JOURNAL OF BACTERIOLOGY, Oct. 2006, p. 6977–6985 Vol. 188, No. 19
0021-9193/06/$08.00+0 doi:10.1128/JB.00506-06

Chromosome Rearrangement and Diversification of *Francisella tularensis* Revealed by the Type B (OSU18) Genome Sequence†

Joseph F. Petrosino,[1,2]* Qin Xiang,[2] Sandor E. Karpathy,[2] Huaiyang Jiang,[2] Shailaja Yerrapragada,[2] Yamei Liu,[3] Jason Gioia,[1] Lisa Hemphill,[2] Arely Gonzalez,[4] T. M. Raghavan,[3] Akif Uzman,[4] George E. Fox,[3] Sarah Highlander,[1,2] Mason Reichard,[5] Rebecca J. Morton,[5] Kenneth D. Clinkenbeard,[5] and George M. Weinstock[1,2]

Department of Molecular Virology and Microbiology[1] and Human Genome Sequencing Center,[2] Baylor College of Medicine, Houston, Texas 77030; Department of Biology and Biochemistry, University of Houston, Houston, Texas 77204[3]; Department of Natural Sciences, University of Houston-Downtown, Houston, Texas 77002[4]; and Department of Veterinary Pathobiology, Center for Veterinary Health Sciences, Oklahoma State University, Stillwater, Oklahoma 74078[5]

Received 9 April 2006/Accepted 17 July 2006

The γ-proteobacterium *Francisella tularensis* is one of the most infectious human pathogens, and the highly virulent organism *F. tularensis* subsp. *tularensis* (type A) and less virulent organism *F. tularensis* subsp. *holarctica* (type B) are most commonly associated with significant disease in humans and animals. Here we report the complete genome sequence and annotation for a low-passage type B strain (OSU18) isolated from a dead beaver found near Red Rock, Okla., in 1978. A comparison of the *F. tularensis* subsp. *holarctica* sequence with that of *F. tularensis* subsp. *tularensis* strain Schu4 (P. Larsson et al., Nat. Genet. 37:153–159, 2005) highlighted genetic differences that may underlie different pathogenicity phenotypes and the evolutionary relationship between type A and type B strains. Despite extensive DNA sequence identity, the most significant difference between type A and type B isolates is the striking amount of genomic rearrangement that exists between the strains. All but two rearrangements can be attributed to homologous recombination occurring between two prominent insertion elements, IS*Ftu1* and IS*Ftu2*. Numerous pseudogenes have been found in the genomes and are likely contributors to the difference in virulence between the strains. In contrast, no rearrangements have been observed between the OSU18 genome and the genome of the type B live vaccine strain (LVS), and only 448 polymorphisms have been found within non-transposase-coding sequences whose homologs are intact in OSU18. Nonconservative differences between the two strains likely include the LVS attenuating mutation(s).

Francisella tularensis, a facultative intracellular pathogen, is the causative agent of tularemia and is among the most infectious pathogens known, both in terms of the number of zoonotic species that it infects (>250 species) and in terms of the number of organisms needed to establish a potentially lethal infection (<10 organisms by the airborne route, with a mortality rate of 5 to 30% when the infection is left untreated) (14). *F. tularensis* is a category A biodefense concern because of its ability to cause incapacitating, potentially fatal illness, because of its ability to spread via aerosolization, and because of the potential burden on the public health system in the event of an outbreak. It was developed as a biological weapon by Japan, the United States, and the Soviet Union during the 20th century, and there is concern that bioweapons containing this organism still exist (3, 10). *F. tularensis* is able to infect rabbits, beavers, muskrats, and other mammals; however, *F. tularensis* reservoirs and route(s) of passage in natural foci have yet to be determined due to the variety of hosts and transmission methods involved, as well as the presence of two subspecies that differ in ecology and virulence but have overlapping niches. Environmental observations and the nutritional requirements for growing *F. tularensis* suggest that the organism is not free living in nature (27). *F. tularensis* replicates in macrophages as well as in a variety of cell types from different animal species, yet despite accelerated research on this organism in recent years, relatively little is understood regarding its lifestyle, virulence mechanisms, and the identity of its natural reservoirs.

There are four subspecies of *Francisella tularensis*: *F. tularensis* subsp. *tularensis* (type A), *F. tularensis* subsp. *holarctica* (type B), *F. tularensis* subsp. *novicida*, and *F. tularensis* subsp. *mediasiatica*. Type A, found exclusively in North America, causes the most severe form of human disease, while type B is less pathogenic and is found throughout North America, Europe, and Asia (9). An attenuated live vaccine strain (LVS) was derived from repeated passage of a type B strain sometime between the 1930s and 1950s in the former Soviet Union (8). LVS provides protection against both type A and type B infection but is not licensed for use in the United States because its immunogenicity in humans is poorly characterized, the mechanism of attenuation for this strain is unknown, and the strain kills mice with a 50% lethal dose of <10 CFU when it is introduced via intraperitoneal injection (4).

The recent publication of the genome sequence of a type A strain (16) shed light on the genetic makeup of *F. tularensis*, yet many questions pertaining to the virulence and lifestyle of this

* Corresponding author. Mailing address: Department of Molecular Virology and Microbiology, Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, BCM280, Houston, TX 77030. Phone: (713)798-7912. Fax: (713)798-7375. E-mail: jpetrosi@bcm.edu.

† Supplemental material for this article may be found at http://jb.asm.org/.

organism remain. The sequence of the attenuated type B strain, LVS, has recently been made available by GenBank (GenBank accession number AM233362). While comparisons of these two strains should begin to show differences between virulent type A strains and an attenuated type B strain, the genome sequence of a virulent type B strain presented here allows identification of genetic differences that underlie differences in the virulence of type A and type B strains, as well as the genetic differences underlying attenuation of LVS.

Here we report the sequence of a virulent type B strain (strain OSU18), which was isolated from a beaver that died of tularemia near Red Rock, Okla., in 1978. A comparison of OSU18 with the virulent type A strain Schu4 revealed almost 99% sequence identity, yet there are numerous rearrangements and pseudogene content disparities that likely distinguish the virulence phenotypes and geographic and host distributions of type A and B strains. In contrast, a relatively limited number of polymorphisms and no rearrangements were found when OSU18 and LVS were compared. The polymorphisms are anticipated to include the mutations that attenuate LVS.

MATERIALS AND METHODS

Genome sequencing and assembly. Sequencing and assembly of the *F. tularensis* subsp. *holarctica* strain OSU18 genome were accomplished by the whole-genome shotgun (WGS) method, similar to a previously described method (22). Briefly, the WGS clones were sequenced using ABI 3730 sequencers, and the sequence bases were called using the Applied Biosystems sequencing analysis software KB Basecaller. The WGS reads were assembled by using Atlas (11) and Phrap (7). The initial WGS assembly resulted in 132 contigs in 33 scaffolds with approximately 26× sequence coverage. Gaps between contigs and scaffolds were closed by sequencing PCR products that spanned gaps or by sequencing small insert libraries generated from the PCR products. Low-quality regions were resequenced using clones or PCR products spanning the regions to ensure that the Phrap quality score for each base was equal to or greater than 30. This relatively deep data set should enable further studies involving new sequencing, comparative genomics, and proteomics strategies and technologies. Included among these strategies and technologies are (i) using sequencing reads to scan for possible phase variation in *Francisella* cultures, (ii) using WGS clones as gene expression constructs for peptide array-based antigen screens, and (iii) using the deep coverage of sequencing reads as a representative data set for comparing existing sequencing methodologies to new technologies in various stages of development.

Gene prediction and annotation. The Glimmer (2) and GeneMark (19) gene prediction programs were used to predict open reading frames (ORFs) in combination with TBLASTX of the GenBank nr protein database to generate an initial ORF list. Each ORF was also evaluated against the GenBank conserved domain database (CDD) (21), the UCSD Transport Classification Database (http://www.tcdb.org/), and the MEROPS peptidase database (29). tRNAs were identified using tRNAscan-SE (18). All results were used to populate a local database that was used for manual annotation to identify the best coordinates and definition for each ORF. Every ORF was double-blind annotated by two different annotators, and differences were reconciled at the end of the annotation by following the rules described previously (22) to maintain annotation consistency. The intergenic regions were scanned for genes and gene fragments that the original analysis may have missed to further ensure an accurate annotation. Manual curation of the annotation was facilitated by using Genboree (http://www.genboree.org/) and other software developed in the Human Genome Sequencing Center at the Baylor College of Medicine.

Genome comparisons. Homologs of OSU18 genes in the LVS and Schu4 genomes were initially identified by BLASTN comparisons (1) with the Schu4 genome (GenBank accession number AJ749949) and with the LVS genome (the LVS sequence used here was produced by the BBRP group at Lawrence Livermore National Laboratory and was obtained with permission from http://bbrp.llnl.gov/bbrp/html/microbe.html; the LVS sequence has since been deposited in the GenBank database under accession number AM233362.). When paralogs or multiple copies of identical proteins were present, regional gene order was used to identify the correct homolog. Genetic differences reported here are based upon the OSU18 annotation. A linear comparison of the OSU18 genes to the genes annotated in the LVS GenBank entry was performed as an additional validation of gene organization as determined by the respective annotation groups. It was found that 16 ORFs in the LVS annotation did not match an annotated ORF within 10 kb of the expected position of the gene in OSU18. Conversely, 86 OSU18 ORFs were not found within 10 kb of the expected position in the LVS annotation. Further investigation revealed that, while the DNA sequences were present in both strains, many of the ORFs were not called by one of the annotation groups or the other and/or were disrupted so as not to necessitate annotation.

TABLE 1. Nucleotide and CDS data for *F. tularensis* OSU18

Parameter	Data
Size	1,895,727 bp
G+C content	32.16%
No. of CDS	1,934
No. of conserved hypothetical proteins	234 (46 pseudogenes)
No. of *Francisella* conserved hypothetical proteins	243 (26 pseudogenes)
No. of hypothetical proteins	71 (1 pseudogene)
No. of pseudogenes or gene fragments	325 (including transposases)
No. of IS elements	
IS*Ftu1*	59 (21 pseudogenes)
IS*Ftu2*	41 (all pseudogenes)
IS*Ftu3*	5 (all pseudogenes)
IS*Ftu4*	1 (pseudogene)
IS*Ftu5*	1 (pseudogene)
IS*Ftu6*	1 (pseudogene)
IS*Sod13* (IS*tron*)	1 (pseudogene)
No. of rRNA operons	3
No. of tRNAs	38
No. of other stable RNAs	1

Nucleotide sequence accession numbers. The annotated *F. tularensis* subsp. *holarctica* strain OSU18 whole-genome shotgun project sequence has been deposited in the DDBJ/EMBL/GenBank database under project accession number CP000437. Supplemental data are also available in the supplemental material at http://jb.asm.org/.

RESULTS

***F. tularensis* OSU18 genome is conserved among sequenced *Francisella* strains and contains many disrupted genes.** The genome of *F. tularensis* OSU18 is comprised of a 1,895,727-bp circular chromosome containing 1,934 predicted coding sequences (CDS) (including pseudogenes) and has an overall G+C content of 32.16% (Table 1). The OSU18 sequence was verified by comparing the assembly to optical map data of OSU18 genomic DNA digested with the XhoI restriction endonuclease (OpGen, Madison, WI). A detailed description of the methods used for sequencing and annotation of OSU18 is given in Materials and Methods.

Of the 1,934 OSU18 CDS, 1,560 are predicted to code for functional proteins, 38 are predicted to code for tRNAs, 10 are predicted to code for rRNAs, and 1 is expected to code for a noncoding transfer mRNA (a unique RNA species involved in ribosome rescue [23]). Gene-disrupting mutations were found in 325 CDS (including 73 transposases) which are classified as pseudogenes. Genes with unknown functions were annotated

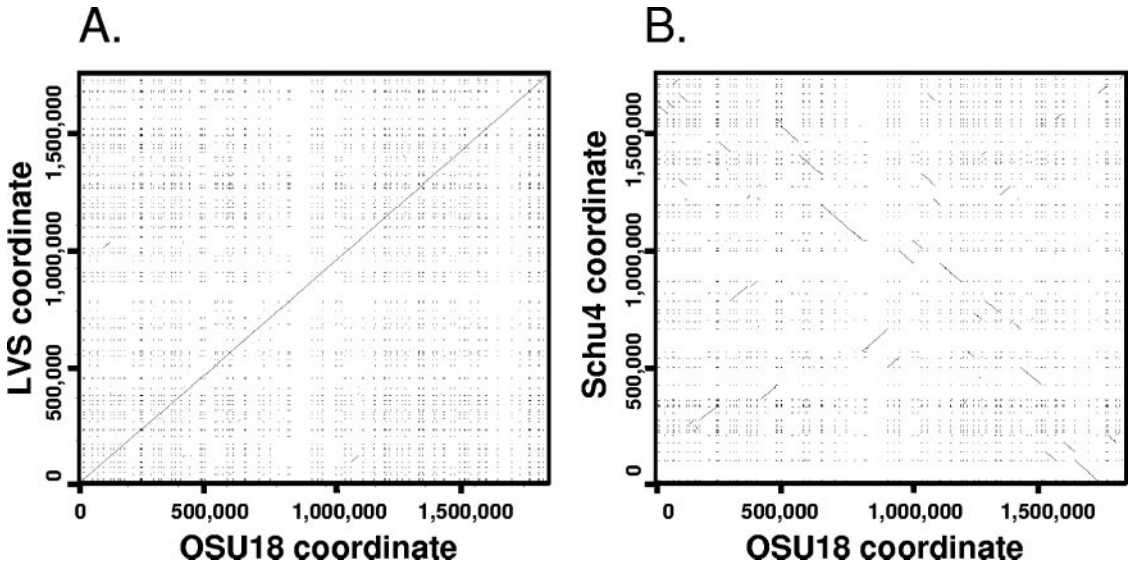

FIG. 1. Nucleotide alignment for OSU18, LVS, and Schu4. Each pair of genomes was aligned using BLASTN, and high-scoring segment pairs ($E \leq 1e^{-10}$) were plotted. Note the many short aligned sequences throughout both genomes. These are the multiple alignments between highly conserved IS elements. All but two syntenic blocks in panel B start and end at these IS elements. (A) OSU18 versus LVS; (B) OSU18 versus Schu4.

as genes encoding hypothetical proteins and account for 28.3% of the annotated CDS (encoding 548 proteins; 73 of these are pseudogenes). Of these proteins, 234 CDS products resemble proteins found in other organisms, which may include other *Francisella* strains (conserved hypothetical proteins), 243 resemble proteins identified only in other *Francisella* strains (*Francisella* conserved hypothetical proteins), and 71 do not resemble any proteins previously deposited in the GenBank nr database (hypothetical proteins) ($E \leq 1e^{-10}$).

Comparison of OSU18 to Schu4: rearrangements and pseudogenes. The OSU18 sequence matches 97.63% of the Schu4 genome and is 98.94% identical in these regions, indicating that the differences between type A virulence and lifestyle and type B virulence and lifestyle are not due to large differences in gene content between the two subspecies. Using nucleotide sequence comparison (BLASTN) (1), only 12 ORFs annotated in OSU18 were found to be neither intact nor pseudogenes in Schu4, whereas 17 Schu4 ORFs were not found in OSU18 (see Tables S1 and S2 in the supplemental material). Genome sequencing data for two additional *F. tularensis* strains, ATCC 6223 (an avirulent type A strain, also known as B38) and OR960246 (a virulent type B strain) (unpublished data), and for the publicly available LVS sequence together reinforce the conclusion that the type A and B subspecies have similar gene contents.

In contrast to the strong nucleotide identity, there are 51 syntenic blocks rearranged between Schu4 and OSU18, but no rearrangements have been observed between the type B OSU18 and LVS genomes (Fig. 1A and B). Inspection of the ends of each syntenic element revealed repeated DNA sequences belonging to the IS*Ftu1* and IS*Ftu2* insertion sequence (IS) elements found in *F. tularensis* type A and type B and *F. tularensis* subsp. *novicida* (16). Rearrangements at 98 of 102 sites (flanking 49 syntenic segments) arose from homologous recombination at an IS*Ftu1* or IS*Ftu2* sequence (Fig. 2; see Table S3 in the supplemental material). The other four rearrangements resulted from recombination at rRNA sequences (recombination events involving rRNA adjacent to positions 424062 and 1128048 in OSU18 and rRNA adjacent to positions 1307578 and 1374701 in Schu4.).

EcoRI endonuclease chromosomal DNA optical maps for Schu4, ATCC 6223, and OSU18, showed that there were rearrangements not only when type A strains were compared to type B strains but also when type A strains were compared to each other (Fig. 3). It appears that genome rearrangements occurred in the type A *Francisella* subspecies after the type A and type B subspecies diverged evolutionarily or that type B strains were derived from one specific type A strain that lost the ability to undergo genomic rearrangement. Additional finished type B genome sequences are needed to verify that further rearrangements have also occurred within type B genomes. Genes required for homologous recombination, such as those in the RecABCD and RecFOR pathways, are intact in OSU18, so the capacity for rearrangement, as a function of recombination potential, in OSU18 does not appear to be impaired. One possible cause for the IS element-based recombination events may be recombination-promoting events (e.g., single-stranded DNA nicks) at the IS sequences. As discussed below, polymorphisms in the IS*Ftu2* transposase (*tnp*) gene could allow this enzyme to stimulate recombination in type A strains but not in type B strains. Extensive genome rearrangement due to recombination events between IS elements and rRNA genes, while unusual, is not unique to *Francisella*. Such events have been postulated to play a large role in the differentiation of both *Yersinia* and *Bordetella* species, and rRNA gene recombination plays a role in chromosomal rearrangement in *Salmonella* sp. (17).

Typically, 1 to 5% of annotated CDS in a given microbial genome are transposase genes. In OSU18, *tnp* genes encompass approximately 5.8% (112 of 1,934) of the total number of CDS, or ~4% of the total nucleotide composition of the genome. Five different types of insertion sequence elements have been described previously in *F. tularensis* (IS*Ftu1* to IS*Ftu5*)

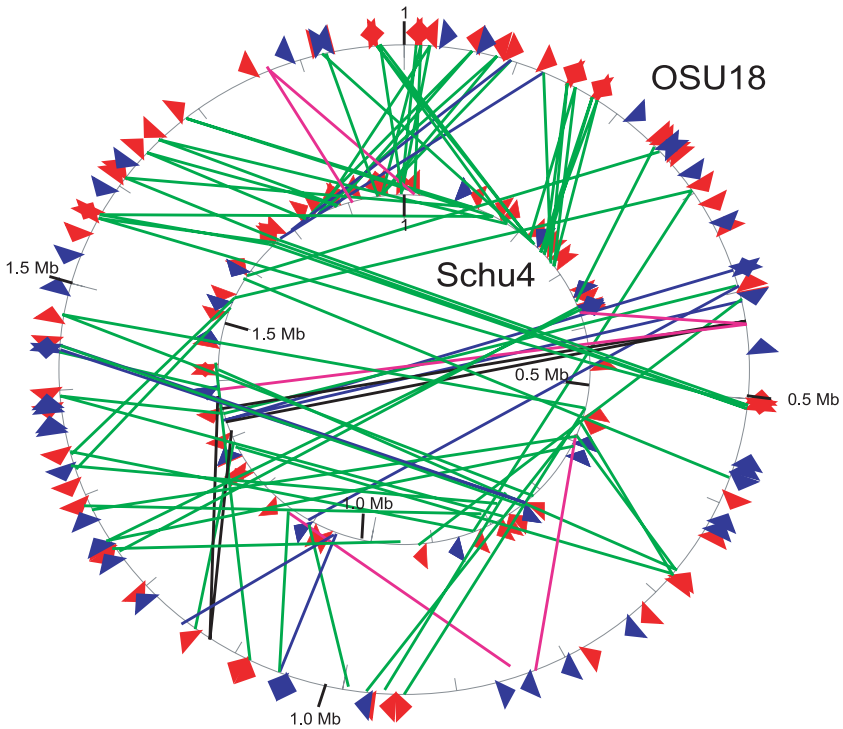

FIG. 2. Gene rearrangements between the OSU18 (outer circle) and Schu4 (inner circle) genomes. Homologous recombination at IS*Ftu1* and IS*Ftu2* sequences results in rearrangement of 49 syntenic blocks of sequence between OSU18 and Schu4. Two additional pairs of rearrangements originate at rRNA genes. Red triangles, IS*Ftu1* genes; blue triangles, IS*Ftu2* genes; pink lines, rearrangement events where an IS*Ftu1* or IS*Ftu2* sequence is still present in only the Schu4 genome; blue lines, rearrangement events where an IS*Ftu1* or IS*Ftu2* sequence is still present in only the OSU18 genome; green lines, rearrangement events where an IS*Ftu1*or IS*Ftu2* sequence is still present in both genomes; black lines, rearrangement events involving rRNA genes.

(16). OSU18 contains these five IS element types, as well as an additional type, IS*Ftu6*. IS*Ftu1* is an IS*630* Tc-1 mariner family IS element, and 59 copies are present in OSU18; 21 of these copies have been designated pseudogenes due to deletions (Table 1). The IS*Ftu1 tnp* gene is composed of two distinct, nonoverlapping ORFs. It has been hypothesized that a programmed ribosomal frameshift is required for translation of the entire functional transposase protein (16). In agreement with this hypothesis, a potential "slippery heptamer" (AAAAAAG) has been identified near the C-terminal end of the first ORF (20).

IS*Ftu2* is an IS*5* family IS element, and 41 copies are present

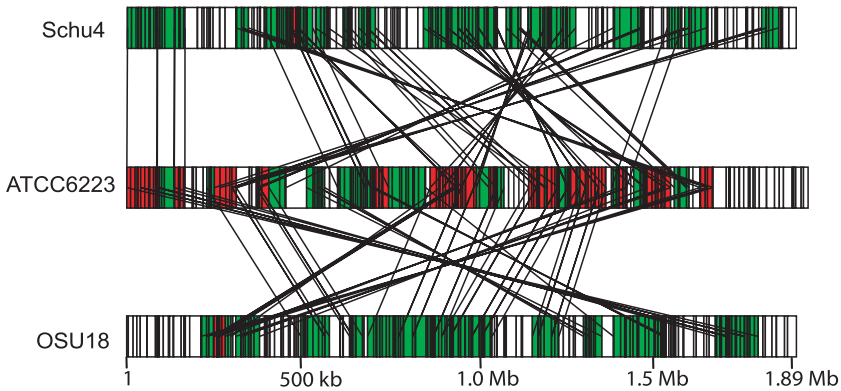

FIG. 3. Optical map comparisons of *Francisella* strains. EcoRI restriction maps for the type A Schu4 and ATCC 6223 and type B OSU18 chromosomes were constructed by OpGen (Madison, WI). Green indicates areas where the restriction maps are homologous for the paired genomes. Red in the ATCC 6223 genome indicates areas where the restriction maps have homologies in all three genomes. Red in the OSU18 and Schu4 genomes indicates regions in these genomes that are homologous to more than one region in ATCC 6223. The comparison of OSU18 to ATCC 6223 shows a level of rearrangement similar to that found when OSU18 was compared to Schu4 (Fig. 1 and 2). The comparison of Schu4 to ATCC 6223 demonstrates that rearrangement between chromosomes is prevalent within the type A subspecies, whereas the two finished type B genomes examined thus far (LVS and OSU18) suggest that rearrangements in this subspecies may be much less prevalent.

in OSU18. While the copy numbers of IS*Ftu1* are essentially identical in Schu4 and OSU18 (50 and 59 copies, respectively), Schu4 contains less than half as many IS*Ftu2* elements as OSU18 (19 and 41 copies, respectively) (16). Examination of the LVS genome further supported the general distribution of IS*Ftu1* and IS*Ftu2* sequences found in type A genomes versus type B genomes: LVS has 59 copies of IS*Ftu1*, which is equal to the number in OSU18 (59 copies) and similar to that in Schu4 (50 copies), and 46 copies of IS*Ftu2*, which is more than double the number found in Schu4 (19 copies) but is similar to the number in OSU18 (41 copies).

The IS*Ftu2 tnp* genes identified in OSU18 differ from those found in Schu4. First, all but three IS*Ftu2* genes in OSU18 have a point mutation resulting in a premature stop at the 20th amino acid compared to the Schu4 sequence (CAA to TAA) (see Fig. S1 in the supplemental material). The remaining three IS*Ftu2* genes are missing the first 66 nucleotides (nt), corresponding to the region upstream of the premature stop found in other OSU18 alleles. Two potential downstream start codons were identified in IS*Ftu2 tnp* that could translate the predicted DDE catalytic domain, although we were unable to identify a potential DNA binding domain in the predicted peptide, so it is not known if initiation from one of the downstream starts would produce a fully functional protein. Preliminary analysis of IS*Ftu2* sequences from the ATCC 6223, LVS, and OR960246 genomes showed that the CAA-to-TAA nonsense mutation at codon 20 appears only in type B strains, while the type A strains retain the CAA codon (unpublished data). The second unique feature of IS*Ftu2 tnp* genes in type B strains is loss of the TGA stop codon in 36 of 41 copies (see Fig. S1 in the supplemental material). A TGA-to-AGA point mutation in this codon results in read-through of the stop codon in the 36 mutant alleles and is followed by 25 bp of conserved sequence. After the conserved sequence, the sequences diverge, and the stop codon for each allele is positioned at various distances downstream. Because of the premature stop codons at position 20, all copies of IS*Ftu2 tnp* have been classified as pseudogenes. Despite these stop codons, only the three IS*Ftu2 tnp* genes with the 5′ deletions have accumulated deletions elsewhere in their coding sequences (27 bp is deleted at five identical locations in two of these *tnp* genes [the IS*Ftu2 tnp* genes starting at positions 248226 and 1824837]; the other *tnp* gene, starting at position 1259232, is missing one of the single-base deletions found in the other two genes, and only one of the five deletion events in each gene is in frame.). The additional mutations found in the three IS*Ftu2 tnp* genes that lack the N terminus suggest that there is selective pressure to retain the nucleotide sequence of IS*Ftu2* genes having an intact N terminus even though they have a nonsense mutation at codon 20 and that this pressure is relaxed if the region upstream of the premature stop is missing. Thus, it seems likely that the region upstream of the premature stop is translated or is otherwise required in type B strains. For these *tnp* genes to be expressed in their entirety, an unrecognized suppressor tRNA is or was present in both OSU18 and LVS, or else low-level misreading of the stop codon may occur. A high level of transposase activity is presumably deleterious for maintenance of chromosomal stability and cell survival; therefore, mechanisms that ensure transposase repression are expected to be present. A search of the OSU18 genome with tRNAscan-SE (18) resulted in no likely candidates for suppressor tRNAs.

Virulence gene candidates. The 33.9-kb duplicated *Francisella* pathogenicity island (FPI) found in Schu4 is present in two ~30-kb regions in OSU18. The FPIs were originally identified in *F. tularensis* subsp. *novicida*, in part because they have a reduced G+C content (26.3%, compared to 32.2% for the OSU18 genome). As in other *Francisella* subspecies studied thus far, the duplicated OSU18 FPI contains the *iglABCD* operon and the *pdpABC* genes. However, the shorter length of the OSU18 FPI is caused by deletion of the portion of the *pdpD* gene encoding the first 980 amino acids and deletion of an upstream pseudogene encoding a conserved hypothetical protein found in Schu4 (corresponding to FTT1716c and FTT1361c in the two FPIs of Schu4). The *iglC* and *pdpD* genes have both been shown to be important for intramacrophage growth of *Francisella* (24) and are regulated by *MglA*, which is located outside the FPI (30). The functions of these and other predicted genes in the island are unknown but are of immediate interest due to their apparent role in virulence.

A comprehensive list of putative OSU18 virulence factors was defined as all CDS with products that have been classified as potential virulence factors in *F. tularensis* or other pathogens. Using this inclusive approach, 268 of the 1,934 (13.9%) annotated ORFs can be hypothesized to have a role in *Francisella* virulence (Table 2; see Table S4 in the supplemental material). Many of the associated physiological functions may have dual roles for survival in tick and other animal hosts, as well as for survival in the intracellular niche that *F. tularensis* occupies. While obligate intracellular parasites such as *Bucherna* spp. have extremely limited metabolic functions that coincide with those obviated by the host supply of metabolites, *F. tularensis* may have retained more of the factors from its free-living ancestors to maintain its broad animal infectivity. Of the 268 putative OSU18 virulence genes, there is currently experimental evidence from mutagenesis, in vitro models, or proteomics to identify 38 of the proteins as virulence factors in *F. tularensis* spp. (see Table S4 in the supplemental material). Pseudogenes of predicted OSU18 virulence genes are likely contributors to the difference between type B virulence and type A virulence. Among the virulence genes experimentally verified in *Francisella*, both copies of *pdpD* and an acid phosphatase gene are pseudogenes in OSU18. In addition, the gene for phospholipase D shown to be involved in *Yersinia pestis* survival in its flea host (13) is also disrupted in OSU18 (12).

In addition to the virulence pseudogenes identified above, seven additional OSU18 pseudogenes were identified within postulated virulence genes. Among these are three genes involved in type IV pilus structure and assembly (*pilT* and two *pilE* homologs, *pilE2* and *pilE3*). Type IV pilus genes have been implicated in the pathogenicity of *Francisella* and other bacterial species (28), and the OSU18 pilus-associated pseudogenes may be at least partially responsible for the pathogenicity differences between type A and type B strains. However, these pseudogenes do not account for the attenuation of LVS as the alleles are identical in OSU18 and LVS and are not essential for pilus assembly as mature pili have been observed on the surface of LVS (6). Two of the other possible virulence-associated, OSU18-specific pseudogenes may be involved in antibiotic efflux, while another, *htpG*, encodes a chaperone

TABLE 2. Classification of potential virulence genes identified in OSU18[a]

Virulence factor classification	No. of intact genes in *F. tularensis*		No. of pseudogenes in *F. tularensis*			Total
	OSU18 and Schu4	OSU18, but no corresponding CDS in Schu4	OSU18 and Schu4	OSU18 but intact in Schu4	Schu4 but intact in OSU18	
Stress response, efflux pumps, and antibiotic resistance	49	1	3	5	3	61
Lipopolysaccharide, capsule, and polysaccharide decoration	50	0	0	0	0	50
Posttranslational processing and secretion	34	0	4	2	0	40
F. tularensis pathogenicity island	30	0	0	4	0	34
Intracellular survival	19	0	2	2	1	24
Iron acquisition, transport, and metabolism	19	0	1	0	1	21
Pili and outer membrane proteins	14	0	0	3	0	17
Transcriptional and translational regulation	11	0	0	2	0	13
Extracellular toxins, enzymes, and effectors	5	0	0	2	1	8
Total	231	1	10	20	6	268

[a] The numbers of genes found intact and as pseudogenes in each strain are indicated. A list of virulence genes can be found in Table S4 in the supplemental material.

induced during oxidative stress. An additional pseudogene for a possible virulence factor is a gene encoding a conserved protein having an unknown function that appears to contain a heme binding domain.

In addition to virulence pseudogenes, the overall pseudogene differences between the type A and type B *F. tularensis* subspecies may also affect their relative pathogenicities and geographic distributions. Type A and B strains have relatively large numbers of pseudogenes; 321 of the 1,934 annotated OSU18 ORFs (16.6%) are pseudogenes, including disrupted *tnp* (there are 253 pseudogenes [13.1% of the total number of ORFs] if *tnp* pseudogenes are not included) (see Table S5 in the supplemental material). This number of pseudogenes is high compared to the numbers in other bacterial genomes (1 to 8%) and reinforces the conclusion that the *Francisella* genome is deteriorating. This may also be evidence of the organism's shift from existence in a free-living state to an intracellular lifestyle. Ninety-eight of the pseudogenes are shared by the two strains. The remaining 155 pseudogenes unique to OSU18 include 28 pseudogenes encoding transporters, 31 pseudogenes encoding enzymes involved in the general metabolic functions of the cell, and 73 pseudogenes whose functions are unknown (see Tables S6 and S7 in the supplemental material).

OSU18 and LVS differ by relatively few polymorphisms. The OSU18 genome sequence was compared to the LVS sequence to identify potential mutations responsible for LVS attenuation. In contrast to the rearrangement observed between OSU18 and Schu4, nucleotide alignment and optical mapping of the LVS and OSU18 genomes showed no major DNA rearrangements (Fig. 1A and data not shown). Further analysis at the nucleotide level showed that both strains have at least a fragment of every gene found in each individual strain. Therefore, smaller sequence variations and insertions and deletions (indels) may be responsible for LVS attenuation. This high level of identity is somewhat surprising given the fact that LVS was derived in Russia from a European type B strain sometime during the 1930s to 1950s and OSU18 was isolated almost 30 years later in Oklahoma. Phylogenic analysis of more than 150 type B isolates using multiple-locus variable-number tandem repeat analysis (15) suggested that there is little genetic diversity among type B isolates globally (5, 15). The authors concluded that the lack of diversity and geographical differentiation in type B isolates is consistent with "rapid transmission of a recently emerged pathogen over great distances" (5). The similarity of the OSU18 and LVS genome sequences supports this conclusion and encourages identification of the attenuating mutations in LVS.

A pairwise analysis of the OSU18 and LVS genomes was performed to catalog all sequence variants for the two strains occurring in genes that are intact in OSU18 (excluding *tnp* genes). A total of 409 base substitutions and 39 indels were found in LVS homologs of intact OSU18 genes. In contrast, over 5,000 substitutions and 500 indels were found when Schu4 and intact OSU18 genes were compared. When translated, 206 of the 409 base substitutions result in conservative amino acid changes (119 of these are silent mutations). The other 203 substitutions are not conservative, and three are nonsense mutations. Of the 39 indels, 10 retain the original reading frame of the encoded gene, and seven of these result in insertion or deletion of more than three amino acids (Table 3).

The 409 substitutions are located in 338 ORFs (see Table S8 in the supplemental material), with no gene containing more than four substitutions and 275 genes having only one altered

TABLE 3. Effects of substitutions and indels in OSU18 and LVS ORFs

Category	No.
Substitutions	
Affected genes	338
Single substitution	274
Double substitution	56
Triple substitution	7
Quadruple substitution	1
ORFs with only conservative (silent) substitutions	158 (97)
ORFs with only nonconservative substitutions[a]	180
Indels	
Affected genes	36
Three-amino-acid or less, in frame deletion[a]	10
Pseudogene in LVS[a]	19

[a] Category of polymorphisms that most likely contains LVS attenuating mutations.

nucleotide. Ninety-seven ORFs encode the original amino acid sequence, while 180 ORFs have undergone a nonconservative substitution(s). The effects of indels on 36 OSU18 and LVS ORFs were also evaluated (Table 3; see Table S9 in the supplemental material). Ten ORFs had in-frame indels, while 19 ORFs resulted in the LVS allele becoming a pseudogene. These 29 indel ORFs, combined with the 180 nonconservatively substituted ORFs, represent all of the intragenic, nonconservative variants for LVS and OSU18 and will be the subject of further studies to determine the LVS attenuating mutations.

One interesting indel involves a five-member cluster of genes encoding hypothetical proteins in the Schu4 genome identified by Larsson et al. (FTT0025, FTT0267, FTT0602, FTT0918, and FTT0919) (16). Homologs of four of these genes are intact in OSU18, while the gene corresponding to FTT0267 is a pseudogene. Disruption of FTT0918, but not disruption of FTT0919, was recently shown to attenuate Schu4 (31), suggesting that at least one, but not all, of the genes in this protein family is necessary for *Francisella* virulence. In LVS, the FTT0918 and FTT0919 homologs have undergone a deletion event where the last 785 nt of FTT0918, the first 679 nt of FTT0919, and the 17 nt between these genes have been deleted, resulting in an FTT0918/FTT0919 fusion protein consisting of amino acids encoded by the first 297 codons of FTT0918 and the last 255 codons of FTT0919. Expression of the FTT0918/FTT0919 fusion protein was also reported for a spontaneous, avirulent type A mutant (FSC043) (31). The function of each member of this family of proteins is of immediate interest in terms of *Francisella* virulence and lifestyle.

Of the 374 OSU18 and LVS ORFs that differ between the two strains, 11 contain nonconservative changes in predicted virulence genes. Included among these ORFs are three genes encoding proteins involved in type IV pilus assembly (*pilB* and *pilF*, each encoding a nonconservative amino acid polymorphism, and *pilE4*, which contains a frameshift approximately 19 codons from the 3′ end of the gene). From studies with *Pseudomonas* and *Neisseria*, PilB is thought to be involved in type IV pilus extension (25), PilF is the ATPase required for pilus assembly, and PilE4 is one of the pilus subunit proteins. As mentioned above, type IV pili have been observed on the surface of LVS by electron microscopy despite the three type B subspecies-specific pseudogenes mentioned above in *pilT*, *pilE2*, and *pilE3* (6). While these pseudogenes do not appear to be responsible for LVS attenuation, the LVS-specific mutations in *pilB*, *pilF*, and *pilE4* could impact LVS virulence through subtle pilus defects that impair the normal function. Other putative OSU18 and LVS virulence genes that contain nonconservative mutations are genes encoding a translational regulator (peptidylprolyl isomerase) and tyrosine phosphoprotein (*typA*) (26), genes encoding two multidrug efflux pumps, a preprotein translocase subunit gene (*yajC*), a glycosyltransferase gene possibly involved in lipopolysaccharide biosynthesis, and a gene encoding an NMC family nicotinamide mononucleotide uptake permease, *pnuC*.

Other OSU18 and LVS variant genes include genes belonging to other categories which may be evaluated for their attenuation potential; however, this list does not include promoter or other intergenic variations that may affect transcriptional activator or repressor binding. Only two sigma factor genes were found in the OSU18 genome (*rpoD* and *rpoH*), but even with this apparent constraint, initial efforts using *Escherichia coli*-trained promoter identification programs and intergenic sequence alignments to identify putative OSU18 promoters gave ambiguous results.

A separate search was performed with OSU18 pseudogenes to determine whether their LVS homologs are also pseudogenes or are intact. A total of 135 of the 253 non-*tnp* OSU18 pseudogenes contained sequence variants compared to their LVS homologs. In these 135 alleles, we found 70 indels and 163 single-nucleotide polymorphisms. Interestingly, 13 of these 135 OSU18 pseudogenes are intact in LVS, and 11 of these 13 are also intact in Schu4. Proteins encoded by three of these genes are associated with virulence in other organisms (phospholipase C, phospholipase D, and a cold shock protein homolog). Whether these differences affect *Francisella* pathogenicity is currently unknown.

DISCUSSION

The genetic factors underlying *Francisella* pathogenicity and lifestyle are largely unknown, although recent studies have revealed several genes required for intracellular survival of the bacterium in murine and human systems. Sequencing and comparison of representative strains of each subspecies, such as Schu4, LVS, and OSU18, should further reveal the genes that impart greater virulence to type A strains in humans and should also help define candidate attenuation alleles for vaccine development. The most striking distinction between the type A and B subspecies is the level of genomic rearrangement that has taken place within them. Optical map and gene organization data for two type A strains, ATCC 6223 and Schu4, and two type B strains from different continents, OSU18 (North America) and LVS (Asia), reinforce the observation that while type B strains maintain their syntenic order, type A strains are highly rearranged compared to type B strains and to each other. It is noteworthy that the *tnp* genes for IS*Ftu2* may be functional in type A strains but not in type B strains. Genome rearrangements may then be stimulated in type A strains but frozen in type B strains as a consequence.

Despite these differences in gene order, overall gene content in both subspecies is highly conserved. The primary difference in gene content between the strains is that while OSU18 and Schu4 have over 300 and 200 pseudogenes, respectively, only 98 of the pseudogenes are shared. Therefore, each strain expresses over 100 unique genes that may influence its individual pathogenicity level and distribution. The FPI-encoded protein PdpD is one protein that is expressed in type A strains and is disrupted in type B isolates. A murine infection study showed that *pdpD* is required for intramacrophage growth of *F. tularensis* subsp. *novicida* (24). Therefore, pseudogenes of *pdpD* and *acpA* (both of which have been studied in *F. tularensis*) and phospholipase D are at least partially responsible for the differences between type A pathogenicity and type B pathogenicity. Validation of virulence phenotypes for the other unique type A and type B alleles requires further study and would be assisted by comparison of additional *Francisella* genome sequences.

The worldwide distribution and limited variation of type B *F. tularensis* strains indicate that this subspecies emerged recently

and spread rapidly (5) compared to the genetically diverse, North America-restricted type A subspecies. The two clades of the type A subspecies have been associated with specific vector species in the United States, while type B isolates have not followed any such pattern (5). The widespread transmission of the type B subspecies suggests that among the differences that distinguish it from the type A subspecies are alleles that facilitate type B subspecies dissemination. These differences may be caused by genes uniquely expressed or upregulated in OSU18 that confer to type B strains the ability to be transmitted with fewer host barriers or nutritional requirements. Alternatively, perhaps the genes that prohibit widespread dissemination of type A strains are disrupted or are regulated differently in type B strains, thus relaxing a growth restriction and allowing type B strains to spread. Candidate alleles like the alleles encoding the OSU18 chitinase gene, a pseudogene in Schu4, may enable type B strains to survive in more diversified natural reservoirs (e.g., a greater number of tick subspecies). Further studies of *Francisella* and its natural reservoirs are needed to better understand the global distribution of the various subspecies.

The relatively low level of diversity among type B strains facilitates a search for candidate attenuation alleles in LVS. The 374 genes showing OSU18/LVS variation represent the primary pool of candidate attenuation alleles. This population can be narrowed to 209 nonconservative mutations which may cause differences in virulence or attenuation. Further type A and B sequences should narrow this list further as polymorphisms naturally present in virulent *Francisella* populations can be identified and eliminated from the attenuation candidate list. Studies are under way to identify which of the candidate alleles are, in fact, attenuating. Related studies, not directly involving LVS, have assisted in this process. For example, the FTT0918/FTT0919 fusion mutation found in a spontaneous, attenuated type A mutant is also present in LVS, but the FTT0918 and FTT0919 homologs are intact in OSU18. Therefore, inactivation of FTT0918 is probably one factor that contributes to LVS attenuation. Other factors will be studied as more *Francisella* genomes become available and as the genetic tools for *Francisella* continue to improve.

ACKNOWLEDGMENTS

This research was supported by a grant from the National Institute of Allergy and Infectious Disease to G.W. (R21 AI061106). Joseph Petrosino was also supported by a career development award from the Western Regional Center of Excellence (RCE VI) for Biodefense and Emerging Infectious Diseases (U54 AI057156). T. M. Raghavan acknowledges the Marine Biological Laboratory's NASA Planetary Biology Internship Program for their financial support for his contribution in this work. Optical mapping was supported by the OSU Center for Veterinary Health Sciences.

We thank Richard Gibbs, Donna Muzny, Christie Kovar-Smith, Lynne Nazareth, Erica Sodergren, David Parker, Aleks Milosavljevic, and the rest of the staff at the Human Genome Sequencing Center for their support during this project.

REFERENCES

1. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.
2. **Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg.** 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. **27:**4636–4641.
3. **Dennis, D. T., T. V. Inglesby, D. A. Henderson, J. G. Bartlett, M. S. Ascher, E. Eitzen, A. D. Fine, A. M. Friedlander, J. Hauer, M. Layton, S. R. Lillibridge, J. E. McDade, M. T. Osterholm, T. O'Toole, G. Parker, T. M. Perl, P. K. Russell, and K. Tonat.** 2001. Tularemia as a biological weapon: medical and public health management. JAMA **285:**2763–2773.
4. **Elkins, K. L., T. R. Rhinehart-Jones, S. J. Culkin, D. Yee, and R. K. Winegar.** 1996. Minimal requirements for murine resistance to infection with *Francisella tularensis* LVS. Infect. Immun. **64:**3288–3293.
5. **Farlow, J., D. Wagner, M. Dukerich, M. Stanley, M. Chu, K. Kubota, J. Petersen, and P. Keim.** 2005. *Francisella tularensis* in the United States. Emerg. Infect. Dis. **11:**1835–1841.
6. **Gil, H., J. L. Benach, and D. G. Thanassi.** 2004. Presence of pili on the surface of *Francisella tularensis*. Infect. Immun. **72:**3042–3047.
7. **Gordon, D., C. Abajian, and P. Green.** 1998. Consed: a graphical tool for sequence finishing. Genome Res. **8:**195–202.
8. **Green, M., G. Choules, D. Rogers, and R. W. Titball.** 2005. Efficacy of the live attenuated *Francisella tularensis* vaccine (LVS) in a murine model of disease. Vaccine **23:**2680–2686.
9. **Gurycova, D.** 1998. First isolation of *Francisella tularensis* subsp. *tularensis* in Europe. Eur. J. Epidemiol. **14:**797–802.
10. **Harris, S.** 1992. Japanese biological warfare research on humans: a case study of microbiology and ethics. Ann. N. Y. Acad. Sci. **666:**21–52.
11. **Havlak, P., R. Chen, K. J. Durbin, A. Egan, Y. Ren, X. Z. Song, G. M. Weinstock, and R. A. Gibbs.** 2004. The Atlas genome assembly system. Genome Res. **14:**721–732.
12. **Hinnebusch, B., A. Rudolph, P. Cherepanov, J. Dixon, T. Schwan, and A. Forsberg.** 2002. Role of *Yersinia* murine toxin in survival of *Yersinia pestis* in the midgut of the flea vector. Science **296:**733–735.
13. **Hinnebusch, J., P. Cherepanov, Y. Du, A. Rudolph, J. D. Dixon, T. Schwan, and A. Forsberg.** 2000. Murine toxin of *Yersinia pestis* shows phospholipase D activity but is not required for virulence in mice. Int. J. Med. Microbiol. **290:**483–487.
14. **Hopla, C. E.** 1974. The ecology of tularemia. Adv. Vet. Sci. Comp. Med. **18:**25–53.
15. **Johansson, A., J. Farlow, P. Larsson, M. Dukerich, E. Chambers, M. Bystrom, J. Fox, M. Chu, M. Forsman, A. Sjostedt, and P. Keim.** 2004. Worldwide genetic relationships among *Francisella tularensis* isolates determined by multiple-locus variable-number tandem repeat analysis. J. Bacteriol. **186:**5808–5818.
16. **Larsson, P., P. C. Oyston, P. Chain, M. C. Chu, M. Duffield, H. H. Fuxelius, E. Garcia, G. Halltorp, D. Johansson, K. E. Isherwood, P. D. Karp, E. Larsson, Y. Liu, S. Michell, J. Prior, R. Prior, S. Malfatti, A. Sjostedt, K. Svensson, N. Thompson, L. Vergez, J. K. Wagg, B. W. Wren, L. E. Lindler, S. G. Andersson, M. Forsman, and R. W. Titball.** 2005. The complete genome sequence of *Francisella tularensis*, the causative agent of tularemia. Nat. Genet. **37:**153–159.
17. **Liu, S. L., and K. E. Sanderson.** 1995. The chromosome of *Salmonella paratyphi* A is inverted by recombination between *rrnH* and *rrnG*. J. Bacteriol. **177:**6585–6592.
18. **Lowe, T., and S. Eddy.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25:**955–964.
19. **Lukashin, A. V., and M. Borodovsky.** 1998. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. **26:**1107–1115.
20. **Mahillon, J., and M. Chandler.** 1998. Insertion sequences. Microbiol. Mol. Biol Rev. **62:**725–774.
21. **Marchler-Bauer, A., J. B. Anderson, P. F. Cherukuri, C. DeWeese-Scott, L. Y. Geer, M. Gwadz, S. He, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, C. A. Liebert, C. Liu, F. Lu, G. H. Marchler, M. Mullokandov, B. A. Shoemaker, V. Simonyan, J. S. Song, P. A. Thiessen, R. A. Yamashita, J. J. Yin, D. Zhang, and S. H. Bryant.** 2005. CDD: a conserved domain database for protein classification. Nucleic Acids Res. **33:**D192—D196.
22. **McLeod, M. P., X. Qin, S. E. Karpathy, J. Gioia, S. K. Highlander, G. E. Fox, T. Z. McNeill, H. Jiang, D. Muzny, L. S. Jacob, A. C. Hawes, E. Sodergren, R. Gill, J. Hume, M. Morgan, G. Fan, A. G. Amin, R. A. Gibbs, C. Hong, X.-J. Yu, D. H. Walker, and G. M. Weinstock.** 2004. Complete genome sequence of *Rickettsia typhi* and comparison with sequences of other rickettsiae. J. Bacteriol. **186:**5842–5855.
23. **Moore, S. D., and R. T. Sauer.** 2005. Ribosome rescue: tmRNA tagging activity and capacity in *Escherichia coli*. Mol. Microbiol. **58:**456–466.
24. **Nano, F. E., N. Zhang, S. C. Cowley, K. E. Klose, K. K. Cheung, M. J. Roberts, J. S. Ludu, G. W. Letendre, A. I. Meierovics, G. Stephens, and K. L. Elkins.** 2004. A *Francisella tularensis* pathogenicity island required for intramacrophage growth. J. Bacteriol. **186:**6430–6436.
25. **Nudleman, E., and D. Kaiser.** 2004. Pulling together with type IV pili. J. Mol. Microbiol. Biotechnol. **7:**52–62.
26. **Owens, R., G. Pritchard, P. Skipp, M. Hodey, S. Connell, K. Nierhaus, and C. O'Connor.** 2004. A dedicated translation factor controls the synthesis of the global regulator Fis. EMBO J. **23:**3375–3385.
27. **Oyston, P. C., A. Sjostedt, and R. W. Titball.** 2004. Tularaemia: bioterrorism defence renews interest in *Francisella tularensis*. Nat. Rev. Microbiol. **2:**967–978.

28. **Paranjpye, R. N., and M. S. Strom.** 2005. A *Vibrio vulnificus* type IV pilin contributes to biofilm formation, adherence to epithelial cells, and virulence. Infect. Immun. **73:**1411–1422.
29. **Rawlings, N. D., F. R. Morton, and A. J. Barrett.** 2006. MEROPS: the peptidase database. Nucleic Acids Res. **34:**D270–D272.
30. **Santic, M., M. Molmeret, K. E. Klose, S. Jones, and Y. A. Kwaik.** 2005. The *Francisella tularensis* pathogenicity island protein IglC and its regulator MglA are essential for modulating phagosome biogenesis and subsequent bacterial escape into the cytoplasm. Cell. Microbiol. **7:**969–979.
31. **Twine, S., M. Bystrom, W. Chen, M. Forsman, I. Golovliov, A. Johansson, J. Kelly, H. Lindgren, K. Svensson, C. Zingmark, W. Conlan, and A. Sjostedt.** 2005. A mutant of *Francisella tularensis* strain SCHU S4 lacking the ability to express a 58-kilodalton protein is attenuated for virulence and is an effective live vaccine. Infect. Immun. **73:**8345–8352.