



Application of DNA Microarrays to Study the Evolutionary Genomics of *Yersinia pestis* and *Yersinia pseudotuberculosis*

Stewart J. Hinchliffe, Karen E. Isherwood, Richard A. Stabler, et al.

Genome Res. 2003 13: 2018-2029

Access the most recent version at doi:[10.1101/gr.1507303](https://doi.org/10.1101/gr.1507303)

References This article cites 37 articles, 22 of which can be accessed free at:
<http://genome.cshlp.org/content/13/9/2018.full.html#ref-list-1>

License

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

An advertisement banner with a teal background. On the left, the text reads "CRISPR and RNAi Genetic Screening. Your new superpower." In the center, there is a white box with the words "LEARN MORE". On the right, there is a photograph of a woman wearing a red and white superhero cape and mask, and the Cellecta logo, which is a green molecular structure with the word "CELLECTA" below it.

To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>

Cold Spring Harbor Laboratory Press

Application of DNA Microarrays to Study the Evolutionary Genomics of *Yersinia pestis* and *Yersinia pseudotuberculosis*

Stewart J. Hinchliffe,¹ Karen E. Isherwood,² Richard A. Stabler,³ Michael B. Prentice,⁴ Alexander Rakin,⁵ Richard A. Nichols,⁶ Petra C.F. Oyston,² Jason Hinds,³ Richard W. Titball,² and Brendan W. Wren^{1,7}

¹The London School of Hygiene and Tropical Medicine, London, WC1E 7HT, UK; ²Defence Science and Technology Laboratories, Salisbury, Wiltshire, SP4 0JQ, UK; ³Department of Medical Microbiology, St. George's Hospital Medical School, London, SW17 0RE, UK; ⁴Department of Medical Microbiology, St. Bartholomew's and the Royal London School of Medicine and Dentistry, London, E1 2AD, UK; ⁵Max von Pettenkofer-Institut für Hygiene und Medizinische Mikrobiologie, 80336 München, Germany; ⁶School of Biological Sciences, Queen Mary, University of London, London, E1 4NS UK

Yersinia pestis, the causative agent of plague, diverged from *Yersinia pseudotuberculosis*, an enteric pathogen, an estimated 1500–20,000 years ago. Genetic characterization of these closely related organisms represents a useful model to study the rapid emergence of bacterial pathogens that threaten mankind. To this end, we undertook genome-wide DNA microarray analysis of 22 strains of *Y. pestis* and 10 strains of *Y. pseudotuberculosis* of diverse origin. Eleven *Y. pestis* DNA loci were deemed absent or highly divergent in all strains of *Y. pseudotuberculosis*. Four were regions of phage origin, whereas the other seven included genes encoding a vitamin B12 receptor and the insect toxin *sepC*. Sixteen differences were identified between *Y. pestis* strains, with biovar Antiqua and Mediaevalis strains showing most divergence from the arrayed CO92 Orientalis strain. Fifty-eight *Y. pestis* regions were specific to a limited number of *Y. pseudotuberculosis* strains, including the high pathogenicity island, three putative autotransporters, and several possible insecticidal toxins and hemolysins. The O-antigen gene cluster and one of two possible flagellar operons had high levels of divergence between *Y. pseudotuberculosis* strains. This study reports chromosomal differences between species, biovars, serotypes, and strains of *Y. pestis* and *Y. pseudotuberculosis* that may relate to the evolution of these species in their respective niches.

[Supplemental material is available online at www.genome.org. The following individuals kindly provided reagents, samples, or unpublished information as indicated in the paper: D. Tsereteli, M. Chu, H. Wolf-Watz, and H. Fukushima.]

Yersinia pestis is a Gram-negative bacterium that is the causative agent of the systemic invasive infectious disease classically referred to as plague (Perry and Fetherston 1997). There have been three recorded human plague pandemics; the Justinian plague (6th to 8th centuries), the Black Death (14th to 19th centuries), and modern plague (19th century to present day; Perry and Fetherston 1997). The Black Death alone is estimated to have claimed one-third of the European population, and this catastrophic event shaped the development of modern civilization. The recent identification of multidrug-resistant strains (Galimand et al. 1997), and the possible use of *Y. pestis* as an agent of biological warfare means that plague still poses a significant threat to human health.

Multi-locus sequence typing (MLST) of housekeeping genes suggests that *Y. pestis* is a clone of the enteropathogen *Yersinia pseudotuberculosis* (Achtman et al. 1999). Current *Y. pestis* strains form a homogeneous group that is estimated to have emerged 1500–20,000 years ago (Achtman et al. 1999). In this short period of evolutionary time, *Y. pestis* has evolved the ability to colonize an insect vector (the flea) and establish a transmission cycle between mammalian hosts by novel subcutaneous and pneumonic

routes of infection. Genetic analysis of these two *Yersinia* species provides an excellent opportunity to study how new and highly virulent pathogens evolve. To date, the acquisition of two plasmids (pMT1/pFra and pPCP1/pPla) and some small regions of chromosomal DNA have been identified as specific for the *Y. pestis* subspecies of *Y. pseudotuberculosis* (Ferber and Brubaker 1981; Parkhill et al. 2001; Deng et al. 2002; Radnedge et al. 2002). Apart from the plasmid-encoded Ymt protein on pMT1 that has been shown to be necessary for the colonization of fleas (Hinnebusch et al. 2002), little else is known about the genetic basis of the recent change in host adaptation and virulence.

Y. pestis strains can be divided into three biovars, Antiqua, Mediaevalis, and Orientalis, that are biochemically distinguished by their abilities to ferment glycerol and to reduce nitrate (Devignat 1951). They can also be differentiated by ribotyping (Guiyoule et al. 1994) and by PFGE analysis of *SpeI* fragments (Lucier and Brubaker 1992). The initial ribotyping studies, in conjunction with data on the geographical sources of the strains, were used to correlate the different biovars with the three plague pandemics (Guiyoule et al. 1994). This has been further validated by restriction fragment length polymorphism (RFLP) analysis of *IS100* element insertions from strains of the three biovars (Achtman et al. 1999) and PCR-based *IS100* genotyping by use of the *Y. pestis* CO-92 genome sequence as a reference (Motin et al. 2002). *Y. pseudotuberculosis* strains are grouped in 21 serotypes,

⁷Corresponding author.

E-MAIL Brendan.Wren@lshtm.ac.uk; **FAX** 44-207-637-4314.

Article and publication are at <http://www.genome.org/cgi/doi/10.1101/gr.1507303>.

on the basis of differences in their lipopolysaccharide (LPS) content. From these studies, it has been proposed that *Y. pestis* originated from a *Y. pseudotuberculosis* serotype O:1b strain (Achtman et al. 1999; Skurnik et al. 2000).

The recent sequencing of the entire genome of a *Y. pestis* Orientalis strain, CO-92 (Parkhill et al. 2001) and a Mediaevalis strain, KIM10+ (Deng et al. 2002), has provided an opportunity for the detailed genetic comparisons of *Y. pestis* strains of different biovars. However, although multiple genetic rearrangements affecting gene order were apparent on comparing the two strains, (Deng et al. 2002) and some rearrangements were seen during laboratory culture of *Y. pestis* CO-92 (Parkhill et al.), >98% of the genome sequence was shared between *Y. pestis* KIM and CO-92 (Deng et al. 2002). In many cases, the boundaries of genome rearrangements were formed by insertion sequences. A 102-kb region of *Y. pestis* DNA, incorporating the pigmentation (*pgm*) and yersiniabactin loci, flanked by insertion sequences, is known to be unstable. Although necessary for full virulence and flea colonization in vivo, this region is lost during in vitro growth at a high frequency (2×10^{-3} ; Hare and McDonough 1999), probably due to recombination between flanking *IS100* elements (Fetherston et al. 1992; Fetherston and Perry 1994; Buchrieser et al. 1998a). It is possible, therefore, that recombination between these sequence elements is responsible for other deletions in the *Yersinia* chromosome with less obvious in vitro phenotypes. A recent subtractive hybridization study by Radnedge et al. (2002) identified six chromosomal regions of *Y. pestis* that varied between strains. Four of these six difference regions (DFRs) were associated with flanking-insertion sequences or other repeats. This study showed that the six DFRs could be used to form profiles of *Y. pestis* strains that can be correlated to biovar and the evolution of individual strains (Radnedge et al. 2002).

Traditional phylogenetic classification of bacteria to study evolutionary relatedness is based on the variation in a limited number of conserved genes, rRNA/rDNA, or signature sequences. However, due to the acquisition of DNA through lateral gene transfer, the differences between closely related bacterial strains, particularly members of the enterobacteriaceae such as the yersiniae, can be significant. Genome comparisons between pathogenic and nonpathogenic strains within a species are particularly useful for identifying determinants important in virulence, transmission, and host specificity. DNA microarray analysis has been utilized recently to investigate genome-wide analysis of several bacterial pathogens including *Helicobacter pylori* (Salama et al. 2000), *Campylobacter jejuni* (Dorrell et al. 2001), *Mycobacterium tuberculosis* (Kato-Maeda et al. 2001), *Staphylococcus aureus* (Fitzgerald et al. 2001), *Vibrio cholerae* (Dziejman et al. 2002), and *Salmonella enterica* (Porwollik et al. 2002). Because *Y. pestis* and *Y. pseudotuberculosis* are highly clonal and closely related (Achtman et al. 1999), microarray analysis represents an ideal methodology for full genome comparisons of the two species.

To further elucidate the genetic differences between *Y. pestis* and *Y. pseudotuberculosis* since their divergence, we designed a *Y. pestis* CO-92 biovar Orientalis gene-specific microarray to probe full genome comparisons of chromosomal DNA between different strains, biovars, serotypes, and species. The data confirms and extends previous studies and identified new loci specifically present in *Y. pestis* and absent in *Y. pseudotuberculosis* strains.

RESULTS AND DISCUSSION

DNA sequences representing all 4221 predicted coding sequences (4012 chromosomal and 209 plasmid encoded) from *Y. pestis* CO-92 (biovar Orientalis) were amplified and spotted onto glass microscope slides to produce a CO-92 gene-specific microarray

(see Methods). Twenty-two strains of *Y. pestis* and ten strains of *Y. pseudotuberculosis* were chosen to be compared at a genomic level with the CO-92 strain by competitive hybridization to the array. *Y. pestis* strains were chosen to cover all three biovars and included a previously sequenced strain (KIM10+) and a strain that has been referred to previously as *Y. pestoides* (G-8786). *Y. pseudotuberculosis* strains were chosen in order to allow a genomic comparison of different serotypes with *Y. pestis*, and also to aid investigation into their differing phenotypes, the results of which are not discussed here. A comparison of the relative levels of hybridization signal across the entire CO-92 chromosome for all strains is shown in Figure 1. *Y. pestis* strains showed relatively few differences compared with the CO-92 array strain, whereas *Y. pseudotuberculosis* strains revealed greater variation.

Microarray Validation

Data obtained from microarray analysis was validated by comparison with limited genomic analysis of the species reported in previous studies. Microarray data for the KIM10+ strain of *Y. pestis* was compared with the published genome sequence (Deng et al. 2002). Our microarray data concur with the genome sequence data, in that four multiple gene regions (YPO0738–YPO0754, YPO1165–YPO1172, YPO2271–YPO2281, and YPO2375–YPO2376) and the *sepC* gene (YPO2380) are absent from the KIM10+ genome.

A genomic signature tag (GST) approach for determining differences between CO-92 and EV76 (biovar Orientalis) was reported recently, and suggested the potential absence of six regions in the EV76 genome (Dunn et al. 2002). Our data reveals that only two of the regions appear to be absent from the EV76 genome, both of which were predicted, but not fully defined, by the GST method. The GST study predicted nucleotides 2,172,627–2,254,447 of the CO-92 sequence, corresponding to YPO1919–YPO1985 in the 102-kb unstable region, to have a high probability of being absent in EV76. Our data reveals that YPO1902–YPO1967 is absent in EV76. A second region predicted to be absent, in the GST study, is nucleotides 1,307,243–1,316,087 of the CO-92 sequence, corresponding to YPO1159–YPO1168. Our data shows that YPO1165–YPO1172 is absent in EV76, as well as F361–F366 and KIM10+. We noted no absences in any of the other four predicted regions. Only one other absence was revealed by our data, that of a single gene, YPO0599, encoding a potential adhesin. These discrepancies in boundaries of deleted regions may be due to the effect of extensive genomic rearrangements in gene order on the probability functions used in the GST method.

To further validate the microarray studies, we also compared our results with the subtractive hybridization data published recently by Radnedge et al. (2002). Thirteen of the strains analyzed by the Radnedge study were tested on the microarray. Our microarray data identified all of the deletions that were reported previously by Radnedge et al. (2002) along with other deletions not reported previously. From the microarray data, a total of 16 differences were identified between the *Y. pestis* strains tested when compared with CO-92. These differences range from a single gene to a region spanning 41 genes. Of the 16 regions, 6 corresponded with the DFRs defined by Radnedge et al. (2002). This finding suggests that subtractive hybridization is a relatively insensitive method for analyzing strain-to-strain variation. In summary, our microarray data concurs with previous *Yersinia* genetic comparisons, validating the use of this methodology for genome-wide comparisons. In comparison with subtractive hybridization or GST analysis of multiple strains, the microarray analysis appeared more comprehensive and identified additional gene and locus differences.

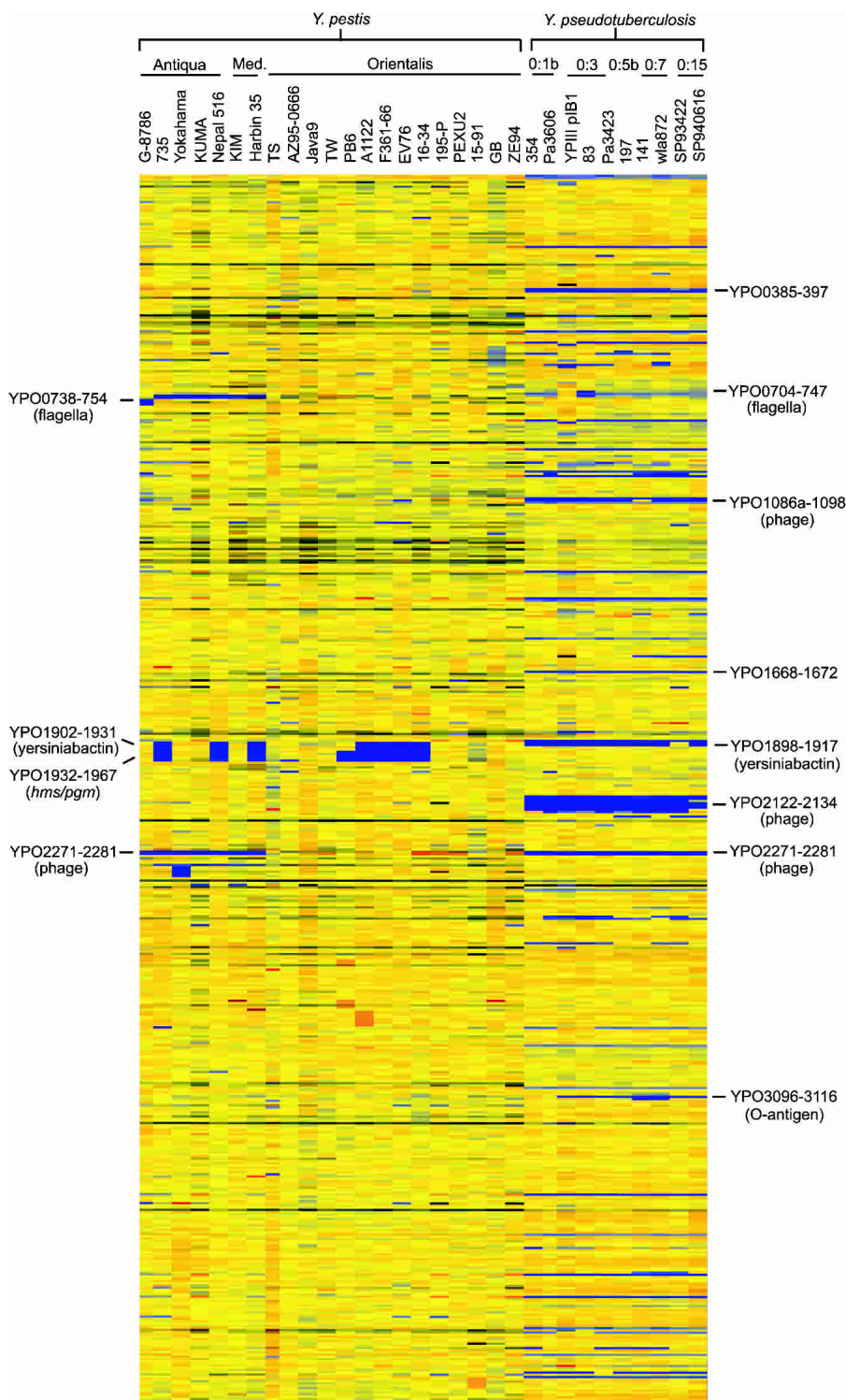


Figure 1 Chromosomal comparison of 22 strains of *Y. pestis* and 10 strains of *Y. pseudotuberculosis* to the sequenced CO-92 strain. This data was generated by Genespring software. *Y. pestis* strains are grouped according to biovar, whereas *Y. pseudotuberculosis* strains are grouped according to serotype. Gene status is color coded according to the Genespring software default colors with reference to the control strain (CO-92). Thus, in this comparison, yellow indicates presence, blue indicates absence or high divergence, and orange/red indicates a duplication.

Y. pestis/*Y. pseudotuberculosis* Strain Comparisons

The most obvious difference between *Y. pestis* and *Y. pseudotuberculosis* was the number of IS elements. The total number of insertion sequences reported in the CO-92 genome exceeds that reported for most other bacterial pathogens (comprising 3.7% of the genome). Of these, *IS1541* elements were the most abundant, with *IS100*, *IS285*, and *IS1661* elements making up the majority of the remainder (Parkhill et al. 2001). All of the *Y. pestis* strains were nearly identical in their complement of IS elements apart from *IS100*. Copy numbers of *IS100* elements ranged from half that of CO-92 in strains G-8786 and 735, to approximately three times the CO-92 level in strains KUMA and ZE94. Elevated copy numbers of *IS100* were also noted in strains 195-P and PEXU2 (two and one-half times) and A1122 (one and one-half times), compared with CO-92. All 10 *Y. pseudotuberculosis* strains appeared to contain single copies of *IS903* and *IS1400* elements identified in the *Y. pestis* CO-92 chromosome. However, they all contained fewer copies of *IS1541* and *IS285* elements compared with *Y. pestis* CO-92, and virtually no *IS100* or *IS1661* elements. The high number of *IS100* elements in the *Y. pestis* genome has been suggested to have played a role in its evolution, as regions flanked by IS elements are often unstable as a result of recombination events. A 102-kb region of *Y. pestis* DNA, incorporating the pigmentation (*pgm*) and yersiniabactin loci flanked by insertion sequences is known to be unstable due to recombination between flanking *IS100* elements (Fetherston et al. 1992; Fetherston and Perry 1994; Buchrieser et al. 1998a). Our data confirmed this instability as the entire 102-kb yersiniabactin, *hms/pgm* locus (encompassing YPO1902–YPO1967) was absent from many of the *Y. pestis* strains. However, passage of three strains, ZE94, 195-9, and Java9 (all biovar Orientalis) through mice enriched for bacteria retaining the 102-kb region on repeat hybridization with the array (P. Oyston and R. Titball, unpubl.). Strain PB6 (biovar Orientalis) was the only strain to have lost the yersiniabactin locus, yet retained most of the *hms* and *pgm* loci, with a deletion spanning YPO1902–YPO1931. Despite the large number of IS elements apparent in *Y. pestis* genomes, no other unstable regions in *in vitro* cultured bacteria that could be stabilized by *in vivo* growth were identified.

The yersiniabactin locus of *Y. pseudotuberculosis* spans the region

YPO1898–YPO1917, and has been reported to be present in only a small number of *Y. pseudotuberculosis* strains (Buchrieser et al. 1998b). Only 1 *Y. pseudotuberculosis* strain SP93422 (serotype O:15) hybridized to the equivalent locus in *Y. pestis* CO-92, whereas all 10 *Y. pseudotuberculosis* strains contained the hemin storage (*hms*) and pigmentation (*pgm*) loci (YPO1918–YPO1967).

Aside from IS element differences, 11 regions of the CO-92 chromosome were absent from all 10 strains of *Y. pseudotuberculosis* (Table 1); of these, only 2 were deemed absent from any of the *Y. pestis* strains tested. Four of these *Y. pestis*-specific regions consisted almost entirely of phage-related coding sequences. The prophage encoded by YPO2271–YPO2281 is absent from all 10 strains of *Y. pseudotuberculosis*, and is also absent from all Antiqua and Mediaevalis strains of *Y. pestis*. These prophage genes have high similarity to a filamentous prophage found in virulent K1 strains of *Escherichia coli* (Gonzalez et al. 2002). One of the genes in this locus, *puvA* (equivalent to YPO2277), was identified in *E. coli* O18:K1:H7 strain RS218 as a potential virulence factor during a signature-tagged mutagenesis (STM) screen of an infant rat model of meningitis (Gonzalez et al. 2001).

Two of the other *Y. pestis*-specific phage regions YPO1085–YPO1088 and YPO1094–YPO1098 flanked another phage-associated region YPO1089–YPO1092a, found in only one of the *Y. pseudotuberculosis* strains tested (strain 141, serotype O:7). The final *Y. pestis* species-specific phage region YPO2084–YPO2119 adjoined another phage region YPO2120–YPO2135, which is absent from 5 of the 10 *Y. pseudotuberculosis* strains. Most of the putative coding sequences in the region YPO2120–YPO2135 were also absent in the other five *Y. pseudotuberculosis* strains, but a few of the sequences hybridized to the array. This indicates that this region may also be absent in these strains, and the presence of other, as yet unidentified phage sequences in these strains may have resulted in cross-hybridization.

The remaining 7 *Y. pestis*-specific regions contained 21 predicted coding sequences, 14 of which are of unknown function. The other seven encode a putative membrane protein, a YjgF-family lipoprotein, a putative DNA-binding protein, the putative transcriptional activator RegC, a methylase enzyme, the vitamin B12 receptor BtuB, and the putative insecticidal toxin SepC. No obvious role in virulence or host adaptation can be ascribed to any of these genes, with the exception of SepC, which could be important in the colonization of the flea. However, this gene is also absent in the KIM10+ genome sequence, a strain that is capable of colonizing the flea (Hinnebusch et al. 2002). The ap-

parent absence of the vitamin B12 receptor was unexpected, as *Y. pseudotuberculosis* is vitamin B12 dependent and uptake is known to occur (M. Prentice, unpubl.). A BLAST search of the *Y. pseudotuberculosis* IP32953 currently being sequenced at The Lawrence Livermore Institute (http://bbrp.llnl.gov/bbrp/bin/y.pseudotuberculosis_blast) using the CO-92 *btuB* gene sequence reveals that the gene is present, but contains regions of diversity separated by regions of near perfect identity. PCR and sequence analysis revealed that *btuB* in all 10 *Y. pseudotuberculosis* strains tested was identical to the *Y. pseudotuberculosis* IP32953 strain (data not shown). *Y. pestis* and *Y. pseudotuberculosis* are so closely related at the genetic level that *Y. pestis* is sometimes classified as a subspecies of *Y. pseudotuberculosis* rather than a distinct species. Thus, the divergence of the *btuB* gene between *Y. pestis* and *Y. pseudotuberculosis* is surprising, as, to date, there have been no reports of variation in *btuB* between strains of the same species. The vitamin B12 receptor may act as a receptor for colicins and bacteriophages (for review, see James et al. 1996), thus, there is the potential for variation of the *btuB* gene due to selective pressure. *Y. pseudotuberculosis* is found widely in the soil environment and only causes human disease when ingested. Thus, it is likely to be in contact with other bacterial species, a source of colicins and bacteriophages. To survive in these habitats, variation in the *btuB* gene may be advantageous. *Y. pestis*, on the other hand, is an obligate blood-borne pathogen of mammals with an insect vector. In these niches, *Y. pestis* would not be in close contact with other bacterial species, possibly resulting in a lack of selective pressure for variation of the *btuB* gene.

Y. pestis Interstrain Comparison

A schematic comparison of the deletions found in all of the *Y. pestis* strains tested, excluding the unstable 102-kb yersiniabactin/*pgm/hms* locus, is shown in Figure 2. Overall, strains belonging to biovars Antiqua or Mediaevalis showed the greatest divergence from *Y. pestis* strain CO-92. Conversely, the other biovar Orientalis strains showed the least divergence from strain CO-92. Three separate regions were identified as being present in all Orientalis strains, but absent in the majority of Antiqua and Mediaevalis strains. Only the filamentous CUS-2 prophage, YPO2271–YPO2281 was deemed absent from all seven Antiqua and Mediaevalis strains, but two further regions were absent from both Mediaevalis strains and four of the five Antiqua strains. These two regions may be considered biovar specific as the only Antiqua strain that encodes these regions, G-8786, has also been re-

Table 1. Regions of the CO-92 Genome That Are Absent or Highly Divergent in All 10 *Y. pseudotuberculosis* Strains Tested

Gene region	DFR	Contents
YPO387–397	—	Hypothetical proteins similar to hypothetical protein from <i>Xylella fastidiosa</i> , <i>Neisseria meningitidis</i> and <i>Deinococcus radiodurans</i>
YPO1087–1088	DFR2	Putative prophage proteins.
YPO1094–1098	DFR2	Putative prophage and hypothetical proteins.
YPO1668–1672	—	Putative membrane protein <i>yihN</i> . YjgF-family lipoprotein. Putative DNA-binding protein.
YPO2084–2130	—	Putative prophage proteins.
YPO2261	—	Hypothetical protein.
YPO2271–2281	DFR5	Putative prophage proteins.
YPO2503	—	Hypothetical protein.
YPO2380	DFR6	<i>sepC</i> : Similar to <i>Serratia entomophila</i> plasmid pADAP virulence protein SepC and to <i>Photobacterium luminescens</i> insecticidal toxin complex protein TccC.
YPO3910	—	Similar to <i>Escherichia coli</i> vitamin B12 receptor protein BtuB.
YPO4031–4032	—	Similar to <i>Serratia marcescens</i> transcriptional activator <i>regC</i> , and hypothetical protein.

List of regions deemed absent from all 10 strains of *Y. pseudotuberculosis*. Regions that correspond to the difference region s(DFRs) identified by Radnedge et al. (2002) are indicated. Gene functions as assigned upon annotation of the CO-92 sequence are also indicated.

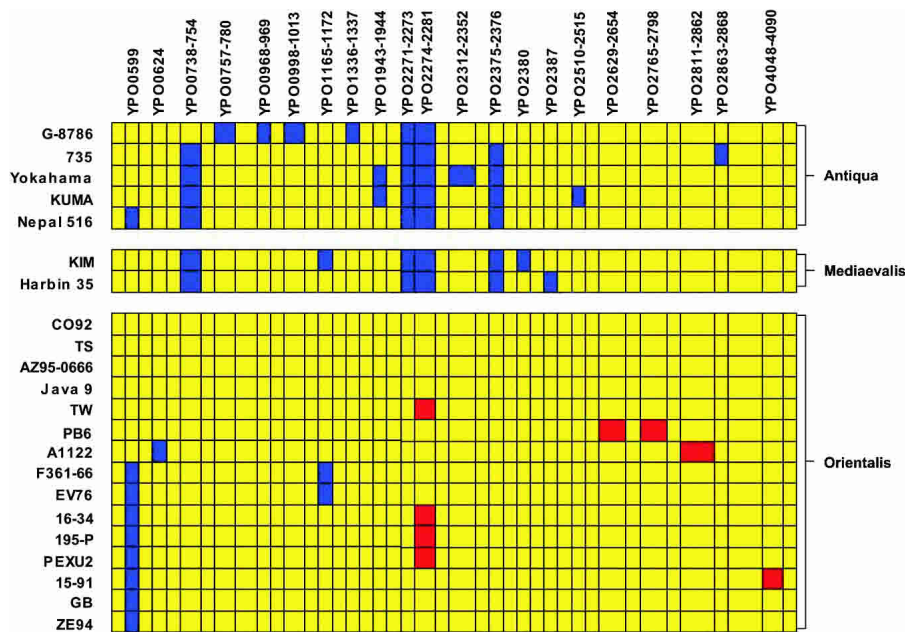


Figure 2 Schematic of chromosomal comparison of 22 strains of *Y. pestis* detailing all of the regions of divergence from CO-92. Strains are grouped by biovar. The 102-kb unstable region (YPO1902–YPO1967) has not been included in this comparison. Gene status is color coded as in Figure 1 for ease of comparison, with yellow indicating presence, blue indicating absence, and red indicating a duplication.

ferred to as *Y. pestoides*, and thus, is considered an atypical Antiqua strain. Strain G-8786 also contains four strain-specific regions of deletion, perhaps reflecting the remote origin of this strain (a rodent from the Caucasian High Mountainous region in Georgia), and further indicating that this is not a typical Antiqua strain. These regions encoded a putative aldo/keto reductase and hypothetical protein, YPO2375–YPO2376, and one of the two cryptic flagella gene clusters, YPO0738–YPO0747, along with the adjacent genes YPO748–YPO754.

In common with some of the *Y. pseudotuberculosis* strains, eight Orientalis strains and one of the Antiqua strains (Nepal 516) did not contain YPO0599 (Fig. 2). This putative adhesin was not enriched by passage through mice in *Y. pestis* strain ZE94 (P. Oyston and R. Titball, unpubl.). Three strains, one Mediaevalis and two Orientalis, had deletions of YPO1165–YPO1172 encoding proteins involved in the choline-glycine betaine pathway, whereas two Antiqua strains had deletions of YPO1943–YPO1944, encoding two putative membrane proteins. All other deletions were strain specific, with the Antiqua strain G-8786 appearing to be the most divergent strain, with four strain-specific deletions. A phylogenetic tree based on the *Y. pestis* differences clearly shows the divergence of strain G-8786 from the other strains (Fig. 3). A clear divide can also be seen between the Orientalis strains and strains from the other two biovars. The 15 *Y. pestis* Orientalis strains are all very similar genetically, and thus, are closely related. The Antiqua and Mediaevalis strains also share a great deal of genetic similarity with each other, but also show several strain-specific variations; thus these two biovars are unable to be easily distinguished by their genetic differences.

There are five regions of the CO-92 chromosome that appear to have been duplicated in some *Y. pestis* strains. One of these regions, YPO2274–YPO2281, corresponds to the invariant part of the *E. coli* CUS-2 prophage, which is identical to CUS-1 (Gonzalez et al. 2001), and a possible explanation for this apparent partial duplication is the presence of a CUS-1 or similar phage in addition to CUS-2. This duplication is seen in 4 of the 15 Orientalis

strains. The other four duplications appear to be strain specific and are all flanked by insertion sequences.

Y. pseudotuberculosis

Strain Comparison

As found previously for housekeeping gene sequences (Achtman et al. 1999), gene complement and some gene sequences are more divergent when compared across *Y. pseudotuberculosis* strains than within the *Y. pestis* group. Microarray analysis revealed high levels of variation throughout the entire genome of *Y. pseudotuberculosis* strains compared with *Y. pestis* CO-92 and with each other (Fig. 1). Analysis of this data revealed that between 6.3%–10.4% of predicted coding sequences, identified from the sequencing of *Y. pestis* strain CO-92, were absent or divergent in individual *Y. pseudotuberculosis* strains. Many of these encode proteins that may be involved in the successful colonization of mammalian and insect hosts (Table 2).

Two regions showed high levels of divergence from the *Y. pestis* CO-92 genome in all 10 of the *Y. pseudotuberculosis* strains. These both contain genes responsible for the structure of potentially

antigenic surface structures, the flagella and lipopolysaccharide (LPS), and thus, they may be under selective pressure, for example, from the mammalian immune system, resulting in high levels of divergence.

All *Y. pseudotuberculosis* strains produce LPS, an important virulence determinant, mediating resistance to complement-mediated and phagocyte killing (Makela et al. 1988; Darwin and Miller 1999; Karlyshev et al. 2001). In contrast, *Y. pestis* strains produce rough LPS, lacking the O-antigen, due to a number of mutations in the O-antigen gene cluster (Skurnik et al. 2000; Parkhill et al. 2001; Deng et al. 2002). Variance in LPS structure is the basis for typing *Y. pseudotuberculosis* strains, with 21 serogroups identified. Microarray analysis revealed significant variation within the O-antigen gene cluster between serotypes, with a high level of divergence of most genes, as has been demonstrated previously by PCR and sequencing (Skurnik et al. 2000). This data provided evidence that *Y. pestis* originated from a serotype O:1b strain of *Y. pseudotuberculosis*. Our results concur with these findings and are summarized in Table 3. The serotype O:7 strains, for example, have no genes within this region with sufficient similarity to CO-92 to hybridize to the array, with the exception of the O-antigen chain-length determinant *wzz*. This was also the only gene within this region that Skurnik et al. (2000) was able to identify using PCR, with the serotype O:7 strains. The O1b serotypes, on the other hand, contain CO-92-homologous counterparts for all of these genes with the exception of *wzx*, the O-unit flippase gene YPO3110, which appears to be divergent. Sequence data on the O-antigen gene cluster of the serotype O:1b strain Pa3606 (GenBank accession no. AJ251712) showed that all genes within this cluster are 98%–100% identical to their CO-92 array PCR products at the nucleotide level, with the exception of *wzx*, which is only 84.5% identical. Many genes in this cluster are apparently absent in some serotypes due to their diversity from the CO-92 sequence. The absence of key genes such as the O-unit flippase and O-unit polymerase would mean that these strains lack high molecular weight LPS, as they are unable to transport

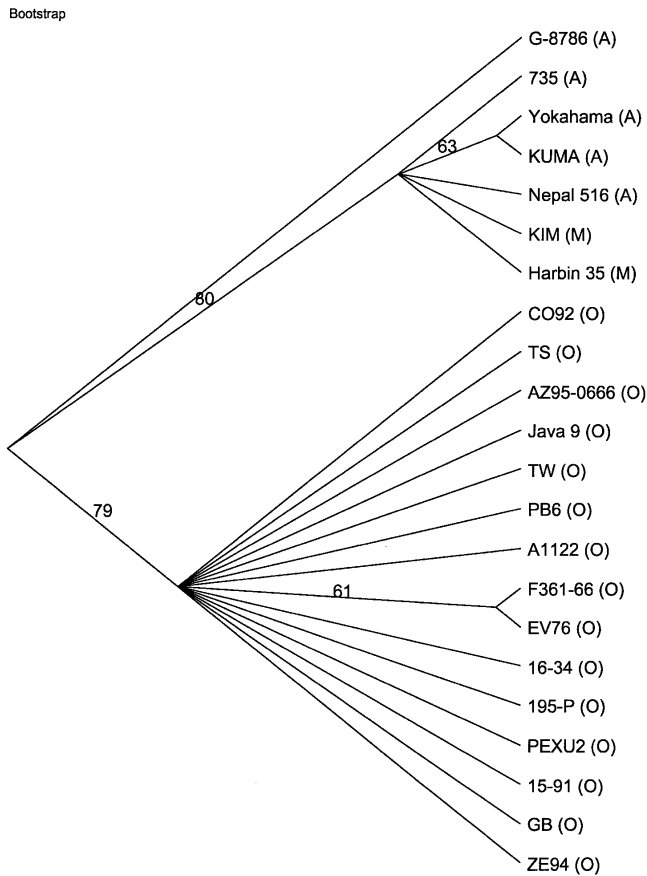


Figure 3 Parsimony analysis of *Y. pestis* strain microarray data. Bootstrap 50% majority-rule consensus tree with bootstrap values (1000 replicates) overlying branch points. A total of 22 characters were used, of which 8 were parsimony informative and given equal weight. Six equally parsimonious trees were found (see Supplemental Data, available at www.genome.org). Strain names are followed by their biovar in parentheses.

O-units to the periplasmic face, or polymerize them into O-antigen. However, these serotypes express LPS, indicating that the genes are present, but are divergent enough not to hybridize to the microarray. This particular region of the genome is important in the virulence of the organism in a murine infection model. Signature-tagged mutagenesis of the serotype O:3 strain YPIII pIB1 revealed that mutation in any of five of the LPS biosynthesis genes caused a decrease in virulence (Karlyshev et al. 2001). Thus, it is possible that the differences in LPS composition between serotypes may account for differences in the virulence of certain strains of *Y. pseudotuberculosis*. The lack of variation in *Y. pestis* may be due to the absence of the selective pressure, because these surface structures are no longer expressed. It has been shown that *Y. pestis* contains mutations in five of the O-antigen cluster genes, and thus produces lipo-oligosaccharide (LOS) lacking O-antigen (Skurnik et al. 2000; Parkhill et al. 2001; Prior et al. 2001).

The other region that showed high levels of divergence from the *Y. pestis* CO-92 genome in all 10 of the *Y. pseudotuberculosis* strains is part of 1 of the 2 cryptic flagella operons (Table 4). In contrast to *Y. pestis*, *Y. pseudotuberculosis* strains are highly motile. However, the CO-92 genome sequence revealed the presence of two cryptic flagella operons, YPO0704–YPO0747 and YPO1790–YPO1842. The YPO0704–YPO0747 operon is highly unlikely to be functional in *Y. pestis* due to multiple frameshift

mutations (Parkhill et al. 2001; Deng et al. 2002). A large part of this operon (YPO0714–YPO0749) shows high levels of divergence in all *Y. pseudotuberculosis* strains tested (Table 4). The serotype O:3 strain, YPIII pIB1, is the only strain that appears to show little divergence from *Y. pestis* CO-92 in this region, with only the *fliF*, *fliG*, and *fliA* genes exhibiting divergence. The other operon (YPO1790–YPO1842) is potentially fully functional in CO-92 and is also present in all of the *Y. pestis* and *Y. pseudotuberculosis* strains tested, with no apparent divergence from CO-92 in any strain. This indicates that this operon is not under any selective pressure to vary and may not be expressed in either species.

Variable Determinants That May Be Important in Host Adaptation and Virulence

To elucidate the mechanisms of bacterial virulence and the evolution of virulence, we focused our analysis on those sequences that encode known or putative virulence factors. These are grouped into three categories, those potentially involved in the colonization of insects, those involved in adhesion and invasion of mammalian cells, and other bacterial toxins.

Insect Colonization

There are seven potential insecticidal toxins encoded by the CO-92 chromosome, which are similar to those produced by the insect pathogen *Photobacterium luminescens* (Waterfield et al. 2001). Three are members of the *tca* gene family (*tcaA*, *tcaB*, and *tcaC*), and four are paralogs of the *tcc* genes. The function of these genes in the pathogenic yersiniae has yet to be elucidated. The putative insect toxins (Tcc) encoded by *Y. pseudotuberculosis* show significant variation between strains. The *tccC* paralog, *sepC* (YPO2380), is absent from all *Y. pseudotuberculosis* strains tested, whereas two of the serotype O:3 strains are also deficient in another *tccC* paralog (YPO3673). Due to the high level of identity of YPO3673 with the adjacent YPO3674, lower levels of hybridization signal were observed for both genes (Table 2). PCR analysis was used to confirm the absence of YPO3673 rather than YPO3674 (data not shown).

The second insect toxin complex *tcaABC*, encoded by genes YPO3678–YPO3681, also showed divergence within *Y. pseudotuberculosis* strains (Table 2). The *tcaA* gene appeared highly conserved in all 10 *Y. pseudotuberculosis* strains. However, only two serotype O:3 strains possessed DNA sequences with high similarity to the CO-92 *tcaB* gene. The other 8 strains give a weakened signal for *tcaB*, whereas all 10 strains give a weakened signal for *tcaC*. PCR analysis confirmed the presence of *tcaB* and *tcaC* paralogs within these *Y. pseudotuberculosis* strains (data not shown). One explanation for this divergence may be that the expression of the *tcaB* and *tcaC* paralogs encoded by *Y. pseudotuberculosis* may be more toxic for the flea vector. *Y. pestis* has to persist in the flea gut for relatively long periods, thus potent insect toxins may be detrimental to its life cycle. The *tcaB* gene has been shown to have a frameshift mutation and *tcaC* an internal deletion in the CO-92 genome sequence (Parkhill et al. 2001).

Another sequence that may be related to the parasitism of insects, YPO0339, a putative enhancin encoded by baculoviral pathogens of insects (Parkhill et al. 2001), appeared conserved among all *Y. pestis* and *Y. pseudotuberculosis* strains.

Adhesion and Invasion of Mammalian Cells

Three putative autotransporter genes, *yapA*, *yapB*, and *yapH* show variation among *Y. pseudotuberculosis* strains. Eight of the ten strains were positive for *yapA*, the two exceptions were both serotype O:3. All strains hybridized to *yapB*, although the hybridization signal was reduced in seven strains, suggesting that this gene

Table 2. Potential Virulence Determinants Which Are Absent or Highly Divergent in Some of the 10 *Y. pseudotuberculosis* Strains Tested

		354 (O:1b)	Pa3606 (O:1b)	YPIII pIB1 (O:3)	83 (O:3)	Pa3423 (O:3)	197 (O:5b)	141 (O:7)	wla872 (O:7)	SP93422 (O:15)	SP940616 (O:15)
YPO0599–602	Putative adhesin.	–	–	+	+	–	–	–	–	+	+
YPO0947	Putative virulence determinant	+	0.65	+	+	0.75	0.5	+	+	0.7	+
YPO1002	Enterotoxin-like protein	0.6	0.6	+	+	+	+	0.65	+	0.5	+
YPO1003	Putative exported protein	+	0.7	+	+	+	+	0.6	+	0.6	0.65
YPO1004	Putative autotransporter <i>yapH</i> .	+	–	+	+	+	+	–	–	–	+
YPO1898–1917	Yersiniabactin locus.	–	–	–	–	–	–	–	–	+	–
YPO2044	Putative hemolysin activator protein.	+	0.3	+	+	+	0.3	0.3	0.25	0.3	0.3
YPO2045	Putative hemolysin.	0.4	0.4	+	+	0.5	0.4	0.2	0.2	0.3	0.45
YPO2150	LysR-family transcriptional regulator.	+	+	+	+	+	–	–	+	–	–
YPO2490	Putative hemolysin	+	–	0.5	–	–	0.2	0.3	+	–	–
YPO2886	Putative autotransporter <i>yapA</i> .	+	+	–	–	+	+	+	+	+	+
YPO2887	Putative autotransporter <i>yapB</i> .	+	+	0.55	0.7	0.7	0.5	0.7	+	0.6	0.7
YPO3673	Putative insecticidal toxin TccC.	+	+	0.35	0.35	+	+	+	+	0.7	0.75
YPO3674	Putative insecticidal toxin TccC.	+	+	0.4	0.5	+	+	+	+	+	+
YPO3678	Putative insecticidal toxin TcaC.	0.4	0.3	0.5	0.6	0.5	0.5	0.4	0.3	0.4	0.45
YPO3679	Putative insecticidal toxin TcaB.	0.4	0.35	+	+	0.65	0.55	0.6	0.4	0.35	0.4
YPO3681	Putative insecticidal toxin TcaA.	+	+	+	+	+	+	+	+	+	+
YPO3682	LysR-family transcriptional regulator.	+	0.3	+	+	0.45	0.4	0.5	0.7	0.5	0.7
YPO3720	Putative hemolysin activator protein	+	+	0.4	0.4	+	+	0.7	+	0.7	0.75
YPO3883–3885	Colicin and immunity protein	–	–	+	+	–	–	+	–	–	–

Selected genes that may be involved in virulence or host adaptation. *Y. pseudotuberculosis* strain names are followed by their serotype in parentheses. The presence of a gene is indicated by a plus sign (+), absence by a minus sign (–), and divergence by the signal to control ratio represented numerically.

may be slightly divergent from the CO-92 ortholog. Attempts at amplifying genomic DNA from strain YPIII pIB1, with two primer sets specific for CO-92 *yapB* failed in any of the four possible combinations, suggesting that the *yapB* gene has diverged sufficiently to affect primer annealing (data not shown). Six of the ten strains hybridized to *yapH*, including the two strains that lacked *yapA*. The other four strains amplified a truncated PCR product with *yapH*-specific primers, indicating a deletion within the gene (data not shown). Autotransporter proteins are capable of being secreted from the bacterial outer membrane without the aid of secretion pathways, or other proteins, and have been implicated in the virulence of bacterial pathogens (Henderson et al. 1998). The *E. coli* autotransporter TibA has strong amino acid similarity to both YapA and YapB and acts as an adhesin and invasin of human epithelial cells (Lindenthal and Elsinghorst 2001). It has been shown that noninvasive *E. coli* can be induced to invade epithelial cells by the expression of TibA. It is possible that divergence of *yapA*, *yapB*, and *yapH* in *Y. pseudotuberculosis* may be important in the spectrum of disease caused by this pathogen.

The putative adhesin, YPO0599, is absent from 7 of the 10 *Y. pseudotuberculosis* strains, along with 3 adjacent genes of unknown function. It appears to be a relatively recent loss from the *Y. pestis* genome, being absent from 8 of the 15 Orientalis strains, and only 1 of the more ancient strains. This indicates that YPO0599 was acquired by *Y. pseudotuberculosis* along with the

three adjacent genes, but has been subsequently lost by some of the *Y. pestis* strains. Other proteins involved in attachment and invasion of mammalian cells such as *inv* (YPO1793), *ail* (YPO1860 and YPO2905), *hmwa* (YPO3247), and the putative invasin encoded by YPO3944 were all present in all strains of *Y. pseudotuberculosis*. Also present was a coding sequence similar to *virK* from *Shigella flexneri*, which is required for intercellular spreading in this organism.

Other Toxins

Several coding regions with similarity to known pore-forming toxins were identified by the sequencing of the two *Y. pestis* strains, CO-92 and KIM10+. These include four putative hemolysins and activator proteins with similarity to Sh1A and Sh1B from *Serratia marcescens*. The Sh1A hemolysins from *S. marcescens* require phosphatidylethanolamine as a cofactor and can form pores in fibroblasts and epithelial cells as well as erythrocytes (for review, see Hertle 2000). Two of these putative hemolysins and two associated activators showed divergence in *Y. pseudotuberculosis* strains (Table 2). Divergence was also seen in a coding region with similarity to the pore-forming protein RtxA of *V. cholerae* (YPO0947). However, no divergence was apparent in genes encoding the putative RTX toxin transporters, RtxB (YPO2249) and RtxD (YPO2250). Thus, the toxin itself shows divergence and, therefore, possible variation in activity between strains, whereas

Table 3. Divergence of the O-Antigen Gene Cluster in the 10 Strains of *Y. pseudotuberculosis*

		354 (O:1b)	Pa3606 (O:1b)	YPIII pIB1 (:3)	83 (O:3)	Pa3423 (O:3)	197 (O:5)	141 (O:7)	wla872 (O:7)	SP93422 (O:15)	SP940616 (O:15)
YPO3096	O-antigen chain length determinant (<i>wzz</i>)	+	+	+	+	+	+	0.6	0.4	+	+
YPO3097	Phosphomannomutase (<i>manB</i>)	+	+	+	+	+	0.7	–	–	+	+
YPO3098	Glycosyltransferase (<i>wbyL</i>)	+	+	0.3	0.3	0.3	0.4	–	–	0.4	0.3
YPO3099	Mannose-1-phosphate guanylyltransferase (<i>manC</i>)	+	+	+	+	+	+	–	–	+	+
YPO3100	GDP-L-fucose synthetase (<i>fcl</i>)	+	+	+	+	+	+	–	–	+	+
YPO3102	GDP-mannose 4,6-dehydratase (<i>gmd</i>)	+	+	+	+	+	+	–	–	+	+
YPO3104	Mannosyltransferase (<i>wbyK</i>)	+	+	0.45	0.4	0.6	–	–	–	–	–
YPO3105	O-unit polymerase protein (<i>wzy</i>)	+	+	–	–	–	–	–	–	–	–
YPO3107	Mannosyltransferase (<i>wbyI</i>)	+	+	–	–	–	–	–	–	–	–
YPO3108	Glycosyltransferase (<i>wbyI</i>)	+	+	–	–	–	–	–	–	–	–
YPO3110	O-unit flippase (<i>wzx</i>)	0.5	0.4	–	–	–	0.5	–	–	0.6	+
YPO3111	Exported protein (<i>wbyH</i>)	+	+	–	–	–	0.7	–	–	0.7	+
YPO3112	CDP-paratose synthetase (<i>prt/rfbS</i>)	+	+	+	+	+	–	–	–	+	+
YPO3113	CDP-4-keto-6-deoxy-D-glycose-3-dehydratase (<i>ddhC</i>)	+	+	+	+	+	–	–	–	+	+
YPO3114	CDP-D-glucose-4,6-dehydratase (<i>ddhBlrfbG</i>)	+	+	+	+	+	–	–	–	+	+
YPO3115	Glucose-1-phosphate cytidyltransferase (<i>ddhA/rfbF</i>)	+	+	+	+	+	+	–	–	+	+
YPO3116	cdp-6-deoxy-delta-3,4-glucoseen reductase (<i>ddhD/rfbI</i>)	+	+	+	+	+	+	–	–	+	+

Hybridization of the highly variable O-antigen gene cluster by *Y. pseudotuberculosis* strains. Strain names are followed by their serotype in parentheses. The presence of a gene is indicated by a plus sign (+), absence by a minus sign (–), and divergence by the signal to control ratio represented numerically.

the toxin delivery system remains constant. The RTX toxin of *V. cholerae* requires an activator protein RtxC, a RTX activator has yet to be identified in *Yersinia* species. The other pore-forming toxin encoded by *Yersinia* species is the antimicrobial toxin colicin E8 along with its immunity protein. These appear to be absent from 7 of the 10 strains.

Nonpore-forming toxins identified in the sequenced *Y. pestis* strains include an enterotoxin-like protein similar to the iron-regulated sHET2 enterotoxin SenA from *S. flexneri* and *E. coli*. These toxins cause diarrhea by interfering with the signal-transduction pathways involved in regulating water and electrolyte fluxes across intestinal mucosa (Fasano et al. 1995). The cytotoxic necrotizing factor (YPO1449) shows no apparent divergence between the 10 *Y. pseudotuberculosis* strains tested, yet this toxin has been shown to be expressed in only a limited number of strains (Lockman et al. 2002).

Y. pestis and *Y. pseudotuberculosis* Evolution

Y. pestis seems to have adapted rapidly from being a mammalian enteropathogen, widely found in the environment, to a blood-borne pathogen of mammals that is able to parasitize insects, and has a limited capability for survival outside these hosts. Some of these changes may be a result of the genome differences identified in this study, in addition to the known importance of the pMT1 and pPCP1 plasmids. However, genome rearrangements, particularly as a result of the recombination of IS elements, and the accumulation of pseudogenes may also have played a significant role in the rapid evolution of *Y. pestis*. These types of genetic differences are difficult to identify by microarray analysis. Analysis of the acquired chromosomal regions has provided no clear

explanation as to the differences between the two species, but the acquisition of phage-related sequences may be significant. However, the data from this study has shown that sequences that may be related to the parasitism of insects (insecticidal toxins and baculovirus enhancer) in *Y. pestis* are found in a wide range of *Y. pseudotuberculosis* strains. This finding suggests that *Y. pestis* did not adapt to the flea gut in a single evolutionary event, but rather that *Y. pseudotuberculosis* had been associated with insect hosts or insect pathogens for a considerable time. *Y. pestis* has had to adapt to colonize, but not kill, the flea in order to be transferred between mammalian hosts.

The microarray data was also used to construct a phylogenetic tree for *Y. pseudotuberculosis* (Fig. 4). Due to differences in scale, all of the *Y. pestis* and *Y. pseudotuberculosis* strains could not be shown on the same figure, but reference *Y. pestis* strains were included in the *Y. pseudotuberculosis* analysis. The phylogenetic tree, on the basis of the genetic differences between *Y. pestis* strains, shows a clear divergence between the Orientalis strains and strains from the other two biovars (Fig. 3). However, no obvious evolutionary pathway from *Y. pseudotuberculosis* can be determined. Only 2 of the 10 regions determined to be absent from all *Y. pseudotuberculosis* strains were absent from any of the *Y. pestis* strains. Thus, it appears that the ancestral *Y. pestis* strain acquired at least eight of these regions before the three biovars diverged. The progenitor of the Antiqua and Mediaevalis biovars subsequently lost regions YPO0738–YPO0754 and YPO2375–YPO2376 before the individual strains diverged by the loss of strain-specific regions, whereas the Orientalis biovar acquired the filamentous CUS-2 prophage (YPO2271–YPO2281). Strain G-8786 probably diverged from the Antiqua and Mediaevalis progenitor before the loss of YPO0738–YPO0754 and YPO2375–YPO2376.

Table 4. Divergence of One of the Two Flagella Gene Clusters in the *Y. pseudotuberculosis* Strains

		354 (O:1b)	Pa3606 (O:1b)	YPIII pIB1 (O:3)	83 (O:3)	Pa3423 (O:3)	197 (O:5)	141 (O:7)	wla872 (O:7)	SP93422 (O:15)	SP940616 (O:15)
YPO0714	Flagellar M-ring protein <i>flif</i>	+	+	–	+	+	+	+	+	+	0.75
YPO0715	Flagellar motor switch protein <i>flig</i>	+	+	–	+	+	+	+	+	+	0.75
YPO0716	Flagellar assembly protein <i>flih</i>	+	+	+	+	+	+	+	+	+	0.75
YPO0717	Flagellum-specific ATP-synthase <i>flil</i>	+	+	+	+	+	+	0.75	+	+	0.65
YPO0718	Hypothetical protein	+	+	+	+	+	+	+	+	+	+
YPO0719	Hypothetical protein	+	0.3	+	+	0.4	0.75	+	+	+	0.75
YPO0720	Putative flagellar regulatory protein	0.7	0.3	+	+	0.3	+	+	0.7	+	0.75
YPO0721	Basal body P-ring formation protein <i>flgA</i>	0.7	0.5	+	+	0.6	+	0.65	+	0.7	0.7
YPO0722	Basal body rod protein <i>flgB</i>	0.7	0.6	+	+	0.7	0.7	0.6	0.6	0.6	0.6
YPO0723	Basal body rod protein <i>flgC</i>	+	0.7	+	+	+	0.75	0.7	+	0.75	0.65
YPO0724	Basal body rod modification protein <i>flgD</i>	0.7	0.6	+	+	0.5	0.6	0.6	0.7	0.7	0.5
YPO0725	Flagellar hook protein <i>flgE</i>	0.5	0.5	+	–	0.35	0.4	0.4	0.45	0.4	0.25
YPO0727f1	Basal body rod protein <i>flgF</i>	+	0.5	+	–	0.55	0.55	+	+	0.5	0.45
YPO0728	Basal body rod protein <i>flgG</i>	+	+	+	–	0.5	0.5	0.6	0.7	0.45	0.4
YPO0729	Flagellar L-ring protein <i>flgH</i>	+	0.6	+	–	0.7	0.7	0.75	0.7	0.6	0.5
YPO0730	Flagellar P-ring protein <i>flgI</i>	0.4	0.55	+	–	0.5	0.6	+	+	0.4	0.4
YPO0731	Putative flagellar protein <i>flgJ</i>	0.5	+	+	–	0.55	0.55	0.6	+	0.55	0.6
YPO0732	Flagellar hook-associated protein <i>flgK</i>	0.3	0.65	+	–	+	0.35	0.5	0.6	0.65	0.6
YPO0733	Flagellar hook-associated protein <i>flgL</i>	0.3	0.55	+	–	0.75	0.4	0.55	0.6	0.55	0.6
YPO0734	Hypothetical protein	+	+	+	0.4	+	0.6	+	+	+	+
YPO0735	Hypothetical protein	+	0.7	+	0.75	0.7	+	0.7	+	+	0.65
YPO0736	Putative regulatory protein	+	+	+	+	+	0.65	0.75	+	+	+
YPO0737	Flagellin <i>flaA1</i>	0.25	0.3	+	0.2	0.3	0.35	0.35	0.5	0.3	0.5
YPO0738	Flagellin <i>flaA2</i>	0.45	0.4	+	0.4	0.4	0.4	0.3	0.5	0.3	0.55
YPO0739	Flagellin <i>flaA3</i>	0.45	0.4	+	0.4	0.45	0.45	0.3	0.45	0.3	0.6
YPO0740	Flagellar hook-associated protein <i>flaD</i>	0.6	0.5	+	–	0.65	0.55	0.5	0.6	0.6	0.6
YPO0741	Putative flagellar protein <i>flaS</i>	+	+	+	–	+	+	+	+	+	+
YPO0742	Hypothetical protein	0.55	0.7	+	–	0.7	0.75	0.7	+	0.65	0.6
YPO0743	Flagellar hook length control protein <i>flaK</i>	+	0.75	+	0.75	0.75	0.75	0.7	0.7	0.7	0.65
YPO0744	Flagellar biogenesis protein	0.6	0.55	+	–	0.65	0.7	0.6	0.75	0.7	0.55
YPO0745	RNA polymerase flagellar sigma factor <i>flaA</i>	0.65	+	–	–	0.7	+	0.7	+	+	0.7
YPO0746	Chemotaxis protein <i>motA</i>	+	+	+	+	+	+	+	+	+	+
YPO0747	Chemotaxis protein <i>motB</i>	+	+	+	+	+	+	+	+	+	0.75
YPO0748	Hypothetical protein	+	+	+	+	+	0.75	0.7	+	0.7	0.7
YPO0749	Hypothetical protein	0.35	0.4	+	+	0.4	0.4	0.35	0.4	0.3	0.3

Hybridization of the highly variable flagellar region by *Y. pseudotuberculosis* strains. Strain names are followed by their serotype in parentheses. The presence of a gene is indicated by a plus sign (+), absence by a minus sign, (–), and divergence by the signal to control ratio represented numerically.

Although *Y. pestis* does not produce LPS with an O-antigen, the vestigial O-antigen cluster is still present, and this has been used to infer the evolution of *Y. pestis*. Previous sequence analysis of the O-antigen gene clusters from different strains of *Y. pseudotuberculosis* has indicated that *Y. pestis* evolved from a serotype O1b strain of *Y. pseudotuberculosis* (Skurnik et al. 2000). Although our studies confirm this finding (Table 3), looking at overall gene complement, we found that, whereas one O:1b strain (Pa3606) was closer to representative *Y. pestis* strains than other *Y. pseudotuberculosis* strains tested, the other O:1b strain (354) was as different as other *Y. pseudotuberculosis* strains from a range of serotypes (Tables 2 and 4; Fig. 4). Thus, although it is likely that *Y. pestis* evolved from a single O:1b strain, the O:1b strains as a whole appear to be no more similar to *Y. pestis* than any other serotype.

Our results provide the first detailed comparisons of differ-

ent strains of *Y. pestis* and *Y. pseudotuberculosis* at a genome-wide level. We show that the genes specifying serotype form a small minority of the chromosomal gene complement differences between *Y. pestis* and *Y. pseudotuberculosis*, and that an approach identifying all *Y. pseudotuberculosis* serotype O:1b strains as immediately ancestral to *Y. pestis* is potentially misleading. Our findings also provide further evidence that biovar typing *Y. pestis* strains based on the ability to reduce nitrate and ferment glycerol consistently correlates with true genetic differences for Orientalis biovar strains, but may not always show correct genetic groupings for Medievalis and Antiqua biovars. Genes that are unique to *Y. pestis* should now be characterized in more detail to determine whether they play a specific role in colonization of the flea, or in disease in mammals. Additionally, genes that are unique to *Y. pestis* might form the basis of alternative tests for the identification and molecular typing of *Y. pestis*.

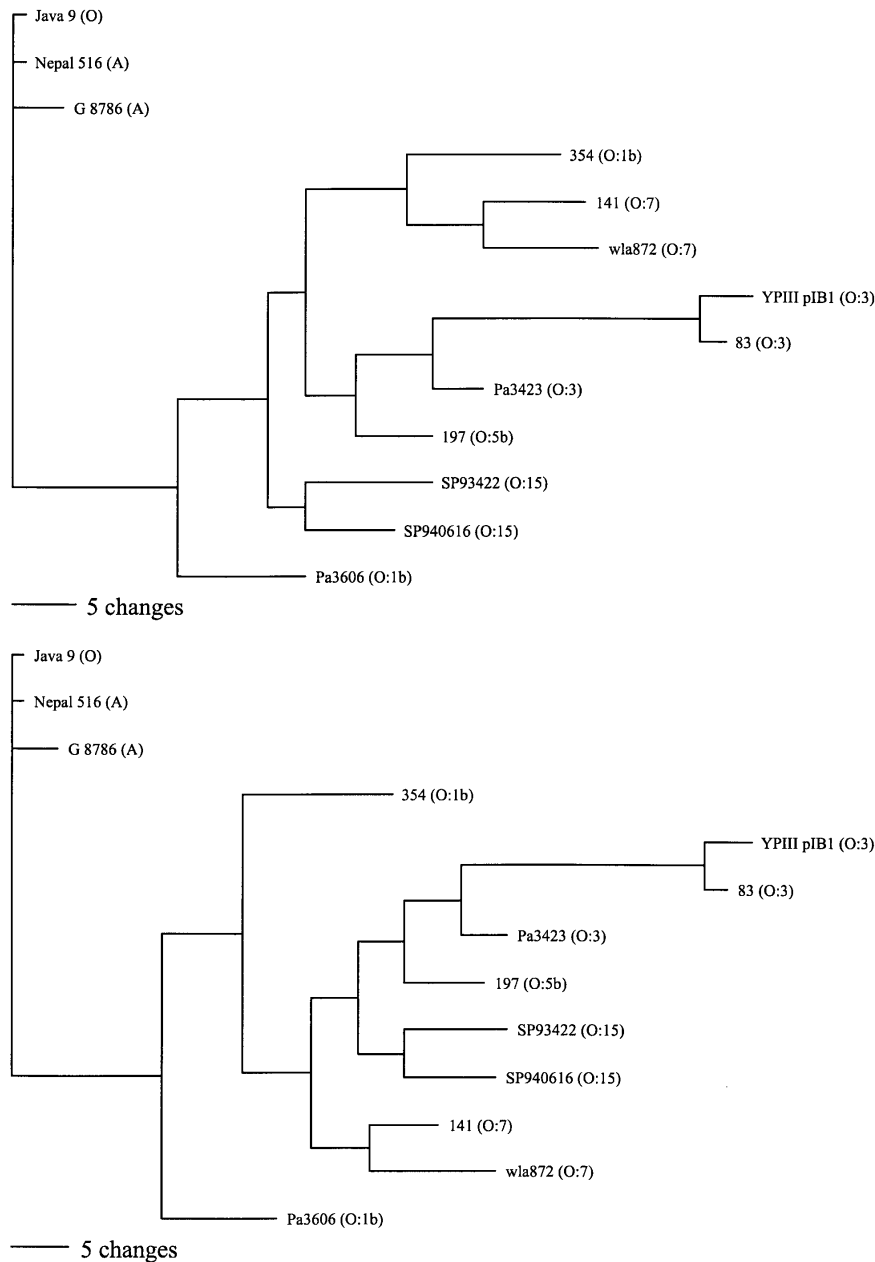


Figure 4 Parsimony analysis of *Y. pseudotuberculosis* and representative *Y. pestis* strain microarray data. The two most parsimonious trees are shown. Of 87 variable characters, 63 were parsimony informative and equally weighted. *Y. pseudotuberculosis* strain names are followed by their serotype in parentheses. Three *Y. pestis* strains are shown, one from each of the main clades in Fig. 3 and are followed by their biovar in parentheses. Horizontal scale bar indicates number of character changes (Reanalysis omitting all characters pertaining to the O-antigen locus gives a similar result, see Supplemental Data).

METHODS

Construction of the *Y. pestis* CO-92 Microarray

The *Y. pestis* CO-92 microarray was constructed from spotted PCR products designed and printed at the Bacterial Microarray facility at St. George's Hospital Medical School. Gene-specific primers were designed using Microarray Design (MAD) software (Hinds et al. 2002). This utilizes an algorithm to select a single PCR product for each gene that was unique and only self-detected in BLAST analysis of the whole genome. If a unique PCR product was not possible for a particular gene, then the algorithm selected a PCR

product from the BLAST analysis that demonstrated minimal cross-hybridization with other nontarget genes. PCR products were amplified in a 96-well plate format using an MWG Biotech RoboAmp 4200. The PCR products were gridded at high density on poly-L-lysine coated glass microscope slides using a BioRobotics MicroGrid II robot.

The finished array consisted of spotted PCR products representing each of the 4011 predicted coding sequences from the CO-92 chromosome together with the 209 predicted coding sequences from the three plasmids found in CO-92 (9 from pPCP1, 97 from pCD1, and 103 from pMT1).

Strains Used in Study

Y. pestis strains were provided by D. Tsereteli (National Centre for Disease Control, Tbilisi, Georgia), A. Rakin (Max von Pettenkofer-Institut für Hygiene und Medizinische Mikrobiologie, Munich, Germany), M. Chu (Centers for Disease Control and Prevention, Atlanta, USA), and R. Titball (DSTL, Porton Down, UK). The *Y. pseudotuberculosis* strain YPIII pIB1 was obtained from H. Wolf-Watz (Umea University, Sweden). All other *Y. pseudotuberculosis* strains were obtained from H. Fukushima (Shimane Prefectural Institute of Public Health and Environmental Science, Matsue, Japan).

Hybridization of Genomic DNA

Genomic DNA extraction and microarray hybridization procedures were performed as described previously (Dorrell et al. 2001). Test strains were labeled with Cy5-dCTP. CO-92 genomic DNA was used as a control on all hybridizations and was labeled with Cy3-dCTP. Labeling of from 2 to 3 μ g of denatured genomic DNA was performed using the Klenow fragment in the presence of 5 mM dATP, 5 mM dGTP, 5 mM dTTP, 2 mM dCTP, and 750 pM Cy5-dCTP or Cy3-dCTP. Klenow fragment was obtained from Invitrogen, and Cy-labeled dCTP was obtained from Amersham Pharmacia. After incubation at 37°C in the dark for 90 min, control and test strain-labeled DNA were mixed and purified together using a single QIAGEN mini-elute column. Microarray slides were prehybridized in $3.5 \times$ SSC, 0.1%SDS, 10 mg/mL BSA at 65°C. The denatured DNA was applied to the microarray slide in hybridization solution at a final concentration of $4 \times$ SSC,

0.3% SDS. Hybridization was for 18 h at 65°C, prior to washing the slides once in $1 \times$ SSC, 0.05% SDS at 65°C, and twice in $0.06 \times$ SSC for 2 min. All hybridizations were performed in duplicate.

Microarray Analysis

All 4221 of the CO-92 predicted coding sequences (4012 chromosomal and 209 plasmid encoded) are represented on the DNA microarray. Extensive analysis of the three plasmids was not undertaken, due to the frequent loss of these plasmids from many isolates. All gene numbers refer to the published CO-92 anno-

tated chromosomal sequence (EMBL Accession No. AL590842). During the construction of the DNA microarray, those predicted coding sequences that are disrupted by IS elements in CO-92 were treated as two separate genetic elements and represented on the array by two spots, one for each part of the sequence. Thus 4042 spots represent the 4012 predicted coding sequences in the CO-92 chromosome. Of these, a total of 73 spots were poor, failing to hybridize with CO-92 genomic DNA. The remaining 3969 genetic elements hybridized with the CO-92 genomic DNA, and were therefore used for the analysis.

Microarray slides were scanned using an Affymetrix 418 scanner (MWG Biotech), and images analyzed using Imagene (BioDiscovery) and Genespring (Silicon Genetics) software. Samples were normalized using the following conditions: All values <0.0 were set to 0.0. All sample signals were divided by the corresponding control signal from the same hybridization. The entire array was normalized to the 50th percentile of all measurements from that array.

Analysis of genomic DNA from multiple strains results in varied signal intensity due to sequence divergence between strains. Previous hybridization studies have shown that there is a sigmoidal relationship between the percentage nucleotide identity of a gene to the PCR product on the microarray, and the signal produced (Dong et al. 2001; Wu et al. 2001). The relative strength of the hybridization signal produced by genes of known percentage identity from genome sequence data can be used to estimate the relative divergence of genes from nonsequenced strains. To estimate the levels of divergence in our study, we analyzed the signal intensity generated by *Y. pseudotuberculosis* strains with previously sequenced O-antigen gene clusters. A signal ratio of 0.82 was observed for YPO3096 (*wzz*) from strain 197, which is 96.8% identical at the nucleotide level to the array PCR product for this gene. Similarly, signal ratios of 0.6 and 0.4 were observed for the YPO3110 (*wzx*) genes of strains SP93422 and Pa3606, which are 94.6% and 84.5% identical at the nucleotide level to the PCR product, respectively. Thus after normalization, genes were determined to be absent if the signal to control ratio was <0.2. Genes with a ratio of above 0.8 were presumed to be present and highly similar to the CO-92 gene. Genes that gave a ratio of between 0.2 and 0.8 were deemed to be present, but slightly divergent from the CO-92 gene. Genes that gave a signal intensity of approximately twice the fluorescence intensity in the Cy5 channel compared with the Cy3 CO-92 control were deemed to have been duplicated.

PCR Analysis

The presence of certain genes deemed absent or divergent was verified by PCR analysis. Gene-specific PCR primers, designed for the array construction, were used to amplify specific products from the genomic DNA of relevant strains, using Promega *Taq* DNA polymerase and standard conditions. If no PCR product was obtained, then a second set of gene-specific primers were designed and PCR amplification was repeated.

Phylogenetic Analysis

Y. pestis

Genes found present by the above microarray analysis were scored as two, absent genes were scored as zero, divergent genes were scored as one, and duplicated genes as D. IS elements were excluded from analysis. For interstrain comparison of *Y. pestis* isolates, all genes were present, absent, or duplicated with no divergent ORFs. Characters adjacent in the *Y. pestis* CO-92 genome sequence showing the same pattern of absence (signal 0) or duplication (D) in one or more of the tested strains compared with *Y. pestis* CO-92 were combined into single characters, which were incorporated into the PAUP* software program version 4.0b10 (PPC; <http://paup.csit.fsu.edu>). Maximum parsimony trees with equal weighting of characters were drawn following a heuristic search.

Y. pestis/Y. pseudotuberculosis

No duplicated genes were found in *Y. pseudotuberculosis* strains, and many genes were divergent and difficult to score as characters for phylogenetic analysis. Therefore, only loci for which at least one *Y. pseudotuberculosis* strain scored zero were incorporated in the dataset. The loci in each strain were scored as either present or absent. The genome contains several large operons (O-antigen, flagella, etc.), with shared patterns of absence, divergence, or presence of multiple ORFs in different strains, and which might be subject to strong selection. To deal with the possible non-independence of adjacent loci, characters that are adjacent in the *Y. pestis* CO-92 genome sequence and show the same pattern of presence and absence across all *Y. pseudotuberculosis* and *Y. pestis* strains were combined into a single character for the analysis. To check for possible biases introduced by selection (e.g., convergent evolution) the analysis was repeated with the omission of loci thought to be candidates for strong selection. Maximum parsimony trees were constructed using PAUP* (4.0b10) with data from the *Y. pseudotuberculosis* strains and three *Y. pestis* strains (one from each main clade).

ACKNOWLEDGMENTS

We acknowledge DSTL for funding this research, and the Wellcome Trust funded Bacterial Microarray facility at St. George's Hospital for the construction of the array. We thank D. Tsereteli, M. Chu, H. Wolf-Watz, and H. Fukushima for their generous gifts of purified genomic DNA or bacterial strains.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

- Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A., and Carniel, E. 1999. *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proc. Natl. Acad. Sci.* **96**: 14043–14048.
- Buchrieser, C., Prentice, M., and Carniel, E. 1998a. The 102-kilobase unstable region of *Yersinia pestis* comprises a high-pathogenicity island linked to a pigmentation segment which undergoes internal rearrangement. *J. Bacteriol.* **180**: 2321–2329.
- Buchrieser, C., Brosch, R., Bach, S., Guiyoule, A., and Carniel, E. 1998b. The high-pathogenicity island of *Yersinia pseudotuberculosis* can be inserted into any of the three chromosomal *asn* tRNA genes. *Mol. Microbiol.* **30**: 965–978.
- Darwin, A.J. and Miller, V.L. 1999. Identification of *Yersinia enterocolitica* genes affecting survival in an animal host using signature-tagged transposon mutagenesis. *Mol. Microbiol.* **32**: 51–62.
- Deng, W., Burland, V., Plunkett III, G., Boutin, A., Mayhew, G.F., Liss, P., Perna, N.T., Rose, D.J., Mau, B., Zhou, S., et al. 2002. Genome sequence of *Yersinia pestis* KIM. *J. Bacteriol.* **184**: 4601–4611.
- Devignat, R. 1951. Varietes de l'espece *Pasteurella pestis*. Nouvelle hypothese. *Bull. W.H.O.* **4**: 247–263.
- Dong, Y., Glasner, J.D., Blattner, F.R., and Triplett, E.W. 2001. Genomic interspecies microarray hybridization: Rapid discovery of three thousand genes in the maize endophyte, *Klebsiella pneumoniae* 342, by microarray hybridization with *Escherichia coli* K-12 open reading frames. *App. Environ. Micro.* **67**: 1911–1921.
- Dorrell, N., Mangan, J.A., Laing, K.G., Hinds, J., Linton, D., Al-Ghusein, H., Barrell, B.G., Parkhill, J., Stoker, N.G., Karlyshev, A.V., et al. 2001. Whole genome comparison of *Campylobacter jejuni* human isolates using a low-cost microarray reveals extensive genetic diversity. *Genome Res.* **11**: 1706–1715.
- Dunn, J.D., McCorkle, S.R., Praisman, L.A., Hind, G., van der Lelie, D., Bahou, W.F., Gnatenko, D.V., and Krause, M.K. 2002. Genomic signature tags (GSTs): A system for profiling genomic DNA. *Genome Res.* **12**: 1756–1765.
- Dziejman, M., Balon, E., Boyd, D., Fraser, C.M., Heidelberg, J.F., and Mekalanos, J.J. 2002. Comparative genomic analysis of *Vibrio cholerae*: Genes that correlate with cholera endemic and pandemic disease. *Proc. Natl. Acad. Sci.* **99**: 1556–1561.
- Fasano, A., Noriega, F.R., Maneval Jr., D.R., Chanasongcram, S., Russell, R., Guandalini, S., and Levine, M.M. 1995. Shigella enterotoxin 1: An enterotoxin of *Shigella flexneri* 2a active in rabbit small intestine in vivo and in vitro. *J. Clin. Invest.* **95**: 2853–2861.

- Ferber, D.M. and Brubaker, R.R. 1981. Plasmids in *Yersinia pestis*. *Infect. Immun.* **31**: 839–841.
- Fetherston, J.D. and Perry, R.D. 1994. The pigmentation locus of *Yersinia pestis* KIM6+ is flanked by an insertion sequence and includes the structural genes for pesticin sensitivity and HMWP2. *Mol. Microbiol.* **13**: 697–708.
- Fetherston, J.D., Schuetze, P., and Perry, R.D. 1992. Loss of the pigmentation phenotype in *Yersinia pestis* is due to the spontaneous deletion of 102kb of chromosomal DNA which is flanked by a repetitive element. *Mol. Microbiol.* **6**: 2693–2704.
- Fitzgerald, J.R., Sturdevant, D.E., Mackie, S.M., Gill, S.R., and Musser, J.M. 2001. Evolutionary genomics of *Staphylococcus aureus*: Insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic. *Proc. Natl. Acad. Sci.* **98**: 8821–8826.
- Galimand, M., Guiyoule, A., Gerbaud, G., Rasoamanana, B., Chanteau, S., Carniel, E., and Courvalin, P. 1997. Multidrug resistance in *Yersinia pestis* mediated by a transferable plasmid. *N. Engl. J. Med.* **337**: 677–680.
- Gonzalez, M.D., Lichtensteiger, C.A., and Vimr, E.R. 2001. Adaptation of signature-tagged mutagenesis to *Escherichia coli* K1 and the infant-rat model of invasive disease. *FEMS Microbiol. Lett.* **198**: 125–128.
- Gonzalez, M.D., Lichtensteiger, C.A., Caughlan, R., and Vimr, E.R. 2002. Conserved filamentous prophage in *Escherichia coli* O18:K1:H7 and *Yersinia pestis* biovar orientalis. *J. Bacteriol.* **184**: 6050–6055.
- Guiyoule, A., Grimont, F., Iteman, I., Grimont, P.A., Lefevre, M., and Carniel, E. 1994. Plague pandemics investigated by ribotyping of *Yersinia pestis* strains. *J. Clin. Microbiol.* **32**: 634–641.
- Hare, J.M. and McDonough, K.A. 1999. High-frequency RecA-dependent and -independent mechanisms of Congo red binding mutations in *Yersinia pestis*. *J. Bacteriol.* **181**: 4896–4904.
- Henderson, I.R., Navarro-Garcia, F., and Nataro, J.P. 1998. The great escape: Structure and function of the autotransporter proteins. *Trends Microbiol.* **6**: 370–378.
- Hertle, R. 2000. *Serratia* type pore forming toxins. *Curr. Protein. Pept. Sci.* **1**: 75–89.
- Hinds, J., Witney, A.A., and Vass, J.K. 2002. Microarray design for bacterial genomes. In *Methods in microbiology: Functional microbial genomics*. (eds. B.W. Wren and N. Dorrell), pp. 67–82. Elsevier Science, London, UK.
- Hinnebusch, B.J., Rudolph, A.E., Cherepanov, P., Dixon, J.E., Schwan, T.G., and Forsberg, A. 2002. Role of *Yersinia* murine toxin in survival of *Yersinia pestis* in the midgut of the flea vector. *Science* **296**: 733–735.
- James, R., Kleanthous, C., and Moore, G.R. 1996. The biology of E colicins: Paradigms and paradoxes. *Microbiology* **142**: 1569–1580.
- Karlyshev, A.V., Oyston, P.C., Williams, K., Clark, G.C., Titball, R.W., Winzler, E.A., and Wren, B.W. 2001. Application of high-density array-based signature-tagged mutagenesis to discover novel *Yersinia* virulence-associated genes. *Infect. Immun.* **69**: 7810–7819.
- Kato-Maeda, M., Rhee, J.T., Gingers, T.R., Salamon, H., Drenkow, J., Smittipat, N., and Small, P.M. 2001. *Genome Res.* **11**: 547–554.
- Lindenthal, C. and Elsinghorst, E.A. 2001. Enterotoxigenic *Escherichia coli* TibA glycoprotein adheres to human intestine epithelial cells. *Infect. Immun.* **69**: 52–57.
- Lockman, H.A., Gillespie, R.A., Baker, B.D., and Shakhnovich, E. 2002. *Yersinia pseudotuberculosis* produces a cytotoxic necrotizing factor. *Infect. Immun.* **70**: 2708–2714.
- Lucier, T.S. and Brubaker, R.R. 1992. Determination of genome size, macrorestriction pattern polymorphism, and nonpigmentation-specific deletion in *Yersinia pestis* by pulsed-field gel electrophoresis. *J. Bacteriol.* **174**: 2078–2086.
- Makela, P.H., Hovi, M., Saxen, H., Valtonen, M., and Valtonen, V. 1988. Salmonella, complement and mouse macrophages. *Immunol. Lett.* **19**: 217–222.
- Motin, V.L., Georgescu, A.M., Elliott, J.M., Hu, P., Worsham, P.L., Ott, L.L., Slezak, T.R., Sokhansanj, B.A., Regala, W.M., Brubaker, R.R., et al. 2002. Genetic variability of *Yersinia pestis* isolates as predicted by PCR-based IS100 genotyping and analysis of structural genes encoding glycerol-3-phosphate dehydrogenase (glpD). *J. Bacteriol.* **184**: 1019–1027.
- Parkhill, J., Wren, B.W., Thomson, N.R., Titball, R.W., Holden, M.T., Prentice, M.B., Sebahia, M., James, K.D., Churcher, C., Mungall, K.L., et al. 2001. Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **413**: 523–527.
- Perry, R.D. and Fetherston, J.D. 1997. *Yersinia pestis*—Etiological agent of plague. *Clin. Microbiol. Rev.* **10**: 35–66.
- Porwollik, S., Wong, R.M.-Y., and McClelland, M. 2002. Evolutionary genomics of *Salmonella*: Gene acquisitions revealed by microarray analysis. *Proc. Natl. Acad. Sci.* **99**: 8956–8961.
- Prior, J.L., Hitchen, P.G., Williamson, D.E., Reason, A.J., Morris, H.R., Dell, A., Wren, B.W., and Titball, R.W. 2001. Characterization of the lipopolysaccharide of *Yersinia pestis*. *Microbiol. Pathog.* **30**: 49–57.
- Radnedge, L., Agron, P.G., Worsham, P.L., and Andersen, G.L. 2002. Genome plasticity in *Yersinia pestis*. *Microbiology* **148**: 1687–1698.
- Salama, N., Guillemin, K., McDaniel, T.K., Sherlock, G., Tompkins, L., and Falkow, S. 2000. A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proc. Natl. Acad. Sci.* **97**: 14668–14673.
- Skurnik, M., Peippo, A., and Ervelä, E. 2000. Characterization of the O-antigen gene clusters of *Yersinia pseudotuberculosis* and the cryptic O-antigen gene cluster of *Yersinia pestis* shows that the plague bacillus is most closely related to and has evolved from *Y. pseudotuberculosis* serotype O:1b. *Mol. Microbiol.* **37**: 316–330.
- Waterfield, N.R., Bowen, D.J., Fetherston, J.D., Perry, R.D., and French-Constant, R.H. 2001. The *tc* genes of *Photobacterium*: A growing family. *Trends Microbiol.* **9**: 185–191.
- Wu, L., Thompson, D.K., Li, G., Hurt, R.A., Tiedje, J.M., and Zhou, J. 2001. Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Appl. Environ. Microbiol.* **67**: 5780–5790.

WEB SITE REFERENCES

- <http://paup.csit.fsu.edu>; Official Web site of PAUP phylogenetic analysis software.
- http://bbrp.llnl.gov/bbrp/bin/y.pseudotuberculosis_blast; BLAST server for *Yersinia pseudotuberculosis* genome sequencing project.

Received May 6, 2003; accepted in revised form July 8, 2003.