Title: Complete genome sequence of *Yersinia pestis* strains Antiqua and Nepal516: evidence of gene reduction in an emerging pathogen.

Running Title: *Yersinia pestis* genome sequences

Patrick S. G. Chain[1,2§], Ping Hu[3§], Stephanie A. Malfatti[1,2], Lyndsay Radnedge[1¥], Frank Larimer[2,4], Lisa M. Vergez[1,2], Patricia Worsham[5], May C. Chu[6], and Gary L. Andersen[3]*

[1]Biosciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA 94550

[2]Joint Genome Institute, Walnut Creek, CA

[3]Center for Environmental Biotechnology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

[4] Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831

[5]United States Army Research Institute of Infectious Diseases, Fort Detrick, MD 21702

[6]Centers for Disease Control and Prevention, Fort Collins, CO 80522

[§]These authors contributed equally to this work.

*Corresponding author.

Mailing address: Lawrence Berkeley National Laboratory, 1 Cyclotron Road, mailstop 70A3317, Berkeley, CA94720, USA

Phone: 519-495-2795

FAX: 510-486-7152

E-mail: GLAndersen@lbl.gov

[¥]Present address: Monogram Biosciences Inc., 345 Oyster Point Boulevard, South San Francisco, CA 94080

ABSTRACT

Yersinia pestis, the causative agent of bubonic and pneumonic plague, has undergone detailed study at the molecular level. To further investigate the genomic diversity among this group and to help characterize lineages of the plague organism that have no sequenced members, we present here the genomes of two isolates of the "classical" Antiqua biovar, strains Antiqua and Nepal516. The genomes of Antiqua and Nepal516 are 4.7 Mb and 4.5 Mb and encode 4,138 and 3,956 open reading frames respectively. Though both strains belong to one of the three classical biovars, they represent separate lineages defined by recent phylogenetic studies. We compare all five currently sequenced *Y. pestis* genomes and the corresponding features in *Y. pseudotuberculosis*. There are strain-specific rearrangements, insertions, deletions, single nucleotide polymorphisms and a unique distribution of insertion sequences. We found 453 single nucleotide polymorphisms in protein coding regions, which were used to assess evolutionary relationships of these *Y. pestis* strains. Gene reduction analysis revealed that the gene deletion processes are under selective pressure and many of the inactivations are probably related to the organism's interaction with its host environment. The results presented here clearly demonstrate the differences between the two Antiqua lineages and support the notion that grouping *Y. pestis* strains based strictly on the classical definition of biovars (predicated upon two biochemical assays) does not accurately reflect the phylogenetic relationships within this species. Comparison of four virulent *Y. pestis* strains with the human-avirulent strain 91001 provides further insight into the genetic basis of virulence to humans.

INTRODUCTION

Plague is a zoonotic disease, endemic throughout the world and is highly infectious in humans. The causative agent, *Yersinia pestis*, primarily infects a wide range of rodents and is transmitted via flea vectors. Throughout history, plague has ravaged human populations in three major pandemic waves: Justinian's plague (541-767 AD), which started in Africa and spread to Mediterranean, the Black death of 1346 to early 19th century may have originated in central Asia and spread from the Caspian Sea to Europe, and modern plague (since 1894), which began in southwest China and spread globally via marine shipping routes from Hong Kong. Although human disease is rare, *Y. pestis* is dangerous, highly infectious and thus has been identified as having potential for use in bioterrorism or as a biological weapon.

It was shown that *Y. pestis* recently diverged from *Y. pseudotuberculosis* – an enteropathogen, and likely comprises a clonal lineage (1, 3, 37, 40). *Y. pestis* strains have historically been classified according to their ability to utilize glycerol and reduce nitrate and have been grouped into three main subtypes, or biovars: Antiqua, Mediaevalis and Orientalis. Isolates from the Orientalis biovar have a worldwide distribution due to spread by steamship beginning 100 years ago. In contrast, isolates of Antiqua and Mediaevalis biovars are generally limited to localized regions containing long-term plague foci from enzootic rodent hosts in Africa and Central Asia. It has been argued that each of the biovars was associated with one of the plague pandemics (14, 20, 34) and recent studies have tried to provide direct evidence whether *Y. pestis* was associated with any of the historical pandemics (15, 44). DNA sequences from ancient human remains dispute this assertion that different biovars were responsible for each of the last three pandemics and suggest that instead, Orientalis-like *Y. pestis* may have been involved in all three (15). This suggestion remains highly controversial.

Isolates from the biovar Antiqua have been thought to represent a more ancestral branch of the plague pathogen, primarily due to their association with long-established plague foci as well as sharing an additional set of genetic regions with *Y. pseudotuberculosis* and "non-classical" (e.g. Microtus biovar) subspecies of *Y. pestis*. Our previous work using suppression subtractive hybridization demonstrated a pattern of difference fragments (DFR profiles), including a 15,603 bp segment of chromosomal DNA that was shared by *Y. pseudotuberculosis* and a portion of both the "non-classical" subspecies of *Y. pestis* and the "classical" biovar Antiqua (38). There are currently three completed genome sequences for *Y. pestis*, one each from the Orientalis, the Mediaevalis, and the "non-classical" Microtus biovars. To get a better understanding of the detailed genetic changes in a pathogen that is adapting to an intracellular lifestyle, we have sequenced two isolates from the classical "Antiqua" biovar. Strain Antiqua is fully virulent and possesses a DFR profile closest to *Y. pseudotuberculosis*. A *pgm*$^-$ derivative of the virulent strain Nepal516 was found to have a different DFR profile and is believed to represent a different lineage of this biovar. Comparison with the genome sequence of the previously sequenced *Y. pestis* strains as well as *Y. pseudotuberculosis* gave further insight into the loci required for the adaptation to an intracellular pathogenic lifestyle. Additional insight into the acquisition of virulence to humans was obtained by the comparison to the human-avirulent isolate, 91001.

MATERIALS AND METHODS

Bacterial strains

Y. pestis Nepal516 was isolated from a human infection in Nepal (possibly from a 1967 outbreak of pneumonic plague), while strain Antiqua was isolated from a human infection in Africa (1965, Republic of Congo). Both have been biochemically characterized to belong to the Antiqua biovar and carry the three previously described "virulence" plasmids found in most classical isolates of *Y. pestis*. Both strains have been used previously in a variety of studies (4, 21, 27, 36, 37, 45). The wild type Antiqua strain and a *pgm*$^-$ version of the Nepal516 strain were available and used in this genome sequencing project. The Nepal516 strain lacks the ~100 kb *pgm* region, including the high pathogenicity island, the pesticin/yersiniabactin complex, and the haemin storage locus, that are normally located between two parallel IS*100* insertion elements (5, 8, 18, 22, 29, 35, 42).

Construction, sequencing, and assembly

Genomic DNA was isolated from *Y. pestis* strains Antiqua and Nepal516. The two genomes were sequenced using the whole-genome shotgun method as previously described (9). Briefly, 3kb- and 8kb-sized, randomly sheared DNA fragments were isolated and cloned into pUC18 and pMCL200 respectively, for amplification in *Escherichia coli*. A larger fosmid library was constructed containing approximately 40kb inserts of sheared genomic DNA cloned into the pCC1Fos cloning vector. Double-ended plasmid sequencing reactions were performed from all three libraries at the Department of Energy Joint Genome Institute using ABI 3730xl DNA Analyzers and MegaBACE 4500 Genetic Analyzers as described on the JGI website http://www.jgi.doe.gov/.

Approximately 110,556 and 113,541 of sequences were assembled for Antiqua and Nepal516, respectively, producing an average of 11 fold coverage across the genomes. Processing of sequence traces, base calling and assessment of data quality were performed with PHRED and PHRAP (P. Green, University of Washington, Seattle), respectively. Assembled sequences were visualized with CONSED. The initial assemblies consisted of 154 and 113 contigs (≥ 20 reads per contig). Gaps in the sequence were primarily closed by resolving the many repetitive regions found within the genome. The remaining gaps were closed by primer walking on gap-spanning library clones or PCR products from genomic DNA. True physical gaps were closed by combinatorial (multiplex) PCR. Sequence finishing and polishing added roughly 300 reads and assessment of final assembly quality was completed as described (9).

For the genome of Nepal516, the ~70 kb pCD plasmid was underrepresented and was not completed as part of the sequencing project. Nepal516 is known to contain the pCD plasmid (Scott Bearden, personal communication). The existence of the pCD plasmid was verified by PCR in our laboratory (data not shown). We believe that the failure of obtaining sufficient quantity of pCD DNA for sequencing is due to particular laboratory conditions and has no biological implications on the sequences of the chromosome, pMT and pPCP plasmids.

Sequence analysis and annotation

Automated gene modeling was completed by combining results from Critica, Generation, and Glimmer modeling packages, and comparing the translations to GenBank's non-redundant (NR) database using basic local alignment search tool for proteins (BLASTP). The protein set was also searched against KEGG Genes, InterPro, TIGRFams, PROSITE, and Clusters of Orthologous Groups of proteins (COGS) databases to further assess function. Manual functional assignments were assessed on individual gene-by-gene basis as needed. Sequence alignment and

protein domain search tools (BLAST, CLUSTALW, Pfam, etc) were applied in various stages of comparison. CO92 gene nomenclature is used in this work when possible, other nomenclature is mentioned and used when no CO92 ortholog is available.

Single nucleotide polymorphism (SNP) analysis

Yersinia genomes are known to harbor extensive rearrangements as well as a large number of insertion sequence elements (IS) and other duplicated regions (10, 13, 33, 41). These repeats and insertion elements were excluded from consideration in SNP analysis. Genome-wide SNP discovery was achieved by whole genome alignments using the software package Mummer3 (28) and by subsequent orthologous gene alignments. For coding regions, pair-wise reciprocal BLASTP analyses were performed with the five sets of *Y. pestis* proteins. An ortholog pair was defined as reciprocal best top hits using a cutoff of 95% sequence identity. If an ortholog was not found in any one of the five genomes, the proteins were removed from further analysis. The sequences of the orthologous genes were used to find SNPs using Mummer3. Whole genome comparisons were also done using Mummer3. SNPs were selected from regions not covered by the ortholog alignment method described above. Synonymous and non-synonymous sites were calculated as follows: for every position in the genome, it was assessed whether it was located in an intergenic or a coding region; if it is in a coding region (excluding coding regions from insertion elements and other repetitive elements) and the nucleotides substitution results in no change in amino acid sequence, it was classified as a potential synonymous SNP site, otherwise it was regarded as a potential non-synonymous site.

Comparative analysis of gene deletions in *Y. pestis* genomes

We analyzed the loss of function patterns in all *Y. pestis* genomes, focusing on the presence and absence of protein functions. Some deletions, such as *tufB* deletion in Nepal516

(described below), were not included because of gene duplication. Complete datasets of proteins for each *Y. pestis* genome were downloaded from published reports or from the final annotations of the newly sequenced genomes. Transposases and enzymes related to insertion elements were removed. The final protein datasets for deletion analysis were 3723, 3909, 3896, 3769, 3777 and 3867 proteins for strains CO92, KIM, Antiqua, Nepal516, 91001 and *Y. pseudotuberculosis* IP32953, respectively. Pairwise alignments of proteins of all five *Y. pestis* genomes were accomplished by BLASTP. A protein function was deemed absent if there was no top hit greater than 95% identity or at least 75% of the query sequence. This analysis focused on the presence or absence of protein and functional representation, therefore, if the protein has a closely related paralog or is duplicated in the genome, it was considered present. Due to the differences in annotation (particularly with smaller gene calls), we applied a cutoff criteria to remove all small proteins since these were more frequently found to be differentially annotated across the *Y. pestis* genomes. With a size filter of 75 amino acids, we are certain to have missed a small number of real proteins smaller than 75 amino acids, such as the 61 amino acid carbon storage regulator *crs*A (YPO3304). The final set of proteins found to be absent in at least one genome was manually inspected with the aid of multiple sequence alignment tools CLUSTALW and nucleotide sequence alignment tool BLASTN. If the deletion was comprised solely of repetitive units, the protein was removed from this analysis, because the deletion mechanism may be different in those cases and may revert frequently. Additionally, these final sets of proteins were inspected in multiple genome alignments to distinguish annotation differences vs. true differences in the genomes. A similar set of criteria was employed to see if homologs of these proteins exist in the *Y. pseudotuberculosis* IP32953 genome.

Nucleotide sequence accession number and locus tag prefixes

The annotated sequences of the complete genomes of *Y. pestis* strains Antiqua and Nepal516 are available at GenBank/EMBL/DDBJ (accession numbers pending). The prefixes YPA and YPN are used for locus tags (gene identifier prefixes) in strains Antiqua and Nepal516, respectively. When referring to specific genes throughout the text, we use the CO92 gene numbers (prefixed with YPO) where possible, unless the gene does not exist in CO92 (or if it is clearer to use a different prefix), then the locus tags for a different genome are used and mentioned.

RESULTS

Genome overviews

The genomes of *Y. pestis* strains Antiqua and Nepal516 each consist of a single circular chromosome and the three virulence plasmids, pMT, pCD and pPCP, which are associated with most classical *Y. pestis* strains. Here we report all replicons but the pCD plasmid of Nepal516 (see Materials and Methods). The salient genomic features of each genome are detailed in Table 1, while gross chromosomal comparisons with those of the CO92, KIM and 91001 strains are summarized in Figure 1. Although the global characteristics of the five genomes are quite similar, a number of strain-specific insertions, deletions, rearrangements, and single nucleotide polymorphisms (SNPs) were identified, along with a unique distribution of insertion sequence (IS) elements (see Tables 2, 3, 4 and Supplemental data).

Strain-specific synonymous SNPs (sSNPs) and non- synonymous SNPs (nsSNPs)

The numbers of sSNPs and nsSNPs specific to one or to two genomes are shown in Figure 2. There are 57 sSNPs (135 nsSNPs) specific to strain 91001 compared with all other *Y. pestis* strains. While 27 of these sSNPs (49 nsSNPs) are shared with the ancestral *Y. pseudotuberculosis* IP32953, the remaining 30 sSNPs (and 86 nsSNPs) differ with respect to *Y. pseudotuberculosis* IP32953, indicating that they likely arose in 91001 since its lineage diverged from the remaining *Y. pestis* sequenced isolates. Likewise, the 27 sSNPs (and 49 nsSNPs) specific to 91001 (and identical to *Y. pseudotuberculosis*) are mutations predicted to have arisen in the other *Y. pestis* lineage which gave rise to the remaining sequenced strains (Figure 2). No SNPs (sSNPs or nsSNPs) are found to be specific to strain pairs Antiqua and KIM, Antiqua and Nepal516, CO92 and KIM, or CO92 and Nepal516. However, 4 sSNPs (and 6 nsSNPs) are found specifically in CO92 and Antiqua (i.e. CO92 and Antiqua share the same SNP state, while

all other *Y. pestis* strains have a different SNP state), and 6 sSNPs (and 11 nsSNPs) are specific to KIM and Nepal516. Taken together, these data suggest a separation of these four strains into two distinct branches: where Antiqua and CO92 belong to one branch and KIM and Nepal516 occupy the other (Figure 2). These 2 sets of sSNP and nsSNP mutations have accumulated in the short period of time after the KIM/Nepal516 and CO92/Antiqua lineages diverged but before each lineage further split into two (Figure 2). Thus, this analysis strongly supports the notion that although strains Antiqua and Nepal516 are grouped into the same biovar (Antiqua), they represent distinct lineages.

The genes harboring sSNPs (cumulatively for all the *Y. pestis* strains) can be distributed into functional gene categories (based on Clusters of Orthologous Groups) as shown in Figure 3. Non-synonymous SNPs are found distributed in 20 COG categories (Figure 3a), while synonymous SNPs were distributed in 19 COGs (Figure 3b). We investigated whether SNPs belonging to the various branches in Figure 2 are biased towards any functional categories, but no unique or biased distribution patterns were found. All strain-specific mutations share approximately 1/3 of the COG categories. Although some strain-specific SNPs are found in functional categories not represented in any other strain (sSNP: three categories are unique to 91001 and one category is unique to Antiqua; nsSNP: two categories are unique to 91001) (Figure 3a, 3b), the number of SNPs are too small to determine whether these were random events. No sSNP vs nsSNP bias was readily apparent (the average nsSNP/sSNP ratio is approximately 2.9) except for the large proportion of nsSNPs vs sSNPs in the cell wall/membrane biogenesis COG of 91001, with a ratio of 9:2.

The gene sequences and gene organization in the plasmids are highly conserved. There is only one synonymous SNP in the plasminogen activator protease and 5 SNPs (including small

deletions) in non-coding regions of all pPCP plasmids. There are 8 and 17 SNPs in the predicted gene sets of all plasmids pCD and pMT, respectively. The majority of these SNPs are in hypothetical proteins and are distributed more or less randomly across strains.

Insertion Sequence (IS) elements and genome rearrangements:

As determined by several groups already (10, 13, 33, 46), IS elements have expanded tremendously in *Y. pestis* since its divergence from *Y. pseudotuberculosis*, and have served as the delimiters for recombination events that have led to genomic deletions and genome rearrangements. Due to its presumed continued transposition activity in the wild, IS*100* elements have been successfully used for typing and grouping strains (32). We thus performed a detailed analysis of the precise locations of IS*100* and the other three major IS elements (IS*1541*, IS*285*, IS*1661*) as well as investigated their contribution to the observed rearrangements.

Each sequenced strain of *Y. pestis* has a unique set of IS elements (Table 2) and a core IS set that is shared among them. It was previously observed that *Y. pestis* shares a set of 12 IS elements with *Y. pseudotuberculosis* (3 of each major IS element) that appear to have been acquired by *Y. pseudotuberculosis* before the evolution of *Y. pestis* based on identical locations of insertion (10). In addition to this set of 12, the core set of IS elements shared among all five sequenced *Y. pestis* strains is 45 IS*1541*, 15 IS*100*, 11 IS*285* and 6 IS*1661*, indicative of elements present in the last common ancestor of all these sequenced strains. Several more (2 IS*1541*, 1 IS*100*, 2 IS*285*, 2 IS*1661*) are predicted to be shared among all 5 *Y. pestis* strains, but in at least one strain, these IS elements have been subsequently involved in a deletion event between two IS copies (leaving only one behind), or have been lost as part of a larger deletion. Similarly, the four classical *Y. pestis* strains (Antiqua, Nepal516, KIM and CO92) also share a subset of the remaining IS elements, distinct from non-classical strain 91001: 5 IS*1541*, 5 IS*100*,

3 IS*285* with four additional IS*100* elements that have likely been lost in a strain-specific manner via deletion as described above. In addition, there are 4 IS elements shared between CO92 and Antiqua, and 3 shared between KIM and Nepal516 (Table 2), supporting the SNP-based phylogeny in Figure 2 as well as that depicted by Achtman and colleagues (2). Interestingly, a small number of IS elements were shared by unexpected groups of strains that disagree with the proposed phylogeny: one IS*100* shared between 91001, Antiqua and CO92, one IS*100* shared between Antiqua and KIM; one IS*100* in Antiqua, CO92 and KIM; one IS*1541* in CO92 and in *Y. pseudotuberculosis* IP32053; one IS*285* in CO92, KIM and Nepal but not Antiqua; and one IS*1541* in Antiqua and in *Y. pseudotuberculosis* IP32053.

While IS*1541* was the most active IS element between the divergence of *Y. pseudotuberculosis* and the most recent common ancestor of the five *Y. pestis* strains, with 45 common IS*1541* insertions among all sequenced isolates, IS*100* has been the most active in more recent times, though not equally among the strains (Table 2). This suggests that insertion sequence activity may be punctuated and does not occur at a constant rate across different strains. With the exception of Nepal516, the number of unique IS*100* is significantly more than for any other IS element (Table 2), though the reason for this is unclear.

Despite their extensive sequence similarity, the *Y. pestis* genomes appear to be in a state of flux with respect to large genome rearrangements. Similar to previous observations, all breaks in colinearity between the *Yersinia* genomes occurs at IS elements or other repeated sequences. Differences in the GC skew patterns in *Y. pestis* genomes, including the many breaks observed in Antiqua (Figure 1) are also the result of rearrangements between IS elements as previously observed (10, 33, 46). Similarly, IS elements have played a large role in deletion events observed in the *Y. pestis* genomes. For example, strains KIM and CO92 have undergone overlapping

IS*1541*-mediated deletions of 32 kb (YP0966-YP0994, using 91001 nomenclature) and 21.5 kb (YP0966-YP0986, using 91001 nomenclature), respectively. This deletion accounted for the DFR patterns observed by suppression subtractive hybridization (37). A similar strain-specific, IS-mediated deletion of a phage region is described further below. Additionally, a IS*1661*-mediated 13 kb deletion in both KIM and Nepal516 has removed a large cluster of flagellar genes (YPO0738- YPO0754) including the flagellar RNA polymerase sigma factor and chemotaxis membrane proteins.

Functional Reduction

It has been postulated that the genomes of *Y. pestis* have undergone functional reduction as it made a transition from an oral-fecal pathogen causing gastroenteritis, to a vector-borne pathogen causing a fatal, invasive, septicemic disease, where genes have been inactivated by various mechanisms, such as deletions/insertions, frameshifts, interruptions by insertion elements, and homologous or even non-homologous recombination. We determined all the genes and associated functions that have been lost since *Y. pestis* emerged from *Y. pseudotuberculosis* (by comparing *Y. pestis* strains and identifying functional losses) to better understand *Y. pestis* diversity and evolution. We have identified a large number of strain-specific gene inactivations/deletions as well as some that are specific to only two of the five genomes (Table 3). Other than hypothetical proteins, there is one dominant category of proteins in these strain- or lineage-specific deletions: proteins contributing to the interaction of bacterium with its environment or host, including membrane proteins, membrane receptors, ABC transporters, flagellar proteins and chemotaxis proteins.

The number of strain-specific lost functions was not equal among all strains. In part, this can be explained by several large deletions that effectively delete many genes (functions) in a

single event. While strain 91001 and Antiqua had the greatest number of strain-specific function losses with 69 and 41 respectively, it is interesting to note that CO92 was found to have the fewest compared to other strains. Though strains Antiqua and 91001 share several gene inactivations (8 proteins), these are the result of independent mutations (homoplasy), while those lost in CO92 and Antiqua (12 proteins), as well as those lost in KIM and Nepal516 (16 proteins) share lineage specific, inherited mutations (function losses) that support the phylogenetic tree.

Of the 69 functional losses specific to 91001, 24 are hypothetical protein, seven are membrane proteins, seven are phage-related proteins, five are regulatory proteins and three are transporters. Some of these 91001-specific losses of function may be related to its human-avirulent phenotype. Twenty-one of the 69 belong to a single IS*285*-mediated deletion event (YPO2108 – YPO2134).

Ten of the 41 Antiqua-specific lost functions were the results of two deletion events. Several inactivated or deleted genes were predicted to be directly or indirectly involved in interactions with environment. For example, a glutathione S transferase, YPO2367, is missing from the Antiqua genome. This family of enzymes routinely responds to oxidative stress or detoxification, which can be encountered during entry into phagocytic cells (12). Since bacteria usually have multiple glutathione S transferases (CO92 has at least 4 based on the annotation), losing one may not have a distinct phenotype. Interestingly, there were several cases where similar functional inactivations/deletions were observed in two strains that affect different genes that may have overlapping functions. For example, a potassium efflux pump (YPO3129) was inactivated in Antiqua and a Na^+/H^+ antiporter (YPO2142) inactivated in KIM. The role of the Na^+/H^+ antiporter is sodium extrusion (24) and both the Na^+/H^+ antiporter and the potassium efflux are involved in pH homeostasis. Since both Antiqua and KIM are geographically limited

to localized regions, one possible explanation for the differential inactivations is local adaptation to selectively maintain one of the two similar functions. An alternate possibility is that neither gene is required, that the differences observed in these sequenced strains were random events, and that both genes are not required and are on their way to being lost from the *Y. pestis* gene pool. Whether the similarity of the deletion profiles mentioned above reflects adaptations to their environmental niches or convergent evolution remains to be investigated.

Of the 20 KIM-specific inactivated genes, seven are concentrated in one deletion, and the inactivation via nonsense mutation of YPO3038 (NapA, a periplasmic nitrate reductase) is proposed to be one of the causes of the nitrate negative phenotype of the Medievalis biovar. Nepal516 has 13 specific lost functions, including two transporters (YPO2835 and YPO1350), the chromosomally-encoded type III secretion system protein YPO0266, and two classical virulence factors (YPO2291 – a putative virulence factor and YPO0599 – a hemolysin/adhesin mentioned further below). Some of the deletions in Nepal516 have been previously demonstrated experimentally by microarray hybridizations (21). There are only four lost functions (pseudogenes YPO1087 and YPO3679, along with y1377 and y2928 using KIM nomenclature) that are CO92-specific losses, yet all are putative proteins without clear functional predictions.

No genetic regions were identified in the genomes of Antiqua or Nepal516 that were not present in at least one of the previously sequenced *Y. pestis* or *Y. pseudotuberculosis* strains. Though there were a number of Antiqua and Nepal516 genes or domains of genes that were found to differ significantly from the other strains, many of these differences consist of varying numbers of degenerate tandem repeat elements within the coding sequences of surface proteins, such as in the invasin YPO3944 described further below, and were not interpreted as losses of function. Surprisingly, our analysis revealed only one example of a strain-specific unique genetic

region with no similar DNA sequence in any of the other four sequenced strains or in the *Y. pseudotuberculosis* genome. This is a ssDNA prophage found inserted in CO92 (YPO2271 – YPO2281). This region has been found by PCR in all tested biovar Orientalis strains as well as a few African strains of the biovar Antiqua (10, 19).

Differences in putative virulence factors

Most characterized and putative classical virulence factors are identical throughout all five *Y. pestis* strains, including those found on the virulence plasmids, such as the pPCP-located plasminogen activator Pla required for successful subcutaneous infection (11), and the pMT-encoded murine toxin (Ymt) (23) and F1 capsular protein (16) (important for *Y. pestis* life cycle and vector-mammal transmission). Similarly, loci on the chromosome are also nearly identical between strains, such as the RTX-like toxin gene YPO0947, the attachment invasion locus *ail* (YPO2905) and two *ail*-like genes (YPO1860, YPO2190) are virtually identical among *Y. pestis* strains and with *Y. pseudotuberculosis* as well. Interestingly, a fourth ail-like gene (YPO2506) has been deleted from the Antiqua genome.

Other loci, known to differ between *Y. pestis* and *Y. pseudotuberculosis*, were also investigated. The *Y. pestis* invasin YPO1793 is interrupted by an IS*1541* in all strains, while putative adhesin YPO1562 interrupted by an IS*285* in all *Y. pestis* strains except for 91001, which harbors a nonsense mutation instead. Both are intact in *Y. pseudotuberculosis*. Similarly, RTX transporter YPO2250 and the TccC-family insecticidal toxin YPO2312 are frameshift pseudogenes in all *Y. pestis* strains but appear intact in *Y. pseudotuberculosis*.

A second TccC insecticidal toxin homolog, YPO2380 has been deleted only in *Y. pestis* KIM. Two additional TccC toxins are found in tandem in all *Y. pestis* strains (YPO3674, YPO3673), while only a single copy is found in *Y. pseudotuberculosis*. In *Y. pestis*, a second

family of insecticidal toxins is found upstream of these two TccC homologs and consists of a complex of three genes (YPO3681, YPO3679, YPO3678). While these are present and highly similar in amino acid sequence in *Y. pseudotuberculosis*, YPO3681 is inactivated by a frameshift only in Antiqua, and YPO3679 harbors a frameshift in CO92 only.

One *Y. pestis* hemolysin/adhesin, YPO0599, located within a possible pathogenicity island (YPO0641a-YPO0590) is different from that of *Y. pseudotuberculosis* at the C-terminus. Several additional CDSs can be seen downstream in *Y. pseudotuberculosis* that appear to be "modules", or adhesin fragments that share high similarity to portions of the C-terminus of this CDS. Different "modules" are found downstream of the *Y. pestis* gene. Interestingly Nepal516 is missing a large section of the C-terminal portion of this protein, likely due to a recombination between one of these modules and the corresponding section in the Nepal516 gene. The remainder of this pathogenicity island is highly similar between *Y. pestis* strains and *Y. pseudotuberculosis*. A similar module-recombination scenario is envisioned to have resulted in a modified C-terminus of hemolysin/adhesin YPO2490 in 91001 compared to the other *Y. pestis* strains and *Y. pseudotuberculosis*. A different mechanism, the expansion or contraction of degenerate repeat units within the putative invasin YPO3944, has resulted in different sized invasins in *Y. pseudotuberculosis* (5623 aa - amino acids), and the various *Y. pestis* strains (91001, 3108 aa; Nepal516, 4270 aa; KIM, CO92 and Antiqua, 3013 aa). Further study is required to understand any phenotypic effect these two classes of differences may have.

Loss of functional TufB in Nepal

The genomes of *Y. pestis* and of several other organisms have two copies of the elongation factor Tu (EF-Tu). Due to its highly conserved function and ubiquitous distribution, elongation factors are considered a valuable phylogenetic marker and have been used in

evolution studies of *Enterococci* (26), *Lactobacillus* (43) and other eubacteria (39). Interestingly, in *Y. pestis*, the two copies of this gene are not as conserved as the two genes of *Escherichia coli*, yet the conservation within each copy (among *Y. pestis* strains) is maintained. In *E. coli*, *tuf*A and *tuf*B gene products only differ by a single amino acid (7) and exhibit identical physical, chemical and catalytic properties (17). We found that all five *Y. pestis* genomes and *Y. pseudotuberculosis* have two copies of *tuf* genes (*tufA*, *tufB*) and the general operon structure is similar to that of *E. coli*, however, the sequence identity between the *Y. pestis tufA* and *tufB* is considerably lower. There are 17 amino acid differences (4%) and a total of 138 nucleotide changes (11.7%) in addition to a large C-terminal deletion in *tufB* of Nepal516. This result is unexpected, since previous studies show that duplicate *tuf* gene within a genome differ on average by 0.7% in nucleotide sequence (30). Whether the two copies of the *tuf* genes in *Y. pestis* have different origins requires further investigation. All six *Yersinia* TufA proteins are 100% identical and all *Y. pestis tuf*A genes have identical nucleotide sequences while *Y. pseudotuberculosis* has a single synonymous SNP (G to A). Although the *tuf*B gene products differ from *tuf*A, a similar conservation is evident. We have found however that the TufB of Nepal516 harbors a large C-terminal deletion which affects 57 aa (or 67%) of the GTP-EFTu-D3 domain, involved in binding charged tRNAs and EF-Ts(6). This deletion is likely to cause loss of functionality and thus we believe TufB is not functional in Nepal. It is not known whether the two copies are expressed under different conditions or have slightly different functions or kinetic properties, however, this deletion leads us infer that *Y. pestis* requires only one functional copy for its life cycle.

Comparison to the human-avirulent strain 91001

A previous comparison between strain 91001 and human virulent strains CO92 and KIM revealed a number of differences, including a 33-kb prophage-like sequence (YPO2096 – YPO2135) that was absent in 91001 but intact in both CO92 and KIM, and suggested that this difference may have contributed to this strain's lack of virulence in human (41). This entire region was also absent in *Y. pseudotuberculosis* (10) consistent with this claim. Our analysis shows that this phage-like region is intact in Nepal516, but is partially deleted in Antiqua (Table 4). While the deletion in 91001 can be attributed to recombination between two parallel IS*100* elements, the smaller deletion in Antiqua (from YPO2087 – YPO2106) is likely due to excision between two IS*285* sequences (Supplemental data). Since Antiqua is a fully virulent strain, the region deleted in Antiqua would not seem to contribute to human pathogenicity, however the remaining portion of the prophage region deleted from 91001 may indeed contain genes that are important for human virulence.

Previous comparisons with CO92 and KIM also revealed a list of 91001-specific pseudogenes that may be related to *Y. pestis* pathogenicity and host range (41). Our gene reduction analysis included two additional virulent strains and confirmed the presence or absence of orthologs in *Y. pseudotuberculosis* based on our cutoffs (refer to Materials and Methods). While all of these 91001-specific pseudogenes were also intact in Antiqua (Supplemental data), only one (YPO2258) was also found to be inactivated in Nepal, suggesting this gene has no impact on the avirulent property of 91001 in humans. Among the 91001-specific pseudogenes, there are only four that are also absent in *Y. pseudotuberculosis*, YPO0733 (flagellar hook-associated protein), YPO0737 (flagellin), YPO0962 (hypothetical protein) and YPO3110 (putative O-unit flippase). In addition, the nsSNP mutations that contribute to changes in single amino acids in many proteins may affect 91001 or other *Y. pestis* strain virulence in subtle ways.

For example, nsSNPs were found in some genes with a possible role in virulence (e.g. the Antiqua *ail* gene YPO2905 carries a nsSNP, as does the 91001 RTX toxin gene YPO0947), but the significance of these substitutions is not known.

DISCUSSION

This work presents the complete genome sequences for the two previously unsequenced *Y. pestis* major lineages (both designated Antiqua using classical nomenclature). Phylogenetic relationships were elucidated clearly with the distribution of synonymous SNPs (Figure 2). Since synonymous mutations do not affect protein functions (unlike nsSNPs or some IS elements), their accumulation is not under selective pressure, making this the least biased method for inferring evolutionary relationships. The distribution of sSNPs convincingly demonstrates that a single biovar Antiqua is an inaccurate phylogenetic representation supporting previous claims that categorize Antiqua strains into two groups (2, 10). Using terminology proposed previously (2), lineage 1.ANT (African strain Antiqua) is closely related to Orientalis strain CO92 while 2.ANT (Asian strain Nepal516) is more closely related to Medievalis strain KIM. These four "classical" isolates fall on a branch separate from the non-classical, human-avirulent Chinese strain 91001. This analysis also revealed a relatively rapid divergence of the four distinctive lineages from two ancestral lines for the "classical" *Y. pestis* strains. Although it is only possible to make very crude estimations of age of descent for these four lineages, the numbers of sSNPs are consistent with all of the lineages being present within the last 1,500 years of the 3 great pandemics (calculation not shown).

Comparison of all five *Y. pestis* sequences reveals extensive DNA sequence rearrangement, widespread gene reduction and strain-specific IS elements, as well as SNPs. It was previously reported that *Y. pestis* strains differ greatly in genome synteny, and that repeated

sequences most often were found at the borders of rearrangements (10, 13). Indeed, most rearrangements occur at IS elements and regardless of which genomes were chosen for two-way comparisons, we identified similar numbers of rearrangements to those previously observed between *Y. pestis* strains (13) and even between *Y. pestis* and *Y. pseudotuberculosis* (10) (data not shown). The question remains whether these observed rearrangements have any effect on transcription or whether this has an overall destabilizing influence on the genome.

The distribution pattern of IS elements in the sequenced strains generally supports the SNP-derived phylogeny with several IS elements shared across all "classical" strains but not in 91001 as well as IS elements found only in the CO92/Antiqua or KIM/Nepal516 pairs of strains. Only a few (5 in total) IS elements found to be shared by two or more strains did not conform to the predicted phylogeny (footnotes in Table 2); similar observations have been reported previously (2). Our analyses suggest that a small number of IS elements may have been precisely excised from their insertion locations, that identical insertion events have occurred in 2 different strains/lineages, or that there may be some limited horizontal transfer between *Y. pestis* strains that have resulted in mobilizing IS elements from one strain to a different strain/lineage (or alternatively, removing an IS element by introduction of wild type sequence). One example is an aminotransferase (YPO3250) that is disrupted by an IS*100* in all sequenced *Y. pestis* strains except Nepal516, which instead has the wild type gene and no trace of an IS*100*. These data also suggest that certain IS elements may not be useful for typing or grouping strains and may explain certain discrepancies in phylogenetic groupings using different methods.

Interestingly, the entire complement of IS*1541* (and almost all IS*1661*) elements in strain 91001 was acquired by the ancestor of all *Y. pestis* strains. In contrast, since 91001 diverged from the other strains, it has acquired a number of strain-specific IS*100* and IS*285* elements,

supporting the idea of actively integrating IS elements within the genome of *Y. pestis*. With the exception of Nepal516, IS*100* appears to have been more active (greater number of new transposition events) than other IS elements, but the reason for this is unknown.

Functional reduction analysis also generally agrees with the SNP-based phylogenetic tree (Figure 3) as well as with a more limited study that identified the loss of gene regions across a panel of *Y. pestis* isolates using a CO92 gene-specific microarray (21). Similar to the IS and SNP data, the four "classical" strains appear to share an evolutionary path distinct from strain 91001 based on functional reduction, and the KIM/Nepal516 and CO92/Antiqua pairs also exhibit a larger number of shared function loss. The exceptions are the result of independent mutations: two shared losses between KIM and Antiqua; one shared loss between Nepal516 and 91001; eight shared losses between Antiqua and 91001 and 16 shared losses between CO92 and KIM (Supplemental data and Table 3). The two shared function losses between KIM and Antiqua are a putative siderophore biosynthetic enzyme and a putative membrane protein. The predicted functions suggest that both of the proteins could be involved in interactions with the environment, therefore these losses may reflect adaptations to the *Y. pestis* microenvironment. Similarly, the single functional loss shared between Nepal516 and 91001 is the arabinose operon regulatory protein. Although the observed shared loss of function between Antiqua and 91001 contained several genes, they are exclusively in the prophage region described above and it is the result of independent deletion events. The shared losses between CO92 and KIM were possibly from a single deletion event.

Strain 91001 has the highest number of strain-specific losses of function with a total of 69. Interestingly, all but four of the 91001-specific pseudogenes have homologs with >90% identity in *Y. pseudotuberculosis*, suggesting that 91001 lost those genes while other virulent *Y.*

pestis strains retain them. It is possible that these genes may be involved in human-virulence and/or fitness in the human host. Some inactivated proteins may be related to pathogenicity, such as hemolysin (YPO2045), sulfatase and sulfatase modifier protein (YPO3046 and YPO3047), UPD-glycosyltransferase (YPO1985) and O-unit flippase-like protein (YPO3110) (Supplemental data). Hemolysin is a toxin that forms transmembrane channels and is involved in heme utilization and adhesion. The precise function of the sulfatase operon (YPO3046 and YPO3047) in *Y. pestis* is not known, however these enzymes belong to a family of proteins that hydrolyze various sulfate esters or catalyze sulphur insertions. In mammalian cells, the oligosaccharide moieties on glycoproteins, glycolipids and proteoglycans are frequently modified with sulfate. Sulfatase from pathogenic bacteria have been shown to interact with mucin (47) and a previous study suggested that mucin-sulphatase activity in *Burkholderia cepacia* and *Pseudomonas aeruginosa* may contribute to their association with airway infection in cystic fibrosis patients by possibly facilitating bacterial colonization (25). Thus, the deletion of the sulfatase and sulfatase modifier protein in strain 91001 may have contributed to its human-avirulent phenotype. Finally, the O-unit flippase is involved in translocating polysaccharide unit across the membrane while UDP-glycosyltransferase (YPO1985) is typically involved in O-antigen biosynthesis. Since *Y. pestis* is known to lack O-antigen, the actual functions of YPO3110 and YPO1985 may not directly involve O-antigen but perhaps other surface polysaccharides.

Antiqua also had a high number of strain-specific losses, even after discounting the deletion events which involved several genes (41 and 31, respectively). Interestingly, we found a correlation with the observed higher IS*100* transposition activity in Antiqua, with 13 of the 31 inactivations due to IS*100* interruptions. The profile of Antiqua-specific loss of function contains a significant amount of proteins which interact with environment, such as glutathione S-

reductase, chemotaxis protein, porin C protein, potassium efflux pump, insecticidal toxin, flagellar motor switch protein and 6 membrane proteins without specific known functions. A possible explanation for this may be that the genome has been adapting to the niche the Antiqua organism occupies.

Discounting those genes lost in a single deletion event, the numbers of KIM-specific (14) and Nepal-specific (13) functional loss are similar. Surprisingly, only three CO92-specific losses of function were identified. It is possible that there was a selective advantage for the Orientalis biovar to maintain a greater repertoire of genes and to maintain flexibility and be able to adapt quickly to new host(s). The world-wide distribution of this group and the small number of CO92-specific putative gene inactivations is consistent with this theory. A 31 amino acid deletion in YPO3937 (473 amino acids) confers the glycerol negative phenotype of biovar Orientalis (33), however since the deletion is below length cutoff threshold, it was not included in our study as a loss of function. Unique to strain CO92 are a hypothetical protein (YPO2469), a hemolysin activator protein (YPO3720) and a prophage that do not exist or have been inactivated in the other sequenced *Y. pestis* strains or *Y. pseudotuberculosis*. These genes may again have been retained by CO92 to maintain its ability to interact with a more variable environment.

Unexpectedly we found that Nepal516 has many exceptions compared with the other sequenced *Y. pestis* strains, including the apparent loss of function of TufB, the much smaller number of Nepal516-specific SNPs compared to that of the other strains (Figure 2), and the fact that IS*100* has not been as active in Nepal516 as in the other strains (Table 2). Since both nsSNPs and sSNPs are equally affected, it is unlikely that this is due to selective pressure which should have a neutral effect on sSNPs, but rather that the mutation rate is responsible, suggesting the rate of mutation or evolution is slower in Nepal. The reason for this is not known, however a

possible explanation may be that this phenomenon is driven by fewer rounds of bacterial division with a relatively cooler local environment and hibernation of the host(s) that fostered fewer opportunities for transmission.

Despite the observed differences between different strains of *Y. pestis*, the sequenced genomes reveal a highly conserved chromosomal backbone, reminiscent of what is observed in *Bacillus anthracis* (31). Within the five genomes of *Y. pestis* compared here, a single region present in strain CO92 was found to be unique (not shared with another *Y. pestis* genome), though independent studies have shown that this region which encodes phage genes is present in most, if not all 1.ORI strains as well as some 1.ANT strains (10, 19). We thus believe that most of the genomic sequence shared among the "classical" *Y. pestis* isolates is represented within this data set, though other sequences of non-classical isolates may harbor novel genomic regions not revealed in these analyses.

Conclusions

The two completed genomes presented here, from the previously unrepresented Antiqua biovar, have provided important references for SNP discovery, for the study of insertion element distribution, genome rearrangement, and reductive evolution in *Y. pestis*. Comparisons of the four virulent "classical" strains to the human-avirulent strain 91001 have also provided further insight into *Y. pestis* human virulence. With sSNPs as the preferred method for elucidating phylogenetic relationships, strains Nepal516 and Antiqua were convincingly placed in two clearly separate branches, with one branch shared by strains KIM (Mediaevalis) and Nepal516, and the other shared by strains CO92 (Orientalis) and Antiqua. While IS element distributions and function loss across the strains generally agreed with such a phylogenetic representation, certain exceptions were found and are thought to be the result of lack of selective pressure in the

Y. pestis strains inhabited niche, of possible horizontal gene exchange between *Y. pestis* strains, or of homoplasy in the reductive processes. Though there is some evidence of convergent evolution, whether this is the primary mechanism underlying the observed discrepancies remains to be investigated. The *Y. pestis* genome is a clear example of one actively undergoing reductive evolution, as its lifestyle has altered from an enteropathogen to an intracellular pathogen. The genome has slowly accumulated inactivations and deletions that result in loss of function, which, for the virulent strains (all strains except 91001), have little effect on pathogenicity. The differences between these strains and the human-avirulent 91001 provide an ideal starting point for future experiments to elucidate the mechanisms involved in *Yersinia* pathogenicity.

ACKNOWLEDGEMENTS

We thank Matt Van Ert, Ryan Easterday, Aubree Hinckley, and Maria Shin for technical assistance. This work was performed under the auspices of the US Department of Energy by the University of California, Lawrence Berkeley National Laboratory under Contract NO. DE-AC02-05CH11231. This work was supported by an Intelligence Technology Innovation Center grant.

REFERENCES

1. **Achtman, M.** 2004. Age, Descent and Genetic Diversity within *Yersinia pestis*, p. 432. *In* E. Carniel and B. J. Hinnebusch (ed.), Yersinia: Molecular and Cellular Biology. Horizon Bioscience, Norwich, UK.

2. **Achtman, M., G. Morelli, P. Zhu, T. Wirth, I. Diehl, B. Kusecek, A. J. Vogler, D. M. Wagner, C. J. Allender, W. R. Easterday, V. Chenal-Francisque, P. Worsham, N. R. Thomson, J. Parkhill, L. E. Lindler, E. Carniel, and P. Keim.** 2004. Microevolution and history of the plague bacillus, *Yersinia pestis*. PNAS **101:**17837-17842.

3. **Achtman, M., K. Zurth, G. Morelli, G. Torrea, A. Guiyoule, and E. Carniel.** 1999. *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. PNAS **96:**14043-14048.

4. **Adair, D. M., P. L. Worsham, K. K. Hill, A. M. Klevytska, P. J. Jackson, A. M. Friedlander, and P. Keim.** 2000. Diversity in a Variable-Number Tandem Repeat from *Yersinia pestis*. J. Clin. Microbiol. **38:**1516-1519.

5. **Bearden, S. W., and R. D. Perry.** 1999. The Yfe system of *Yersinia pestis* transports iron and manganese and is required for full virulence of plague. Molecular Microbiology **32:**403-414.

6. **Berchtold, H., L. Reshetnikova, C. O. A. Reiser, N. K. Schirmer, M. Sprinzl, and R. Hilgenfeld.** 1993. Crystal structure of active elongation factor Tu reveals major domain rearrangements. Nature **365:**126-132.

7. **Blattner, F. R., G. Plunkett, III, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H.**

A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao. 1997. The Complete Genome Sequence of *Escherichia coli* K-12. Science **277:**1453-1462.

8. **Buchrieser, C., C. Rusniok, L. Frangeul, E. Couve, A. Billault, F. Kunst, E. Carniel, and P. Glaser.** 1999. The 102-Kilobase pgm Locus of *Yersinia pestis*: Sequence Analysis and Comparison of Selected Regions among Different *Yersinia pestis* and *Yersinia pseudotuberculosis* Strains. Infect. Immun. **67:**4851-4861.

9. **Chain, P., J. Lamerdin, F. Larimer, W. Regala, V. Lao, M. Land, L. Hauser, A. Hooper, M. Klotz, J. Norton, L. Sayavedra-Soto, D. Arciero, N. Hommes, M. Whittaker, and D. Arp.** 2003. Complete Genome Sequence of the Ammonia-Oxidizing Bacterium and Obligate Chemolithoautotroph *Nitrosomonas europaea*. J. Bacteriol. **185:**2759-2773.

10. **Chain, P. S. G., E. Carniel, F. W. Larimer, J. Lamerdin, P. O. Stoutland, W. M. Regala, A. M. Georgescu, L. M. Vergez, M. L. Land, V. L. Motin, R. R. Brubaker, J. Fowler, J. Hinnebusch, M. Marceau, C. Medigue, M. Simonet, V. Chenal-Francisque, B. Souza, D. Dacheux, J. M. Elliott, A. Derbise, L. J. Hauser, and E. Garcia.** 2004. Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. PNAS **101:**13826-13831.

11. **Cowan, C., H. A. Jones, Y. H. Kaya, R. D. Perry, and S. C. Straley.** 2000. Invasion of Epithelial Cells by *Yersinia pestis*: Evidence for a *Y. pestis*-Specific Invasin. Infect. Immun. **68:**4523-4530.

12. **De Groote, M. A., U. A. Ochsner, M. U. Shiloh, C. Nathan, J. M. McCord, M. C. Dinauer, S. J. Libby, A. Vazquez-Torres, Y. Xu, and F. C. Fang.** 1997. Periplasmic

superoxide dismutase protects *Salmonella* from products of phagocyte NADPH-oxidase and nitric oxide synthase. PNAS **94:**13997-14001.

13. **Deng, W., V. Burland, G. Plunkett III, A. Boutin, G. F. Mayhew, P. Liss, N. T. Perna, D. J. Rose, B. Mau, S. Zhou, D. C. Schwartz, J. D. Fetherston, L. E. Lindler, R. R. Brubaker, G. V. Plano, S. C. Straley, K. A. McDonough, M. L. Nilles, J. S. Matson, F. R. Blattner, and R. D. Perry.** 2002. Genome Sequence of *Yersinia pestis* KIM. J. Bacteriol. **184:**4601-4611.

14. **Devignat, R.** 1951. Varietes de l'especе *Pasteurella pestis*: nouvelle hypothese. Bull. W. H. O **4:**247-263.

15. **Drancourt, M., V. Roux, L. V. Dang, L. Tran-Hung, D. Castex, V. Chenal-Francisque, H. Ogata, P. Fournier, E. Crubézy, and D. Raoult.** 2004. Genotyping, Orientalis-like *Yersinia pestis*, and Plague Pandemics. Emerg Infect Dis. **10:**1585-1592.

16. **Du, Y., R. Rosqvist, and A. Forsberg.** 2002. Role of Fraction 1 Antigen of *Yersinia pestis* in Inhibition of Phagocytosis. Infect. Immun. **70:**1453-1460.

17. **Furano, A. V.** 1977. The elongation factor Tu coded by the tufA gene of *Escherichia coli* K-12 is almost identical to that coded by the tufB gene. J. Biol. Chem. **252:**2154-2157.

18. **Gehring, A. M., E. DeMoll, J. D. Fetherston, I. Mori, G. Mayhew, F. R. Blattner, C. T. Walsh, and R. D. Perry.** 1998. Iron acquisition in plague: modular logic in enzymatic biogenesis of yersiniabactin by *Yersinia pestis*. Chem Biol. **5:**573-586.

19. **Gonzalez, M. D., C. A. Lichtensteiger, R. Caughlan, and E. R. Vimr.** 2002. Conserved Filamentous Prophage in *Escherichia coli* O18:K1:H7 and *Yersinia pestis* Biovar orientalis. J. Bacteriol. **184:**6050-6055.

20. **Guiyoule, A., F. Grimont, I. Iteman, P. Grimont, M. Lefevre, and E. Carniel.** 1994. Plague pandemics investigated by ribotyping of *Yersinia pestis* strains. J. Clin. Microbiol. **32:**634-641.

21. **Hinchliffe, S. J., K. E. Isherwood, R. A. Stabler, M. B. Prentice, A. Rakin, R. A. Nichols, P. C. F. Oyston, J. Hinds, R. W. Titball, and B. W. Wren.** 2003. Application of DNA Microarrays to Study the Evolutionary Genomics of *Yersinia pestis* and *Yersinia pseudotuberculosis*. Genome Res. **13:**2018-2029.

22. **Hinnebusch, B. J., R. D. Perry, and T. G. Schwan.** 1996. Role of the *Yersinia pestis* Hemin Storage (hms) Locus in the Transmission of Plague by Fleas. Science **273:**367-370.

23. **Hinnebusch, B. J., A. E. Rudolph, P. Cherepanov, J. E. Dixon, T. G. Schwan, and A. Forsberg.** 2002. Role of *Yersinia* Murine Toxin in Survival of *Yersinia pestis* in the Midgut of the Flea Vector. Science.1069972. Science **296:**733-735.

24. **Ito, M., A. Guffanti, J. Zemsky, D. Ivey, and T. Krulwich.** 1997. Role of the nhaC-encoded Na+/H+ antiporter of alkaliphilic *Bacillus firmus* OF4. J. Bacteriol. **179:**3851-3857.

25. **Jansen, H. J., C. A. Hart, J. M. Rhodes, J. R. Saunders, and J. W. Smalley.** 1999. A novel mucin-sulphatase activity found in *Burkholderia cepacia* and *Pseudomonas aeruginosa*. J Med Microbiol. **48:**551-557.

26. **Ke, D., M. Boissinot, A. Huletsky, F. J. Picard, J. Frenette, M. Ouellette, P. H. Roy, and M. G. Bergeron.** 2000. Evidence for Horizontal Gene Transfer in Evolution of Elongation Factor Tu in *Enterococci*. J. Bacteriol. **182:**6913-6920.

27. **Klevytska, A. M., L. B. Price, J. M. Schupp, P. L. Worsham, J. Wong, and P. Keim.** 2001. Identification and Characterization of Variable-Number Tandem Repeats in the *Yersinia pestis* Genome. J. Clin. Microbiol. **39:**3179-3185.

28. **Kurtz, S., A. Phillippy, A. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. Salzberg.** 2004. Versatile and open software for comparing large genomes. Genome Biology **5:**R12.

29. **Kutyrev, V. V., A. A. Filippov, O. S. Oparina, and O. A. Protsenko.** 1992. Analysis of *Yersinia pestis* chromosomal determinants Pgm super(+) and Pst super(s) associated with virulence. Microbial Pathogenesis **12:**177-186.

30. **Lathe, I., Warren C., and P. Bork.** 2001. Evolution of tuf genes: ancient duplication, differential loss and gene conversion. FEBS Letters **502:**113-116.

31. **Medini, D., C. Donati, H. Tettelin, V. Masignani, and R. Rappuoli.** 2005. The microbial pan-genome. Current Opinion in Genetics & Development. **15:**589-594.

32. **Motin, V. L., A. M. Georgescu, J. M. Elliott, P. Hu, P. L. Worsham, L. L. Ott, T. R. Slezak, B. A. Sokhansanj, W. M. Regala, R. R. Brubaker, and E. Garcia.** 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.

33. **Parkhill, J., B. W. Wren, N. R. Thomson, R. W. Titball, M. T. G. Holden, M. B. Prentice, M. Sebaihia, K. D. James, C. Churcher, K. L. Mungall, S. Baker, D. Basham, S. D. Bentley, K. Brooks, A. M. Cerdeno-Tarraga, T. Chillingworth, A. Cronin, R. M. Davies, P. Davis, G. Dougan, T. Feltwell, N. Hamlin, S. Holroyd, K. Jagels, A. V. Karlyshev, S. Leather, S. Moule, P. C. F. Oyston, M. Quail, K.**

Rutherford, M. Simmonds, J. Skelton, K. Stevens, S. Whitehead, and B. G. Barrell. 2001. Genome sequence of *Yersinia pestis*, the caus*ative agent of plague. **413:**523-527.*

34. ***Perry, R., and J. Fetherston.*** *1997.* Yersinia pestis--etiologic agent of plague. Clin. Microbiol. Rev. **10:**35-66.

35. **Perry, R. D., A. G. Bobrov, O. Kirillina, H. A. Jones, L. Pedersen, J. Abney, and J. D. Fetherston.** 2004. Temperature Regulation of the Hemin Storage (Hms+) Phenotype of *Yersinia pestis* Is Posttranscriptional. J. Bacteriol. **186:**1638-1647.

36. **Prentice, M. B., K. D. James, J. Parkhill, S. G. Baker, K. Stevens, M. N. Simmonds, K. L. Mungall, C. Churcher, P. C. F. Oyston, R. W. Titball, B. W. Wren, J. Wain, D. Pickard, T. T. Hien, J. J. Farrar, and G. Dougan.** 2001. *Yersinia pestis* pFra Shows Biovar-Specific Differences and Recent Common Ancestry with a *Salmonella enterica* Serovar Typhi Plasmid. J. Bacteriol. **183:**2586-2594.

37. **Radnedge, L., P. G. Agron, P. L. Worsham, and G. L. Andersen.** 2002. Genome plasticity in *Yersinia pestis*. Microbiology **148:**1687-1698.

38. **Radnedge, L., S. Gamez-Chin, P. M. McCready, P. L. Worsham, and G. L. Andersen.** 2001. Identification of Nucleotide Sequences for the Specific and Rapid Detection of *Yersinia pestis*. Appl. Environ. Microbiol. **67:**3759-3762.

39. **Sela, S., D. Yogev, S. Razin, and H. Bercovier.** 1989. Duplication of the tuf gene: a new insight into the phylogeny of eubacteria. J Bacteriol. **171:**581-584.

40. **Skurnik, M.** 2003. The genus *Yersinia*: entering the functional genomic era (advances in experimental medicine and biology. Plenum US.

41. **Song, Y., Z. Tong, J. Wang, L. Wang, Z. Guo, Y. Han, J. Zhang, D. Pei, D. Zhou, H. Qin, X. Pang, Y. Han, J. Zhai, M. Li, B. Cui, Z. Qi, L. Jin, R. Dai, F. Chen, S. Li, C.**

Ye, Z. Du, W. Lin, J. Wang, J. Yu, H. Yang, J. Wang, P. Huang, and R. Yang. 2004. Complete genome sequence of *Yersinia pestis* strain 91001, an isolate avirulent to humans. DNA Res. **11:**179-197.

42. **Une, T., and R. R. Brubaker.** 1984. In vivo comparison of avirulent Vwa- and Pgm- or Pstr phenotypes of *yersiniae*. Infect Immun. **43:**895-900.

43. **Ventura, M., C. Canchaya, V. Meylan, T. R. Klaenhammer, and R. Zink.** 2003. Analysis, Characterization, and Loci of the tuf Genes in *Lactobacillus* and *Bifidobacterium* Species and Their Direct Application for Species Identification. Appl. Environ. Microbiol. **69:**6908-6922.

44. **Wiechmann, I., and G. Grupe.** 2005. Detection of Yersinia pestis DNA in two early medieval skeletal finds from Aschheim (Upper Bavaria, 6th century A.D.). American Journal of Physical Anthropology **126:**48-55.

45. **Wilmoth, B., M. Chu, and T. Quan.** 1996. Identification of *Yersinia pestis* by BBL Crystal Enteric/Nonfermenter Identification System. J. Clin. Microbiol. **34:**2829-2830.

46. **Wren, B. W.** 2003. The *yersiniae*--a model genus to study the rapid evolution of bacterial pathogens. Nat Rev Microbiol. **1:**55-64.

47. **Wright, D. P., C. G. Knight, S. G. Parkar, D. L. Christie, and A. M. Roberton.** 2000. Cloning of a Mucin-Desulfating Sulfatase Gene from *Prevotella* Strain RS2 and Its Expression Using a Bacteroides Recombinant System. J. Bacteriol. **182:**3002-3007.

Table 1. General genome features for *Yersinia pestis* strains Antiqua and Nepal516

	Antiqua	Nepal516
Chromosome Size (bp)	4,702,289	4,534,590
G+C content (%)	47.70	47.58
Coding sequences	4138	3956
Average gene length (bp)	953	958
Coding Density (%)	83.8	83.6
16S-23S-5S rRNAs	7	7
Transfer RNAs	68	72
pMT Size (bp)	96,471	100,918
G+C content	50.24	50.16
Coding sequences	99	104
Average gene length (bp)	832	820
Coding Density (%)	85.3	84.5
pCD Size (bp)	70,299	-*
G+C content	44.83	
Coding sequences	89	
Average gene length (bp)	601	
Coding Density (%)	76.1	
pPCP Size (bp)	10,777	10,778
G+C content	45.44	45.44
Coding sequences	9	9
Average gene length (bp)	573	573
Coding Density (%)	47.8	47.8

*pCD of Nepal516 was not completed in this study, see Materials and Methods

Table 2. Chromosome comparison between the sequenced *Y. pestis* strains

	Antiqua	Nepal516	91001	KIM	CO92
Molecular grouping (1)	1.ANT	2.ANT	0.PE4	2.MED	1.ORI
Size (bp)	4,702,289	4,534,595	4,595,065	4,600,755	4,653,728
Total IS elements					
IS*100*	75	32	30	34	44
IS*285*	24	25	23	19	21
IS*1541*	67	64	47	55	65
IS*1661*	10	8	8	8	8
Unique IS elements*					
IS*100*	39	4	15	6	13
IS*285*	2	4	11	1	1
IS*1541*	5	4	0	0	8
IS*1661*	1	0	0	0	0

*Some IS elements were shared between expected partners such as three IS*100* and one IS*285* shared between Antiqua and CO92, as well as one IS*100*, one IS*1661* and one IS*1541* shared between Nepal516 and KIM. However, five exceptions were observed: one IS*100* shared between 91001, Antiqua and CO92 but not present in Nepal or KIM; one IS*100* in 91001, Antiqua, CO92 and KIM but not in Nepal; one IS*285* in 91001, CO92, KIM and Nepal but not in Antiqua; one IS*1541* in CO92 and in *Y. pseudotuberculosis* IP32053; one IS*1541* in Antiqua and in *Y. pseudotuberculosis* IP32053

Table 3. Genome specific inactivation of genes

Deletion specific to the genome(s)	Number of proteins inactivated
CO92	4
KIM	20
Antiqua	41
Nepal516	13
91001	69
CO92, KIM	0
CO92, Antiqua	11
CO92, Nepal516	0
CO92, 91001	0
KIM, Antiqua	2
KIM, Nepal516	16
KIM, 91001	0
Antiqua, Nepal516	0
Antiqua, 91001	8
Nepal516, 91001	1

* For rows with 2 strains, the data indicate inactivations in the same CDS of both strains

Table 4. Prophage-like fragment specific to virulent *Y. pestis* strains

GENE	*Y. pestis*				*Y. pseudo-tuberculosis*	COG	Product
	CO92	KIM	ANTIQUA	NEPAL	IP32953		
YPO2095	+	+	-	+	-	-	hypothetical protein
YPO2096	+	*+	-	*+	-	-	hypothetical protein
YPO2097	+	*+	-	*+	-	-	hypothetical protein
YPO2098	+	+	-	+	-	R	putative phage lysozyme
YPO2099	+	+	-	+	-	-	putative prophage endopeptidase
YPO2100	+	+	-	+	-	S	phage regulatory protein
YPO2101	+	+	-	+	-	-	hypothetical protein
YPO2102	+	+	-	+	-	-	hypothetical protein
YPO2104	+	+	-	+	+	L	transposase for the IS*285* insertion element
YPO2108	+	+	+	+	-	-	hypothetical protein
YPO2109	+	+	+	+	-	-	hypothetical protein
YPO2110	+	+	+	+	-	-	hypothetical protein
YPO2111	+	+	+	+	-	-	hypothetical protein
YPO2112	+	+	+	+	-	-	hypothetical protein
YPO2113	+	+	+	+	-	-	hypothetical protein
YPO2114	+	+	+	+	-	-	hypothetical protein
YPO2115	+	+	+	+	-	-	hypothetical protein
YPO2116	+	+	+	+	-	-	hypothetical protein
YPO2117	+	+	+	+	-	-	hypothetical protein
YPO2118	+	*+	*+	*+	-	-	hypothetical protein
YPO2119	+	+	+	+	-	S	putative phage tail protein
YPO2120	+	+	+	+	-	S	hypothetical protein
YPO2122	+	+	+	+	-	S	hypothetical

YPO2123	+	+	+	+	-	R	protein putative phage minor tail protein
YPO2124	+	+	+	+	-	-	hypothetical protein
YPO2125	+	+	+	+	-	-	putative phage regulatory protein
YPO2126	+	+	+	+	-	K	hypothetical protein
YPO2127	+	+	+	+	-	-	putative phage-related membrane protein
YPO2128	+	*+	*+	*+	-	-	putative phage-related lipoprotein
YPO2129	+	+	+	+	-	S	putative phage tail assembly protein
YPO2130	+	*+	*+	*+	-	-	hypothetical protein
YPO2131	+	+	+	+	-	S	putative phage host specificity protein
YPO2132	+	+	+	+	-	-	hypothetical protein
YPO2133	+	+	+	+	-	-	hypothetical protein
YPO2134	+	+	+	+	-	-	putative phage tail fiber assembly protein
YPO2135	+	*+	+	+	-	-	hypothetical protein
YPO2487	+	*+	*+	*+	*+	-	putative membrane protein
YPO2488	+	+	+	+	+	-	hypothetical protein
YPO2489	+	+	+	+	+	S	hypothetical protein

* The protein sequence was not found in the genome, however, the DNA fragment did exist in the intergenic region.

Figure 1. Circular representation of the strain Antiqua (A) and strain Nepal516 (B) chromosomes. The different rings represent (from outer to inner): 1 and 2, all genes color-coded by functional category; 3 and 4, IS elements (IS*100*, IS*285*, IS*1541*, IS*1661*); 5, deviation from average G+C content; 6, GC skew.

Figure 2. Phylogenetic ordering of *Yersinia* by SNP analysis. The number of sSNPs and the number of nsSNPs (in parentheses symbols) are illustrated at the corresponding positions.

Figure 3. Functional distribution of genes bearing SNPs. The number of genome-specific nsSNPs (Fig. 3a) and sSNPs (Fig. 3b) were grouped into COG functional classes. These were sub-categorized based on what genome(s) they were found in (light orange: 91001; red: CO92; dark red: KIM; blue: Antiqua; light blue: Nepal516; green: KIM and Nepal516; yellow: Antiqua and CO92). COG functional classes: C, energy production; D, cell division; E, amino acid metabolism; F, nucleotide metabolism; G, carbohydrate metabolism; H, coenzyme metabolism; I, lipid metabolism; J, translation; K, transcription; L, DNA replication or repair; M, cell wall/membrane biogenesis; N, cell motility; O, posttranslational modification; P, inorganic ion metabolism; Q, secondary metabolites biosynthesis, transport and catabolism ; R, general function prediction only; S, function unknown; T, signal transduction; U, intracellular trafficking and secretion; V, defense mechanism

Figure 1

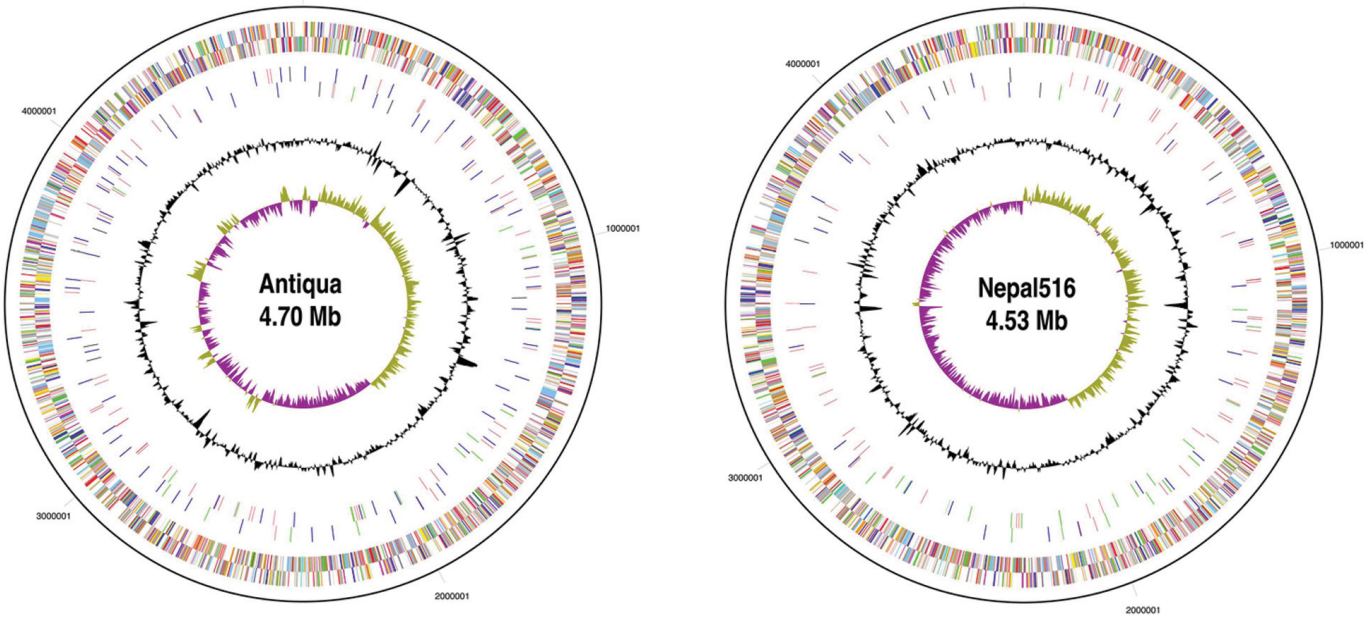
Antiqua
4.70 Mb
4000001
1000001
3000001
2000001
Nepal516
4.53 Mb
4000001
1000001
3000001
2000001

Figure 2

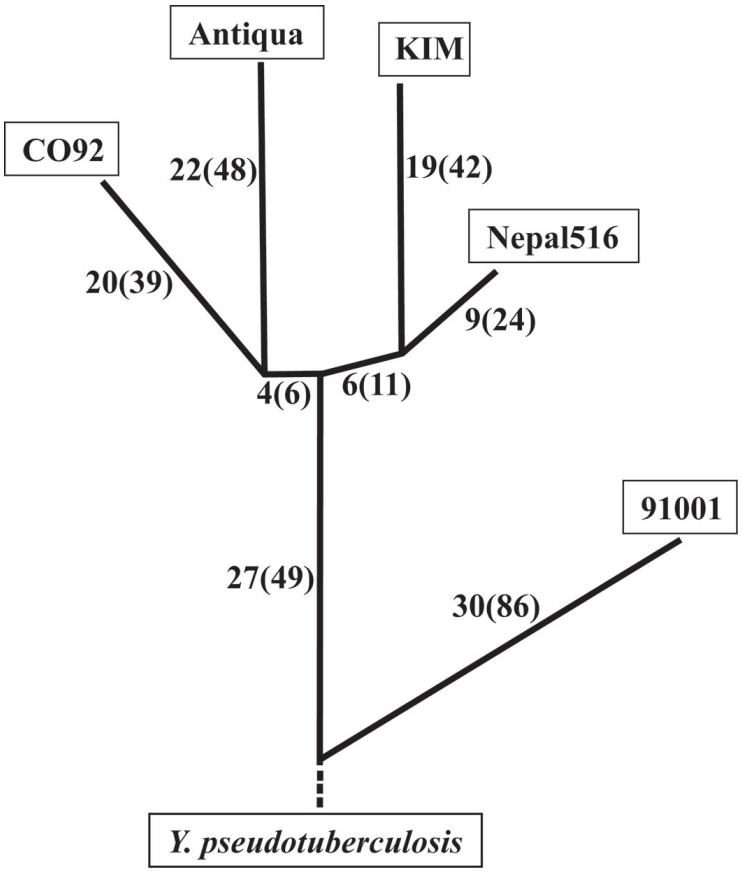
Antiqua
KIM
CO92
22(48)
19(42)
Nepal516
20(39)
9(24)
4(6)
6(11)
91001
27(49)
30(86)
Y. pseudotuberculosis

Figure 3.

(a)

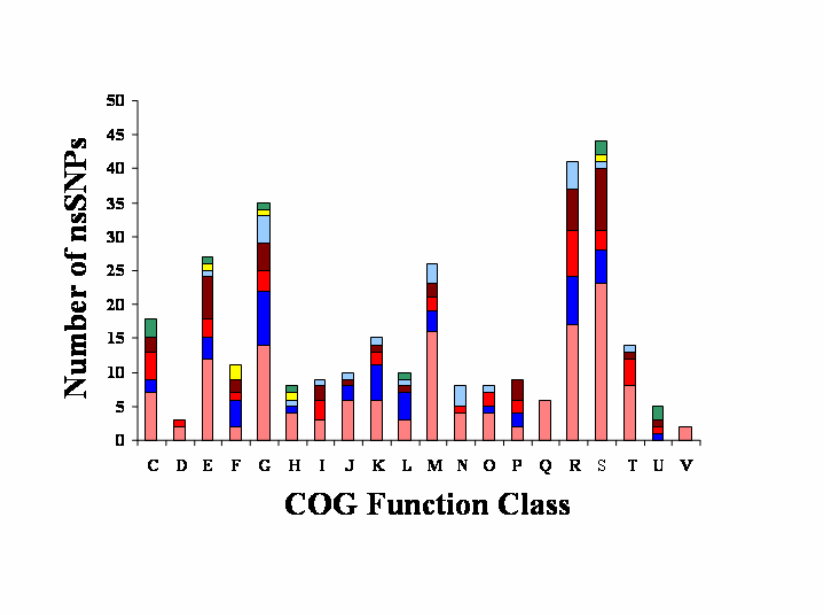

Figure 3
(b)