon usages in different gene classes of the *E. coli* genome. *Mol. Microbiol.* 29, 1341–1355

b Karlin, S. *et al.* (1998) Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.* 32, 185–225

c Karlin, S. and Mrázek, J. (2000) Predicted highly expressed genes of diverse prokaryotic genomes. *J. Bacteriol.* 182, 5238–5250

d Karlin, S. and Macken, C. (1991) Some statistical problems in the assessment of inhomogeneities of DNA sequence data. *J. Am. Stat. Assoc.* 86, 27–35

e Karlin, S. *et al*. (1996) Frequent oligonucleotides and peptides of the *Haemophilus influenzae* genome. *Nucleic Acids Res.* 24, 4263–4272

characteristics of virulence factors and information pertinent to drug discovery.

PAIs can differ from the rest of the genome in several ways. Five methods are proposed for detecting PAIs and GIs: (1) G+C content anomalies (the standard method). Take a sliding window of 10, 20...50 kb in length and ascertain the average G+C frequency of each, singling out windows significantly deviant in G+C content. (2) Genomic signature contrasts (dinucleotide bias) (Box 2). The genome signature of each individual window is compared with the average genome signature over all windows of the genome. A highly disparate genome signature might imply a potential PAI. (3) Extremes of codon bias (Box 3). For each window *W*, collect all genes of the window and calculate the codon bias of F_W with respect to the average gene of the genome. An outlying gene with high codon bias could be a potential PAI. (4) Divergence in amino acid usages (Box 4). For window *W* and genes F_W, a window with strongly altered amino acid frequencies from the average could represent a PAI (Box 4). (5) pA gene clusters. Means for discriminating anomalous gene regions in many (but not all) cases are synonymous to identifying clusters of pA genes. For example, pA gene clusters identified using these methods include the Mu prophage region (1559–1593 kb) of *Haemophilus influenzae*, antigenic LPS biosynthesis gene clusters on the surface of many prokaryotic genomes, and the toxin-coregulated pili (*tcp*) gene cluster of *V. cholerae* (Fig. 1).

These five methods are largely independent. The genome signature (i.e. the vector of dinucleotide relative abundances, Box 2) factors out the embedded mononucleotide frequencies. Codon-usage differences and genome signature evaluations are effectively uncorrelated assessments. Codon bias and amino acid differences are essentially complementary measures of gene composition. The methods outlined in Box 1 can be extended to higher-order oligonucleotide contrasts; for example, tetranucleotide relative abundances factor out mono-, di- and trinucleotide abundances, and we can also deal with dicodon biases.

Putative GIs and/or PAIs in diverse bacterial genomes

We have applied the procedures outlined in Box 1 to detect potential PAIs or pA gene clusters in several genomes. Not every pA gene cluster is a PAI and, conversely, not every PAI is a pA gene cluster.

H. pylori

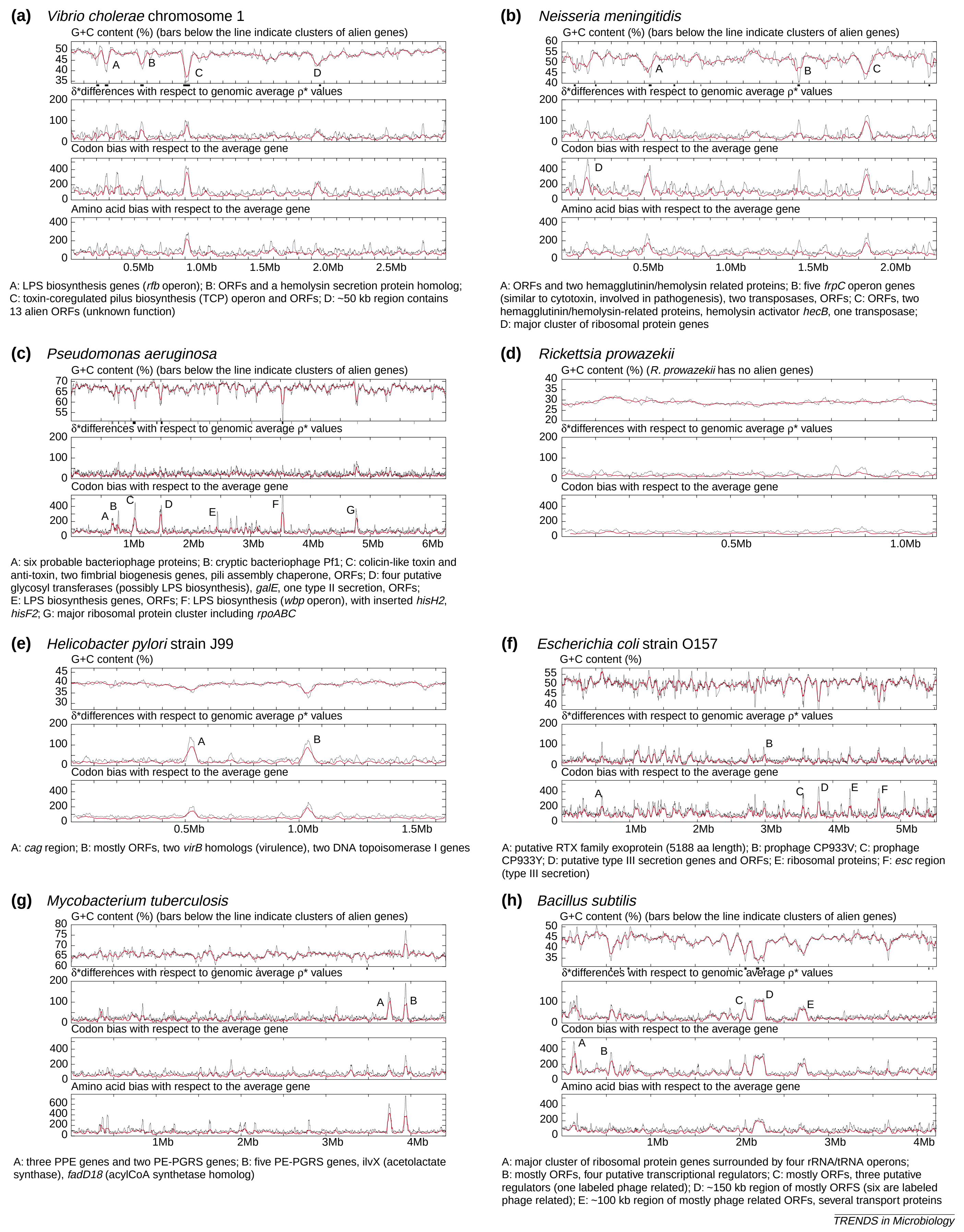
H. pylori is a Gram-negative spiral-shaped bacterium that colonizes the human stomach. Approximately

Box 4. Extremes of amino acid usages

For the genes F_W, we compare the amino acid frequencies of F_W with the average amino acid frequencies of the genome. Windows containing genes with significant deviations in overall amino acid usages might signify a PAI. For two classes of genes, *F* and *G*, the amino acid bias is calculated by the formula:

$$A(F|G) = (1/20)\sum_{i=1}^{20} |a_i(F) - a_i(G)|$$

where $\{a_i(F)\}$ is the average amino acid frequency of a_i in *F*.



(a) A: LPS biosynthesis genes (*rfb* operon); B: ORFs and a hemolysin secretion protein homolog; C: toxin-coregulated pilus biosynthesis (TCP) operon and ORFs; D: ~50 kb region contains 13 alien ORFs (unknown function)

(b) A: ORFs and two hemagglutinin/hemolysin related proteins; B: five *frpC* operon genes (similar to cytotoxin, involved in pathogenesis), two transposases, ORFs; C: ORFs, two hemagglutinin/hemolysin-related proteins, hemolysin activator *hecB*, one transposase; D: major cluster of ribosomal protein genes

(c) A: six probable bacteriophage proteins; B: cryptic bacteriophage Pf1; C: colicin-like toxin and anti-toxin, two fimbrial biogenesis genes, pili assembly chaperone, ORFs; D: four putative glycosyl transferases (possibly LPS biosynthesis), *galE*, one type II secretion, ORFs; E: LPS biosynthesis genes, ORFs; F: LPS biosynthesis (*wbp* operon), with inserted *hisH2*, *hisF2*; G: major ribosomal protein cluster including *rpoABC*

(e) A: *cag* region; B: mostly ORFs, two *virB* homologs (virulence), two DNA topoisomerase I genes

(f) A: putative RTX family exoprotein (5188 aa length); B: prophage CP933V; C: prophage CP933Y; D: putative type III secretion genes and ORFs; E: ribosomal proteins; F: *esc* region (type III secretion)

(g) A: three PPE genes and two PE-PGRS genes; B: five PE-PGRS genes, ilvX (acetolactate synthase), *fadD18* (acylCoA synthetase homolog)

(h) A: major cluster of ribosomal protein genes surrounded by four rRNA/tRNA operons; B: mostly ORFs, four putative transcriptional regulators; C: mostly ORFs, three putative regulators (one labeled phage related); D: ~150 kb region of mostly ORFS (six are labeled phage related); E: ~100 kb region of mostly phage related ORFs, several transport proteins

TRENDS in Microbiology

Fig. 1. The contrasts are displayed for 20 and 50 kb sliding window plots of G+C content, genomic signature profiles, codon usage biases and amino acid anomalies for eight different bacterial species. Abbreviations: LPS, lipopolysaccharide; ORF, open reading frame.

50% of all humans are infected with *H. pylori* but only 10% exhibit clinical diseases including gastric carcinoma and peptic ulcer[10]. Three regions stand out in the G+C plot of the *H. pylori* strain 26659 genome: A (from 445 to 475 kb), B (545–585 kb) and C (1048–1068 kb). Patients with peptic ulcer often express the genes in region B, referred to as the cytotoxin-associated gene (*cagA*) island. The genomes of two *H. pylori* strains, 26695 and J99, have been sequenced in their entirety[11,12]. Virulent *H. pylori* differ from less-virulent strains by the presence of the ~40-kb segment of genes corresponding to the *cagA* locus. The *cagA* region of *H. pylori* deviates the most in genomic signature from the rest of the genome; comparison of the codon bias of the genes in each 40 kb interval with respect to the average *H. pylori* gene shows that genes in the *cagA* region carry the highest codon bias. Strains including this region cause cultured gastric epithelial cells to secrete the proinflammatory cytokine interleukin (IL)-8; this ability is abolished by specific mutation of many of the 26 ORFs found in the PAI (Ref. 13).

The two isolates of *H. pylori* strains 26695 and J99 (Fig. 1e) feature three and two outlying domains, respectively, with a G+C frequency that is approximately 35%, roughly 4% lower than in the rest of the genome. In the J99 genome, this 'plasticity zone' unites regions A and C of the 26695 genome and qualifies as a GI when analyzed by three methods (G+C content, δ* deviants and codon biases). This region was named the plasticity zone as it contains nearly 50% of the genes that are unique to strains 26695 and J99. The plasticity zone contains two *virB* homologs (virulence factors) and two DNA topoisomerase I genes[14].

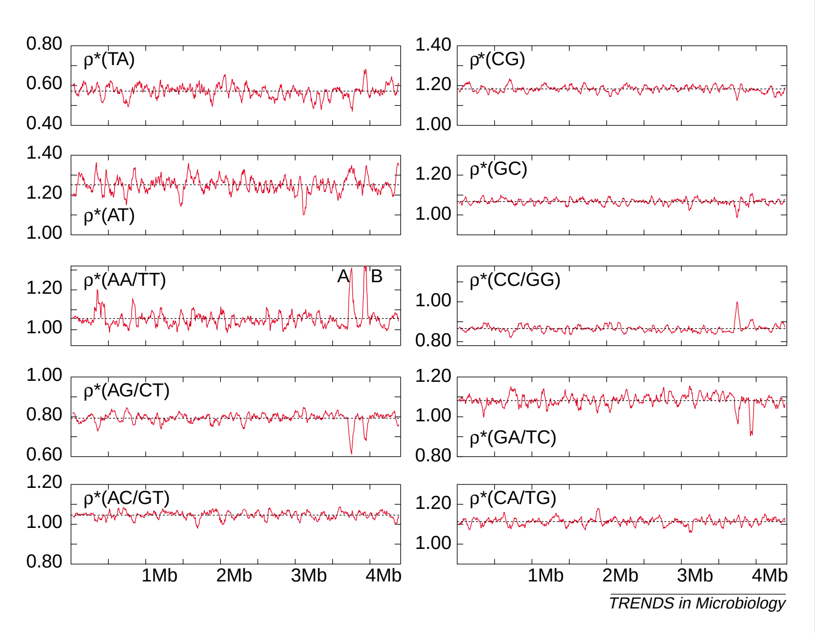


Fig. 2. The ten individual component signature contrasts in *Mycobacterium tuberculosis* over a 50-kb sliding window. The distinguished regions A and B identified earlier are detectable for several individual signature component plots, strengthening confidence in their pathogenicity island (PAI) character. Dotted lines in the plots indicate genomic average ρ* values.

V. cholerae

In the large chromosome of *V. cholerae*[15] a massive pA gene cluster of ~50 kb spans a PAI (C in Fig. 1a) that contains the *tcp* region. *V. cholerae* Tcp is a type IV pilus that is an essential intestinal epithelium adherence factor. Two other pA gene clusters occur at 197–220 kb and 265–295 kb. Four of the methods discussed in Box 1 identify a significant pA cluster (D in Fig. 1a), highlighting 13 alien ORFs. The genes encoding the cholera enterotoxins CtxA and CtxB, both around position 1500 kb in chromosome I, were not detected by our methods but are both pA genes. The receptor for entry of the Ctx prophage into the *V. cholerae* cell is contained within the *tcp* gene conglomerate. Genes encoding potential toxins include secreted hemolysins (region B), proteases and lipases. A gene cluster (covering 17 kb length) contains several LPS biosynthesis genes (region A). The *V. cholerae* large chromosome contains two significantly large segments in antipodal positions:

$$\overline{H}_1 = 43-327\,\text{kb and } \overline{H}_2 = 1657-1985\,\text{kb}$$

devoid of highly expressed genes, and contains intervening regions devoid of pA genes:

$$\overline{A}_1 = 2196-2573\,\text{kb}, \ \overline{A}_2 = 2621-2885\,\text{kb and}$$
$$\overline{A}_3 = 1168-1437\,\text{kb}$$

These indicate that highly expressed and alien genes are irregularly distributed in the *V. cholerae* chromosomes.

Neisseria meningitidis

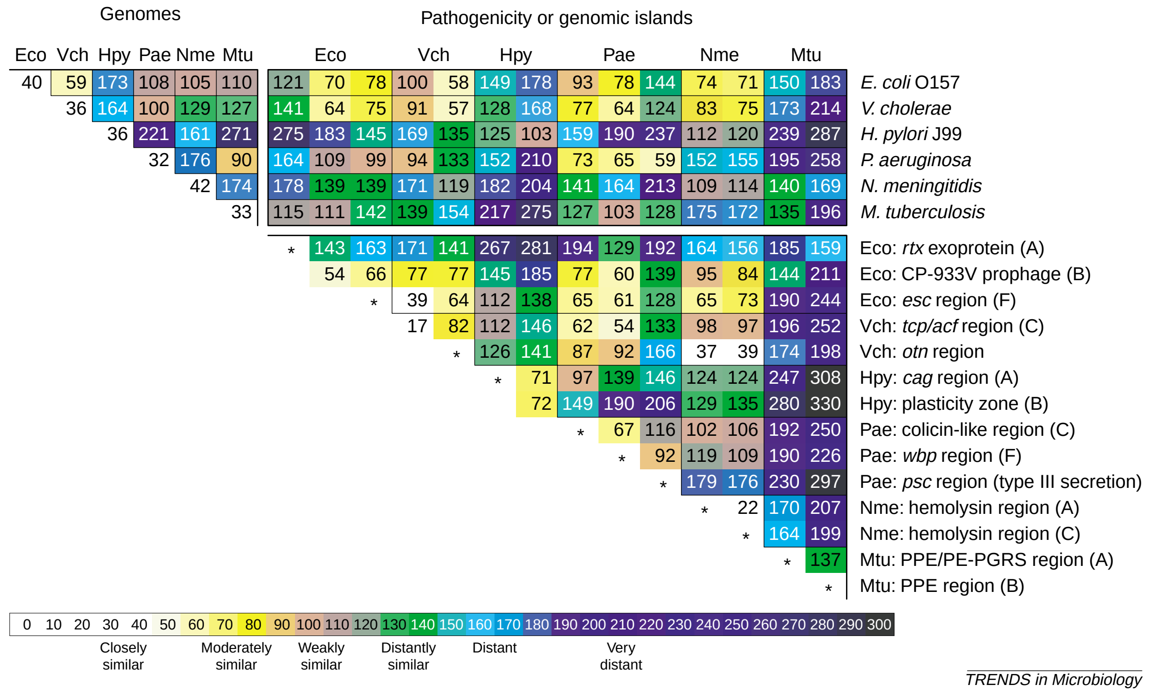
All methods confirm two potential PAI peaks (A and C) in the *N. meningitidis* genome, emphasizing pairings of hemagglutinin and hemolysin-related genes (Fig. 1b)[16].

P. aeruginosa *strain PAO1*

At 6.3 Mb, the genome of this Gram-negative bacterium is the largest completely sequenced to date[17] and is viewed as an opportunistic human pathogen. Examination of its codon bias (Fig. 1c) shows two cryptic regions (A and B) of bacteriophage genes. Peak C corresponds to genes encoding a colicin-like toxin and anti-toxin. Three distinct pA gene clusters (D, E and F) encode clusters of LPS biosynthesis genes (see also Fig. 3).

Mycobacterium tuberculosis

M. tuberculosis, the causative agent of tuberculosis, is an intracellular pathogen of mononuclear phagocytes especially adapted to humans[18]. Genome signature differences and amino acid biases for 50 kb windows of the *M. tuberculosis* genome (Fig. 1f) show two outstanding segments, A (3730–3770 kb) and B (3925–3950 kb). There are around 80 unusual genes labeled PE-PGRS in the *M. tuberculosis* genome[19], which encode a preponderance of glycine–glycine doublets, are significantly low in charged (acidic and basic) residues, aliphatic and aromatic residues, and virtually devoid of cysteine residues, thereby



Genomes						Pathogenicity or genomic islands														
Eco	Vch	Hpy	Pae	Nme	Mtu	Eco			Vch		Hpy		Pae			Nme		Mtu		
40	59	173	108	105	110	121	70	78	100	58	149	178	93	78	144	74	71	150	183	*E. coli* O157
	36	164	100	129	127	141	64	75	91	57	128	168	77	64	124	83	75	173	214	*V. cholerae*
		36	221	161	271	275	183	145	169	135	125	103	159	190	237	112	120	239	287	*H. pylori* J99
			32	176	90	164	109	99	94	133	152	210	73	65	59	152	155	195	258	*P. aeruginosa*
				42	174	178	139	139	171	119	182	204	141	164	213	109	114	140	169	*N. meningitidis*
					33	115	111	142	139	154	217	275	127	103	128	175	172	135	196	*M. tuberculosis*
						*	143	163	171	141	267	281	194	129	192	164	156	185	159	Eco: *rtx* exoprotein (A)
							54	66	77	77	145	185	77	60	139	95	84	144	211	Eco: CP-933V prophage (B)
								*	39	64	112	138	65	61	128	65	73	190	244	Eco: *esc* region (F)
									17	82	112	146	62	54	133	98	97	196	252	Vch: *tcp/acf* region (C)
										*	126	141	87	92	166	37	39	174	198	Vch: *otn* region
											*	71	97	139	146	124	124	247	308	Hpy: *cag* region (A)
												72	149	190	206	129	135	280	330	Hpy: plasticity zone (B)
													*	67	116	102	106	192	250	Pae: colicin-like region (C)
														*	92	119	109	190	226	Pae: *wbp* region (F)
															*	179	176	230	297	Pae: *psc* region (type III secretion)
																*	22	170	207	Nme: hemolysin region (A)
																	*	164	199	Nme: hemolysin region (C)
																		*	137	Mtu: PPE/PE-PGRS region (A)
																			*	Mtu: PPE region (B)

TRENDS in Microbiology

Fig. 3. Average δ* differences among complete genomes and genomic islands based on 30-kb sequence samples. Letters in parentheses next to pathogenicity island descriptions indicate corresponding peaks in Fig. 1. The *Vibrio cholerae otn* region is associated with some pathogenic strains but missing from the current genome. The *psc* region of *Pseudomonas aeruginosa* is not associated with a strong peak in Fig. 1c. Abbreviations: Eco, *Escherichia coli* O157; Hpy, *Helicobacter pylori* J99; Mtu, *Mycobacterium tuberculosis.*; Nme, *Neisseria meningitidis*; Pae, *Pseudomonas aeruginosa*; Vch, *Vibrio cholerae.*

precluding electrostatic, hydrophobic and disulfide-bridge interactions. The two *M. tuberculosis* PAIs mainly contain PE-PGRS genes. There is evidence to suggest that the proteins encoded by these genes, which have anomalous repetitive structures, are surface exposed and can obstruct the host immune system. Only region B significantly deviates in G+C content. To our knowledge, neither A nor B has been investigated as a PAI (Fig. 2).

Obligate intracellular parasites

The *Rickettsia prowazekii*, *Chlamydia trachomatis* and *Chlamydia pneumoniae* genomes, when analyzed by any of the five methods outlined in Box 1, do not contain a PAI or even a pA gene cluster (Fig. 1d). This might be as expected as the host putatively protects these bacteria from lateral gene transfer. Also, by our methods, *Ureaplasma urealyticum* does not contain a pA gene cluster.

E. coli *strainO157:H7*

This bacterium is responsible for bloody diarrhea in many countries. *E. coli* O157:H7 has about 1.34 Mb of DNA that is absent from the laboratory strain K12, which is essentially non-pathogenic, and K12 has about 0.53 Mb that is absent from O157:H7 (Ref. 20). It is thought that most differences are due to lateral transfer. The plots of Fig. 1f show a peak (A) of repeats in toxin (RTX) genes, which are active after post-translational fatty acid acylation and transported to the cell surface[21]. Two prophage regions can be distinguished: CP-933V (B) and CP-933Y (C), and two peaks representing putative type III secretion genes. As expected, the K12 plots show no peaks (not shown).

Statistics of pA gene clusters in several bacterial genomes

As shown in Table 1, the highest percentage of pA genes is detected in *Deinococcus radiodurans* (254; 9% of all its genes) and the highest overall count is in *E. coli* (272; 7%). Strong contrasts in the numbers of pA genes are found when comparing the mycoplasma species *Mycoplasma genitalium* (5; 1%) and *Mycoplasma pneumoniae* (50; 8%). Not surprisingly, *R. prowazekii* is totally devoid of pA genes, consistent with its protected status as an obligate intracellular (human) parasite. Similarly, *C. trachomatis* and *C. pneumoniae* possess only 1% pA genes, but no pA gene cluster. Significant proportions of pA genes are detected in the genomes of *Synechocystis* PCC6803 (7%) and *Pyrococcus horikoshii* (7%). *Borrelia burgdorferi* contains the second-smallest number of pA genes (3; 0.3%). There are no pA transposases or IS genes in small parasitic or archaeal genomes, with the exception of *Archaeoglobus fulgidus*. By contrast, *M. tuberculosis* has 10 pA transposases. Intriguingly, *Bacillus subtilis* appears to have no transposases. Genomes with LPS biosynthesis alien clusters include, to date, *E. coli*, *V. cholerae*, *D. radiodurans*, *P. aeruginosa*, *Synechocystis* spp., *Aquifex aeolicus*, *Thermotoga maritima*, *Methanobacterium thermoautotrophicum*, *Thermoplasma acidophilum*, *P. horikoshii* and *Holobacterium* spp.

Table 1. Some statistics on putative alien genes ≥100 amino acids[a]

Genome	Alien genes				Alien gene clusters of ≥5 genes
	All (%)	Known function[b]	ORFs	Transposon+IS[c]	
E. coli	272 (7%)	94	178	27	10
H. influenzae	60 (4%)	34	26	5	1
V. cholerae	175 (5%)	56	119	2	10
B. subtilis	172 (5%)	59	113	0	8
D. radiodurans	254 (9%)	82	172	25	4
R. prowazekii	0	0	0	0	0
C. trachomatis	9 (1%)	3	6	0	0
C. pneumoniae	14 (1%)	4	10	0	0
M. genitalium	5 (1%)	4	1	0	0
M. pneumoniae	50 (8%)	32	18	0	0
B. burgdorferi	3 (0.3%)	0	3	0	0
T. pallidum	26 (3%)	10	16	0	0
H. pylori	27 (2%)	5	22	0	0
M. tuberculosis	158 (4%)	59	99	10	4
Synechocystis	189 (7%)	78	111	56	4
A. aeolicus	74 (5%)	31	43	0	3
T. maritima	64 (4%)	37	27	1	2
M. jannaschii	28 (2%)	9	19	0	0
M. thermoautotrophicum	77 (5%)	21	56	0	3
A. fulgidus	133 (6%)	19	114	4	4
P. horikoshii	135 (7%)	3	132	0	0
P. abyssi	89 (5%)	17	72	0	0

[a] Abbreviations: IS, insertion sequence; ORF, open reading frame.

[b] Counts of known genes include putative identification.

[c] Transposon and IS counts are included in the counts of known alien genes.

pA gene clusters in E. coli *K12*

The pA genes of *E. coli* include at least four classes of IS and/or transposase families – IS*2*, IS*5*, IS*30* and IS*186* – accounting for a total of 27 genes. These genes feature in >10 clusters, with each cluster including at least four consecutive genes. A four-gene cluster, in one orientation, extending approximately 4 kb over positions 567–571 kb is encoded on the leading strand. This might represent a laterally transferred operon. A cluster of eight genes (of mixed orientations) includes genes similar to those encoding the e14 integrase, the bacteriophage P21 integrase and the phage P22 repressor protein c2. Apparently, these genes arose as prophage incorporations. A pA gene cluster of six genes extending over positions 1416–1427 kb corresponds to the Rac-defective prophage[22]. A contiguous 11 pA gene cluster (2100–2111 kb) features five genes (*rfbX*, *C*, *A*, *D* and *B*) functioning in LPS biosynthesis. A ten pA gene cluster (3791–3804 kb) has similarities to the surface antigen LPS core biosynthesis or glucosyltransferase proteins carrying the designations *rfaL*, *K*, *Z*, *Y*, *J*, *I*, *B*, *S* and *P*. Another pA cluster of eight genes (4311–4321 kb), encoded from the lagging strand, encodes the PhnP, M, L, K, J, F and E proteins involved in alkylphosphonate uptake. There are several pA genes encoding restriction-modification enzymes, namely *merA*, *merC* and *hsdS*. Other pA genes have been identified by similarity to transposons of the Tn*2501* resolvase family.

B. subtilis *pA genes*

The pA genes of *B. subtilis* feature contiguous genes similar to DNA-methyltransferase (cytosine specific) genes, extending over positions 654–661 kb: a DNA restriction enzyme, a DNA-entry nuclease (*nucA*) and a pA two-component response regulator (*citT*). Several *B. subtilis* pA genes are similar to some encoding various ATP-binding cassette (ABC) transporters of glutamine, of dipeptides, unspecified transporters and the spore coat protein B (*cotB*).

H. influenzae *pA genes*

In *H. influenzae*, the 60 pA genes detected from the 1529 genes[23] (lengths ≥100 amino acids) include 26 ORFs of unknown function and 34 genes of assigned function[24]. Eleven genes found in the region spanning 1557–1593 kb contain deposits of a cryptic Mu-like prophage. This region is detectable by all five methods and by the absence of the uptake signal sequence AAGTGCGGT, which is abundant in the rest of the *H. influenzae* genome[25].

Putative alien genes in *H. influenzae* include *Hin*dIII, *Hin*cII and *Hgi*DI, genes of the restriction-modification systems that might have been acquired through plasmid integration. Bracketing these restrictase/methylase genes are homologs of cryptic phages incorporating fragments of phages λ and P22 (whose natural habitats are enterobacteria), Mu phage, the L5 phage of *Mycobacterium smegmatis* and phages resident in other bacteria that are capable of invading *H. influenzae*[26]. These phage fragments flank the Trp operon. pA genes are present in two significant clusters in the region of the defective prophage Mu. In addition to these clusters, there are five pairs of overlapping genes, including three restriction systems, the *thiM* and *thiD* genes involved in thiamin biosynthesis and the hypothetical ABC transporter HI0354, which is adjacent to the hypothetical protein HI0355.

D. radiodurans *pA genes*

Among the pA genes of *D. radiodurans*, a cluster of enzymes contributing to LPS biosynthesis stands out. LPS biosynthesis proteins are replete with repeat structures that could help confound host immune responses. Six Par proteins (which help in cell partitioning) are encoded by pA genes and three pA pilin type IV genes (DR0548, DR1232 and DR1233) have been identified. In *D. radiodurans*, we can identify 41 individual transposases (nine families) of which 28 qualify as pA genes[27]. The total number of pA genes in *D. radiodurans* is 282 (9% of all genes), of which 190 are ORFs of unknown function. A single extensive pA gene cluster on chromosome I consisting almost exclusively of ORF genes extends 27 kb and covers positions 527 kb to 554 kb (DR0526–DR0548). This region behaves as a GI.

Signature differences among PAIs and their hosts

The average differences between 30 kb sequence samples of several genomes and PAIs of the same host and different PAIs are given in Fig. 3. It is striking that the genome signature differences of potential PAI regions and their host DNA are generally moderately similar to each other (the same degree of similarity as between plasmids and their hosts[28]). This similarity suggests that relatively close genome signatures promote PAI compatibility or that during their residence, the PAIs have changed relatively rapidly to their hosts' signature.

This analysis also indicates that the PAI sequences of *E. coli* O157:H7 and the two PAIs of *V. cholerae* are moderately similar in their genome signature (Fig. 3). Actually, the *E. coli* O157:H7 sequence and the Tcp gene cluster are remarkably close, $\delta^* = 39$. Strikingly, the genome signature of the PAI potential regions of *P. aeruginosa* is moderately similar, $\delta^* = 59–73$, to its host and to *E. coli* or *V. cholerae* but clearly distant or very distant from the other proteobacterial genomes. Similarly, the PAI *cag* region of *H. pylori* is weakly similar, $\delta^* = 103$, to its host. The two PAIs of *N. meningitidis* are close to the *V. cholerae* Otn region. The possible PAIs of *M. tuberculosis* are distant or very distant from all PAIs.

Concluding comments

Microbial genomes are substantially homogeneous in G+C content and genome signature[29] but the composition of GIs differs, albeit not strongly, from that of the overall genome. Five criteria for detecting GIs and PAIs are highlighted in Box 1. When a genomic segment deviates sufficiently from the average genome composition in G+C, and/or genome signature (δ* distances), and/or codon usage, and/or amino acid frequencies and/or contains a cluster of pA genes, we view this gene segment as a GI, putatively derived by lateral gene transfer. If the same segment is identified using all five criteria we are more confident in the PAI status or GI character of that segment.

Current studies of molecular evolution underscore lateral gene transfer as a major evolutionary mechanism[7,30–33]. For some genomes, it is estimated that 10–20% of genes have been laterally transferred within the last 100 million years[34]. Established vehicles for lateral transfer include transposition, conjugative plasmids, phage hitchhiking, free DNA and natural transformation. Ochman *et al.*[7] emphasize the dynamic character of genomes in balancing gene acquisition with gene loss. Accordingly, lateral gene transfer provides a means of bacterial diversification[7]. One would expect that genes beneficial to the recipient organism for its defense or in furnishing nutrients and metabolic addenda have potential for successful interspecies gene flow. However, successful incorporations are rare. We have previously verified that plasmids almost always have a genome signature moderately similar to that of any potential host[28]. I propose the following hypothesis. Genome signature similarity is essential for compatibility and transfer between two genomes or between gene domains and a host genome. In a parallel manner, we expect GIs to be close or moderately similar in genome signature to the host genome signature.

One of the requirements for the successful acquisition of a laterally transferred gene is its utility to the recipient organism. For example, the heat shock (HSP60) proteins (GroEL and thermosomes) facilitate folding of a great variety of proteins and act on substrates whose selection seems to be solely constrained by their size[35]. The non-specificity of HSP60 proteins relative to their targets makes them potential gene acquisitions. Along these lines, α-proteobacteria possess multiple features that suggest they can be donors of HSP60 genes: (1) they possess unusual facility for HSP60 gene duplication and transposition, as indicated by a plethora of HSP60 sequence replicates among classical α-proteobacteria, in contrast to the absence of paralogs of HSP60 sequences in the γ, β, δ and ε proteobacterial clades[36]; (2) α-proteobacteria establish close spatial and functional relationships, often endosymbiotic, with plant eukaryotic organisms[37]. Lateral gene transfer provides a means of rapid response to strong selection pressures induced by virulence factors ranging from toxin production to immune evasion. Examples include Ti plasmids of *Agrobacterium tumefaciens*, nodulation plasmids of *Rhizobium* and virulence plasmids of *Shigella* and *Yersinia*; (3) copies of α-proteobacterial HSP60 genes have been found to reside on extra chromosomal elements, for example the megaplasmids pSyma or *Rhizobium meliloti*.

Ochman *et al.*[7] describe examples of gene clusters acquired by lateral transfer that are near a host promoter and are expressed in the recipient cells. Although most PAIs are probably laterally transferred and have a compatible genome signature, it is not necessary that every GI is the same.

Acknowledgements
I gratefully acknowledge Drs J. Mrázek, A. Campbell and B.E. Blaisdell of Stanford and H. Shuman of Columbia University Microbiology for valuable discussions. Supported in part by NIH grants 5R01GM10452-35 and 5R01HG00335-12.

References

1 Lawrence, J.G. and Roth, J.G. (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* 143, 1843–1860

2 Rakine, A. *et al*. (1999) The high pathogenicity island of *Yersiniae*. In *Pathogenicity Islands and Other Mobile Virulence Elements* (Kaper, J.B. and Hacker, J., eds), pp. 77–90, ASM Press

3 Karlin, S. and Mrázek, J. (2000) Predicted highly expressed genes of diverse prokaryotic genomes. *J. Bacteriol.* 182, 5238–5250

4 Finley, B.B. and Falkow, S. (1997) Common themes in microbial pathogenicity II. *Mol. Biol. Microbiol. Rev.* 61, 136–169

5 Hacker, J. and Kaper, J. (2000) Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol*. 54, 641–679

6 Kaper, J.B. and Hacker, J., eds (1999) *Pathogenicity Islands and Other Mobile Virulence Elements*, ASM Press

7 Ochman, H. *et al.* (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* 405, 299–304

8 Parsot, C. and Sansonetti, P.J. (1999) The virulence plasmid of *Shigellae*: archipelagoe of pathogenicity islands. In *Pathogenicity Islands and Other Mobile Virulence Elements* (Kaper, J.B. and Hacker, J., eds), pp. 151–165, ASM Press

9 Iriarte, M. and Cornelis, G.R. (1999) The 70 kb virulence plasmid of *Yersiniae*. In *Pathogenicity Islands and Other Mobile Virulence Elements* (Kaper, J.B. and Hacker, J., eds), pp. 91–126, ASM Press

10 Blaser, M.J. and Parsonnet, J. (1994) Parasitism by the 'slow' bacterium *Helicobacter pylori* leads to altered gastric homeostasis and neoplasia. *J. Clin. Invest*. 94, 4–8

11 Tomb, J.F. *et al.* (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* 388, 539–547

12 Alm, R.A. *et al*. (1999) Genomic sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* 397, 176–180

13 Censini, S. *et al*. (1996) A *cag*, a pathogenicity island of *Helicobacter pylori*, encodes type I-specific and disease-associated virulence factors. *Proc. Natl. Acad. Sci. U. S. A.* 93, 14648–14653

14 Liu, G. *et al.* (1999) Sequence anomalies in the *cag7* gene of the *Helicobacter pylori* pathogenicity island. *Proc. Natl. Acad. Sci. U. S. A.* 96, 7011–7016

15 Heidelberg, J.F. *et al*. (2000) DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* 406, 477–483

16 Tettelin, H. *et al*. (2000) Complete genome sequence of *Neisseria meningitidis* Serogroup B Strain MC58. *Science* 287, 1809–1815

17 Stover, C.K. *et al.* (2000) Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature* 406, 959–964

18 Orme, I.M. (1998) The immunopathogenesis of tuberculosis: a new working hypothesis. *Trends Microbiol*. 6, 94–97

19 Cole, S.T. *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393, 537–544

20 Perna, N.T. *et al*. (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* 409, 529–533

21 Lally, E.T. *et al*. (1999) The interaction between RTX toxins and target cells. *Trends Microbiol*. 7, 356–361

22 Snyder, L. and Champness, W. (1997) *Molecular Genetics of Bacteria*, ASM Press

23 Fleischmann, R.D. *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512

24 Mrázek, J. and Karlin, S. (1999) Detecting alien genes in bacterial genomes*. Ann. New York Acad. Sci.* 870, 314–329

25 Karlin, S. *et al*. (1996) Frequent oligonucleotides and peptides of the *Haemophilus influenzae* genome. *Nucleic Acids Res.* 24, 4263–4272

26 Hendrix, R.W. *et al*. (1999) Evolutionary relationships among diverse bacteriophage and prophage: all the world's a phage. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2192–2197

27 Karlin, S. and Mrázek, J. (2001) Predicted highly expressed and putative alien genes of *Deinococcus radiodurans* and implications for resistance to ionizing radiation damage. *Proc. Natl. Acad. Sci. U. S. A.* 98, 5240–5245

28 Campbell, A. *et al.* (1999) Genome signature comparisons among prokaryote, plasmid and mitochondrial DNA. *Proc. Natl. Acad. Sci. U. S. A.* 96, 9184–9189

29 Karlin, S. (1998) Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr. Opin. Microbiol.* 1, 598–610

30 Campbell, A.M. (2000) Lateral gene transfer in prokaryotes. *Theor. Popul. Biol.* 57, 71–77

31 Brocchieri, L. (2001) Phylogenetic inferences from molecular sequences: review and critique. *Theor. Popul. Biol.* 59, 27–40

32 de la Cruz, I. and Davies, I. (2000) Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol.* 8, 128–133

33 Doolittle, W.F. (1998) You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet.* 14, 307–311

34 Lawence, J.G. and Ochman, H. (1998) Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. U. S. A.* 95, 9413–9417

35 Houry, W.A. *et al.* (1999) Identification of *in vivo* substrates of the chaperonin GroEL. *Nature* 402, 147–154

36 Gupta, R.S. (1998) Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. *Microbiol. Mol. Biol. Rev.* 62, 1435–1491

37 Ogawa, J. and Long, R. (1995) The *Rhizobium meliloti* groELc locus is required for regulation of early *nod* genes by the transcription activator NodD. *Genes Dev.* 9, 714–729