INFECTION AND IMMUNITY, Dec. 2006, p. 6895–6906 Vol. 74, No. 12
0019-9567/06/$08.00+0 doi:10.1128/IAI.01006-06

Potential Source of *Francisella tularensis* Live Vaccine Strain Attenuation Determined by Genome Comparison▽†

Laurence Rohmer,[1] Mitchell Brittnacher,[1] Kerstin Svensson,[5,6] Danielle Buckley,[4] Eric Haugen,[4] Yang Zhou,[4] Jean Chang,[4] Ruth Levy,[4] Hillary Hayden,[4] Mats Forsman,[5] Maynard Olson,[1,3,4] Anders Johansson,[5,6] Rajinder Kaul,[3,4] and Samuel I. Miller[1,2,3]*

Departments of Genome Sciences,[1] Microbiology,[2] and Medicine[3] and University of Washington Genome Center,[4] University of Washington, Seattle, Washington 98195; NBC-Analysis, Division of NBC-Defence, Swedish Defence Research Agency, SE-901 82 Umeå, Sweden[5]; and Department of Clinical Microbiology, Infectious Diseases, Umeå University, SE-901 85 Umeå, Sweden[6]

Received 26 June 2006/Returned for modification 15 August 2006/Accepted 15 September 2006

***Francisella tularensis* is a bacterial pathogen that causes the zoonotic disease tularemia and is important to biodefense. Currently, the only vaccine known to confer protection against tularemia is a specific live vaccine strain (designated LVS) derived from a virulent isolate of *Francisella tularensis* subsp. *holarctica*. The origin and source of attenuation of this strain are not known. To assist with the design of a defined live vaccine strain, we sought to determine the genetic basis of the attenuation of LVS. This analysis relied primarily on the comparison between the genome of LVS and *Francisella tularensis holarctica* strain FSC200, which differ by only 0.08% of their nucleotide sequences. Under the assumption that the attenuation was due to a loss of function(s), only coding regions were examined in this comparison. To complement this analysis, the coding regions of two slightly more distantly related *Francisella tularensis* strains were also compared against the LVS coding regions. Thirty-five genes show unique sequence variations predicted to alter the protein sequence in LVS compared to the other *Francisella tularensis* strains. Due to these polymorphisms, the functions of 15 of these genes are very likely lost or impaired. Seven of these genes were demonstrated to be under stronger selective constraints, suggesting that they are the most probable to be the source of LVS attenuation and useful for a newly defined vaccine.**

The zoonotic disease tularemia is caused by the gram-negative bacterium *Francisella tularensis* (40). *F. tularensis* is often transmitted to humans through the bite of ticks or mosquitoes, inhalation of hay dust, ingestion of infected food or water, or physical contact with infected animals. The severity of the disease depends on the route and dose of infection as well as bacterial subtype (41). *F. tularensis* is currently divided into several subspecies: *tularensis*, *holarctica*, and *mediasiatica* (53). The species *Francisella novicida* is genetically very close to *Francisella tularensis* and causes a disease similar to that caused by *F. tularensis* subsp. *tularensis* and *holarctica* in mice (53). The subspecies *tularensis* and *holarctica* are often referred to as type A and type B, respectively, and are the two clinically dominant subspecies. Type A strains are among the most infectious human pathogens known with an infectious dose as low as 10 to 50 CFU (19, 51, 52). *F. tularensis* is considered a potential biological weapon and classified as a category A biological agent by the Centers for Disease Control and Prevention, implying a risk to national security if it is intentionally spread (48). Type A strains are endemic to North America and have been fatal in some cases of infection (41). Type B strains cause a similar but milder form of disease in humans and are found in North America, Europe, and Asia (42). Previous analyses of *F. tularensis* indicate a predominantly clonal population structure of the species and suggest that all type B strains form a recently emerged clonal lineage exhibiting very restricted genetic diversity among strains (26, 55).

No vaccine against this intracellular pathogen is currently approved by the Food and Drug Administration in the United States or the European Medicines Agency in the European Union (9, 14). Early studies of the efficacy of whole killed cells as a crude tularemia vaccine demonstrated a low level of protection in humans (6, 51, 52). Because *Francisella tularensis* is an intracellular pathogen, a genetically defined live vaccine, rather than a component vaccine, may be the best approach to vaccinating against tularemia (40, 56).

Consistent with this idea, the only effective tularemia vaccine available is an attenuated *F. tularensis* subsp. *holarctica* (type B) strain designated as the live vaccine strain (LVS). It was developed by the U.S. Department of Defense by selection of immunogenic bacterial colonies from an ampoule of Russian tularemia live vaccine that was imported in 1956 (57). Batch-to-batch variations in immunogenicity and adverse effects of LVS have been observed (50, 57). Russian and U.S. scientists also showed that LVS cultures easily dissociate into two colony types, of which one is immunogenic and the other is poorly immunogenic (13, 57). Several U.S. studies showed that LVS afforded protective immunity in humans, although the protection was insufficient upon exposure to large doses of highly virulent type A strains by the respiratory route (12). For these reasons as well as because of the unclear genetic basis of its attenuation, LVS is considered an experimental vaccine with restricted use in the United States (9). The Russian tularemia

* Corresponding author. Mailing address: Department of Medicine, University of Washington, 1959 NE Pacific St., Campus Box 357710, Seattle, WA 98195. Phone: (206) 616-5110. Fax: (206) 616-5109. E-mail: millersi@u.washington.edu.

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

▽ Published ahead of print on 25 September 2006.

vaccines were produced from fully virulent type B strains by classical bacteriological methods, including selection of individual bacterial colonies and repeated subcultures under conditions stressful to the bacterium (44). These laboratory procedures presumably led to mutations in the LVS genome that were both incidental and responsible for virulence attenuation.

This paper describes our comparative genomics approach to determine the genetic basis of the attenuation of LVS and presents a list of candidate genes that could be inactivated to develop new genetically defined live vaccines against *F. tularensis*. Prior to this study only two genetic regions that may relate to the attenuation of LVS, designated RD18 and RD19, have been described (49, 55). *F. tularensis* subsp. *tularensis* mutants with defective genes in each of these regions (the FTT0918 gene in RD18 and *pilA* in RD19) were shown to have attenuated virulence in a mouse model (18, 58). The genome of LVS (NCBI accession number NC_007880; GenBank accession number AM233362) was recently sequenced and made available to the public, presenting an opportunity to determine the causes of its attenuation. Ideally, LVS should be compared with its closest virulent progenitor to identify what changes are likely to be relevant for attenuation, since a virulent parent strain of LVS is not available. Because all *F. tularensis* type B isolates are highly conserved at the DNA level and represent a single clonal lineage, genome comparisons using another *F. tularensis* type B strain should reveal the smallest number of genetic differences. A previous single-nucleotide analysis of seven *F. tularensis* genes showed that type B strains of Eurasian origin, including strains from Russia, were all identical at the nucleotide level while North American type B strains exhibited a few nucleotide substitutions (55). Therefore, comparison of LVS with a European isolate would likely maximize the chance of identifying genetic differences that are relevant to the attenuation while minimizing the number of natural evolutionary differences. *F. tularensis* subsp. *holarctica* strain FSC200, which is currently being sequenced at the University of Washington Genome Center (UWGC), is a fully virulent strain of European origin and is genetically similar to LVS. FSC200 (*Francisella* Strain Collection, Swedish Defense Research Agency, Umeå, Sweden) is a clinical type B isolate obtained during a late-summer human epidemic of tularemia in Ljusdal, Sweden, in 1998. FSC200 has been stored at −70°C and was not subjected to laboratory passage before genome sequencing. The isolate tested fully virulent in mice and is genetically characterized as *F. tularensis* subsp. *holarctica* (18, 55). We sought to determine all nucleotide variations in the coding regions of LVS compared to the genome of FSC200. Due to the preliminary stage of the FSC200 genome sequence and to track natural variations between LVS and FSC200, we added to our analysis the genomes of two other *Francisella* strains for which the sequences were available (*F. tularensis* subsp. *tularensis* Schu S4 and *F. tularensis* subsp. *holarctica* OSU18). To identify genes for which the polymorphism specific to LVS is more likely to contribute to the attenuation, we separated the candidates into two classes based on predictions of the effect of the mutation on the gene function in the compared strains. To further this analysis, we used the assumption that genes in which a mutation would be detrimental for the bacterium are subjected to strong evolutionary selective pressure (i.e., are more conserved than highly dispensable genes). Under this assumption, we assessed the evolutionary pressures for some of the candidates and predicted genes under strong selective pressure, which allowed us to identify genes more likely to contribute to the attenuation of LVS.

MATERIALS AND METHODS

Whole-genome shotgun sequencing and assembly. The genome of *Francisella tularensis* subsp. *holarctica* FSC200 was sequenced using the standard sequencing protocols and data collection tools. Initially, a total of 41,127 paired-end sequence reads were collected from randomly picked small-insert plasmid clones. In addition, 960 paired-end sequence reads were collected from randomly picked fosmid clones. The sequences were assembled and viewed using phred/phrap/consed software (15, 16, 21). The shotgun data provided 12-fold sequence coverage for the FSC200 genome. The FSC200 genome underwent four rounds of autofinishing and is currently being finished by an expert in high-quality genome assembly. We used version 1.0 of the genome assembly software for current comparative genome analysis, including 45 contigs of 2 kb or more.

Genomic and protein sequences used to determine the nonsilent sequence variations. The *F. tularensis* subsp. *holarctica* strains used in this study were LVS (NC_007880), the draft sequence (45 contigs) of FSC200 (University of Washington) and the draft sequence of OSU18 (GenBank accession number NC_008369). Protein coding sequences for these three strains were predicted using Glimmer 2.13 (11). The *F. tularensis* subsp. *tularensis* Schu S4 protein sequences used were those of the published annotation (NC_006570).

Genome-wide comparisons. Genomic sequence comparisons were performed with the program Nucmer from the package MUMmer (33), using a minimum cluster length of 650 bp.

Identification and comparison of orthologous genes. Orthologous proteins in strains LVS and FSC200 were first determined by the reciprocal best hit method using the blastp algorithm (1, 47). Potential candidates with a role in LVS attenuation are the genes for which the protein sequences are not identical between the two strains. All sequences homologous to known IS elements were not considered in the analysis. The protein candidates were associated with their counterpart orthologous proteins in the strains Schu S4 and OSU18 using the same reciprocal best hit method as above. For the few cases where orthologous genes were not predicted by Glimmer or a gene had undergone significant deletion or frameshifting in LVS, the blastn algorithm was used to search for a matching sequence in the genome of LVS.

The location and type of polymorphisms (single-nucleotide polymorphism [SNP], indel) was determined from sequence alignments generated by Dialign2 (36).

Prediction of the effect of the variations on protein function. Protein sequence variations specific to LVS were investigated for potential loss or impairment of protein function using the sorting intolerant from tolerant (SIFT) algorithm (37). The SIFT algorithm predicates its decision as to whether a particular amino acid substitution is null on a measure of conservation in aligned homologous sequences from a nonredundant database. For cases in which an insufficient number of homologous sequences were available in protein databases (i.e., SIFT could not be used), a potential loss of function was inferred from similarity of the substituted amino acids using BLOSUM90 matrix scores. Positive scores for a substitution were counted as conservative. We considered the BLOSUM90 similarity matrix to be the most appropriate to our analysis because the average homology between the candidates and their counterparts in the other strains was greater than 90%.

Pseudogenes were predicted when the orthologous protein sequences differed in length by more than 20% (35). For all cases of putative pseudogenes observed in the set of candidates, the alignment of orthologs was examined manually and validated by alignments with orthologs in other bacterial species.

Functional annotation of the candidate proteins. Protein function was manually determined based on homologies to domains found in the Pfam database (http://pfam.wustl.edu/), the Prosite database (http://www.expasy.org/prosite/), the cdd database (ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd/), homologies to proteins of the nr database, and the TCDB database (http://www.tcdb.org/). Protein locations were predicted using PSORTB (20). Signal peptide cleavage sites were predicted using the software SignalP 3.0 (5).

Sequencing of selected loci in multiple strains. To assess the evolutionary pressure on candidate genes, selected loci were sequenced in 15 *F. tularensis* isolates from diverse biological and geographical sources, obtained from the *Francisella* Strain Collection at the Swedish Defense Research Agency, Umeå, Sweden (see Table 3). The DNA from the live vaccine strain ATCC 29684, lot 11 (American Type Culture Collection, Rockville, MD), was kindly provided by

Tina Guina (University of Washington) and sequenced as a control. PCR amplification and sequencing primers were designed based upon the consensus sequences from three finished reference genomes: *F. tularensis* subsp. *tularensis* SchuS4, *F. tularensis* subsp. *holarctica* LVS, and *Francisella novicida* U112 (UWGC, unpublished results). Primer sequences, conditions for PCR amplification, and expected product sizes are listed in Tables S1 and 2 in the supplemental material. DNA sequencing of PCR-amplified products was carried out in both directions using standard sequencing protocols at the University of Washington Genome Center. The locus-specific sequence data from all 16 strains were assembled using PHRED and PHRAP (15, 16).

Assessment of the selective pressure for each candidate locus. For this analysis, the genes considered pseudogenes were eliminated, and sequences were trimmed to cover only the sequence of the coding region of the gene of interest. Gene sequences from the *Francisella* complete genomes (*F. tularensis* subsp. *holarctica* FSC200, *F. tularensis* subsp. *holarctica* OSU18, *F. tularensis* subsp. *tularensis* Schu S4, and *F. novicida* U112 [University of Washington Genome Center]) were also used in this analysis. Nucleotide alignments were generated with DIALIGN2 on the basis of the translation of nucleotide diagonals into peptide diagonals (36). Based on these alignments, the phylogenetic distance between sequences was determined by using Dnadist in the Phylip software package (http://evolution.genetics.washington.edu/phylip.html), using the Kimura two-parameter model (31). Neighbor-joining phylogenetic trees were then built with the program neighbor based on those distances. The alignments and derived phylogenetic trees were used in Codeml from the PAML software package (61) to calculate the w-ratio of nonsynonymous to synonymous substitutions (dN/dS) for each codon position. The following evolutionary models were examined (62): Model M0 assumed a constant w-ratio, models M1 and M7 assumed that amino acid substitutions were either neutral ($w \sim 1$) or conservative ($w \sim 0$), and models M3 and M8 allowed the occurrence of positively selected sites ($w > 1$). Models M7 and M8 assume a β-distribution for the w-value between 0 and 1. The likelihood ratio of two models is compared (M3 versus M1 or M0 and M8 versus M7) to test which model fits the data significantly better. Twice the difference in log likelihood between the two models is compared with a chi-square distribution with n degrees of freedom, n being the difference between the numbers of parameters of the two models compared. The cutoff chosen was a P value of 0.05, which is acceptable because this likelihood ratio test is very conservative (2). The w-ratio values generated by the best-fit model were examined to determine whether the gene was subjected to purifying selection, was evolving neutrally, or was subjected to positive selection. Nucleotide variation rates were calculated by averaging the sum of the number of variations of each sequence compared to the consensus sequence over the number of sequences.

Nucleotide sequence accession numbers. The sequences of the PCR products obtained for this study have been deposited in GenBank with the accession numbers DQ863333 to DQ863497. The FSC200 whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the project accession number AASP00000000. The version described in this paper is the first version, AASP01000000.

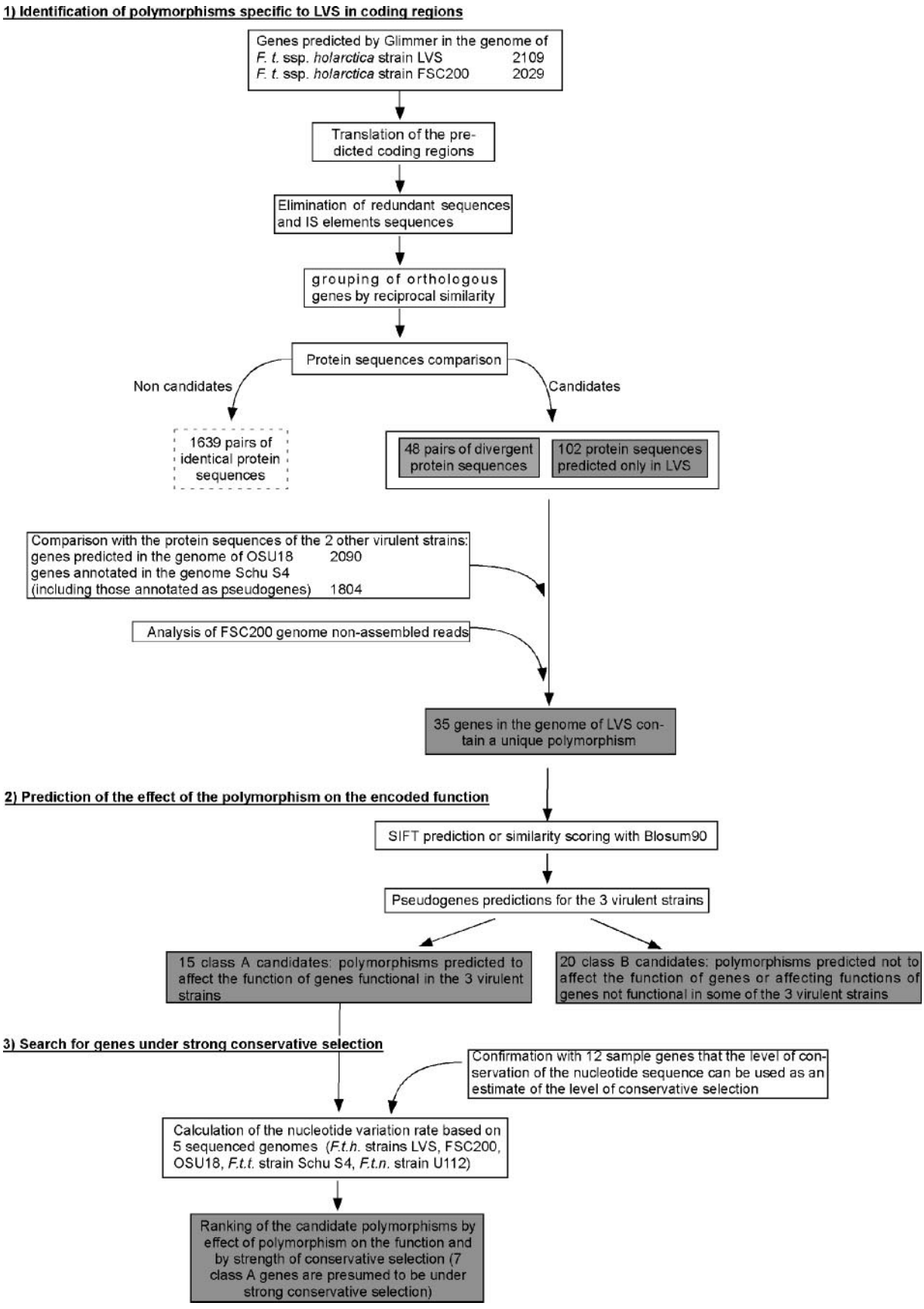
RESULTS

The estimated genetic variation between *Francisella tularensis* subsp. *holarctica* strain LVS and *Francisella tularensis* subsp. *holarctica* strain FSC200 is 0.08%. To identify the genetic causes of LVS attenuation, we sought to compare its genome to that of a closely related strain. We sequenced the genome of *F. tularensis* subsp. *holarctica* FSC200, and the draft genome assembly for FSC200 (1,816,540 bp in 45 contigs) was compared to the LVS genome (1,895,994 bp). A comparison of the DNA sequences of the LVS and FSC200 genomes showed them to be almost identical with an average identity of 99.92% (and conversely, 0.08% divergence). A closer examination revealed that much of the sequence variation was due to incomplete matching to IS elements. The percent identity will likely be higher as the final assembly of the FSC200 genome is completed.

The estimated genetic variations between LVS and two additional virulent strains are 0.11% and 0.7%, respectively. To identify the mutations that occurred in the genome of FSC200 during its natural evolution relative to LVS and potential sequencing errors, we added to our analysis the comparison with the genomes of two other *F. tularensis* strains: the partial draft sequence of the genome of *F. tularensis* subsp. *holarctica* OSU18 (1,845,916 bp) and the published sequence of *F. tularensis* subsp. *tularensis* SchuS4 (1,892,819 bp). OSU18 and Schu S4 are both virulent strains isolated in the United States. If a position diverges between FSC200 and LVS but not between LVS and one or the two other virulent strains, this polymorphism is not specific to LVS and therefore does not play a role in the attenuation of this strain. The average identity between LVS and Schu S4 over the entire genome sequence is 99.30%. The identity between the sequence of LVS and the draft sequence of OSU18 was found by Nucmer to be 99.89%. About 22 kb of unique sequence in the genome of Schu S4 was not found in LVS. These sequences were not found in the draft version of FSC200 or OSU18 either. Hence, the sequence missing in LVS seems to be missing in general in *F. tularensis* subsp. *holarctica*. The sequences of these two genomes are slightly more divergent from the sequence of LVS, so more polymorphisms than those found by comparison with FSC200 are expected.

Thirty-five genes show sequence variations predicted to alter the protein sequence in LVS compared to the other *Francisella tularensis* strains. In an attempt to identify induced alterations in the genome of LVS that could explain the attenuation of LVS, we sought to determine all the nonsilent modifications in the coding regions of the LVS genome relative to the other *F. tularensis* strains. Silent mutations in the coding regions have no impact on the protein sequences and therefore cannot alter protein function. The procedure we followed to identify the coding regions with nonsilent mutations in LVS is depicted in Fig. 1. We first compared the sequence of the proteins encoded in the genome of LVS with the sequence of the proteins encoded in the genome of FSC200. Open reading frames (ORFs) were first predicted for both genomes using the software tool Glimmer version 2.13. Gene prediction for LVS yielded 2,109 predicted genes, while 2,029 genes were predicted for FSC200. Then, gene translation products were paired in the two genomes using reciprocal best hits from blastp sequence alignments. The protein sequences of 1,639 pairs were found to be identical and were eliminated from the list of potential candidate genes. A total of 145 proteins homologous to known IS element-encoded transposases were also eliminated. The average identity of the remaining proteins of the two strains was 99.95%. Forty-eight encoded proteins were found to differ between the two strains. In addition, 104 LVS-encoded proteins were not found in the preliminary sequence of the genome of FSC200. ORFs unique to LVS were essentially repeated sequences (IS elements) and small ORFs unlikely to encode proteins. Together, the 152 LVS genes constituted our original set of candidates.

The proteins for which a difference between the LVS and FSC200 strains was observed were then compared to the protein sequences of *F. tularensis* subsp. *tularensis* strain Schu S4 (including those annotated as pseudogene products) and of *F. tularensis* subsp. *holarctica* strain OSU18 (Fig. 1). When the variant positions between FSC200 and LVS in a candidate protein were conserved between LVS and the highly virulent


1) Identification of polymorphisms specific to LVS in coding regions
Genes predicted by Glimmer in the genome of
F. t. ssp. holarctica strain LVS 2109
F. t. ssp. holarctica strain FSC200 2029
Translation of the predicted coding regions
Elimination of redundant sequences and IS elements sequences
grouping of orthologous genes by reciprocal similarity
Protein sequences comparison
Non candidates
Candidates
1639 pairs of identical protein sequences
48 pairs of divergent protein sequences
102 protein sequences predicted only in LVS
Comparison with the protein sequences of the 2 other virulent strains:
genes predicted in the genome of OSU18 2090
genes annotated in the genome Schu S4
(including those annotated as pseudogenes) 1804
Analysis of FSC200 genome non-assembled reads
35 genes in the genome of LVS contain a unique polymorphism
2) Prediction of the effect of the polymorphism on the encoded function
SIFT prediction or similarity scoring with Blosum90
Pseudogenes predictions for the 3 virulent strains
15 class A candidates: polymorphisms predicted to affect the function of genes functional in the 3 virulent strains
20 class B candidates: polymorphisms predicted not to affect the function of genes or affecting functions of genes not functional in some of the 3 virulent strains
3) Search for genes under strong conservative selection
Confirmation with 12 sample genes that the level of conservation of the nucleotide sequence can be used as an estimate of the level of conservative selection
Calculation of the nucleotide variation rate based on 5 sequenced genomes (F.t.h. strains LVS, FSC200, OSU18, F.t.t. strain Schu S4, F.t.n. strain U112)
Ranking of the candidate polymorphisms by effect of polymorphism on the function and by strength of conservative selection (7 class A genes are presumed to be under strong conservative selection)

strain Schu S4, or the virulent subsp. *holarctica* strain OSU18, the protein was eliminated from the list of candidates. The remaining 35 proteins constituted the final set of candidates (Table 1). The polymorphisms in this set are 27 SNPs, including three nonsense mutations, five nucleotide deletions leading to two predicted protein fusions, and three nucleotide insertions leading to frameshift mutations. Two large sequence deletions (at the loci FTL_0391-FTL_0392 and FTL_0439) were also found by other groups using different experimental methods and could likely contribute to the attenuation of LVS (49, 55).

Fifteen genes with polymorphisms unique to LVS have an increased probability of lost or impaired function. To identify which of these differences between LVS and the other strains were more likely to result in loss of function, each sequence was examined individually. The two large deletions led to protein fusions (at the loci FTL_391-FTL_0392 and FTL_0439) with presumable loss of function. The variations in the loci FTL_1291-FTL_1292 (nonsense mutation) and FTL_0168 (frameshift mutation leading to a premature stop codon) result in the truncation of over 90% of the encoded proteins. A total loss of function is assumed for these LVS proteins. In the case of SNPs and moderate protein truncations, the SIFT software was utilized to predict the impact of the sequence changes on the protein function. SIFT relies on alignments with homologous proteins in a nonredundant database. Out of 19 proteins for which relevant homologous proteins could be found, 9 were predicted to contain a deleterious alteration (partial or total loss of function) and 10 contained a predicted tolerated alteration (no loss of function) (Table 1).

SIFT analysis could not be performed for proteins with few or very distant homologous proteins in the nonredundant database. In this case, the potential loss of function was investigated with the similarity score of the substituted amino acid at the variable position using the BLOSUM90 similarity matrix scores (see Materials and Methods). When the substitution was nonconservative according to the score, the mutation in the LVS protein was predicted to result in a loss of function. By this method, more proteins are predicted to have lost their function than by the SIFT method, since every position in the protein is considered equally. Out of five proteins, two were predicted to carry deleterious substitutions, and three were predicted to carry tolerated substitutions (Table 1).

In addition to using this approach, we examined the counterparts of the 35 candidates in the other available genomes to see whether they were functional in these genomes. Indeed, the genome of *F. tularensis* subsp. *tularensis* strain Schu S4 contains at least 200 pseudogenes (>11% of all identified genes) (34). This strain is very virulent, so a mutation in the genome of LVS in a gene that has become a pseudogene in Schu S4 is not likely to play an important role in the attenuation of the virulence of the LVS strain. A similar assumption can be made about pseudogenes in one of the *F. tularensis* subsp. *holarctica* virulent strains. Out of 35 genes, 13 were predicted to be nonfunctional in one or several of the sequenced strains. Three were identified as pseudogenes in *F. tularensis* subsp. *tularensis* Schu S4, two (FTT0358 and FTT0880) had been annotated as such, and one was found to be a pseudogene in this study because its product is 50% shorter than its counterpart in *F. novicida* U112.

Fifteen genes in total are functional in all strains, and their variation in LVS is predicted to be deleterious for the function of the encoded protein. These 15 genes are therefore the candidates most likely to explain LVS attenuation and were assigned to candidate class A. Another 20 candidates were termed class B and are considered less likely to explain the attenuation of LVS. Seven class B genes are functional in all strains other than LVS, but the mutation is predicted not to affect protein function, and 10 class B gene polymorphisms are predicted pseudogenes in *F. tularensis* subsp. *holarctica*. Finally, three class B genes had counterparts that are predicted to be pseudogenes in Schu S4.

Evolutionary analysis of 12 loci reveals genes among the candidates that are subjected to strong purifying selection. In general, genes that play a major role in bacterial survival or fitness are subjected to strong purifying selection (23, 27, 39). The substitution rate ratio for these genes, dN/dS, is predicted to be low. We sought to identify among the candidates genes that would be subjected to strong purifying selection. As a first attempt, we investigated the evolutionary pressures of a subset of candidate loci: nine class A candidates and three class B candidates (Table 2). The loci were sequenced in 15 representative *Francisella* strains from *F. tularensis* subsp. *holarctica*, *F. tularensis* subsp. *mediasiatica*, *F. tularensis* subsp. *tularensis*, and *Francisella novicida* (Table 3). The mutations in LVS were confirmed by resequencing. The coding regions from the loci in the different strains were aligned, and the nucleotide sequences were compared to assess the dN/dS ratio using Codeml of the PAML package (61). The dN/dS ratio for the best-fit PAML model is shown in Table 2 (see Materials and Methods). All the results of the PAML analysis are displayed in Table S3 in the supplemental material. Five loci (FTL_1773, FTL_1066, FTL_1611, FTL_1141, and FTL_0391-FTL_0392) seem to be under strong purifying selection consistent with an essential role.

Two loci among the five, FTL_1773 and FTL_1066, are subjected to high purifying selection over the entire open reading frame. FTL_1773 (dN/dS = 0.062, with model M0) is homologous to numerous proteins annotated as dyp-type peroxidases, and the Pfam domain Dyp-perox is found in the protein sequence with a level of significance of 1.7e-16. The LVS locus has undergone a deletion of 90 nucleotides encoding several residues conserved across bacterial species; hence, it is likely to be inactivated, as predicted by SIFT.

FTL_1066 (dN/dS = 0.060, with model M0) is altered by one SNP (R82I), which is predicted by SIFT to be deleterious. The

FIG. 1. Flowchart of the search for nonsilent polymorphisms in coding regions of LVS that are potentially relevant for the attenuation of the strain. In a first step, all the polymorphisms were collected. The second and third steps represent an attempt to assess the relevance of each polymorphism for the attenuation of LVS based on the predicted effect of the polymorphism for the function of the gene and on the selective pressures to which the gene is predicted to be subjected.

TABLE 1. All loci in LVS for which a nonsynonymous mutation has been detected by comparison to other sequenced *Francisella* strains

ORF locus tag[a]	Predicted product encoded by the locus	Type of mutation	Effect of mutation on protein translation[b]	Predicted impact on protein function[c]	Status of gene in the other four sequenced genomes[d]	Candidate class
FTL_0039	Hypothetical membrane protein	SNP	G68R	Deleterious (SIFT)	Functional in all strains	A
FTL_0101-FTL_0102	Cl^-:H^+ antiporter	1-bp deletion	107-aa chimeric protein resulting from a frameshift mutation	Deleterious	Functional in all strains	A
FTL_0168	Hypothetical membrane protein	1-bp insertion	15-aa chimeric protein resulting from a frameshift mutation	Deleterious	Functional in all strains	A
FTL_0180	Acyltransferase	SNP	I164S	Deleterious (SIFT)	Functional in all strains	A
FTL_0391-FTL-0392	Pilus assembly protein pilA	0.5-kb deletion	55-aa chimeric protein resulting in the fusion of two ORFs	Deleterious	Functional in all strains	A
FTL_0439	Hypothetical protein	1.5-kb deletion	551-aa chimeric protein resulting from a gene fusion	Deleterious	Functional in all strains	A
FTL_0806	Amino acid transporter family protein	SNP	F219V	Deleterious (SIFT)	Functional in all strains	A
FTL_1066	Fumaryl acetoacetate hydrolase family protein	SNP	R82I	Deleterious (SIFT)	Functional in all strains	A
FTL_1141	3-Oxoacyl-(acyl carrier protein) synthase III	SNP	N297H	Deleterious (SIFT)	Functional in all strains	A
FTL_1246	Conserved hypothetical protein	SNP	V144F	Deleterious (SIFT)	Functional in all strains	A
FTL_1517	Hypothetical protein	SNP	A79V	Deleterious (BLOSUM90)	Functional in all strains	A
FTL_1521	Chitinase family 18 protein	12-bp insertion	Additional NNDQ repeat at position 658	Deleterious (SIFT)	Functional in all strains	A
FTL_1611	Glycosyltransferase, group 2 family protein	SNP	Q80L	Deleterious (SIFT)	Functional in all strains	A
FTL_1773	Conserved hypothetical protein	90-bp deletion	30-aa truncation, positions 132 to 262	Deleterious (SIFT)	Functional in all strains	A
FTL_1860	Phosphoribosylformylglycinamidine synthase	SNP	T1038I	Deleterious (SIFT)	Functional in all strains	A
FTL_0062	Transcriptional regulator	SNP	E68* (position in predicted FSC200 ORF), protein truncated by 163 aa	Deleterious	Pseudogene in *F. tularensis* subsp. *holarctica*	B
FTL_0108	Major facilitator superfamily transport protein	SNP	V248I	Tolerated (SIFT)	Pseudogene in *F. tularensis* subsp. *holarctica*	B
FTL_0171	Tetrapyrrole methyltransferase family protein	SNP	L220I	Tolerated (SIFT)	Functional in all strains	B
FTL_0212	Hypothetical protein	1-bp insertion	308-aa chimeric protein resulting from a frameshift mutation at position 305 in a variable region	Tolerated (SIFT)	Missing in U112	B
FTL_0341-FTL_0342	Deoxyribodipyrimidine photolyase	SNP	G143V	Deleterious (SIFT)	Pseudogene in *F. tularensis* subsp. *holarctica*	B
FTL_0381	Conserved hypothetical transmembrane protein	SNP	L14F	Tolerated (SIFT)	Pseudogene in Schu S4	B
FTL_0524a[e]	Hypothetical protein	SNP	V65F	Tolerated (SIFT)	Functional in all strains	B
FTL_0847	Preprotein translocase family protein	SNP	A26S	Tolerated (SIFT)	Functional in all strains	B
FTL_0864	Predicted sugar isomerase	SNP	T2K	Tolerated (SIFT)	Functional in all strains	B
FTL_0997a[e]	Hypothetical protein	SNP	T36* (position in predicted FSC200 ORF), protein truncated by 232 aa	Deleterious	Pseudogene in *F. tularensis* subsp. *holarctica*	B
FTL_1242	ThiJ/Pfpl family protein	SNP	D46N	Tolerated (BLOSUM90)	Pseudogene in *F. tularensis* subsp. *holarctica*	B
FTL_1265	2-Amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase/ dihydropteroate synthase	SNP	D232N	Tolerated (SIFT)	Functional in all strains	B
FTL_1291-FTL_1292	Conserved hypothetical protein	SNP	Q106*, protein truncated by 222 aa	Deleterious	Pseudogene in Schu S4	B
FTL_1325	Hypothetical protein	SNP	T786N	Tolerated (SIFT)	Pseudogene in *F. tularensis* subsp. *holarctica*	B
FTL_1408	Chitin binding protein	SNP	S192I	Tolerated (SIFT)	Pseudogene in Schu S4 and potentially in *F. tularensis* subsp. *holarctica*	B
FTL_1552	Hypothetical membrane protein	SNP	A233T	Tolerated (BLOSUM90)	Functional in all strains	B
FTL_1619	Conserved hypothetical membrane protein	SNP	T141K	Deleterious (BLOSUM90)	Pseudogene in *F. tularensis* subsp. *holarctica*	B
FTL_1697-FTL_1698	Metal ion transporter	2-bp deletion	25-aa chimeric protein resulting from a frameshift mutation	Deleterious	Pseudogene in *F. tularensis* subsp. *holarctica*	B

FTL_1731	licB-like transmembrane protein	SNP	S12F	Tolerated (SIFT)	Functional in all strains	B
FTL_1959	Conserved hypothetical protein	SNP	H7Q	Tolerated (BLOSUM90)	Pseudogene in F. tularensis subsp. holarctica	B

[a] Locus tags from the published sequence NC_007880; when two ORFs resulted from the mutation, the locus tags for both are cited.
[b] SNPs are represented by the amino acid and its position in the protein sequence of the virulent F. tularensis subsp. holarctica strains, followed by the substituted amino acid in LVS.; stop codon; aa, amino acids.
[c] The method used to assess the impact of the mutation is indicated in parentheses. When no method is mentioned, the mutation was predicted to be deleterious because it truncated the protein by more than 20% of its length.
[d] F. tularensis subsp. tularensis Schu S4, F. tularensis subsp. holarctica FSC200, F. tularensis subsp. holarctica OSU18, and F. novicida U112.
[e] The ORF in which the mutation took place has no locus tag assigned and is downstream of the locus mentioned.

protein is similar to proteins involved in various aromatic compound catabolism pathways. For example, FTL_1066 matches the profile of COG0179 (2-keto-4-pentenoate hydratase/2-oxohepta-3-ene-1,7-dioic acid hydratase [significance, 1e-23]) and the Pfam domain FAA_hydrolase (fumaryl acetoacetate hydrolase family [2.4e-11]). On the basis of these homologies FTL_1066 is likely involved in aromatic compound catabolism, but its specific enzyme function cannot be identified. The FTL_1066 gene is predicted to be part of an operon that also encodes an amino acid permease. This suggests that FTL_1066 participates in the catabolism of an aromatic amino acid.

Three additional loci are predicted to be under strong purifying selection over most of their open reading frame. The locus FTL_1611 (dN/dS = 0.056, with model M3) encodes a predicted glycosyltransferase. FTL_1611 is similar to COG0463 (glycosyltransferases involved in cell wall biogenesis [1e-14]) and the Pfam domain Glycos_transf_2 (glycosyltransferase family 2 [7.3e-21]), suggesting a role in surface lipopolysaccharide (LPS) biosynthesis. Consistent with this role, the protein is predicted to be located in the inner membrane. It does not, however, belong to a characterized LPS biosynthesis operon. This could suggest that the gene may be adding some component to LPS in a condition-specific or regulated manner.

The locus FTL_0391-FTL_0392 has undergone a 0.5-kb deletion that led to the fusion of two genes. The gene at the 3′ end of the fusion is predicted to be a pseudogene in the sequenced *F. tularensis* subsp. *holarctica* strains. To assess the importance of the gene at the 5′ end (its counterpart in Schu S4 is FTT0890), the ORF was sequenced in the 15 strains. The sequence in strain U112 was discarded from this analysis, because evidence points to the fact that it is not the direct ortholog of the pilin gene being investigated. Most of the locus is predicted by PAML to be under strong purifying selection: dN/dS is ~0.061 with the model M3 for 86% of the positions, while about 5% of positions are subjected to positive selection ($P < 0.05$), and the rest are subjected to neutral genetic drift. The protein is predicted to be a pilin, based on its similarity to COG4969 (Tfp pilus assembly protein, major pilin PilA) and to the Pfam prokaryotic N-terminal methylation motif (N_methyl), found at the N termini of pilins and other proteins involved in secretion. It is part of a predicted operon of three genes, all of them predicted to be pilins from type IV pili.

The last protein, FTL_1141, is predicted to be under conservative selection (dN/dS is equal to 0.134 for 97% of the sequence, with model M3). However, it is inactivated in *F. tularensis* subsp. *tularensis* FSC054 by a nonsense mutation and in *F. tularensis* subsp. *tularensis* FSC046 and *F. novicida* FSC595 by unique genomic rearrangements at the 5′ end of the gene. It is possible that the mutations in the FTL_1141 locus have been acquired in the laboratory following the strain's isolation (Table 3). Even though this gene may be dispensable in rich medium growth conditions, it may be important to some steps of the bacterium's natural life cycle, including virulence. This gene is predicted to be part of the elongation pathway of saturated fatty acid, based on strong homologies to 3-oxoacyl-(acyl carrier protein) synthases III, encoded by *fabH*, of numerous bacterial species. The *Francisella* gene is located in an operon upstream of genes that have been found essential in *Escherichia coli* K-12, but its ortholog in K-12 is itself not essential (4).

TABLE 2. Predicted levels of nucleotide variability and purifying selective pressure for the candidate genes of class A and three sample genes of class B

Candidate class	ORF locus tag[a]	Predicted product encoded by the locus	Nucleotide variation rate[b]	dN/dS ratio derived from best-fit model in PAML analysis[c]	Predicted level of purifying selection[d]	Sequence data from the 15 additional *Francisella* strains[e]
A	FTL_0039	Hypothetical membrane protein	0.0123	NA[f]	Weaker	−
A	FTL_0102-FTL0101	Cl^-:H^+ antiporter	0.0188	0.4920	Weaker	+
A	FTL_0168	Hypothetical membrane protein	0.0044	NA	High	−
A	FTL_0180	Acyltransferase	0.0080	NA	Weaker	−
A	FTL_0391-FTL_0392	Pilus assembly protein pilA	0.0049	0.3095	High	+
A	FTL_0439_ORF1	Hypothetical protein	0.0040	NA	High	−
A	FTL_0439_ORF2	Hypothetical protein	0.0069	0.1037	Weaker	+
A	FTL_0806	Amino acid transporter family protein	0.0056	NA	Weaker	−
A	FTL_1066	Fumaryl acetoacetate hydrolase family protein	0.0027	0.0891	High	+
A	FTL_1141	3-Oxoacyl-(acyl carrier protein) synthase III	0.0039	0.3606	High	+
A	FTL_1246	Conserved hypothetical protein	0.0331	0.6701	Weaker	+
A	FTL_1517	Hypothetical protein	0.0082	NA	Weaker	−
A	FTL_1521	Chitinase family 18 protein	0.0406	NA	Weaker	−
A	FTL_1611	Glycosyltransferase, group 2 family protein	0.0031	0.1242	High	+
A	FTL_1773	Conserved hypothetical protein	0.0040	0.0615	High	+
A	FTL_1860	Phosphoribosylformylglycinamidine synthase	0.0079	0.0810	Weaker	+
B	FTL_0997a	Hypothetical protein	NA	0.0848	Weaker	+
B	FTL_1291-FTL_1292	Conserved hypothetical protein	NA	1.1930	Weaker	+
B	FTL_1697-FTL_1698	Metal ion transporter	NA	0.7753	Weaker	+

[a] Locus tags from the published sequence NC_007880; when two ORFs resulted from the mutation, the locus tags for both are cited.
[b] Based on the alignments of the ORFs from the five sequenced genomes and/or PAML analysis.
[c] Results for all models are available in Table S2 in the supplemental material.
[d] Inferred from the nucleotide variation rate, relative to the rate values for genes known to be under purifying selection.
[e] +, sequences available; −, no sequence available.
[f] NA, not available.

Seven variant loci may be more likely to contribute to the attenuation of LVS based on their high level of nucleotide conservation. It has been shown for *E. coli* that essential genes tend to evolve at a slower rate than dispensable genes (27). In their study, Jordan et al. observed less overall nucleotide variation in *E. coli* essential genes than in *E. coli* nonessential genes, synonymous and nonsynonymous substitutions alike. It is plausible that *Francisella tularensis* genes that determine

TABLE 3. Additional *Francisella* strains in which 12 loci were sequenced

Strain	Alternate designation	Source	Location of isolation	Year of isolation
F. tularensis subsp. *tularensis* (avirulent)				
FSC043	SCHU, avirulent	Type AI human ulcer	Ohio	1941
FSC230	ATCC 6223	Type AII human lymph node	Utah	1920
F. tularensis subsp. *tularensis* AI				
FSC041	Vavenby	Tick	British Columbia, Canada	1941
FSC046	Fox Downs	Human pleural fluid	Ohio	1940
F. tularensis subsp. *tularensis* AII				
FSC054	Nevada 14	Rabbit	Nevada	1953
FSC604	BA8859	Foal	Montana	1958
F. tularensis subsp. *holarctica*				
FSC035	BA423A	Beaver	Montana	1976
FSC398		Human ulcer	Orebro, Sweden	2003
FSC354		Avirulent derivative (pilA mutant) of FSC074, hare	Nås, Sweden	1974
F. tularensis subsp. *holarctica* (Japonica)				
FSC021	Ebina	Human	Japan	1958
FSC017	S-2	Human lymph node	Japan	1926
F. tularensis subsp. *mediasiatica*				
FSC149	120	Hare	Central Asia	1965
FSC148	240	Tick	Central Asia	1982
F. novicida				
FSC454	FNSp-1	Human blood	Spain	2003
FSC595	F58	Human	United Kingdom	2003

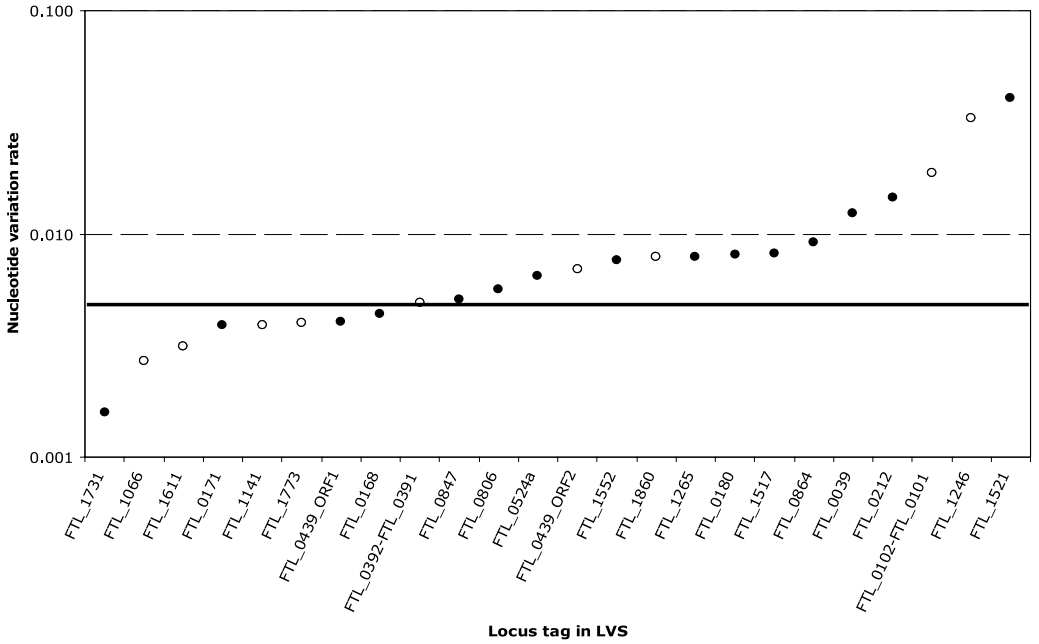


FIG. 2. Nucleotide variation rates of the candidate loci based on the alignment of their sequences in *F. tularensis* subsp. *tularensis* Schu S4, *F. novicida* U112, *F. tularensis* subsp. *holarctica* FSC200, and *F. tularensis* subsp. *holarctica* OSU18, displayed on a logarithmic scale. Open circles indicate loci that were sequenced in 15 additional *Francisella* strains, and the even distribution of their variation rates shows that they are a representative sample of the complete set of loci. The bar represents the threshold below which the nucleotide variation rate suggests strong conservative pressure. The threshold is the value of the highest nucleotide variation rate among those for genes predicted by PAML to undergo strong conservative selection.

growth or survival in the host evolve more slowly than dispensable ones. The nucleotide variation rate of nine class A loci and three class B loci across the sequences of the 15 strains (Table 3) and the five sequences from the genomes available (FSC200, OSU18, U112, Schu S4, and LVS) was compared to the PAML dN/dS ratio for the loci. The three class B loci were not included in the analysis because they were predicted to contain pseudogenes in too many of the strains (Table 2). The dN/dS ratio and the level of nucleotide variation are correlated with a Pearson correlation coefficient of 0.82. If the sequences from the five sequenced genomes alone are used to calculate the nucleotide variation rate, the correlation coefficient between dN/dS ratio and nucleotide variation rate is 0.84. Therefore, we used the nucleotide variation rate based on the sequences from the five sequenced genomes to approximately estimate the level of conservative selection to which each gene is subjected.

The nucleotide variation rate was calculated for all class A loci and the class B loci containing no pseudogenes based on the alignment between the five genome sequences available (FSC200, OSU18, U112, Schu S4, and LVS). Figure 2 shows the distribution of loci plotted against their respective nucleotide variation rates. As expected, the loci for which strong purifying selection was predicted (FTL_1773, FTL_1066, FTL_1611, FTL_0391-FTL_0392, and FTL_1141) have low nucleotide variation rates. If we apply as a conservative threshold the highest nucleotide variation rate among the loci predicted to undergo strong purifying selection (FTL_0391-FTL_0392, 0.0050), two additional class A candidate loci are predicted to undergo strong purifying selection: the ORF in the 5′ end of the loci FTL_0439_ORF1 (nucleotide variation rate, 0.0040) and FTL_0168 (nucleotide variation rate, 0.0044). The locus FTL_0439 is the remnant of two ORFs (ORF1 [FTT0918] and ORF2 [FTT0919] in Schu S4) that have been fused by a deletion of 1.5 kb.

No function could be predicted for FTL_0439_ORF1 and FTL_0168. However, they each contain a signal peptide cleavage site, suggesting that they may be secreted. Additionally, FTL_0439_ORF1 is predicted by PSORTB to be located in the outer membrane. A portion of FTL_0439_ORF1 is weakly similar to the PROSITE domain Cystatin, which represents a cystein protease inhibitor signature. The function of a cystein protease inhibitor in the context of an infection is not clear. Some cystein proteases are involved in apoptosis, so such a bacterial factor may contribute to control host cell death. The second conserved protein, FTL_0168, is homologous exclusively to eukaryotic proteins, grouped in COG5184 (alpha-tubulin suppressor and related RCC1 domain-containing proteins [5e-12]). It also matches the PROSITE domain RCC1_3, regulator of chromosome condensation (RCC1) repeat profile, twice. It is possible that FTL_0168 interacts with eukaryotic factors present in the host cell.

Two loci from class B (FTL_1731, FTL_0171) also show evidence of strong conservation, suggesting that they are important for the bacteria. However, the variations in LVS for these two genes are predicted to have no effect on the function of the protein. They could be important genes for the bacteria. In total, seven class A loci seem to undergo stronger conservative selection than the other candidates. The reason why these genes would be subjected to stronger selective pressures

may be that they confer an advantage in some standard conditions to *Francisella tularensis* and their alteration could result in reduced fitness in these conditions. Under this assumption, these candidates, altered in LVS, are more likely to partly explain LVS attenuation.

DISCUSSION

The goal of this study was to identify mutations in the genome of LVS that are responsible for its attenuation. We sought this information to provide researchers with candidate genes for use in the design of a new attenuated live vaccine strain. We worked under the assumption that the attenuation of LVS was due to the loss of one or several functions through the alteration of proteins. By comparing the coding regions of LVS with the coding regions of the close strain *F. tularensis* subsp. *holarctica* FSC200, and with the coding regions of two other close relatives, *F. tularensis* subsp. *tularensis* Schu S4 and *F. tularensis* subsp. *holarctica* OSU18, we uncovered 35 variations in predicted coding regions. Among them, 15 variations were predicted to significantly alter the function encoded by the gene in which they are located, and these genes were assigned to the candidate class A. The genes containing polymorphisms less likely to have an impact on the gene's function were assigned to class B. The 15 class A variations include three deletions (1.5 kb, 0.5 kb, and 90 bp, respectively), several indels, and SNPs. It is likely that not all of the 15 mutations are responsible for LVS attenuation. In the list of candidates, we sought to predict the genes potentially impacting the growth and/or virulence of strain LVS by identifying genes subjected to purifying selection. We identified nine genes that are subjected to stronger purifying selection than the other candidates, which indicate that loss of function may have a greater significance.

Seven class A genes were subjected to stronger purifying selection, indicating strong selective constraints and leading to high conservation of their gene sequence. Two have already been investigated in virulent *Francisella tularensis* strains. The deletion of the locus FTL_0439_ORF1 in *F. tularensis* subsp. *tularensis* strains (FTT0918) attenuates their virulence in mice (58). Bioinformatics predictions suggest that this protein could be secreted to the outer membrane and potentially act as a cystein protease inhibitor. The locus FTL_0391-FTL_0392 underwent a 0.5-kb deletion that inactivated a gene encoding a type IV pilus assembly protein. Its deletion in *F. tularensis* subsp. *holarctica* strains was found to attenuate the virulence of these strains in mice (18). The pilin could be important for the attachment of the bacteria to a host cell as has been demonstrated for *Pseudomonas aeruginosa* (22). Alternatively, this protein may be involved in protein secretion, and its inactivation could potentially impair the delivery of virulence factors from the bacterium to the host cell. These experimental studies indicate that some genes under strong selective constraints play an important role in the virulence of *Francisella tularensis*, which adds validity to our analysis. Additionally, they suggest that several mutations are involved in the attenuation of LVS.

For a successful infection, *Francisella tularensis* must be able to enter the host cell, survive, and replicate there. Two highly conserved genes (FTL_0168 and FTL_0439_ORF1) are potential virulence factors since they are predicted to be secreted and exhibit domains potentially involved in eukaryotic processes. It is consistent with the fact that strains in which FTL_0439_ORF1 is inactivated are attenuated in virulence. However, we could not accurately predict their mode of action.

The replication of *Francisella tularensis* in macrophages could be a major step towards virulence since the strains compromised in their ability to replicate in macrophages are attenuated in virulence (54). The ORF FTL_1773, which underwent a 90-nucleotide deletion in LVS and is highly conserved among the other strains (one of the seven most conserved in class A), is homologous to multiple peroxidases. After phagocytosis by the macrophage, some pathogens use peroxidases to counter oxidative stress (7), and without this protection LVS may be less viable in the host cells. Another candidate, FTL_0101-FTL_0102, altered by a frameshift mutation in LVS, may contribute to survival at low pH, another stress encountered in macrophage phagosomes. This locus, when not mutated, is predicted to encode a Cl^-:H^+ antiporter. A homolog of this gene in *E. coli* has been shown to be important for bacterial survival in low pH (gastric) environments. It works in concert with one of the extreme acid resistance systems (25). Although an extreme acid resistance system is present in *Francisella*, one of the three genes encoding this system in *Francisella tularensis* subsp. *holarctica* is predicted to be a pseudogene. It is therefore not clear how or whether the gene from locus FTL_0101-FTL_0102 contributes to the ability of the bacterium to cope with acidic stress.

FTL_1611, one of the seven most conserved of our candidates in class A, is predicted to encode a glycosyltransferase. The LVS gene contains a SNP predicted to be deleterious to gene function. Glycosyltransferases contribute to the virulence of numerous pathogens either by contributing to the synthesis of a polysaccharide capsule (10) or by alteration of the carbohydrate component of lipopolysaccharide (29). It is therefore possible that the locus may contribute to the attenuation of LVS by loss of extracellular carbohydrate. It has been observed that the LPS of LVS lacks a galactosamine-1-phosphate, in contrast to the LPS of another type B strain (43). The difference observed in LVS may well be due to the inactivation of FTL_1611. FTL_1521 is predicted to encode a chitinase family 18 protein. Recently, the counterpart of FTL_1521 in a highly virulent *F. tularensis* strain (encoded by FTT_0715) was identified as the single most upregulated protein during tularemia infection in a mouse model (59). The protein in LVS differs by an additional four-amino-acid repeat (NNDQ) compared to virulent strains of *F. tularensis* subsp. *holarctica* and *F. tularensis* subsp. *tularensis*. SIFT predicted this variation to be deleterious for the function of the chitinase, but the reason remains unclear (change in substrate specificity or affinity?). It has been observed in another intracellular human pathogen, *Leishmania mexicana*, that a chitinase family 18 member may represent a virulence determinant (28). It is interesting that mutations were found in a gene encoding a chitinase (FTL_1521), a gene encoding an acyltranferase immediately upstream of a gene encoding a pilus assembly protein (FTL_0180), a gene encoding a glycosyltransferase (FTL_1611), and a gene encoding a pilus assembly protein (FTL_0391-FTL_0392), since these proteins may be involved in pathogen attachment processes (17, 32, 45, 46, 60) A defect in the attachment process could impair the ability of the LVS to spread efficiently within a human host.

The inactivation of genes in metabolic systems is considered

a good strategy to develop a rationally attenuated vaccine since it could leave the bacteria deficient in some key components in the circumstances of an infection. Four genes in class A seem to belong to metabolic pathways: FTL_1066, FTL_0806, FTL_1141, and FTL_1860. Among the highly conserved candidates in class A, the FTL_1066 protein shows similarity to a domain shared by proteins of various enzymatic functions, including fumaryl acetoacetate hydrolases. Most of these proteins are predicted to be involved in catabolism of aromatic compounds. It is therefore a strong likelihood that this protein is involved in catabolism of some aromatic compound although the specific pathway cannot be identified. Under starvation conditions aromatic compounds might serve as an alternative carbon source, and loss of the capacity to catabolize these compounds may limit bacterial growth in the host. Another possible consequence of the disruption of a catabolic pathway is the accumulation of a toxic intermediate (3). Disruption of the degradation of an aromatic compound leading to starvation and/or toxicity for the bacterium would limit its growth. Similarly, loss of an amino acid transporter could potentially limit bacterial uptake of a necessary source of carbon and/or amino acids under starvation conditions. The candidate gene FTL_0806 is predicted to encode an amino acid transporter, and its inactivation could possibly impair the growth of LVS in the context of starvation and/or infection. FTL_1141 is predicted to encode 3-oxoacyl-(acyl carrier protein)-synthase III, a protein involved in the elongation of saturated chains of fatty acids. It is possible that within the host, the synthesis of some important lipid requires saturated-chain fatty acids. Finally, the phosphoribosylformylglycinamidine synthase PurL (FTL_1860) functions in an early stage of purine biosynthesis. Genes belonging to the same pathway have already been shown to attenuate growth/virulence of human bacterial pathogens. For example, a *Brucella melitensis purE* mutant is attenuated in growth in human cells (8). Purine pathway mutants in *Francisella tularensis* (30) as well as *Bacillus anthracis* (24) and *Salmonella enterica* serovar Typhimurium (38) have shown promise for induction of immunological protection against wild-type challenge. The two latter genes, FTL_1141 and FTL_1860, have been inactivated in some virulent strains. However, it is possible that these inactivations took place in laboratory conditions, following the isolation of the strains. Growth in laboratory conditions may not require all the metabolic pathways required in vivo, and strains undergoing these modifications could have an attenuating phenotype in the context of infection.

Ideally, a live vaccine strain should retain the ability to infect and migrate to the right niche within the host, so that immune responses against this bacterium can adequately develop. Additionally, it should contain mutations that would not impair its growth in vitro, so it could be easily produced. To allow an effective control of infection, a live vaccine strain should grow and/or spread slower within the human host than its wild-type counterpart. It is possible that the candidates we identified as being involved in metabolic pathways would be especially good for this purpose. Bacteria that are defective in metabolic pathways still retain all virulence attributes, their surface antigens, and their ability to colonize the right niche, but their growth rate is reduced compared to that of the wild type, which prevents an efficient infection of the host. Mutations impairing growth of a live vaccine strain within the host should be clear deletions so that the strain could not regain the lost function through further genomic modifications. SNPs, for example, may easily revert to the wild-type sequence and may contribute to the instability of LVS. It is not clear whether the mutations identified in this study, in particular the SNPs, lead to a total loss or only a partial loss of function. It is possible that some of the genes identified in this study are only diminished in efficiency instead of being completely inactivated. In this case, these candidates may not be suitable since complete deletion of the genes is required in the vaccine strain. This type of genomic analysis of attenuated strains has the potential to rapidly develop vaccine candidates. We believe that the results of the present study can be used in future work aiming at a rational design of a new live tularemia vaccine that is defective in causing severe infection but effective in eliciting an immune response.

ACKNOWLEDGMENTS

We thank Tina Guina (University of Washington) for providing material and insight in this project and the UWGC staff for producing the sequence data. Preliminary sequence data for the strain OSU18 were obtained from the Baylor College of Medicine Human Genome Sequencing Center website (http://www.hgsc.bcm.tmc.edu).

The DNA sequencing of the OSU18 genome was supported by grant 1R21AI061106-01 from NIH/NIAID to George Weinstock and Joseph Petrosino at the BCM-HGSC. A.J., K.S., and M.F. were funded by Swedish MoD project no. A4854 and the Medical Faculty, Umeå, Sweden. M.B., L.R., R.K., D.B., E.H., Y.Z., J.C., R.L., H.H., M.O., and S.I.M. are funded by WWAMI RCE grant U54AIO57141.

REFERENCES

1. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389–3402.
2. **Anisimova, M., J. P. Bielawski, and Z. Yang.** 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol. Biol. Evol. **18:**1585–1592.
3. **Arias-Barrau, E., E. R. Olivera, J. M. Luengo, C. Fernández, B. Galán, J. L. García, E. Díaz, and B. Miñambres.** 2004. The homogentisate pathway: a central catabolic pathway involved in the degradation of L-phenylalanine, L-tyrosine, and 3-hydroxyphenylacetate in *Pseudomonas putida*. J. Bacteriol. **186:**5062–5077.
4. **Baba, T., T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori.** 21 February 2006, posting date. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol. Syst. Biol. [Online.] doi:10.1038/msb4100050.
5. **Bendtsen, J. D., H. Nielsen, G. von Heijne, and S. Brunak.** 2004. Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. **340:**783–795.
6. **Burke, D. S.** 1977. Immunization against tularemia: analysis of the effectiveness of live Francisella tularensis vaccine in prevention of laboratory-acquired tularemia. J. Infect. Dis. **135:**55–60.
7. **Cianciotto, N. P.** 2001. Pathogenicity of Legionella pneumophila. Int. J. Med. Microbiol. **291:**331–343.
8. **Crawford, R. M., L. Van De Verg, L. Yuan, T. L. Hadfield, R. L. Warren, E. S. Drazek, H. H. Houng, C. Hammack, K. Sasala, T. Polsinelli, J. Thompson, and D. L. Hoover.** 1996. Deletion of *purE* attenuates *Brucella melitensis* infection in mice. Infect. Immun. **64:**2188–2192.
9. **Darling, R. G., and J. B. Woods (ed.).** 2004. USAMRIID's medical management of biological casualties handbook, 5th ed., vol. 1. USAMRIID, Fort Detrick, Md.
10. **DeAngelis, P. L.** 2002. Evolution of glycosaminoglycans and their glycosyltransferases: implications for the extracellular matrices of animals and the capsules of pathogenic bacteria. Anat. Rec. **268:**317–326.
11. **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.
12. **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.
13. **Emel'Ianova, O. S.** 1957. [Characteristics of tularemia vaccinal strains according to laboratory indices.]. Zh. Mikrobiol. Epidemiol. Immunobiol. **28:**125–129. (In Russian.)

14. **European Agency for the Evaluation of Medicinal Products.** 31 July 2002, posting date. EMEA/CPMP/4048/01. Guidance document on use of medicinal products for the treatment and prophylaxis of biological agents that might be used as weapons of bioterrorism. [Online.] http://www.emea.eu.int/htms/human/bioterror/bioterror.htm.
15. **Ewing, B., and P. Green.** 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. **8:**186–194.
16. **Ewing, B., L. Hillier, M. C. Wendl, and P. Green.** 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. **8:**175–185.
17. **Feldman, M. F., M. Wacker, M. Hernandez, P. G. Hitchen, C. L. Marolda, M. Kowarik, H. R. Morris, A. Dell, M. A. Valvano, and M. Aebi.** 2005. Engineering N-linked protein glycosylation with diverse O antigen lipopolysaccharide structures in Escherichia coli. Proc. Natl. Acad. Sci. USA **102:** 3016–3021.
18. **Forslund, A. L., K. Kuoppa, K. Svensson, E. Salomonsson, A. Johansson, M. Bystrom, P. C. Oyston, S. L. Michell, R. W. Titball, L. Noppa, E. Frithz-Lindsten, M. Forsman, and A. Forsberg.** 2006. Direct repeat-mediated deletion of a type IV pilin gene results in major virulence attenuation of Francisella tularensis. Mol. Microbiol. **59:**1818–1830.
19. **Forsman, M., and A. Johansson.** 2005. Tularemia (Francisella tularensis), p. 483–488. *In* R. F. Pilch and R. A. Zilinskas (ed.), Encyclopedia of bioterrorism defense, vol. 1. John Wiley & Sons, Inc., Hoboken, N.J.
20. **Gardy, J. L., M. R. Laird, F. Chen, S. Rey, C. J. Walsh, M. Ester, and F. S. Brinkman.** 2005. PSORTb v. 2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics **21:**617–623.
21. **Gordon, D., C. Desmarais, and P. Green.** 2001. Automated finishing with Autofinish. Genome Res. **11:**614–625.
22. **Hambrook, J., R. Titball, and C. Lindsay.** 2004. The interaction of Pseudomonas aeruginosa PAK with human and animal respiratory tract cell lines. FEMS Microbiol. Lett. **238:**49–55.
23. **Hirsh, A. E., and H. B. Fraser.** 2001. Protein dispensability and rate of evolution. Nature **411:**1046–1049.
24. **Ivanovics, G., E. Marjai, and A. Dobozy.** 1968. The growth of purine mutants of Bacillus anthracis in the body of the mouse. J. Gen. Microbiol. **53:**147–162.
25. **Iyer, R., C. Williams, and C. Miller.** 2003. Arginine-agmatine antiporter in extreme acid resistance in *Escherichia coli*. J. Bacteriol. **185:**6556–6561.
26. **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.
27. **Jordan, I. K., I. B. Rogozin, Y. I. Wolf, and E. V. Koonin.** 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. **12:**962–968.
28. **Joshi, M. B., M. E. Rogers, A. M. Shakarian, M. Yamage, S. A. Al-Harthi, P. A. Bates, and D. M. Dwyer.** 2005. Molecular characterization, expression, and in vivo analysis of LmexCht1: the chitinase of the human pathogen, Leishmania mexicana. J. Biol. Chem. **280:**3847–3861.
29. **Kahler, C. M., R. W. Carlson, M. M. Rahman, L. E. Martin, and D. S. Stephens.** 1996. Two glycosyltransferase genes, *lgtF* and *rfaK*, constitute the lipooligosaccharide ice (inner core extension) biosynthesis operon of *Neisseria meningitidis*. J. Bacteriol. **178:**6677–6684.
30. **Karlsson, J., R. G. Prior, K. Williams, L. Lindler, K. A. Brown, N. Chatwell, K. Hjalmarsson, N. Loman, K. A. Mack, M. Pallen, M. Popek, G. Sandstrom, A. Sjostedt, T. Svensson, I. Tamas, S. G. Andersson, B. W. Wren, P. C. Oyston, and R. W. Titball.** 2000. Sequencing of the Francisella tularensis strain Schu 4 genome reveals the shikimate and purine metabolic pathways, targets for the construction of a rationally attenuated auxotrophic vaccine. Microb. Comp. Genomics **5:**25–39.
31. **Kimura, M.** 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Biol. **16:**111–120.
32. **Kirn, T. J., B. A. Jude, and R. K. Taylor.** 2005. A colonization factor links Vibrio cholerae environmental survival and human infection. Nature **438:**863–866.
33. **Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg.** 2004. Versatile and open software for comparing large genomes. Genome Biol. **5:**R12.
34. **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.
35. **Lerat, E., and H. Ochman.** 2005. Recognizing the pseudogenes in bacterial genomes. Nucleic Acids Res. **33:**3125–3132.
36. **Morgenstern, B.** 1999. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics **15:**211–218.
37. **Ng, P. C., and S. Henikoff.** 2003. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. **31:**3812–3814.
38. **O'Callaghan, D., D. Maskell, J. Tite, and G. Dougan.** 1990. Immune responses in BALB/c mice following immunization with aromatic compound or purine-dependent Salmonella typhimurium strains. Immunology **69:**184–189.
39. **Ohta, T.** 1992. The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. **23:**263–286.
40. **Oyston, P. C., A. Sjöstedt, and R. W. Titball.** 2004. Tularaemia: bioterrorism defence renews interest in Francisella tularensis. Nat. Rev. Microbiol. **2:**967–978.
41. **Penn, R. L.** 2005. Francisella tularensis (Tularemia), p. 2674–2685. *In* G. L. Mandell, J. E. Bennet, and R. Dolin (ed.), Mandell, Douglas and Bennett's Principles and practice of infectious diseases, 6th ed., vol. 2. Churchill Livingstone, Oxford, United Kingdom.
42. **Petersen, J. M., and M. E. Schriefer.** 2005. Tularemia: emergence/re-emergence. Vet. Res. **36:**455–467.
43. **Phillips, N. J., B. Schilling, M. K. McLendon, M. A. Apicella, and B. W. Gibson.** 2004. Novel modification of lipid A of *Francisella tularensis*. Infect. Immun. **72:**5340–5348.
44. **Pollitzer, R.** 1967. History and incidence of tularemia in the Soviet Union; a review. Institute for Contemporary Russian Studies, Fordham University, Bronx, N.Y.
45. **Power, P. M., L. F. Roddam, K. Rutter, S. Z. Fitzpatrick, Y. N. Srikhanta, and M. P. Jennings.** 2003. Genetic characterization of pilin glycosylation and phase variation in Neisseria meningitidis. Mol. Microbiol. **49:**833–847.
46. **Reguera, G., and R. Kolter.** 2005. Virulence and the environment: a novel role for *Vibrio cholerae* toxin-coregulated pili in biofilm formation on chitin. J. Bacteriol. **187:**3551–3555.
47. **Rivera, M. C., R. Jain, J. E. Moore, and J. A. Lake.** 1998. Genomic evidence for two functionally distinct gene classes. Proc. Natl. Acad. Sci. USA **95:** 6239–6244.
48. **Rotz, L. D., A. S. Khan, S. R. Lillibridge, S. M. Ostroff, and J. M. Hughes.** 2002. Public health assessment of potential biological terrorism agents. Emerg. Infect. Dis. **8:**225–230.
49. **Samrakandi, M. M., C. Zhang, M. Zhang, J. Nietfeldt, J. Kim, P. C. Iwen, M. E. Olson, P. D. Fey, G. E. Duhamel, S. H. Hinrichs, J. D. Cirillo, and A. K. Benson.** 2004. Genome diversity among regional populations of Francisella tularensis subspecies tularensis and Francisella tularensis subspecies holarctica isolated from the US. FEMS Microbiol. Lett. **237:**9–17.
50. **Sandström, G.** 1994. The tularaemia vaccine. J. Chem. Technol. Biotechnol. **59:**315–320.
51. **Saslaw, S., H. T. Eigelsbach, J. A. Prior, H. E. Wilson, and S. Carhart.** 1961. Tularemia vaccine study. I. Intracutaneous challenge. Arch. Intern. Med. **107:**689–701.
52. **Saslaw, S., H. T. Eigelsbach, J. A. Prior, H. E. Wilson, and S. Carhart.** 1961. Tularemia vaccine study. II. Respiratory challenge. Arch. Intern. Med. **107:** 702–714.
53. **Sjöstedt, A.** 2001. Family XVII. Francisellaceae, Genus I. Francisella. *In* D. J. Brenner (ed.), Bergey's manual of systematic bacteriology. Springer-Verlag, New York, N.Y.
54. **Sjöstedt, A.** 2006. Intracellular survival mechanisms of Francisella tularensis, a stealth pathogen. Microbes Infect. **8:**561–567.
55. **Svensson, K., P. Larsson, D. Johansson, M. Bystrom, M. Forsman, and A. Johansson.** 2005. Evolution of subspecies of *Francisella tularensis*. J. Bacteriol. **187:**3903–3908.
56. **Tärnvik, A.** 1989. Nature of protective immunity to Francisella tularensis. Rev. Infect. Dis. **11:**440–451.
57. **Tigertt, W. D.** 1962. Soviet viable *Pasteurella tularensis* vaccines. A review of selected articles. Bacteriol. Rev. **26:**354–373.
58. **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.
59. **Twine, S. M., N. C. Mykytczuk, M. D. Petit, H. Shen, A. Sjöstedt, J. W. Conlan, and J. F. Kelly.** 2006. In vivo proteomic analysis of the intracellular bacterial pathogen, Francisella tularensis, isolated from mouse spleen. Biochem. Biophys. Res. Commun. **345:**1621–1633.
60. **Warren, M. L., L. F. Roddam, P. M. Power, T. D. Terry, and M. P. Jennings.** 2004. Analysis of the role of pglI in pilin glycosylation of Neisseria meningitidis. FEMS Immunol. Med. Microbiol. **41:**43–50.
61. **Yang, Z.** 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13:**555–556.
62. **Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen.** 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155:**431–449.

Editor: V. J. DiRita