REDEFINING BACTERIAL POPULATIONS: A POST-GENOMIC REFORMATION

Elizabeth A. Joyce, Kaman Chan*, Nina R. Salama‡ and Stanley Falkow**

Sexual reproduction and recombination are essential for the survival of most eukaryotic populations. Until recently, the impact of these processes on the structure of bacterial populations has been largely overlooked. The advent of large-scale whole-genome sequencing and the concomitant development of molecular tools, such as microarray technology, facilitate the sensitive detection of recombination events in bacteria. These techniques are revealing that bacterial populations are comprised of isolates that show a surprisingly wide spectrum of genetic diversity at the DNA level. Our new awareness of this genetic diversity is increasing our understanding of population structures and of how these affect host–pathogen relationships.

Delimiting unambiguous boundaries that define pathogenic bacterial populations is a daunting task. Current methodologies for doing so have given rise to a great deal of confusion, as pathogenic bacteria often have varying abilities to cause infection and disease. To better understand the origin and spread of bacterial disease, it is imperative to determine the basis for this apparent discrepancy and to define the relationships that exist between virulent isolates and their non-pathogenic counterparts.

Until the middle of the twentieth century, the greatest advances in the scientific study of microorganisms coincided with the development of more powerful microscopes and the ability to obtain bacterial isolates in pure culture. These techniques provided the fundamental basis for studying and describing microbes. Historically, bacteria have been systematically classified into genus and species categories on the basis of their microscopic and physiological attributes. However, despite the myriad phenotypes that are used to classify bacteria, there are numerous examples of isolates in a single species that show considerable variation in metabolic capacity, adaptation to ecological niches, antigenic variation and resistance to noxious compounds, which underscores the difficulty of having a classification based on 'phenotypically fuzzy' traits. A particularly thorny example of this is the identification of infectious disease agents, in which different isolates from the same bacterial species can show significant variation in their ability to promote and to cause infection and disease. The difficulty in discriminating between bacteria arises from grouping bacteria on the basis of phenotypic characteristics, which does not always reflect the evolutionary history or genetic relatedness of the organisms in a given group (for an extensive review, see REF. 1). The success of this kind of typing requires that bacteria reproduce in a strictly asexual manner. In this way, a set of phenotypic characteristics would be propagated vertically from mother to daughter cell, which establishes a population with a discrete lineage. However, it is becoming increasingly clear that most bacterial populations are also subject to the complex processes of genome diversification that involve horizontal transmission of genetic material (BOX 1). The shuffling and sharing of genetic information among bacteria can profoundly affect the structure of a bacterial population. In the absence of the diversifying influences of horizontal influx and efflux of DNA, the chromosomal variations observed in one isolate (such as those due to point mutations) are linked, and will be maintained and propagated in all isolates. At the other extreme, high levels of horizontal DNA acquisition and loss lead to an increase in genetic complexity and diversity at the population level, and

**Department of Microbiology and Immunology, Stanford University School of Medicine, 299 Campus Drive, Fairchild D 037, Stanford, California 94305-5402, USA.*
‡Division of Human Biology, Frederick Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, USA.
Correspondence to E.A.J.
e-mail: ejoyce@stanford.edu
doi:10.1038/nrg820

CONJUGATION
The transfer of DNA from a donor cell to a recipient cell that is mediated by direct cell–cell contact.

TRANSFORMATION
The uptake of DNA by a bacterium from the surrounding environment.

TRANSDUCTION
Virus- or phage-mediated introduction into a cell of a DNA fragment that is derived from a different cell.

ARCHAEA
A kingdom of unicellular microorganisms, many members of which can survive extreme environmental conditions, such as temperatures >100 °C, extremely alkaline or acid environs, and highly osmotic conditions.

RIBOTYPING
A technique used to determine genetic and evolutionary relationships between organisms. Oligonucleotide probes targeted to highly conserved domains of coding sequences of ribosomal RNA are amplified and the products are visualized by gel electrophoresis banding patterns are compared with known species and strains to determine organism relatedness.

continuous chromosomal resorting at the level of individual bacterial chromosomes. This results in the unlinked segregation of alleles at different loci in the population. As Feil and Spratt aptly note, "this blurring of gene-pool boundaries has significant long-term evolutionary consequences, most notably the acquisition of novel virulence determinants or metabolic properties that can result in a major shift in the pathogenicity or ecological niche of a subset of a bacterial population"[2].

The advent of whole-genome sequencing and comparative sequence analyses supports and extends the growing body of evidence for intraspecies genetic variability, and has made it increasingly clear that variation mediated by horizontal gene transfer has an important impact on the structure of bacterial populations. But, it is the development and use of adjunct genomic tools, such as DNA microarray technology, that are providing compelling evidence for the extent of this genetic diversity. Therefore, to accurately characterize bacteria into phylogenetically congruent groups, we are challenged to detect and assess the extent of the genetic heterogeneity that exists in a bacterial species. To fully appreciate the significance of this diversity, this information about population structure needs to be incorporated into the existing microbial epidemiology and evolutionary data. Here, we show the power of microarray technology to assess the extent of genetic variation in bacterial populations and offer our insight into its biological significance by discussing the results of several recent studies for a select set of organisms. We propose that the fine level of genetic detail that is provided by microarray analysis will help to redefine the boundaries that are used to distinguish bacterial populations.

Box 1 | **Horizontal gene transfer**

Horizontal gene transfer, or the acquisition of exogenous genetic material and its subsequent stable incorporation into a recipient genome, has been, and continues to be, a central force that drives bacterial evolution. The lateral movement of genes between bacterial genomes in nature was not recognized until the 1950s, when drug-resistant bacterial pathogens first emerged[95]. The significance of this genetic mobilization was unrealized because of the commonly held belief that horizontal gene transfer was the exception rather than the rule, and was relegated to a subset of genes with products that conferred a selective survival advantage, such as antibiotic resistance. In fact, there is a growing body of evidence that the type of gene that can be delivered by these mechanisms is not limited to drug-resistance determinants, but also includes genes that encode core cellular functions, such as HMGCoA (high-mobility group coenzyme A) reductase, glutamine synthetase, Hsp70, ATPases and aminoacyl-transfer RNA synthetases[96–100].

In bacteria, genetic transfer can occur by three distinct processes: CONJUGATION, TRANSFORMATION and TRANSDUCTION. The frequency with which these events occur, which varies considerably among bacterial species, influences the magnitude of the genetic impact of horizontal gene transfer. Gene-transfer events have been revealed through analyses of genome sequences, which differ in guanine and cytosine (G+C) content and codon usage at chromosomal locations that have recently acquired foreign DNA. On the basis of a comparative G+C composition analysis between *Escherichia coli* and *Salmonella*, Lawrence and Ochman concluded that 18% of *E. coli K12* genes have been acquired by horizontal transfer in the past 100 million years[101]. Results from a genome-sequence analysis of the bacterial hyperthermophile *Thermotoga maritima* indicate that 25% of this genome might have been horizontally acquired from ARCHAEAL hyperthermophiles[102] (for more information, see REFS 99,103,104).

Molecular 'genomotyping'

Many molecular typing techniques have been developed during the past 30 years to examine the degree of genetic diversity in bacteria (reviewed in REFS 1,2). The results of DNA fingerprinting, RIBOTYPING and restriction-fragment length polymorphism (RFLP) analysis have been used as coarse measures of change in the DNA sequence between strains. Another technique — multilocus enzyme electrophoresis (MLEE) — offers a way of indirectly examining genomic diversity at several housekeeping loci[3,4] and can detect even slight variations at the amino-acid level. Analysis of several proteins from each bacterial isolate produces a highly specific allelic variation profile, which establishes a basis for comparing strains and provides an estimate of the extent to which alleles at various loci are linked. These tools have been used effectively to examine the genetic relatedness of isolates from several bacterial groups[5–14]. Although these techniques give qualitative information about changes at the DNA-sequence level, a more precise picture of the actual differences between bacterial strains requires knowledge of the nucleotide sequence itself. With improvements in sequencing technology in the mid 1970s, Carl Woese and collaborators promoted the comparison of small subunit ribosomal RNA (SSU rRNA) sequences as the molecular standard for measuring the degree of genetic relatedness between organisms. The use of this molecule, which is universally conserved in structure and function across kingdoms, has been very successful in determining phylogenetic relationships[15] (reviewed in REF. 16). More recently, multilocus sequence typing (MLST; see online links box) was developed as a high-resolution method for indexing genetic variation[17,18]. Using this technique, a specific genetic profile for a given strain can be generated using 500 bp of nucleotide sequence from as few as seven housekeeping genes. MLST has proven to be an invaluable tool for typing bacteria, and large databases of MLST-derived data for several microbes are available[17,18] (see online links box). Both SSU rRNA and MLST are highly discriminatory and quantitative, as the loci under scrutiny are defined directly by their DNA sequence. However, only a limited number of loci are typically examined using these techniques, which gives neither a sense of genome-wide variation nor of the full impact of horizontal gene transfer.

Not until the advent of high-throughput genome sequencing and genome annotation could we truly appreciate the extent of genetic complexity and diversity among the Prokarya. So far, the genomes of more than 75 bacterial species have been sequenced (see the link to the Genome Entry Database at NCBI). For several of these species, such as *Helicobacter pylori*, *Streptococcus pneumoniae* and *Salmonella enterica*, the genome sequences of multiple strains have been determined, thereby launching the field of comparative bacterial genomics. The resulting plethora of sequence data has revealed unforeseen details about the putative functional similarities, as well as differences, among bacteria. It has also allowed the precise identification of genetic differences between isolates of the same species. Our

ATROPHIC GASTRITIS
Chronic inflammation of the stomach, accompanied by atrophy of the mucous membrane and destruction of the peptic glands.

DUODENUM
The first portion of the small intestine, extending from the pylorus (the posterior end of the stomach) to the jejunum (the next portion of the small intestine).

ADENOCARCINOMA
A form of malignant cancer that arises from the glandular epithelium.

SEROVAR/SEROTYPE
A group of intimately related microorganisms distinguished by a common set of antigenic determinants that are expressed on the cell surface.

understanding of the genetic diversity in, and the evolution of, bacteria has been transformed by several recent genome comparison studies[20–26].

The improvements in sequencing technology have fuelled the concomitant development of analytical tools that take advantage of the extensive genome-sequence data. One such innovation that holds great promise for dissecting and defining genetic variability on a genome-wide scale is DNA-array technology[27]. Conceptually, array technology shares the principles that underlie Southern and northern blot analysis. Both DNA and oligonucleotide microarrays are physically ordered collections of generally known DNA sequences that are immobilized on a glass surface (microarrays) or on nylon membranes. When hybridized with pools of differentially labelled RNA or DNA, these sequences serve as highly specific probes that capture complementary nucleotides present in the hybridization solution. The resulting hybridization signal at each spot reflects the relative levels of gene expression if the array is hybridized with labelled mRNA. If the array is hybridized with a sample that is generated from genomic DNA, the signal indicates the presence of a homologous DNA sequence in the hybridized sample. So, this technology is highly versatile, which allows genome-wide examination of gene expression and determination of the genetic content relative to the arrayed reference DNA. Microarrays have already proven invaluable for the study of gene expression in biological systems that were previously thought to be intractable (reviewed in REFS 28–32). However, there is also a growing body of literature that describes using DNA microarrays for genomic comparisons of whole bacterial genomes (see below)[33–48]. The ability to consider the whole genome in this kind of analysis increases the likelihood of identifying genetic variability, including the probability of tracking horizontal gene transfer. Importantly, as many probes with different fluorescent labels can be hybridized at the same time, several bacterial isolates can be examined simultaneously. This provides an excellent method for identifying and estimating the number of strain-specific genes. Before the development of microarray technology, this could only have been achieved by sequencing several isolates (a laborious, time-consuming and expensive alternative). As an adjunct to genome sequencing, microarray analysis is providing valuable genetic detail that can be used to define bacterial populations more accurately. To illustrate the power of this technology in discerning genome-wide variability, we discuss the work on *H. pylori*, *S. pneumoniae* and the species and SEROVARS of the *Salmonella* genus — for example, *S. enterica* and *S. bongori*, the genetic histories of which have been affected differently by horizontal gene transfer. We discuss how the results of these microarray studies are influencing the choices for establishing boundaries that define bacterial populations and offer our critical perspective on the promise of a more in-depth understanding of the evolutionary complexities that are apparent in bacterial populations.

Box 2 | ***Helicobacter pylori***

Infection with the gastric pathogen *Helicobacter pylori* is associated with a broad spectrum of clinical outcomes including ATROPHIC GASTRITIS, DUODENAL or gastric ulceration, gastric ADENOCARCINOMA and mucosa-associated lymphoma. Although the existence of spiral-shaped organisms in the human stomach was first described in the late 1800s[105], it was not until the early 1980s that *H. pylori* was successfully cultured[106]. Marshall's finding took the research community by surprise. The stomach was previously thought to be a sterile environment owing to its high acidity, so reports of organisms at this site had been dismissed as transients that enter with contaminated food. Additionally, the spiral organisms described resembled species of *Campylobacter*, which are fastidious organisms that require a specific growth media and microaerobic growth conditions, but which grow quite quickly, particularly at high temperature (42 °C). *H. pylori*, conversely, does not grow at 42 °C and only forms colonies after 3–5 days. So, the first successful culture came about fortuitously, after leaving the plates out over a long holiday weekend.

Partly due to its recent isolation, the way in which we describe *H. pylori* — by the presence or absence of genetic elements such as the *cag* pathogenicity island — reflects the more recent widespread use of genetic and molecular tools for microbial characterization and classification (recently reviewed in REF. 107). The outcome of the more recent studies has led not only to an increased awareness of the extent of genetic diversity that can be tolerated in a bacterial species, but also to new ways of exploring and thinking about the relationships between genomic heterogeneity and epidemiology and disease. The image is reproduced with permission from REF. 108 © (2001) Society for General Microbiology.

Strain-specific genes: prevalence and disease

Helicobacter pylori. The most severe *H. pylori*-mediated disease states[49,50] can be attributed to infection with strains that have both a gene that encodes a cytotoxin, *vacA*[51,52], and the *cag* pathogenicity island (*cag* PAI)[53], which is a 40-kb DNA element that encodes a bacterial type IV secretion system that translocates the bacterial protein CagA into host cells[54–57] (BOX 2). The results of DNA fingerprinting and RFLP analysis, which were carried out to identify other genetic elements that correlate with more severe disease, indicated extensive variation at the DNA-sequence level among strains[58,59]. Although these methods reveal sequence variation on a genome-wide level, they do not reveal the identity or location of such variable elements on the chromosome, so the relevance of this variation could not be fully appreciated. It was not until the publication of the genome sequences of two *cag* PAI-containing clinical isolates of *H. pylori* (strains 26695 and J99)[60–62] that scientists were able to analyse rigorously the extent and significance of DNA variation on a genome-wide scale. Analysis of these two genomic blueprints shows that most genes are highly conserved between the two strains[60]. Surprisingly, for each strain, 6% of all genes were specific to one strain and absent from the other (73 genes in J99 and 105 genes in 26695). These strain-specific genes tend to be located at two highly polymorphic locations — plasticity zones —on each chromosome. In both genomes,

these zones contain a remarkable number of restriction-modification genes — transposases — and many strain-specific genes. The ability to compare two genome sequences has provided a tremendous and perhaps unexpected opportunity to identify a core set of genes, which are detected in all strains, that begin to define the biology of *H. pylori*. It has also allowed the identification of new genetic elements, such as strain-specific loci, that might allow adaptation to specific hosts or relate to the different virulence potential of individual strains.

In our own lab, we capitalized on these findings by designing and constructing a whole-genome DNA microarray that was specific for genes that are present in both of the sequenced strains of *H. pylori* [34]. We examined the genome composition of 15 clinical isolates of *H. pylori* that either contain or lack the *cag* PAI gene by using labelled probes that are based on the genomic DNA of each clinical isolate in comparative hybridizations. This work extends the previous genome-sequence analysis in two important ways. First, this analysis of genetic diversity identified 362 genes (22% of all *H. pylori* genes) that are not conserved among all 15 strains, including 189 genes that are present in the two sequenced strains. Although some genes share significant homology with other bacterial genes found in the NCBI GenBank database (7%), most of these coding sequences have no predicted function and 58% have no database homology. In addition to revealing the many strain-specific genes that are found in the chromosome of *H. pylori*, this analysis provides a more precise picture of the complement of genes that is likely to define the core of the *H. pylori* genome than was originally imagined on the basis of comparing the 26695 and J99 sequences.

Second, and perhaps more compelling, our analysis allowed us to identify gene groups that seem to co-vary and are likely to be co-inherited, as well as a set of genes that anti-vary and tend not to be co-inherited. Because there is strong selective pressure for co-inheritance of genes, the products of which are involved in common pathways, this observation has important implications for their possible function[63]. To determine if any of the 362 strain-specific genes were co-inherited, we used a hierarchical CLUSTER program (see online links box) to group genes on the basis of their absence or presence among all 15 strains. As expected, the genes that constitute the *cag* PAI locus were found to cluster together. Intriguingly, ten other genetically unlinked genes seem to be co-inherited with the *cag* PAI. One of them, *babA* (Hp1243), which is a virulence factor that was identified in strains that cause more severe disease[64], has been previously shown to co-vary with the *cag* PAI[65]. There were also four open reading frames (ORFs) that tended not to be co-inherited with the *cag* PAI; two of them are predicted to code for restriction-modification proteins, which are used by bacteria to protect against invasion by foreign DNA. The mechanism for these patterns of co-inheritance is unknown, but might imply a stepwise process in the evolution of more (or less) virulent strains. The strong selective pressures of the host environment undoubtedly influence whether such genetically unlinked strain-specific genes are maintained or lost according to how their expression affects the fitness of the organism in its niche.

The above finding leads to an exciting hypothesis: it might be possible to identify genetic variation that contributes to the spectrum of clinical diseases that is seen in different patients. The initial study involved 15 isolates for which little or no clinical information was available, which makes it impossible to test any such correlation. Analysis of more strains has begun to reveal patterns of strain similarity on the basis of gene composition (FIG. 1). Interestingly, different isolates from the same patient or strains that were re-isolated after infection of a new host had measurable alterations at the whole-gene level. Nonetheless, strains that originated from the same patient were more similar to each other than strains from different patients. Additionally, all strains of *H. pylori* that were isolated from rhesus monkeys fell into a single group. Strains that caused the same clinical symptoms did not fall into a single cluster, perhaps owing to incomplete clinical information, imprecise diagnosis (for example, gastric ulcer versus duodenal ulcer) or a lack of strains from patients with no overt disease.

Proving the above hypothesis will require testing a large collection of strains for which information on the pathogenic processes and their clinical outcome is known. These data will then need to be compared with the data from strains taken from asymptomatic individuals.

The impact of genetic diversity in *H. pylori* on disease development and severity was explored in greater depth in two other studies[38,41,42]. Israel *et al.*[42] examined two strains, with similar genotypic markers of virulence (*vacA*$^+$ and *cagA*$^+$), but isolated from two patients: one suffering from a gastric ulcer and one from a duodenal ulcer. These two disease states, which result from distinct types of gastric inflammation, are associated with different cancer risks — patients who suffer from gastric ulcers are far more likely to develop gastric cancer than patients with duodenal ulcers[66]. To determine whether the genetic attributes of the infecting strains contributed to disease outcome, Israel *et al.* used two models of *H. pylori* infection — an *in vivo* gerbil model and an *in vitro* cell-culture model — to determine if the strains elicited different cellular responses. The two strains did, indeed, vary in their ability to induce severe inflammation. Hybridization of DNA from these two strains to the previously described *H. pylori* microarray led to the identification of certain genetic differences between these strains that might contribute to different host responses. Interestingly, the strain that induced less inflammation in both models of infection had lost 19 out of the 26 genes at the *cag* PAI locus. (Several studies have previously shown that this locus has to be intact and functional to induce gastric epithelial cells to secrete high levels of pro-inflammatory CHEMOKINES[53,67,68].) Although *cagA* is a genotypic marker for virulence, its presence alone does not always correlate with a functional PAI island, which shows how useful microarrays are for detecting the presence of all genes.

CHEMOKINES
Small molecules that have a central role in inflammatory responses and trigger migration and activation of phagocytic cells and lymphocytes.

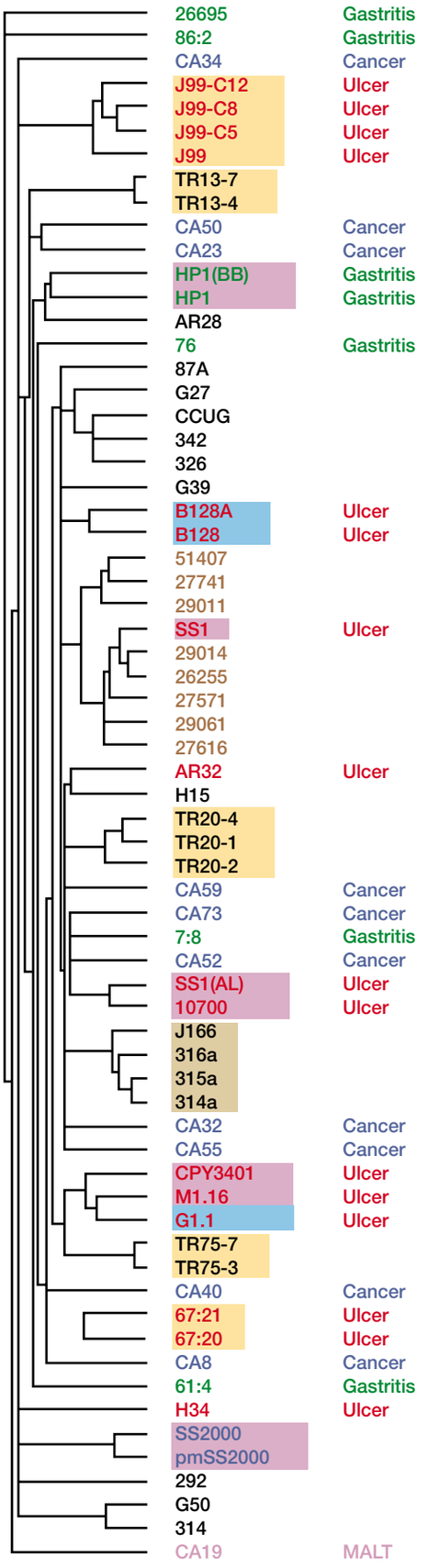


Figure 1 | **Dendrogram of *Helicobacter pylori* strains grouped using the CLUSTER program.** The dendrogram is based on the presence or absence of strain-specific genes. We used the CLUSTER ANALYSIS to group strains that seem most similar according to their gene content. The lengths of the arms of the dendogram are proportional to the level of divergence between strains. When known, the clinical state of the patient from whom the strain was obtained is indicated next to the strain name. Adjacent strains that are shaded in yellow were isolated from the same patient at different times, or from different locations. The human isolates that were genotyped before and after passage through an animal infection model are shaded in pink (mouse), blue (mongolian gerbil) or brown (rhesus monkey). For example, strain CPY3401 was used to infect mice, generating strain M1.16, which was then used to infect mongolian gerbils, generating strain G1.1. Strains from the same patient, which were isolated either from different body locations six years later (J99 series) or after passage through various animal models, have undergone changes, but they still group together. All the strains that were isolated from rhesus monkeys also grouped together. Clinical phenotype, however, was dispersed across several clusters. MALT, mucosa associated lymphoid tissue.

CLUSTER ANALYSIS
A mathematical algorithm that organizes a set of items according to their similarity. For example, genes can be clustered according to their similarity in pattern of expression.

The focus of the subsequent two independent studies was to understand how genetic changes that occur during long-term host colonization might affect the clinical outcomes of infection[38,41]. Bjorkholm *et al.*[38] used microarray analysis to examine two isolates of *H. pylori* from a 90-year-old patient with a gastric ulcer. These isolates had nearly identical gene profiles except for a deletion of the entire *cag* PAI in one strain. Both strains could colonize a transgenic mouse model in germ-free conditions. However, in the presence of the normal mouse microflora, only the strain that contained the *cag* PAI could colonize the stomach. The fact that strains isolated from the same patient had different phenotypic characteristics as a result of a different genetic complement alludes to the potential importance of subspeciation during *H. pylori* infection. Therefore, previous attempts to correlate a particular bacterial genotype with clinical outcome of the disease need to be re-interpreted.

Israel *et al.* examined the extent and types of genetic diversity present during chronic colonization of a single patient[41]. Because the patient, who was the original source of the sequenced strain J99, had refused antibiotic treatment, the *H. pylori* infection persisted for six years after the original diagnosis of a duodenal ulcer. This finding presented an unusual opportunity to study a dynamic population in its natural *in vivo* niche and to determine the sort of genetic changes that occurred over time.

The strains that were isolated during the second course of endoscopy proved to be closely related to the original archival J99 strain, which indicates that, in fact, the original strain had persisted over the preceding six years. Microarray analysis of the archival J99 isolate and of several of the recent isolates showed a remarkably close genetic relationship among all strains, with the maximum divergence between any two strains in the J99 lineage at ~1.9% — compared to the 6% difference between the two unrelated

sequenced strains J99 and 26695. In addition, cluster analysis of microarray data from 14 isolates from the patient who carried the J99 strain and 14 *H. pylori* isolates from other individuals provided striking evidence that the J99 strains all clustered on a branch that was distinct from the other strains (FIG. 2a). As in the Bjorkholm study, a few areas were found, including the plasticity zone, in which several contiguous genes had been lost from some of the recent isolates, although, notably, the *cag* PAI locus remained intact. Remarkably, this analysis also revealed areas in which DNA had been gained — the authors identified a 674-bp insertion that was present in the region of the plasticity zone in almost half of the recently isolated strains[41]. Subsequent sequence analysis revealed that this insertion is a new sequence that is absent in both sequenced strains. Furthermore, a six-gene region that is specific to the 26695 strain was found in all of the recent isolates. These results provide tangible support for the extraordinary genetic variation in the *H. pylori* chromosome that, previously, had only been hinted at by DNA fingerprinting and RFLP analysis. Additionally, analysis of the strains from the same source but at different time points provided a unique opportunity to evaluate critically genetic variation as a continuing dynamic process as it occurs *in vivo*.

Streptococcus pneumoniae. Very little is known about how individual *S. pneumoniae* factors contribute to the different disease states (BOX 3). Recently, the Institute for Genomic Research released the annotated genome sequence for a serotype 4 isolate (TIGR 4) of *S. pneumoniae*[43], and Eli Lilly released the annotated genome sequence for R6, which is an unencapsulated descendant of the serotype 2 D39 strain[44,69]. The availability of these sequences provides the means to explore and to develop a more complete understanding of *S. pneumoniae* pathogenesis. Genome comparisons between TIGR 4 and R6 reveal that, like *H. pylori*, there is a great deal of evidence for horizontal gene acquisition. Both sequences have a remarkable number of INSERTION SEQUENCES (ISs), three types of repetitive sequences and transposons. In addition, Tettelin *et al.* identified 40 genes that have the most significant homology to the genes in Gram-negative organisms[44]. There are also significant differences between the strains: R6 is predicted to contain 2,038 ORFs[69], whereas TIGR 4 contains 2,236 ORFs[44]. Other differences at the DNA level were detected using a TIGR 4-specific DNA microarray. This analysis revealed nine gene clusters that were present in the TIGR 4 strain but seemed to be absent from both R6 and D39 DNA. These clusters were mainly comprised of genes with atypical G+C content, which indicates that they might have been acquired by horizontal transfer. Most of the genetic elements that differ between the strains encode proteins that are predicted to be surface exposed or to be involved in aspects of sugar or lipid metabolism. These genetic elements might contribute to determination of the capsular serotype, and are potentially involved in pathogenesis[44] (see online link to the supplementary online Table 6 from REF. 44).

As an extension of these analyses, Hakenbeck *et al.* examined the genetic relatedness of 20 clinical *S. pneumoniae* and 9 oral streptococci isolates using an Affymetrix high-density oligonucleotide array that is specific for the sequenced serotype 4 of *S. pneumoniae*[43]. Most of the *S. pneumoniae* strains differed from the serotype 4 reference strain by 8–11%, and preliminary results from our own genome comparisons using a D39-specific spotted microarray, are in agreement with these findings (FIG. 2b; E.A.J. and S.F., unpublished data). The *S. pneumoniae* strains chosen in both studies belong to only a few clonal populations and, therefore, might not reflect the full extent of genetic diversity of *S. pneumoniae* isolates that has been suggested by other studies[70,71]. By contrast, the nine oral streptococci strongly diverged from the reference strain — only 15–61% of their DNAs hybridized to the *S. pneumoniae*-specific array. Essentially, none of the known or putative *S. pnuemoniae* virulence factors hybridized with oral streptococci DNA and, out of the elements that are conserved, none are predicted to encode surface features, which indicates that these organisms have distinct antigenic profiles.

In the course of these experiments, 470 *S. pneumoniae* strain-specific genes (24% of those represented on the array) were identified, in which variation was detected in at least one of the *S. pneumoniae* strains that were examined[43]. The strain-specific genes fell into three categories: 50% had no known function, 35% encoded proteins that were involved in sugar or carbon metabolism and 15% encoded surface-localized proteins. When the Affymetrix high-density oligonucleotide array was used, sequence variation in the penicillin-binding protein genes — *pbp2x* and *pbp1a* — and the gene that codes for dihydrofolate reductase — *dhfr* — was detected in strains that were resistant to penicillin and trimethoprim. This was in contrast to the results from experiments using a spotted DNA array. Typically, spotted DNA arrays are not sensitive enough to detect the variability in genes with a single or a limited nucleotide polymorphism. Unlike *H. pylori*, in which *babA* and the 27 genes of the *cag* PAI locus were identified as variable elements, the only confirmed virulence genes that showed any degree of variation among the *S. pneumoniae* isolates were those that encode IgA1 protease and capsular polysaccharides[43] (E.A.J. and S.F., unpublished data). Although pneumococcal virulence depends on the presence of a capsule, other bacterial determinants clearly contribute to the virulence of this organism. Only a fraction of the known capsular serotypes are strongly associated with disease. Furthermore, epidemiological data indicate that, even in this fraction, there are specific serotypes that are more often associated with a particular disease syndrome or site of infection[72–74]; therefore, additional factors must be involved in pathogenicity. The sequencing of other genomes and microarray analysis of a broader strain collection will help to address this epidemiological observation.

INSERTION SEQUENCES
Small, mobile nucleotide sequences found in the genomes of many bacterial populations.

ADHESIVE FIMBRIAE
Hair-like structures that project from the surface of some bacteria. They are involved in adhesion of bacterial cells to surfaces, and can be important in bacterial virulence.

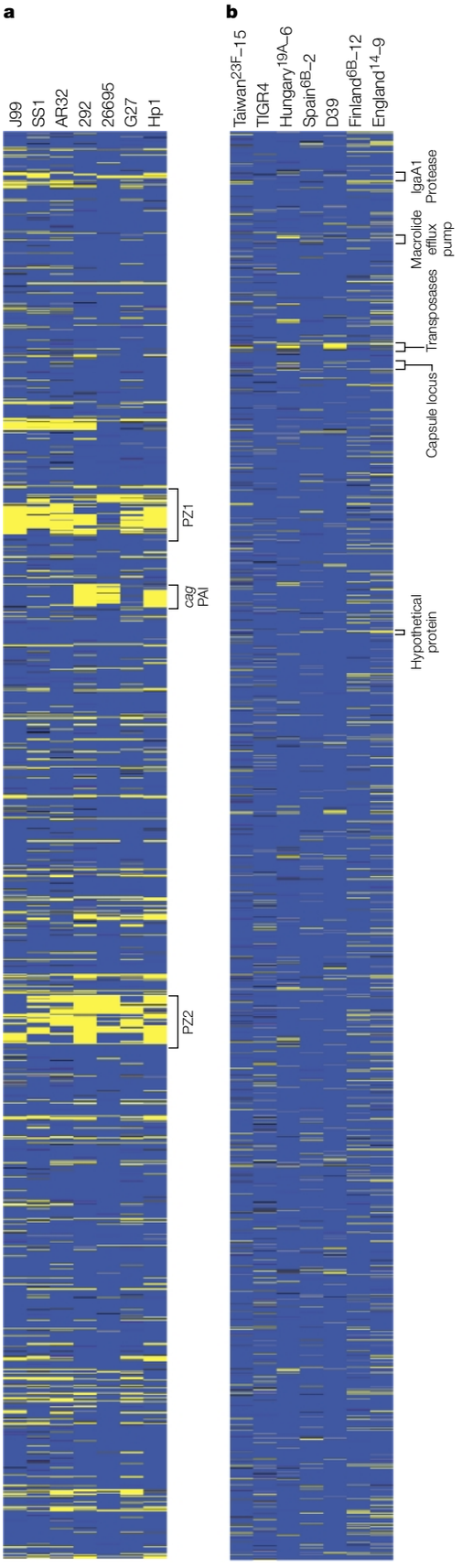
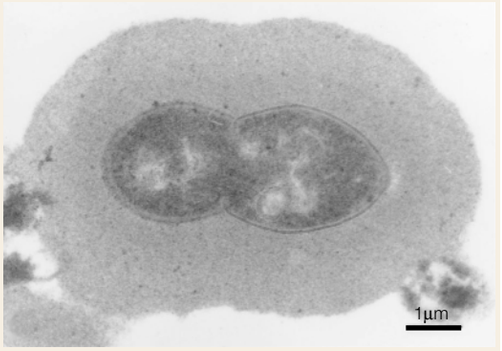


Figure 2 | **Distribution of strain-specific variation observed in two bacterial species.** The panels show results of microassay chromosomal profiles of strain-specific genes in genome order for (**a**) *Helicobacter pylori* (adapted from REF. 34) and (**b**) *Streptococcus pneumoniae*. The *H. pylori* genome was ordered using a combined J99/26695 map, in which the J99-specific genes are placed at the appropriate sites in the 26695 chromosomal map[62]. The *S. pneumoniae* genome was ordered on the basis of chromosomal order described in the TIGR 4 map[44]. Cut-offs for divergence or conservation were determined on an array-specific basis using an algorithm that calculates the likelihood of divergence. Highly homologous (blue) or divergent (yellow) genes are shown according to their position on the chromosome for each test strain. Data that could not be assigned to either category with high confidence are shown in black. Bracketed genes indicate those with highest variation. The *H. pylori* strains shown are described in REF. 34; *S. pneumoniae* strains are described in REFS 44,69,118.

Salmonella. One of the most striking observations to come from genome sequencing and microarray analyses is how much genetic heterogeneity is supported at the species level. Nowhere is this more apparent than in *Salmonella enterica*[26] (BOX 4). This single species comprises more than 2,300 serotypes that differ vastly in their ability to infect different hosts and that cause a spectrum of clinical disease, which ranges from gastroenteritis to systemic infection[75]. Although a few factors have been implicated in host adaptation[76–78], the genes and molecular mechanisms that are involved in this process remain largely unexplored, but are absolutely central to understanding the origin and spread of bacterial disease[75,79–82]. Recently, the genomes of two *Salmonella* serovars, *S. enterica* serovar Typhimurium LT2[83] and *S. enterica* serovar Typhi[84], have been published. A comparison of the two fully sequenced genomes and four other partially sequenced genomes[26] confirms the previously reported high degree of sequence similarity between serovars[5,7,8,85]. But the comparison also highlights loci that vary significantly among serovars. Numerous operons that encode ADHESIVE FIMBRIAE and various mobile genetic elements are interspersed in each genome. Of particular interest is that each serovar contains ~500 kb of unique sequence that is inserted or deleted throughout the chromosome[26]. There is evidence that it might contribute to serotype-specific phenotypes, such as the ability (or inability) to infect and cause disease in one or more host species[76].

We have recently constructed a spotted DNA microarray for *S. enterica* serovar Typhimurium (K.C. and S.F., unpublished data), which is based on the published serovar Typhimurium LT2 sequence[84,86]. This tool is providing a means for us to compare directly the genomic content of many serovars, including those with genome sequences that have not yet been determined. Our initial genomic comparison studies involved four isolates from the SARB and SARC reference collections (see link to *Salmonella* Reference Collections B and C), which have been extensively characterized from biochemical and MLEE analyses[6,85,87]. The data revealed distinct regions of the genome that are heterogeneous and correspond to the various prophages and fimbrial

Box 3 | **Streptococci**



At the beginning of the twentieth century, *Streptococcus pneumoniae*, also known as pneumococcus, killed more people worldwide than any other bacterial pathogen. As the most common bacterial cause of acute respiratory infection and OTITIS MEDIA **100 years later, *S. pneumoniae* remains a formidable infectious agent in much of the world[109]. Although more than 95 serotypes of this** GRAM-POSITIVE**, encapsulated** DIPLOCOCCUS **have been identified, only seven are responsible for most of the disease seen worldwide[110]. Like *Helicobacter pylori*, *S. pneumoniae* occupies a narrow niche — in this case, the human nasal–pharyngeal cavity — and colonization with the organism usually does not result in disease; strains are typically carried asymptomatically for weeks to months. Occasionally, however, they gain access to privileged host sites, such as the lung, blood or cerebral spinal fluid, with devastating consequences.**

A taxonomy scheme for the genus *Streptococcus* (Greek: *streptos*, chain; *kokhos*, berry) was first established in the late 1800s to describe chains of Gram-positive, bulbous, haemolytic organisms that were isolated from various diseased individuals[111]. However, this genus includes numerous organisms with very different biochemical properties, which makes it difficult and laborious to classify its members into discrete groups. Through the efforts of Rebecca Lancefield, a reknowned microbiologist, a test based on the antibody reaction to specific bacterial cell-wall antigens was developed and, by using this, many streptococci could be subdivided into conspicuous groups. However, this test did not distinguish between members of the *S. pneumoniae* species, because a thick CAPSULE **that surrounds clinical isolates masks the cell-wall antigens. A serotyping system based on antibody reaction to capsular antigens, known as the Quellung reaction, was therefore developed for this species[112]. Specific serotypes are identified by their reaction with type-specific antisera, during which the capsule swells and changes its refractive index, both of which can be seen under a light microscope. With more than 95 serotypes identified so far, this method is an important clinical tool for identifying specific serovars. Coupling this typing method with new genomic-based tools, such as microarray technology, will enable careful tracking of pneumococcal infection with increased discriminatory power. The image is reproduced with permission from** REF. 113 **© (1999) American Society for Microbiology.**

operons in the serovar Typhimurium genome (K.C. and S.F., unpublished data). Most of the prophages that are encoded in serovar Typhimurium are unique to this serovar and there is evidence that they encode pathogenic determinants[88]. In agreement with previous analyses and with the genome-sequence comparisons, our data also illustrate the close genetic relationship that exists among the serovars (K.C. and S.F., unpublished data) and support evidence showing that *S. bongori* is distinct from *S. enterica*, indicating that *S. bongori* should be considered as a separate species. Our results also accurately identify the virulence-plasmid-associated genes in serovar Typhimurium, and their absence in serovars Typhi, Choleraesuis and *S. bongori* [83,89–91]. In addition, we detect the presence of the SPI-1 PAI in all of the serovars and the absence of SPI-2 in *S. bongori* [92]. In combination with the data from the active sequencing projects of six additional *Salmonella* serovars (see the link to *Salmonella*.org), the stage is set to use microarray technology in large-scale exploration of genomic heterogeneity in *Salmonella*. This will undoubtedly lead to a more in-depth understanding of the genes and processes that contribute to host adaptation and disease development.

New frontiers in bacterial population biology

The above studies illustrate the efficacy of microarray technology as a tool for studying the genetic relatedness of bacterial isolates. Unlike random amplified polymorphic DNA (RAPD) PCR, RFLP and DNA fingerprinting, microarray analysis exposes genetic variation on a genome-wide scale. A principal strength of this technology is that it specifically identifies the genetic loci that vary from a sequenced strain, thereby overcoming a considerable obstacle that previously prevented us from understanding the basis of, and attributing biological significance to, observed genetic variation. The resolution with which we can now scrutinize entire bacterial genomes allows us to make a quantitative estimate of the core genome content and to study genome structure. Although such information will undoubtedly improve our understanding of bacterial population biology and will help us to define the boundaries that we ascribe to specific bacterial groups, microarray technology has its limitations. The use of microarrays for analysing genomic diversity is in its infancy, and as such, accepted standards for microarray design, and conditions for hybridization, as well as data selection, analysis and interpretation, have not yet been established. Researchers who are interested in constructing microarrays for organisms for which there is at least a completely sequenced and annotated genome are in an excellent position to design microarrays that are comprehensive and highly specific, with a minimum potential for within-genome cross-hybridization. For the moment, this only applies to a handful of microbes. For other microbes, this technology poses greater challenges.

In terms of analysis, for most of the studies published so far, the decision to classify genes into divergent or conserved categories is based on a constant cut-off value — genes are classed as divergent if the ratio of intensity of their signals falls below this cut-off. In some cases, the cut-off value has been empirically determined by comparing the reference strain to a similar strain that is known to be missing certain genes but, in other cases, this value cannot be empirically determined (because a reference strain is not available). If this occurs, genes might be erroneously assigned to divergent or conserved categories. At present, the computational repertoire for addressing these problems is limited (C. C. Kim *et al.*, unpublished data), which emphasizes the need for developing better analytical tools.

Even when great care is taken to design and build a microarray, and genes can be accurately assigned to their correct categories, it is simply not yet known how microarray data obtained by one group compares with that obtained by another, if similar arrays were designed

OTITIS MEDIA
Infection and inflammation of the middle ear space and ear drum.

GRAM REACTION
A differential stain that separates bacteria into two groups, Gram positive and Gram negative, on the basis of the biochemical composition of their cell wall.

DIPLOCOCCUS
Any of a variety of encapsulated bacteria (as the pneumococcus) that usually occur in pairs.

CAPSULE
A thick gel-like material generally composed of hydrophilic polysaccharide that surrounds the cell wall of Gram-positive or Gram-negative bacteria. It can contribute to pathogenicity by inhibiting phagocytosis of the bacteria by the macrophages of the host.

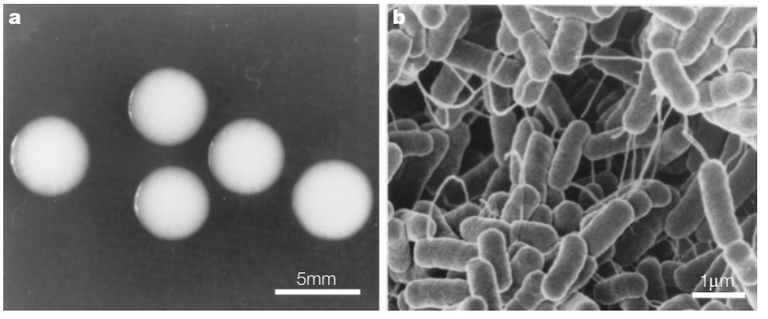
independently. These technical difficulties have important implications for the reproducibility of microarray data both within and between laboratories. As more laboratories pursue these types of genetic comparison, priority must be given to thoughtful consideration of these issues to improve the technology and validate its use as a highly reproducible, reliable and robust method for population-based studies.

In addition to the technical matters, there are important issues about the nature of the bacterial population under study and the selection of isolates that are representative of that population. For example, in *Escherichia coli* and *Salmonella*, the isolates of which are essentially clonal owing to relatively low rates of horizontal gene transfer and recombination, identifying changes in gene content or DNA sequence that might be correlated with differences in biological properties, such as virulence, is relatively straightforward. Indeed, this endeavour has been undertaken with some success[19]. However, in species that experience high rates of horizontal gene transfer and recombination, such as *H. pylori* and *S. pneumoniae*, it is much more challenging to establish phylogenetic relationships between members of a population and specifically to correlate genetic differences with biological processes. In these situations, appropriate population sampling is imperative. In the case of *H. pylori*, the presence of a particular DNA element, the *cag* PAI, is highly correlated with more severe disease and, therefore, acts as a marker for virulence potential. However, in organisms, such as *S. pneumoniae*, which only rarely causes disease, and in which no obvious PAIs have been identified, strain selection is more problematic. For these pathogens, virulence can be a matter of degree — one strain of *S. pneumoniae* might cause disease in 1 in 1,000 carriers, whereas another might cause disease in 1 in 50,000. Without knowing the ratio of the disease incidence to pathogen carriage, the differences in virulence potential between two such strains might be either missed or greatly exaggerated. Indeed, despite the genetic differences between the *S. pneumoniae* strains (see above), little is known about the disease or carriage potential of the strains, which makes it difficult to conclude anything about their virulence. To begin to address which bacterial factors are strongly correlated with a disease that is caused by an organism such as pneumococcus, a detailed understanding of bacterial clones that are associated with a particular disease outcome, as well as clones that are only very rarely associated with disease, is required.

Box 4 | **Salmonellae**

Historically, biochemical assays that took advantage of subtle metabolic differences were used to distinguish isolates of *Salmonella* from each other and to differentiate the Salmonellae from other ENTEROBACTERIACEAE**. Many of the growth media that were developed more than half a century ago are still being used in clinical laboratories for the initial identification of bacterial species[114]. Subsequently, a test — known as the Widal reaction — was developed, which classified the Salmonellae according to their ability to be** AGGLUTINATED **by a panel of antibodies[114]. Taking advantage of the fact that specific serovars have generally immutable antigenic profiles, the Kauffman–White schema was developed for serological diagnosis. It led to the identification of more than 2,300 serovars and to their classification into several serogroups. The members of a specific serovar can be further identified by their ability to resist infection by a known panel of bacteriophages — a process known as phage typing. This is a rather unwieldy means of classifying the Salmonellae, especially as the ability of a specific serovar to cause disease is often independent of the serogroup into which it is assigned.**

Results from multilocus enzyme electrophoresis (MLEE)[6,85,87,115] and chromosomal DNA-hybridization studies that were carried out in the early 1970s showed that most of the Salmonellae serovars were so closely related as to be considered a single species, resulting in a restructuring of the serotypes into several subspecies of the genus *Salmonella*[116]. In MLEE, whole bacterial lysate is electrophoresed on a gel, and the presence of specific enzymes is directly assayed on the gel. The mobility of each of these enzymes is assessed and any differences are assumed to be a result of non-synonymous changes at the nucleotide level. By assaying a panel of enzymes, each serovar was related to all other profiles of enzyme mobility. This work provided further evidence for the clonal nature of the *Salmonella* populations, and has led to a classification scheme that increasingly emphasizes the genetic relationships among the serovars. The Salmonellae are now organized into eight subspecies — I, II, IIIa, IIIb, IV, V, VI and VII — and two species, *S. bongori* (which contains all of subspecies V) and *S. enterica* (which consists of the other seven subspecies). Of all the serovars identified so far, more than 60% belong to subspecies I, which comprises 99% of the serovars that can cause disease in warm-blooded animals. In each subspecies, the Kauffman–White serogroup convention is still used. As a consequence of the taxonomic and phylogenetic heritage of *Salmonella*, the appropriate nomenclature for the Salmonellae is correspondingly cumbersome. For example, *S. typhimurium* (which causes a typhoid-like disease in mice and a gastrointestinal disease in several other animals) should appropriately be referred to as *S. enterica* serovar Typhimurium. The images are reproduced with permission from REF. 117 **© (2001) American Society for Microbiology.**



Uncovering the genetic basis of virulence

Consideration of the above-mentioned limitations will improve the existing microarray technology and its application in bacterial population studies. And there is good reason to do so. The alarming prevalence of drug resistance among bacteria is seriously undermining our ability to combat the spread of infectious disease and portends a return to the pre-antibiotic era. Identifying new bacterial targets for antimicrobial development is, therefore, imperative. However, members of bacterial populations can show significant heterogeneity in a variety of phenotypic traits, which makes identification of antimicrobial targets painstakingly slow and difficult. Nowhere has this been more apparent than in the study and identification of bacterial virulence factors, in which a combination of unlinked genes is often required for full virulence. The development of microarray technology provides a way to examine comprehensively the genetic content of many bacterial isolates. It, therefore, provides an efficient and unbiased approach for classifying genes as either core (detected in all strains) or auxiliary (strain

specific). An analysis of virulent and avirulent isolates will facilitate the identification of bacterial factors that are involved in, for example, processes that determine host specificity, TISSUE TROPISM, and disease and/or transmission of disease, as the genes that are required for these processes will be enriched in the genome sequences of virulent isolates. The microarray analyses discussed above clearly illustrate this point. Our challenge is to determine to what degree, if any, these factors correlate with virulence. The significance of this is far reaching: identifying genetic loci with definable connections to disease will facilitate the development of diagnostic tests to differentiate patients who are infected with benign or virulent strains, and will, therefore, inform treatment decisions. Additionally, these loci identify potential targets for vaccine design or improvement, as well as helping to define pathogen populations.

Horizontal DNA transfer and population biology

Horizontal DNA transfer is becoming more widely appreciated as an important method for the dissemination of genes, including virulence factors, between bacteria. Once these factors are identified, they can be used as specific markers to identify emerging pathogens and to track existing ones by using microarray technology. The possibility for this has already been alluded to in the literature on *Salmonella* and *H. pylori*, in which certain strain- or serovar-specific sequences that are associated with mobile genetic elements have been shown to be important for disease[53,78]. Microarrays, and particularly high-density oligonucleotide array technology, should prove to be particularly sensitive and useful for tracking not only strain-specific genes, but also genes that show more subtle allelic variations of single, or a few, nucleotide polymorphisms. These slight changes can have a profound impact on pathogenesis, as is seen for certain antibiotic resistances — for example, *S. pneumoniae* resistance to penicillin and trimethoperin, as described above. The ability to detect these kinds of polymorphism would facilitate monitoring the spread of antibiotic resistance in a bacterial species and even between species and genera[45].

As illustrated by the above studies, microarray technology is important in determining how DNA polymorphisms correlate with specific host responses to infection and/or with disease outcomes. Pathogenesis is a dynamic relationship that involves both bacterial factors and host responses. An improved ability to assign biologically relevant functions to strain-specific loci will probably provide important insight into the molecular basis of the relationships that exist between bacteria and their hosts. For instance, microarray analysis has provided evidence for the emergence of genetically distinct *H. pylori* subclones in an individual human host[38,41]. Genetic variation was detected at several loci, including several restriction-modification genes, several genes that encode putative surface-localized virulence factors, and other virulence-associated loci, including the *cag* PAI locus and *babA*. What is the significance of these genetic polymorphisms? *H. pylori* is unique in its ability to colonize chronically the gastric lumen of humans — an extremely challenging environment. The dynamic nature of the gastric milieu and the host immune response continuously places strong selective pressures on the *H. pylori* population. Perhaps in response to this pressure, and because of its relative isolation, *H. pylori* must undergo extensive intra-genome and intra-species recombination to maintain population diversity and fitness in its host. In support of this, it has been reported that different *H. pylori* isolates have specific repertoires of active restriction-modification systems[93], even in a single host[94]. Establishing specific restriction-modification activity profiles in a population might be a way for such a specialized organism to strike the delicate balance between sampling DNA from its environment to maintain genetic diversity and establishing barriers against foreign DNA. The result is the selection of genotypes that are tailored to suit each individual host.

The mosaic quality of the *S. pneumoniae* genome is indicative of recombination events that involve small segments of DNA with different evolutionary histories, as well as an impressive number of elements that are hallmarks of horizontal DNA transfer. Although little variation is observed in most of the known virulence determinants, genetic variation is detected at loci that are predicted to encode surface-localized proteins and/or proteins that are involved in sugar metabolism. Drawing broad conclusions on the basis of these data would be misguided, as the analysis was based principally on highly invasive disease-causing bacterial isolates that represent only a small and atypical segment of the *S. pneumoniae* population. From a population biology viewpoint, these *S. pneumoniae* isolates are of little consequence, as strains that have invaded into deeper tissue are rarely transmitted to new hosts, and so have little reproductive future. By including *S. pneumoniae* isolates that are carried asymptomatically, in future, genomic comparisons will probably reveal areas of genomic variation that are important for carriage and transmission, both of which precede and are necessary for pneumococcal virulence.

Final remarks

The post-genomic era has brought with it tools, such as microarray technology, that are allowing us to examine the genome content and structure of bacterial populations at a level that was never before possible. An immense amount of detail on genetic variation that is specific to individual isolates is accumulating at an astonishing rate. But the whole organism is greater than the sum of its genetic parts, and using microarrays solely to identify the differences and similarities among pathogenic isolates would be harnessing but a fraction of the power that this technology offers. Encoded in every genome is a story about the lifestyle of an organism — what it eats, where it can live and how it responds to and survives in its environment. For bacterial pathogens, how they respond to and survive in the host environment is of particular interest. To begin to understand this immensely complex and dynamic relationship, our

ENTEROBACTERIACEAE
A large family of Gram-negative bacilli that inhabit the large intestine of mammals.

AGGLUTINATED
The aggregation of particulate antigen by antibodies.

TISSUE TROPISM
Tissue-specific bacterial adherence and colonization due to a restricted distribution of receptor structures on certain host-cell surfaces and not on others.

challenge is not only to identify the genes that are involved in host–pathogen interactions, but also to understand how they operate. So, proteomic and gene-expression studies will add the next layer in genome analysis — an understanding of the phenotypes created by specific genotypes. Only then will we imbue the types of boundary used to distinguish bacterial populations with biological significance.

1. Van Belkum, A. *et al.* Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology. *Clin. Microbiol. Rev.* **14**, 547–560 (2001).
A thorough review that details current microbial nomenclature and concepts of evolutionary and population-based genetics. It reviews various molecular techniques and makes recommendations for their application in epidemiology, taxonomy and evolutionary studies.
2. Feil, E. J. & Spratt, B. G. Recombination and the population structures of bacterial pathogens. *Annu. Rev. Microbiol.* **55**, 561–590 (2001).
This review addresses the impact of recombination on bacterial populations and gives specific focus to multilocus sequence typing.
3. Milkman, R. Electrophoretic variation in *Escherichia coli* from natural sources. *Science* **182**, 1024–1026 (1973).
4. Selander, R. K. Population genetics of pathogenic bacteria. *Microb. Pathog.* **3**, 1–7 (1987).
5. Beltran, P. Toward a population genetic analysis of *Salmonella*: genetic diversity and relationships among strains of serotypes *S. choleraesuis*, *S. derby*, *S. dublin*, *S. enteritidis*, *S. heidelberg*, *S. infantis*, *S. newport*, and *S. typhimurium*. *Proc. Natl Acad. Sci. USA* **85**, 7753–7757 (1988).
6. Reeves, M. W. Clonal nature of *Salmonella* typhi and its genetic relatedness to other Salmonellae as shown by multilocus enzyme electrophoresis, and proposal of *Salmonella bongori* comb. nov. *J. Clin. Microbiol.* **27**, 313–320 (1989).
7. Selander, R. K. Genetic population structure, clonal phylogeny, and pathogenicity of *Salmonella* paratyphi B. *Infect. Immun.* **58**, 1891–1901 (1990).
8. Selander, R. K. Evolutionary genetic relationships of clones of *Salmonella* serovars that cause human typhoid and other enteric fevers. *Infect. Immun.* **58**, 2262–2275 (1990).
9. Lefevre, J. C. DNA fingerprinting of *Streptococcus pneumoniae* strains by pulsed-field gel electrophoresis. *J. Clin. Microbiol.* **31**, 2724–2728 (1993).
10. Li, J. *et al.* Recombinational basis of serovar diversity in *Salmonella enterica*. *Proc. Natl Acad. Sci. USA* **91**, 2552–2556 (1994).
11. Spratt, B. G. Resistance to antibiotics mediated by target alterations. *Science* **264**, 388–393 (1994).
12. Musser, J. M. Molecular population genetic analysis of emerged bacterial pathogens: selected insights. *Emerg. Infect. Dis.* **2**, 1–17 (1996).
13. Schloter, M. *et al.* Ecology and evolution of bacterial microdiversity. *FEMS Microbiol. Rev.* **24**, 647–660 (2000).
14. Maynard Smith, J., Feil, E. J. & Smith, N. H. Population structure and evolutionary dynamics of pathogenic bacteria. *Bioessays* **22**, 1115–1122 (2000).
An excellent review that discusses how MLEE, MLST and advances in nucleotide sequencing technology give a quantitative estimate of the impact of recombination in bacteria and how this contributes to our understanding of bacterial species definition.
15. Woese, C. R. Bacterial evolution. *Microbiol. Rev.* **51**, 221–271 (1987).
16. Ludwig, W. Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. *FEMS Microbiol. Rev.* **15**, 155–173 (1994).
17. Maiden, M. C. *et al.* Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl Acad. Sci. USA* **95**, 3140–3145 (1998).
This paper introduces multilocus sequence typing as a method for typing microorganisms on the basis of the nucleotide sequences of a limited number of genetic loci.
18. Feil, E. J. Estimating the relative contributions of mutation and recombination to clonal diversification: a comparison between *Neisseria meningitidis* and *Streptococcus pneumoniae*. *Res. Microbiol.* **151**, 465–469 (2000).
19. Reid, S. D. *et al.* Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature* **406**, 64–67 (2000).
20. Wren, B. W. Microbial genome analysis: insights into virulence, host adaptation and evolution. *Nature Rev. Genet.* **1**, 30–39 (2000).
21. Dobrindt, U. & Hacker, J. Whole genome plasticity in pathogenic bacteria. *Curr. Opin. Microbiol.* **4**, 550–557 (2001).
22. Fitzgerald, J. R. & Musser, J. M. Evolutionary genomics of pathogenic bacteria. *Trends Microbiol.* **9**, 547–553 (2001).
23. Graham, M. R. Toward a genome-scale understanding of group A *Streptococcus* pathogenesis. *Curr. Opin. Microbiol.* **4**, 65–70 (2001).
24. Musser, J. M. Pneumococcal research transformed. *N. Engl. J. Med.* **345**, 1206–1207 (2001).
25. Schaechter, M. *Escherichia coli* and *Salmonella* 2000: the view from here. *Microbiol. Mol. Biol. Rev.* **65**, 119–130 (2001).
26. Edwards, R. A., Olsen, G. J. & Maloy, S. R. Comparative genomics of closely related Salmonellae. *Trends Microbiol.* **10**, 94–99 (2002).
27. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
28. Cummings, C. A. & Relman, D. A. Using DNA microarrays to study host–microbe interactions. *Emerg. Infect. Dis.* **6**, 513–525 (2000).
29. Diehn, M. & Relman, D. A. Comparing functional genomic datasets: lessons from DNA microarray analyses of host–pathogen interactions. *Curr. Opin. Microbiol.* **4**, 95–101 (2001).
30. Lucchini, S., Thompson, A. & Hinton, J. C. Microarrays for microbiologists. *Microbiology* **147**, 1403–1414 (2001).
31. Sassetti, C. M., Boyd, D. H. & Rubin, E. J. Comprehensive identification of conditionally essential genes in mycobacteria. *Proc. Natl Acad. Sci. USA* **98**, 12712–12717 (2001).
32. Schoolnik, G. K. The accelerating convergence of genomics and microbiology. *Genome Biol. Online* **2**, REPORTS4009 (2001).
33. Behr, M. A. *et al.* Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* **284**, 1520–1523 (1999).
34. Salama, N. *et al.* A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proc. Natl Acad. Sci. USA* **97**, 14668–14673 (2000).
This study provided the first evidence of the extent of genetic diversity in bacteria using microarray analysis.
35. Wu, L. *et al.* Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Appl. Environ. Microbiol.* **67**, 5780–5790 (2001).
This crucial study describes conditions for examining gene composition in natural microbial communities using microarray analysis.
36. Cho, J. C. & Tiedje, J. M. Bacterial species determination from DNA–DNA hybridization by using genome fragments and DNA microarrays. *Appl. Environ. Microbiol.* **67**, 3677–3682 (2001).
A method based on random genome fragments and DNA microarray technology that reveals taxonomic relationships between bacterial strains.
37. 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).
38. Bjorkholm, B. *et al.* Comparison of genetic divergence and fitness between two subclones of *Helicobacter pylori*. *Infect. Immun.* **69**, 7832–7838 (2001).
By using microarray analysis, the authors detected genetic changes between clinical isolates collected from a single patient, alluding to the potential for *in vivo* sub-species development.
39. Hurt, R. A. *et al.* Simultaneous recovery of RNA and DNA from soils and sediments. *Appl. Environ. Microbiol.* **67**, 4495–4503 (2001).
40. 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).
41. Israel, D. A. *et al. Helicobacter pylori* genetic diversity within the gastric niche of a single human host. *Proc. Natl Acad. Sci. USA* **98**, 14625–14630 (2001).
This study used microarrays to examine the extent and types of genetic change in a bacterial pathogen that occur during long-term host colonization and how these correlate with clinical outcomes of infection.
42. Israel, D. A. *et al. Helicobacter pylori* strain-specific differences in genetic content, identified by microarray, influence host inflammatory responses. *J. Clin. Invest.* **107**, 611–620 (2001).
Microarray technology was used to investigate *H. pylori* genetic diversity and to correlate this with disease development and severity.
43. Hakenbeck, R. *et al.* Mosaic genes and mosaic chromosomes: intra- and interspecies genomic variation of *Streptococcus pneumoniae*. *Infect. Immun.* **69**, 2477–2486 (2001).
These authors used an Affymetrix high-density oligonucleotide array to examine the genetic relatedness of 20 clinical *S. pneumoniae* isolates, 5 *Streptococcus mitis* isolates and 4 *Streptococcus oralis* isolates.
44. Tettelin, H. *et al.* Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science* **293**, 498–506 (2001).
45. Fitzgerald, J. R. *et al.* Evolutionary genomics of *Staphylococcus aureus*: insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic. *Proc. Natl Acad. Sci. USA* **98**, 8821–8826 (2001).
46. Murray, A. E. *et al.* DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes. *Proc. Natl Acad. Sci. USA* **98**, 9853–9858 (2001).
47. Dziejman, M. *et al.* Comparative genomic analysis of *Vibrio cholerae*: genes that correlate with cholera endemic and pandemic disease. *Proc. Natl Acad. Sci. USA* **99**, 1556–1561 (2002).
48. Smoot, J. C. *et al.* Genome sequence and comparative microarray analysis of serotype M18 groupA *Streptococcus* strains associated with acute rheumatic fever outbreaks. *Proc. Natl Acad. Sci. USA* **99**, 4668–4673 (2002).
49. Wotherspoon, A. C. *et al. Helicobacter pylori*-associated gastritis and primary B-cell gastric lymphoma. *Lancet* **338**, 1175–1176 (1991).
50. Montecucco, C. Living dangerously: how *Helicobacter pylori* survives in the human stomach. *Nature Rev. Mol. Cell Biol.* **2**, 457–466 (2001).
51. Tummuru, M. K. *et al.* Cloning and expression of a high-molecular-mass major antigen of *Helicobacter pylori*: evidence of linkage to cytotoxin production. *Infect. Immun.* **61**, 1799–1809 (1993).
52. Covacci, A. *et al.* Molecular characterization of the 128-kDa immunodominant antigen of *Helicobacter pylori* associated with cytotoxicity and duodenal ulcer. *Proc. Natl Acad. Sci. USA* **90**, 5791–5795 (1993).
53. Censini, S. *et al. cag*, a pathogenicity island of *Helicobacter pylori*, encodes type I-specific and disease-associated virulence factors. *Proc. Natl Acad. Sci. USA* **93**, 14648–14653 (1996).
54. Backert, S. *et al.* Translocation of the *Helicobacter pylori* CagA protein in gastric epithelial cells by a type IV secretion apparatus. *Cell. Microbiol.* **2**, 155–164 (2000).
55. Odenbreit, S. *et al.* Translocation of *Helicobacter pylori* CagA into gastric epithelial cells by type IV secretion. *Science* **287**, 1497–1500 (2000).
56. Segal, E. D. *et al.* Altered states: involvement of phosphorylated CagA in the induction of host cellular growth changes by *Helicobacter pylori*. *Proc. Natl Acad. Sci. USA* **96**, 14559–14564 (1999).
57. Stein, M., Rappuoli, R. & Covacci, A. Tyrosine phosphorylation of the *Helicobacter pylori* CagA antigen after *cag*-driven host cell translocation. *Proc. Natl Acad. Sci. USA* **97**, 1263–1268 (2000).
58. Akopyanz, N. *et al.* PCR-based RFLP analysis of DNA sequence diversity in the gastric pathogen *Helicobacter pylori*. *Nucleic Acids Res.* **20**, 6221–6225 (1992).
59. Akopyanz, N. *et al.* DNA diversity among clinical isolates of *Helicobacter pylori* detected by PCR-based RAPD fingerprinting. *Nucleic Acids Res.* **20**, 5137–5142 (1992).
60. Alm, R. A. *et al.* Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen

Helicobacter pylori. *Nature* **397**, 176–180 (1999).
61. Alm, R. A. & Trust, T. J. Analysis of the genetic diversity of *Helicobacter pylori*: the tale of two genomes. *J. Mol. Med.* **77**, 834–846 (1999).
62. Tomb, J. F. *et al.* The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**, 539–547 (1997).
63. Pellegrini, M. *et al.* Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA* **96**, 4285–4288 (1999).
64. Gerhard, M. *et al.* Clinical relevance of the *Helicobacter pylori* gene for blood-group antigen-binding adhesin. *Proc. Natl Acad. Sci. USA* **96**, 12778–12783 (1999).
65. Ilver, D. *et al. Helicobacter pylori* adhesin binding fucosylated histo-blood group antigens revealed by retagging. *Science* **279**, 373–377 (1998).
66. Parsonnet, J. *et al. Helicobacter pylori* infection and the risk of gastric carcinoma. *N. Engl. J. Med.* **325**, 1127–1131 (1991).
67. Sharma, S. A. *et al.* Interleukin-8 response of gastric epithelial cell lines to *Helicobacter pylori* stimulation *in vitro*. *Infect. Immun.* **63**, 1681–1687 (1995).
68. Tummuru, M. K., Sharma, S. A. & Blaser, M. J. *Helicobacter pylori picB*, a homologue of the *Bordetella pertussis* toxin secretion protein, is required for induction of IL-8 in gastric epithelial cells. *Mol. Microbiol.* **18**, 867–876 (1995).
69. Hoskins, J. *et al.* Genome of the bacterium *Streptococcus pneumoniae* strain R6. *J. Bacteriol.* **183**, 5709–5717 (2001).
70. Hall, L. M. *et al.* Genetic relatedness within and between serotypes of *Streptococcus pneumoniae* from the United Kingdom: analysis of multilocus enzyme electrophoresis, pulsed-field gel electrophoresis, and antimicrobial resistance patterns. *J. Clin. Microbiol.* **34**, 853–859 (1996).
71. Spratt, B. G. & Maiden, M. C. Bacterial population genetics, evolution and epidemiology. *Phil. Trans. R. Soc. Lond.* B **354**, 701–710 (1999).
72. Gray, B. M. *et al.* Serotypes of *Streptococcus pneumoniae* causing disease. *J. Infect. Dis.* **140**, 979–983 (1979).
73. Gray, B. M. *et al.* Clinical and epidemiologic studies of pneumococcal infection in children. *Pediatr. Infect. Dis.* **5**, 201–207 (1986).
74. Orange, M. & Gray, B. M. Pneumococcal serotypes causing disease in children in Alabama. *Pediatr. Infect. Dis. J.* **12**, 244–246 (1993).
75. Baumler, A. J. *et al.* Evolution of host adaptation in *Salmonella enterica*. *Infect. Immun.* **66**, 4579–4587 (1998).
76. Conner, C. P. *et al.* Differential patterns of acquired virulence genes distinguish *Salmonella* strains. *Proc. Natl Acad. Sci. USA* **95**, 4641–4645 (1998).
77. Folkesson, A. *et al.* Multiple insertions of fimbrial operons correlate with the evolution of *Salmonella* serovars responsible for human disease. *Mol. Microbiol.* **33**, 612–622 (1999).
78. Townsend, S. M. *et al. Salmonella enterica* serovar Typhi possesses a unique repertoire of fimbrial gene sequences. *Infect. Immun.* **69**, 2894–2901 (2001).
79. Baumler, A. J., Hargis, B. M. & Tsolis, R. M. Tracing the origins of *Salmonella* outbreaks. *Science* **287**, 50–52 (2000).
80. Kingsley, R. A. & Baumler, A. J. Host adaptation and the emergence of infectious disease: the *Salmonella* paradigm. *Mol. Microbiol.* **36**, 1006–1014 (2000).
81. Rabsch, W. *et al.* Competitive exclusion of *Salmonella enteritidis* by *Salmonella gallinarum* in poultry. *Emerg. Infect. Dis.* **6**, 443–448 (2000).
82. Woolhouse, M. E., Taylor, L. H. & Haydon, D. T. Population biology of multihost pathogens. *Science* **292**, 1109–1112 (2001).
83. McClelland, M. *et al.* Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. *Nature* **413**, 852–856 (2001).
84. Parkhill, J. *et al.* Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* **413**, 848–852 (2001).
85. Boyd, E. F. *et al.* Molecular genetic relationships of the Salmonellae. *Appl. Environ. Microbiol.* **62**, 804–808 (1996).
86. Smith, B. P. *et al.* Aromatic-dependent *Salmonella* typhimurium as modified live vaccines for calves. *Am. J. Vet. Res.* **45**, 59–66 (1984).
87. Boyd, E. F. *et al. Salmonella* reference collection B (SARB): strains of 37 serovars of subspecies I. *J. Gen. Microbiol.* **139**, 1125–1132 (1993).
88. Figueroa-Bossi, N. *et al.* Variable assortment of prophages provides a transferable repertoire of pathogenic determinants in *Salmonella*. *Mol. Microbiol.* **39**, 260–271 (2001).
89. Woodward, M. J., McLaren, I. & Wray, C. Distribution of virulence plasmids within Salmonellae. *J. Gen. Microbiol.* **135**, 503–511 (1989).
90. Tinge, S. A. & Curtiss, R. Conservation of *Salmonella* typhimurium virulence plasmid maintenance regions among *Salmonella* serovars as a basis for plasmid curing. *Infect. Immun.* **58**, 3084–3092 (1990).
91. Chiu, C. H. *et al.* Prevalence of the virulence plasmids of nontyphoid *Salmonella* in the serovars isolated from humans and their association with bacteremia. *Microbiol. Immunol.* **43**, 899–903 (1999).
92. Ochman, H. & Groisman, E. A. Distribution of pathogenicity islands in *Salmonella* spp. *Infect. Immun.* **64**, 5410–5412 (1996).
93. Xu, Q. *et al.* Identification of type II restriction and modification systems in *Helicobacter pylori* reveals their substantial diversity among strains. *Proc. Natl Acad. Sci. USA* **97**, 9671–9676 (2000).
94. Aras, R. A. *et al.* Regulation of the *Hpy*II restriction-modification system of *Helicobacter pylori* by gene deletion and horizontal reconstitution. *Mol. Microbiol.* **42**, 369–382 (2001).
95. Davies, J. Origins and evolution of antibiotic resistance. *Microbiologia* **12**, 9–16 (1996).
96. Brown, J. R. & Doolittle, W. F. Archaea and the prokaryote-to-eukaryote transition. *Microbiol. Mol. Biol. Rev.* **61**, 456–502 (1997).
97. Doolittle, R. F. & Handy, J. Evolutionary anomalies among the aminoacyl-tRNA synthetases. *Curr. Opin. Genet. Dev.* **8**, 630–636 (1998).
98. Doolittle, W. F. & Logsdon, J. M. Jr. Archaeal genomics: do Archaea have a mixed heritage? *Curr. Biol.* **8**, R209–R211 (1998).
99. Doolittle, W. F. Phylogenetic classification and the universal tree. *Science* **284**, 2124–2129 (1999).
100. Koonin, E. V. *et al.* Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the Archaea. *Mol. Microbiol.* **25**, 619–637 (1997).
101. Lawrence, J. G. & Ochman, H. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl Acad. Sci. USA* **95**, 9413–9417 (1998).
102. Nelson, K. E. *et al.* Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**, 323–329 (1999).
103. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).
104. Doolittle, W. F. Lateral genomics. *Trends Cell Biol.* **9**, M5–M8 (1999).
105. McNulty, C. A. The discovery of *Campylobacter*-like organisms. *Curr. Top. Microbiol. Immunol.* **241**, 1–9 (1999).
106. Marshall, B. J. & Warren, J. R. Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. *Lancet* **1**, 1311–1315 (1984).
107. Suerbaum, S. Genetic variability within *Helicobacter pylori*. *Ijmm Int. J. Med. Microbiol.* **290**, 175–181 (2000).
108. Kelly, D. Infectious ulcers: not hurry, worry and curry? *Microbiol. Today* **28**, 188–189 (2001).
109. Musher, D. M. Infections caused by *Streptococcus pneumoniae*: clinical spectrum, pathogenesis, immunity, and treatment. *Clin. Infect. Dis.* **14**, 801–807 (1992).
110. Hausdorff, W. P. *et al.* Which pneumococcal serogroups cause the most invasive disease: implications for conjugate vaccine formulation and use, part I. *Clin. Infect. Dis.* **30**, 100–121 (2000).
111. McCarty, M. in *Microbiology, Including Immunology and Molecular Genetics* (eds Davis, B. D., Dulbecco, R., Eisen, H. N. & Ginsberg, H. S.) 607–622 (Harper & Row, Philadelphia, 1980).
112. Neufeld, F. Uber die agglutina der pneumokokken und uber die theorien der agglutination. *Z. Hyg. Infekt-Kr* **40**, 54–72 (1902).
113. Kim, J. O. *et al.* Relationship between cell surface carbohydrates and intrastrain variation on opsonophagocytosis of *Streptococcus pneumoniae*. *Infect. Immun.* **67**, 2327–2333 (1999).
114. White, P. B. *Great Britain Medical Research Council Special Report: No. 103* (Her Majesty's Stationery Office, London, 1926).
115. Crosa, J. H. *et al.* Molecular relationships among the Salmonellae. *J. Bacteriol.* **115**, 307–315 (1973).
116. Scherer, C. A. & Miller, S. I. in *Principles of Bacterial Pathogenesis* (ed. Groisman, E. A.) 266–316 (Academic, San Diego, 2001).
117. Anriany, Y. A. *et al. Salmonella enterica* serovar Typhimurium DT104 displays a rugose phenotype. *Appl. Environ. Microbiol.* **67**, 4048–4056 (2001).
118. McGee, L. *et al.* Nomenclature of major antimicrobial-resistant clones of *Streptococcus pneumoniae* defined by the Pneumococcal Molecular Epidemiology Network. *J. Clin. Microbiol.* **39**, 2565–2571 (2001).

Acknowledgements

We thank our colleagues G. Dougan and S. Baker from the Imperial College of Science, Technology and Medicine, London, for *Salmonella* strains and sharing data; A. Covacci, Chiron-Biocine, Italy; J. Gordon, Washington University, Missouri; J. Parsonette, Stanford University, California; R. Peek Jr, Vanderbilt University, Tennessee; J. Solnick, University of California at Davis, for providing *H. pylori* clinical isolates; L. McGee from the Pneumococcal Diseases Research Unit at the South African Institute for Medical Research, Johannesburg, for the *S. pneumoniae* strains; and A. Kawale and S. Censini for technical assistance, C. Kim for developing analytical tools for microarray analysis and thoughtful discussion, and S. Reid, J. R. Fitzgerald and members of the Falkow Laboratory for critical reviews of the manuscript.

Online links

FURTHER INFORMATION

CLUSTER program: http:\\www.microarrays.org/ software.html
Genome Entry Database at NCBI:
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome
Multilocus sequence typing: http://www.mlst.net
NCBI GenBank Database:
http://www.ncbi.nlm.nih.gov/Genbank/index.html
***Salmonella*.org:** http://www.salmonella.org
***Salmonella* Reference Collections B and C:**
http//:www.ucalgary.ca/~kesander/
Supplementary Table 6 from REF. 44:
http://www.sciencemag.org/cgi/content/full/293/5529/498/DC1#top
The Institute for Genomic Research (TIGR): http://www.tigr.org
Access to this interactive links box is free online.