

# **Final Report for LDRD Project 02-ERD-069: “Discovering the Unknown Mechanism(s) of Virulence in a BW, Class A Select Agent”**

*P. Chain, E. Garcia*

**February 6, 2003**

*U.S. Department of Energy*

Lawrence  
Livermore  
National  
Laboratory

## DISCLAIMER

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This work was performed under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

This report has been reproduced directly from the best available copy.

Available electronically at <http://www.doc.gov/bridge>

Available for a processing fee to U.S. Department of Energy  
And its contractors in paper from  
U.S. Department of Energy  
Office of Scientific and Technical Information  
P.O. Box 62  
Oak Ridge, TN 37831-0062  
Telephone: (865) 576-8401  
Facsimile: (865) 576-5728  
E-mail: [reports@adonis.osti.gov](mailto:reports@adonis.osti.gov)

Available for the sale to the public from  
U.S. Department of Commerce  
National Technical Information Service  
5285 Port Royal Road  
Springfield, VA 22161  
Telephone: (800) 553-6847  
Facsimile: (703) 605-6900  
E-mail: [orders@ntis.fedworld.gov](mailto:orders@ntis.fedworld.gov)  
Online ordering: <http://www.ntis.gov/ordering.htm>

OR

Lawrence Livermore National Laboratory  
Technical Information Department's Digital Library  
<http://www.llnl.gov/tid/Library.html>

UCRL-ID-151379

**“Discovering the Unknown Mechanism(s) of Virulence in a BW, Class A Select Agent”**

**PI:** Patrick Chain, Biology and Biotechnology Research Program, L-441, x45492

**Co-Investigators (BBRP):** Emilio Garcia

**Final Report for LDRD Project 02-ERD-069**

**Background**

The post-September 11 anthrax scare has forced us to reexamine our BW vulnerabilities and recognize the shortcomings of our current understanding of the most serious biothreat organisms. Four microbial organisms (*Yersinia pestis* – plague, *Bacillus anthracis* – anthrax, *Francisella tularensis* – tularemia, and *Clostridium botulinum* – botulism), along with the smallpox virus (*Variola major*) and viral hemorrhagic fever virus families (Filoviruses and Arenaviruses) are currently categorized by the CDC as the most likely agents (Category A) to be used in a biological terrorist attack. Unlike the other CDC Category A microbial agents, *F. tularensis* is not thoroughly studied at the biochemical, genetic and genomic levels and thus remains one of the least characterized potential BW agents.

Much of the scientific research in *Francisella* has focused on classification and typing. The genus *Francisella* is composed of two or three species, depending on whom you talk to and is in a constant state of flux. The current description in Bergey’s Manual of Systematic Bacteriology (1984) states that there are two species in the *Francisella* genus: *F. tularensis* and *F. novicida*. However, with the advent of more sophisticated molecular techniques, it appears that *F. novicida* could be considered a subspecies of *F. tularensis* and that the strains formerly known as *Yersinia philomiragia* should be classified as a species of *Francisella*. We will adopt the nomenclature found on NCBI’s Taxonomy Browser where *F. philomiragia* and *F. tularensis* are the two commonly recognized species of *Francisella* and that *F. tularensis* is subdivided into four subspecies: *F. tularensis* subsp. *tularensis* (Type A), *F. tularensis* subsp. *holartica* (Type B), *F. tularensis* subsp. *novicida*, and *F. tularensis* subsp. *mediasiatica*. *Francisella tularensis* are very small (0.2-0.5 um by 0.7-1 um) gram negative, facultative intracellular pathogens, the causative agents of tularemia, also known as Deer-fly or Rabbit Fever, and are named after Tulare County, California, where they are prevalent. Though this microorganism does not represent a public health threat in developed countries, *F. tularensis* have recently gained recognition as a CDC Select A Agent and a potential bioweapon due to its potential infectivity and lethality (up to 80%). Inhalation or inoculation of as few as 10 *F. tularensis* organisms can cause disease, which makes this one of the most infectious pathogenic bacteria known. With the majority of *Francisella* research still focusing on typing methods such as fingerprinting, and various PCR techniques, there remains a general lack of understanding that extends itself to the

evolution and the particular mechanism(s) of virulence used by *F. tularensis* to infect its mammalian hosts.

### **Purpose**

The goal of this proposed effort was to assess the difficulty in identifying and characterizing virulence candidate genes in an organism for which very limited data exists. This was accomplished by first addressing the finishing phase of draft-sequenced *F. tularensis* genomes and conducting comparative analyses to determine the coding potential of each genome; to discover the differences in genome structure and content, and to identify potential genes whose products may be involved in the *F. tularensis* virulence process.

### **Activities**

The project was divided into three parts: 1) Genome finishing: This part involves determining the order and orientation of the consensus sequences of contigs obtained from Phrap assemblies of random draft genomic sequences. This tedious process consists of linking contig ends using information embedded in each sequence file that relates the sequence to the original cloned insert. Since inserts are sequenced from both ends, we can establish a link between these paired-ends in different contigs and thus order and orient contigs. Since these genomes carry numerous copies of insertion sequences, these repeated elements “confuse” the Phrap assembly program. It is thus necessary to break these contigs apart at the repeated sequences and individually join the proper flanking regions using paired-end information, or using results of comparisons against a similar genome. Larger repeated elements such as the small subunit ribosomal RNA operon require verification with PCR. Tandem repeats require manual intervention and typically rely on single nucleotide polymorphisms to be resolved. Remaining gaps require PCR reactions and sequencing. Once the genomes have been “closed”, low quality regions are addressed by resequencing reactions. 2) Genome analysis: The final consensus sequences are processed by combining the results of three gene modelers: Glimmer, Critica and Generation. The final gene models are submitted to a battery of homology searches and domain prediction programs in order to annotate them (e.g. BLAST, Pfam, TIGRfam, COG, KEGG, InterPro, TMhmm, SignalP). The genome structure is also assessed in terms of G+C content, GC bias (GC skew), and locations of repeated regions (e.g. IS elements) and phage-like genes. 3) Comparative genomics: The results of the various genome analyses are compared between the finished (or almost finished) genomes. Here, we have compared the *F. tularensis* genomes from the extremely lethal strain Schu4 (subsp. *tularensis*), the vaccine strain LVS (subsp. *holartica*), and strain UT01-4992 of the less virulent, opportunistic subsp. *novicida*. Regions present in the highly virulent strain that are absent from the other less virulent strains may provide insight into what factors are required for the high level of virulence.

## **Technical Outcome and Accomplishments**

*Finishing the genomes of several F. tularensis strains.* Approximately 3000 finishing reactions were performed to close most gaps in the assemblies of *F. tularensis* strains Schu4, LVS and UT01-4992. These consisted of primer-walking experiments to extend sequence from pre-sequenced clones and many PCR reactions to distinguish between possible contig joins and to verify computational assembly. Several repeated regions were identified and these sequences were masked after having identified the proper flanking regions for each repeat. The masking and flanking information was incorporated into all subsequent assemblies. All repeats were resolved for the Schu4 and LVS strains, approximately 90% for the UT01-4992 strain. Several regions recalcitrant to cloning were amplified using PCR and were sequenced and closed. Although 99% of all contigs are ordered and oriented, there remain <10 gaps for each genome assembly.

*Annotation and analysis of the genomes.* Preliminary analyses reveal a genome size for all three strains near 1,800,000 - 1,900,000 base pairs (bp) with 32-33 % G+C content. Though the G+C content was expected, the very small genome size is even smaller than expected and is not characteristic of an opportunistic pathogen capable of surviving in the environment as well as promoting disease in intracellular infections. However, similar to what is found in other genomes, the gene density is approximately 1 gene per 1,000 bp (kb), resulting in approximately 1980 genes. Thirty percent of the predicted proteome cannot be ascribed a function, as these products have only loosely conserved domains (near five percent of total gene models), are similar to hypothetical proteins with no known function (also near five percent) or have no relevant similarity to anything in NCBI's GenBank database (near ten percent). With such a small genome size already, and with many members of the Proteobacteria division (including many members of *Francisella's* gamma subdivision), this large amount of hypothetical proteins is rather surprising. The coding regions account for approximately 90% of the genome, similar to other microbial genomes.

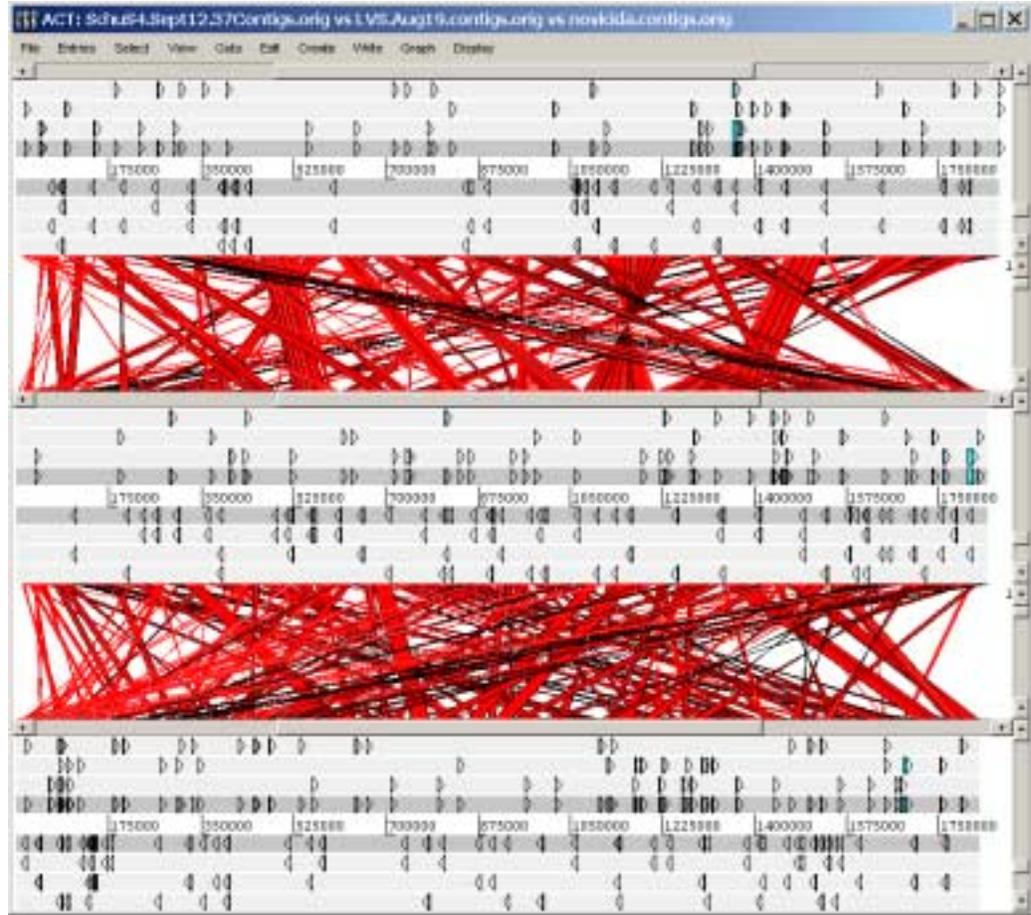
We have discovered several repeated elements in the genome, including two highly repeated IS elements (ISFtu1 and ISFtu2) present in all the *F. tularensis* strains: 53 and 16 respectively in the Schu4 strain, 52 and 39 respectively in the LVS strain, 61 and 39 respectively in the UT01-4992 strain. All three genomes have three 16S-5S-23S rRNA operons. Few cryptic phages or phage remnants were discovered in any of the three genomes.

*Comparative genomics.* Genome-genome comparisons reveal approximately 90% identity between all three pairwise whole-genome alignments. However, this identity does not extend itself to whole genome synteny (genome structure). Presumably as a consequence of the high level of IS elements in the *F. tularensis* genomes, the genomes are only colinear or syntenous for small stretches of 10 - 100 kb (on average). We suggest that the IS elements serve as recombinational hot spots, disrupting colinearity between the *F. tularensis* strains (see Figure 1). This genomic feature lends itself well for classification and phylogenetic analyses. Similar to what has been applied to other sets of bacteria, we can use IS-typing and with the finished genomes of three strains from different subspecies, makes this an even more powerful tool.

*F. tularensis* subsp.  
*tularensis* SchuS4

*F. tularensis* subsp.  
*holarctica* LVS

*F. tularensis* subsp.  
*novicida* UT01-4992



**Figure 1. Whole genome alignments of three *F. tularensis* strains.**

This screen dump is an ACT representation of modified BLASTn alignments between *F. tularensis* SchuS4 vs. *F. tularensis* LVS vs. *F. tularensis* UT01-4992. This graphic is divided into five sections: three sections (gray areas) representing the forward and reverse DNA strands (dark grey horizontal blocks/lines) and the three forward and three reverse reading frames (above and below the DNA representations – lighter grey horizontal blocks/lines); and two sections in between the three genome representations linking homologous regions with red bars. The repeat regions (the IS elements and the three ssRNA operons) are displayed in the reading frames as well as the DNA strands as the triangular arrows. This picture displays both the level of genome rearrangements between these genomes as well as the number and location of the repeated elements.

We have also located 16 regions that are present in the SchuS4 strain, that are not present in the LVS or UT01-4992 genome assemblies. Twelve of these regions are greater than 2 kb, and encode genes with putative gene calls. Among these are predicted type I and III restriction modification systems, helicase, phosphoribosylpyrophosphate, sucrose-6-phosphate hydrolase, and interestingly several genes possibly involved in host interactions (e.g. integral and outer membrane proteins, transport systems). One example region of approximately 10 kb is shown in Figure 2.



**Figure 2. A region unique to *F. tularensis* subsp. *tularensis* strain SchuS4.**

This screen dump is an Artemis representation of a 10 kb region that was found only in the SchuS4 genome and not in the genomes of strains LVS or UT01-4992. This particular region carries a putative type III restriction modification system protein, a putative helicase, an antirestriction protein and an atypical mobilization protein (red) that is interrupted by an IS element (in green), along with several hypothetical proteins with no GenBank hits. This region, similar to several of the other “unique” regions is flanked by two IS elements. Further experiments are required to determine the function of these regions and their possible influence on *F. tularensis* virulence.

**Future Directions**

The results of our analyses and our observations encourage us to investigate the genomes of the fourth subspecies of *F. tularensis*. With all the genomes, we will be in a unique position to be able to construct a *Francisella* chip, one where we would perform comparative genomic hybridizations (CGH) using a collection of strains from geographically and taxonomically distinct locations – a method that is becoming increasingly popular in the microarray field. This would allow us to characterize the diversity among the different populations of *F. tularensis* from anywhere in the world and would complement the comparative analyses we performed by allowing us to focus on the regions that are common to all members of a subgroup that are not present in any of the other strains.

Perhaps more importantly, we plan to ascribe function to the current regions we have identified as unique to strain SchuS4, the more lethal variety of *F. tularensis*. This effort requires the construction of deletion mutants of *F. tularensis* SchuS4, followed by virulence tests in a eukaryotic model (e.g. mouse).

This work was performed under the auspices of the U. S. Department of Energy, contract no. W-7405-Eng-48.

University of California  
Lawrence Livermore National Laboratory  
Technical Information Department  
Livermore, CA 94551

