APR 0 4 2006

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	30.Mar.06	MAJOR REPORT

4. TITLE AND SUBTITLE
PAIRED-END SEQUENCE MAPPING DETECTS EXTENSIVE GENOMIC REARRANGEMENT AND TRANSLOCATION DURING DIVERGENCE OF FRANCISELLA TULARENSIS AND FRANCISELLA TULARENSIS SUBSPECIES

5. FUNDING NUMBERS

6. AUTHOR(S)
MAJ DEMPSEY MICHAEL P

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSITY OF NEBRASKA MEDICAL CENTER

8. PERFORMING ORGANIZATION REPORT NUMBER
CI04-1752

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
THE DEPARTMENT OF THE AIR FORCE
AFIT/CIA, BLDG 125
2950 P STREET
WPAFB OH 45433

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Unlimited distribution
In Accordance With AFI 35-205/AFIT Sup 1

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited

14. SUBJECT TERMS	15. NUMBER OF PAGES
	24
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

Paired-end sequence mapping detects extensive genomic rearrangement and translocation during divergence of *Francisella tularensis* subspecies *tularensis* and *Francisella tularensis* subspecies *holarctica* populations.

Michael Dempsey[1,2], Joseph Nietfeldt[3], Jaques Ravel[4], Steven Hinrichs[1], Robert Crawford[2], and Andrew K. Benson*[3].

[1]Dept. of Pathology and Microbiology, University of Nebraska Medical Center, Omaha, NE, 68198, [2]Division of Microbiology, Armed Forces Institute of Pathology, Washington, D.C., 20306-6000, [3]Department of Food Science and Technology, University of Nebraska, Lincoln, NE 68583-0919, and [4]The Institute for Genome Research, Rockville, MD 20850.

*Corresponding author
Dept. of Food Science and Technology
University of Nebraska
330 Food Industry Complex
Lincoln, NE 68583-0919
Voice (402) 472-5637
Fax (402) 472-1693
Email: abenson1@unl.edu

THE VIEWS EXPRESSED IN THIS ARTICLE ARE THOSE OF THE AUTHOR AND DO NOT REFLECT THE OFFICIAL POLICY OR POSITION OF THE UNITED STATES AIR FORCE, DEPARTMENT OF DEFENSE, OR THE U.S. GOVERNMENT.

Abstract

Comparative genome analyses of the *Francisella tularensis* subspecies *tularensis* and subspecies *holartica* populations show that genome content is highly conserved and only a relatively small number of genes within the *F. tularensis* subsp. *tularensis* genome are absent in other *F. tularensis* subspecies. To catalogue differences in genome organization that could contribute to the unique virulence characteristics and geographic distributions of the *tularensis* and *holartica* subspecies, we have used Paired-End Sequence Mapping (PESM) to identify regions of the genome that are non-contiguous between thse two subspecies. Using PESM, the physical distances between paired-end sequencing reads from a library of a wildtype reference *F. tularensis* subsp. *holartica* strain were compared to the predicted lengths between the reads based on map coordinates of the reads from the subsp. *tularensis* strain Schu S4 and subsp. *holartica* strain LVS genome sequences. A total of 17 different continuous regions were identified in the subsp. *holartica* genome ($CR_{holartica}$) that are non-contiguous in the subsp. *tularensis* genome. At least six of the seventeen different $CR_{holartica}$ are positioned as adjacent pairs in the subspecies *tularensis* genome sequence but are translocated in holartica, implying that arrangements of the $CR_{holartica}$ are ancestral in the *tularensis* subspecies and derived in *holartica*. Using nested PCR assays, the conservation of the events was further assessed by testing 87 additional *tularensis* and *holartica* subspecies isolates. The PCR results showed that the arrangements of the $CR_{holartica}$ are highly conserved, particularly in the *holartica* subspecies, consistent with the hypothesis that holartica populations have recently experienced a periodic selection event or they have emerged from a recent clonal expansion. Two unique *tularensis*-like strains were also observed to share some $CR_{holartica}$ with the *holartica* subspecies and others with the *tularensis* subspecies, implying that these strains may represent a new taxonomic unit.

Deleted: ¶

Introduction

Francisella tularensis is a non-motile, Gram-negative coccobacillus originally isolated from ground squirrels in 1911 during a plague investigation in Tulare County, CA [1]. The geographic distribution of the organism spans the entire Northern Hemisphere, with only a very recent isolated recovery of the organism occurring in the Southern Hemisphere [2, 3]. The organism is a facultative intracellular pathogen and is believed to affect more animal species than any other known zoonotic pathogen [4, 5]. It has been isolated from as many as 250 species of wildlife (reviewed by Oysten, Sjostedt, *et. al*) [6] including various birds, amphibians, fish and many mammalian species. The organism can also be found in invertebrates species, including arthropod vectors such as mosquitoes and ticks (reviewed by Petersen and Schriefer) [7]. Human infection occurs most often through direct exposure to infected animals or by bites from infected arthropod vectors. Recently, terrestrial and aquatic life cycles have been described for *F. tularensis* [8, 9]; and protozoa, such as *Acanthamoeba castellanii*, may also serve as a host for maintenance of *F. tularensis* in the aquatic cycle [10].

The species *F. tularensis* is comprised of four recognized subspecies: Subsps. *tularensis* (Type A), *holarctica* (Type B), *novicida*, and *mediaasiatica*, the two former of which are considered clinically significant in humans [11, 12] and by far have been the most studied. *F. tularensis* subsp. *tularensis* is believed to be more virulent in humans than *F. tularensis* subsp. *holarctica* based on epidemiological data and its higher infectivity in animals. *F. tularensis* subsp. *tularensis* and *F. tularensis* subsp. *holarctica* also show striking geographic differences in their distribution, with both the *tularensis* and *holarctica* subsp. being found in North America but only the *holarctica* subsp. being found in Europe and Asia [7]. Populations of subsp. *mediaasiatica* may be even more geographically limited since, as its name suggests, this

subspecies has only been isolated from the Asian subcontinent. The *novicida* subspecies has been found primarily in the U.S. but was recently detected in Australia [11, 13].

Despite the unique geographic and virulence characteristics, known genetic and phenotypic differences distinguishing the *tularensis* and *holarctica* subpopulations seem to be more limited. Biochemically, the two subspecies have classically been differentiated primarily on the basis of glycerol fermentation, production of citrulline ureidase, and erythromycin resistance [11]. High resolution genotyping methods such as pulsed field gel electrophoresis (PFGE) [14], restriction-fragment length polymorphism (RFLP) [15], Amplified Fragment Length Polymorphisms (AFLP) [14], and Multi-Locus Variable Number Tandem Repeat analysis (MLVA) [4, 16, 17], also distinguish the subspecies genotypically and show that they are divergent, but clonally related.

Given the unique geographical and virulence characteristics, there is tremendous interest in understanding the genetic basis for these characteristics. Recent comparative genome hybridization studies identified limited differences in genome content between the two subspecies, but did include deletion in the *pdpD* region which is associated with virulence [18, 19]. Comparative genome sequencing efforts are also underway and promise to provide detailed information with regard to specific strains. To provide a more complete catalogue of the genomic events which arose early during divergence of the subspecies (true supsbepcies-specific genomic differences as opposed to strain-level differences), we have applied Paired End Sequence Mapping (PESM) to identify candidate regions of genomic difference and further used Comparative Genome PCR (CG-PCR) on a large set of strains to identify regions of genomic difference that are conserved across multiple isolates. PESM was originally developed as a method to identify genomic islands of *Shigella dysenteriae* [20]. The PESM strategy measures

the physical distance between paired-end reads from a clone library, specifically searching for clones whose physical distance is incongruent with the predicted distance based on available genome sequences. In this application, we constructed a library from an *F. tularensis* subsp. *holarctica* strain and compared physical distances with the *F. tularensis* SHU S4 genome sequence. Cloned segments with incongruent lengths compared to the map position were further distinguished as strain-specific versus potentially subspecies-specific by comparison to the *F. tularensis* subsp. *holarctica* strain LVS genome sequence. In instances where the length difference was conserved in the reference strain and the LVS *holarctica* strain, the segments were further tested among a panel of *holarctica* and *tularensis* strains to confirm that the genome difference was broadly conserved across the subspecies. Using this strategy, we identified seventeen regions in the genome that are continuous in 66 of 67 subspecies *holarctica* strains examined, but which are discontinuous in *tularensis* strains. These regions, termed $CR_{holarctica}$, have arisen through massive insertion/deletion, translocation, and rearrangement events and their conservation among *holarctica* strains of distinct temporal and geographic origin implies that this subspecies has likely been through a recent periodic selection event.

Materials & Methods

Bacterial strains and growth conditions. Cultures for this study were propagated on chocolate agar at 37°C in 5% CO_2. Glycerol fermentation of the limited number of isolates was performed either as described [19], or by using Biolog® (Biolog, Inc., Hayward, CA) according to manufacturer's instructions. A summary of spatial, temporal, host, and other pertinent demographic information, as well as prior subspecies determinations of all strains and/or DNA used in this study, is listed in supplemental Table 1 (Table s1).

Subspecies-specific PCR. To confirm the subspecies designation of all the strains in our collection, we tested them using an RD1 PCR assay previously shown to differentiate among all four *F. tularensis* subspecies [18]. All RD1 results are included in Table s1. Primers for RD1 were those same as published by Broekhuijsen *et al.* [18]. All RD1 PCR reactions were performed in 25 µl volumes, each containing 5 mM $MgCl_2$ and 160 µM of each dNTP (Idaho Technology, Salt Lake, UT), 500 nM each of forward and reverse primer (Invitrogen, Carlsbad, CA), and 2.5 Units of Platinum Taq (Invitrogen). Each reaction was conducted on 1.5 µl DNA samples either prepared by using a standard large-scale bacterial genomic DNA extraction protocol [23] or using PUREGENE® DNA isolation kits (Gentra Systems, Inc., Minneapolis, MN). Thermocycling conditions were optimized and performed on a Dyad (MJ Research, Reno, NV) thermocycler according to the following cycling parameters: Initial hold at 95°C for 2 min, 30 sec; 30 cycles of 95°C for 30 sec, 64°C for 1 min, and 72°C for 1 min; final extension at 72°C; and a final indefinite hold at 4°C.

Construction of λ phage library. *Francisella tularensis* subsp. *holarctica* strain MS304 is a human isolate obtained in 2002 from the State of Missouri. A library was constructed from MS304 genomic DNA by partial digestion with *Sau*3A1. After optimization of the partial digestion for 10-15 Kb

fragments, 7 ug of genomic DNA was digested with *Sau*3A1 at in separate reactions with 0.0625, 0.0312, and 0.0156 units per microgram for 1 hour at 37°C. The partially cut DNA was electrophoresed on a 0.7% agarose gel along with molecular weight markers and the regions containing 10-15 kb fragments were excised from the gel. The fragments were electroluted, pooled, and then precipitated to concentrate. Size distribution of the gel-purified fragments was confirmed by agarose gel electrophoresis of a small portion of the fragments alongside molecular weight standards. The remaining purified fragments were then ligated into Lambda DASH II *Bam* HI (Stratagene, La Jolla, CA). The ligations were packaged using Stratagene's Gigapack III Gold Extract according to the manufacturer's recommendations. The packaged phage were titered on XL1-Blue MRA P2 host bacteria (Stratagene). The titer from the packaging was approximately 5 X 10^6 PFU/ml. Library diversity was confirmed by restriction digestion and DNA sequence analysis of inserts from 10 independent plaques. The library was then amplified once using XL1-Blue MRA P2 host bacteria and DMSO was added to 7% final concentration in the clarified supernatant. The amplified library had a titer of 2.5 x 10^7 pfu/ml. One ml aliquots were stored at –80°C.

Formatted: Font: Italic

Direct-PCR Amplification (DPA) of Cloned Fragments. For high throughput PCR amplification of individual plaques, the library was diluted and plated onto mid-logarithmic phase P-2 cells (OD_{600} ~0.6) grown in NZYM Broth with 0.2% maltose. Dilutions of the library stock were made in Lambda suspension medium (SM) buffer and dilutions yielding approximately 120-150 plaques per plate were subsequently plated onto several 80 cm Petri dishes using P2 host cells. After solidifying, the plates were inverted in the incubator overnight at 37°C.

To amplify plaques for length measurement and paired-end sequencing, plaques were chosen from dilutions giving 120-150 well isolated plaques. Plates with the appropriate number of plaques were first wrapped with parafilm and inverted at 4°C for a minimum of 4 hours and a maximum of 4 days. Clearly isolated plaques were further processed by direct-PCR amplification (DPA). DPA was performed in 96-well PCR plates (Applied Biosystems, Foster City, CA) pre-loaded with PCR master mix. Plaques

for DPA were loaded by gouging each candidate plaque with a sterile 20 ul pipette tip (such that the tip contained visible plaque material) and then mixing the tip in a single well of the PCR plate. Long-range PCR was then performed using a *TaKaRa Ex Taq*™ Hot Start master mix kit (TaKaRa Mirus Bio, Madison, WI) containing T3 (5′-AATTAACCCTCACTAAAGGG-3′) and T7 (5'-TAATACGACTCACTATAGGG-3') primers, which prime amplification from the T3 and T7 promoter regions present in the arms of the lambda cloning vector. PCR reaction mixes were prepared according to the manufacturer's recipe with each single 25 ul reaction containing 500 nM of each primer. Thermocycling conditions were optimized and performed on a Dyad (MJ Research, Reno, NV) thermocycler according to the following cycling parameters: Initial hold at 95°C for 2 mins, 30 sec; 36 cycles of 95°C for 50 sec, 55°C for 50 sec, and 72°C for 15 mins; final extension at 72°C for 5 mins; and a final indefinite hold at 4°C.

Following each PCR run, clone-amplicon purification was accomplished using Montage® $PCR_{\mu 96}$ Plates (Millipore, Billerica, MA) which were processed on a SAVM 384 Vacuum Manifold (Millipore) according to the manufacturer's instructions. The final elution was performed using 30 ul of Invitrogen Distilled DNase-/RNase-free water and plates were sealed with adhesive sealing lids (Bio-Rad, Hercules, CA) and stored at 4°C.

Clone-Amplicon Sizing Experiments. Amplicon-size determinations were performed by agarose gel electrophoresis. 15 cm x 25 cm 0.65% agarose gels containing 51-wells were made in 1x Tris-acetate-EDTA (TAE). Each gel accommodated 47 clone-amplification reactions along with 4 separate lanes of 1-15 kb Molecular Ruler (Bio-Rad) for length standardization. Each gel was electrophoresed at 85 V for approximately 4.5 hours. Ethidium bromide-stained gels were imaged using a Syngene GeneGenius (Synoptics, Frederick, MD) imaging system, and the image was analyzed using the GeneTools (Synoptics) software package. The band sizes for all clones and DNA standards were exported from GeneTools into a Microsoft Excel spreadsheet and loaded into the PESM pipeline.

Deleted: G

Paired-End Sequencing. DNA sequencing reactions for all clones were carried out using BigDye® Terminator v3.1 Cycle Sequencing Kits (Applied Biosystems) with pGEM DNA serving as the sequence reaction controls. Each sequence reaction experiment was setup in 96-well reaction plates, with each sample being divided into two reactions – one reaction with T3 primer and the other with T7 primer (both primers the same as from the initial PCR reactions). Sequence reactions were carried out on the Dyad (MJ Research) thermocycler, and the sequence reaction plate was cleaned-up using a Montage SEQ_{96} Kit on the SAVM 384 Vacuum Manifold (Millipore). The labeled DNA was then transferred into new reaction plates and loaded onto an ABI 3100 automated capillary-electrophoresis (CE) sequencer (Applied Biosystems), which was configured with an 80 cm capillary array and loaded with Performance Optimized Polymer (POP)-4 (Applied Biosystems). Control sequencing reactions were performed on the two pGEM vector reactions in each set of 94 sample reactions, and the pGEM control sequences were evaluated to ensure the sequence reaction and sequence run were successful. The .abi sequence files were trimmed in Sequencher V. 4.0.5 (Gene Codes, Ann Arbor, MI), merged, and output as a single FASTA file for further analysis.

Fragment Length and Paired-End Sequence Pipeline. The sizing and sequence files were input into a Perl-based program, referred to as the Paired-End Sequence Mapping Pipeline (PESMP), to identify coordinates of the paired end reads from the draft *F. tularensis* live vaccine strain (LVS) whole genome sequence (ftp://bbrp.llnl.gov/pub/cbnp/F-tularensis/F.tularensis.html) [21] and the completed SCHU S4 genome sequence [22] and then to compare the predicted lengths to the physical length of the cloned segment. The PESMP input files consisted of a composite FASTA file from the trimmed sequences resulting from a single 96-well plate along with a corresponding Microsoft Excel file containing the Long-PCR amplicon sizing data for each clone. The PESMP algorithm then outputs the coordinates and predicted length of the fragment corresponding to the paired end reads from the two genome sequences along with the physical measurement of the long PCR amplicon.

Deleted: (WGS)

To identify clones with discrepancies between physical length between paired-end reads and predicted length, the output from the PESM pipeline was loaded into an Excel spreadsheet and sorted according to the paired-end coordinates from the LVS genome sequence. Clones in which the physical size showed a > 2 Kb discrepancy from the predicted size of the SCHU S4 genome but not the LVS genome were chosen for further characterization. Sequence of these putative regions of genomic difference was obtained by electronic PCR, using the coordinates of the paired-end reads to extract the intervening sequence between the reads from the LVS genome sequence. These electronic clone sequences were then aligned by contig analysis using Sequencher to further delimit the physical boundaries of the $CR_{holarctica}$.

Fine-structure genome mapping. To characterize the extent and nature of the events associated with the $CR_{holarctica}$, each $CR_{holarctica}$ was used for BLAST analysis against the SCHU S4 genome. Query coordinates for all segments/subsegments (shown in Table s2), except for multiple repeats of IS elements were then used to map the location of the events in the SCHU S4 genome corresponding to the $CR_{holarctica}$. $CR_{holarctica}$ maps were assembled using the SeqBuilder module of Lazergene V.6.0 (DNASTAR, Madison, WI) to identify corresponding gene/pseudogene content of the $CR_{holarctica}$ from the SCHU S4 annotation. All genes from SCHU S4 annotation found to be truncated at the flanks due to the beginning or ending of the $CR_{holarctica}$ clones, as well as those found internally due to rearrangements within the $CR_{holarctica}$, were used in BLAST analyses against the LVS sequence for homology comparisons. TIGR in-house Perl scripts were run on a Linux platform to generate graphical representations of all combined $CR_{holarctica}$ mapped on SCHU S4 shown in Fig. 3, and each of 6 individual $CR_{holarctica}$ mapped on SCHU S4 shown in Fig. 5 and supplemental sFig. 1 through sFig. 5. The final figures as shown were assembled in Adobe Illustrator 10.

Deleted: These values were then

Deleted: of

Deleted: ¶

Deleted: To obtain sequence data from the cloned segments, paired end reads from clones with physical lengths corresponding to length predicted from the LVS sequence but not the SCHU S4 sequence were next used to

Deleted: corresponding

Deleted: of the clone

Deleted: "

Deleted: "

Deleted: genome

Deleted: detected as

Formatted: No underline

Deleted: BLASTed

Deleted: **Subspecies-specific PCR**. To confirm the subspecies designation of all the strains in our collection, we tested them using an RD1 PCR assay previously shown to differentiate among all four *F. tularensis* subspecies [18]. All RD1 results are included in Table s1. Primers for RD1 were those same as published by Broekhuijsen *et al.* [18]. All RD1 PCR reactions were performed in 25 µl volumes, each containing 5 mM $MgCl_2$ and 160 µM of each dNTP (Idaho Technology, Salt Lake, UT), 500 nM each of forward and reverse primer (Invitrogen, Carlsbad, CA), and 2.5 Units of Platinum Taq (Invitrogen). Each reaction was conducted on 1.5 µl DNA samples either prepared by using a standard large-scale bacterial genomic DNA extraction protocol [23] or using PUREGENE® DNA isolation kits (Gentra Systems, Inc., Minneapolis, MN). Thermocycling conditions were optimized and performed on a Dyad (MJ Research, Reno, NV) thermocycler according to the following cycling parameters: Initial hold at 95°C for 2 min, 30 sec; 30 cycles of 95°C for 30 sec, 64°C for 1 min, and 72°C for 1 min; final extension at 72°C; and a final indefinite hold at 4°C. ¶

Comparative Genome PCR (CG-PCR). Confirmation of the $CR_{holarctica}$ was conducted by PCR analysis on multiple isolates of the *holarctica* and *tularensis* subspecies. For each $CR_{holarctica}$, a three-primer nested-PCR assay was designed based on the relative coordinates of the corresponding junctions of the $CR_{holarctica}$ maps. Primers for all assays were designed using Primer3 software (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi). The assays were designed by first identifying a primer common to both LVS and SCHU S4, either forward or reverse (designated C-F or C-R), immediately adjacent to a breakpoint in the SCHU S4 genome sequence where synteny of the adjacent segment changed or was translocated leaving a SCHU S4-specific region for a SCHU S4-specific (S) primer, and which likewise left a target for an LVS-specific (L) primer in the adjacent-contiguous LVS sequence. The design of all assays was such to produce "A"-type bands across all 17 CR for SCHU S4, and "B"-type bands across all 17 CR for LVS. Since either intact or truncated IS elements or their corresponding repeated elements were present near all breakpoints, care was taken to avoid placement of primers within IS elements (see Table s3 for all primer coordinates, sequences, and expected amplicon sizes). All PCR assays were conventional by design and conducted on 1.5 µl of the DNA samples used for RD1 PCR. All CG-PCR reactions were performed in 25 µl volumes, each containing 5 mM $MgCl_2$ and 160 µM of each dNTP (Idaho Technology, Salt Lake, UT), 500 nM each of common primer, LVS-specific primer, and SCHU S4-specific primer (Invitrogen, Carlsbad, CA), and 2.5 Units of Platinum Taq (Invitrogen). Thermocycling conditions were optimized and performed on a Dyad (MJ Research, Reno, NV) thermocycler according to the following cycling parameters: Initial hold at 95°C for 2 min, 30 sec; 32 cycles of 95°C for 30 sec, 60°C for 1 min, and 72°C for 1 min; final extension at 72°C; and a final indefinite hold at 4°C. Each assay was first tested against SCHU S4 and LVS, and then against a panel of 91 additional global *Francisella* strains representing both subspecies as well as subsp. *novicida* and a unique *holarctica* strain from Japan [13, 15, 24], tentatively called subsp. *japonica* [4], but for our studies referred to as subsp. *holarctica*-japan, or "*holarc*-jap" (see Table s1 for panel composition). The amplicons were visualized on 0.8% - 1% agarose gels run in 1X TAE at 85 V for

Formatted: Font: Italic

Formatted: Font: Italic

Deleted: at

Deleted: into sequence encoding them, and BLAST searches were performed for all primers to limit their placement to their intended location

Deleted: intended

approximately 2 hours, or until adequate size-discrimination was accomplished. A 100-bp PCR Molecular Ruler ranging from 100 bp to 3 kb (Bio-Rad) was used for size determinations.

Results

Deleted: ¶ ¶ ¶ ¶ ¶ ¶

Paired-End Sequencing. A total of 752 plaques were picked and subjected to DPA with 551 of the DPA yielding amplicons >8 Kb in length that were of sufficient quality and quantity for size determination and DNA sequence analysis (DPA success rate of 73.3%). The mean amplicon/insert size from the 551 successful DPA reactions was 14,174 bp, which corresponds to approximately 7.8 Mb of coverage, or an estimated 4.1 X coverage of the 1.89 Mb *F. tularensis* subsp. *tularensis* genome [22].

Mapping of Paired-End Sequence Reads. Of the 551 clones with quality paired-end reads, 66 clones had physical lengths that were not congruent with distance between the paired-end reads relative to the SCHU S4 genome, but were congruent with distances predicted from the LVS genome. These clones were further considered as candidates for subspecies-specific genomic events. Alignment of the sequences from the paired-end reads of candidate clones grouped the 66 cloned segments into 17 different contiguous regions ($CR_{holarctica}$) which align with the *F. tularensis* subsp *holarctica* strain LVS genome but are non-contiguous or otherwise altered in the *F. tularensis* subsp. *tularensis* SCHU S4 genome sequence. Plotting of the number of $CR_{holarctica}$ identified versus the total number of clones sequenced (Fig. 1), showed that the number of new $CR_{holarctica}$ began to decrease sharply after 250 clones were sequenced. Of the last 301 clones sequenced, only three new $CR_{holarctica}$ were identified, suggesting that the library was nearly saturated.

Comparative Genome PCR (CG-PCR) confirmation of $CR_{holarctica}$. Conservation of the $CR_{holarctica}$ in the MS304 reference strain—which was isolated in 2002, and is temporally and geographically distinct

from the LVS strain isolated in 1941—leads to the simple hypothesis that these CR likely arose early during divergence of the *tularensis* and *holarctica* subspecies and therefore should be conserved across most *holarctica* strains. To confirm this, CG-PCR assays were developed for each $CR_{holarctica}$ using nested primer sets at the junctions of the CR (Table s3). The different CG-PCR reactions for all 17 $CR_{holarctica}$ were then run on a panel of DNA samples from SCHU S4 and LVS, and also from 19 different subsp. *tularensis* strains, 67 subsp. *holarctica* strains, 3 subsp. *novicida* strains, and a single strain each of subsp. *holarctica*-japan and *F. philomiragia*. The results for all 17 CG-PCR panels are shown in Fig. 2. The colors correspond to different size amplicons produced from each CG-PCR. Overall, the subsp. *holarctica* strains produced homogenous results across all 17 CR, with 66 of the strains (98.5%) producing the expected amplicon based on the LVS genome sequence. The only deviation occurred in *F. tularensis* subsp. *holarctica* strain Tu-42, which produced a subsp. *tularensis* A-type band. Thus, excluding this one exception, the $CR_{holarctica}$ identified through the PESM pipeline are indeed highly conserved.

Unlike the 67 subsp. *holarctica* strains, the 19 subsp. *tularensis* strains displayed significantly more heterogeneity in the CG-PCR assays. At least four different subgroups of the *tularensis* subspecies can be resolved. All share the RD1 region in common, implying that they are taxonomically true subsp. *tularensis* derivatives. SCHU S4 and nine other subsp. *tularensis* strains comprise one subgroup and have identical genome organization, producing A-type bands (red squares), across all 17 CR. A second subgroup is represented by A88R160, 88R52, 88R144, AK-1133496, AK-1100558, AK-1100559, with each of these strains sharing an amplicon from the CR10 nested PCR reaction that was unique in size (denoted by orange squares). All six of these strains were isolated from rabbits or hares, with the three AK strains derived from Alaska in 2003 and 2004 and the three others isolated from the contiguous United States. The ATCC 6223 strain was likewise shown to be identical the previous six strains, but it was originally isolated from a human patient in 1920 and has since lost its virulence [4]. A third subgroup is represented by the single strain OK-98041035 which matches the SCHU S4 subgroup except that it failed to produce a CR14 PCR amplicon (denoted by a yellow square). The fourth subgroup

Formatted: Font: Italic
Deleted: .
Deleted: ,

comprises a very unique set of two isolates, strains WY-00W4114 and WY-WSVL02. These strains both produced an A-type band at CR3, CR4, CR5, CR6, CR8, CR11, CR15, and CR17 (red squares), were negative at CR1 and CR13 (yellow squares), and produced B-type bands at CR7, CR9, CR12, and CR14 (green squares). Unlike any other strains, they also produced unique bands at CR16 (blue-grey) and CR10 (blue-grey, different in size from orange). These two strains were also distinguishable from one another in that WY-00W4114 produced a B-type band at CR2 (green) whereas WY-WSVL02 was negative (yellow). These two strains were also only slightly capable of fermenting glycerol [19]. Collectively, the genetic and biochemical data strongly suggest these two strains represent a new taxanomic unit. If indeed this is a new taxon, then the population is likely to be virulent since one of the isolates was obtained from a human clinical sample [19].

As would be expected, the *F. tularensis*, subsp. *novicida*, subsp. *holarctica*-japan, and *F. philomiragia* strains showed heterogenous CG-PCR results. *F. philomiragia* was negative (yellow) across each of the 17 CR. The three subsp. *novicida* strains were negative (yellow) for CR1, CR4, CR14, and CR16; and they produced a unique size amplicon from (blue-grey) CR3, CR5, CR7, CR8, CR9, CR10, CR12, CR13, and CR17. All three strains produced a "B"- type allele (green) across CR6 and CR15 and they all produced an "A"-type allele (red) across CR11. CR2 differentiated between the Tu-43 strain, which produced a unique amplicon while the other two *novicida* strains (from ATCC and USAMRIID) which were both negative. Consistent with its classification as a separate subspecies [4], the single subsp. *holarctica*-japan strain was also distinct from all other strains in this study; it produced a "B"-type allele (green) across CR1, CR2, CR3, CR5, CR6, CR7, CR9, CR10, CR11, CR12, CR15, CR16, a unique allele (orange) across CR13, an "A"-type allele (red) across CR17, and was negative (yellow) across CR4, CR8, and CR14.

Fine Structure Analysis of $CR_{holarctica}$. Fine-structure mapping and annotation of the $CR_{holarctica}$ was next conducted by alignment of the $CR_{holarctica}$ contigs from strain LVS genome sequence with the SCHU S4 genome sequence. The corresponding locations of the aligned regions are shown in Fig. 3. The

combined DNA represented by the 17 CRs corresponds to nearly 230 genes/pseudogenes and over 30 IS-elements, mainly a combination of ISftu1 and ISftu2 elements. Most of the of the rearrangements and translocations are juxtaposed to IS elements, suggesting that many of the events were likely mediated by these elements, resulting in remarkably large changes in the location of specific genome segments between SCHU S4 and LVS, but with little effect on the corresponding content of the transposed/translocated regions. Indeed nearly all of the genes within the $CR_{holarctica}$ are present in both the subsp. *holarctica* LVS and subsp. *tularensis* SCHU S4 genome.

The distribution of the CR1-CR17 segments around the LVS genome and the relative positions of the corresponding regions in the SCHU S4 genome have some remarkable characteristics. First, the CR in the LVS genome show some positional bias, with thirteen of the seventeen $CR_{holarctica}$ being present in roughly one-half of the genome (the region between 8 o'clock and 2 o'clock extending from 1.3 Mb to 0.3 Mb). Secondly, there are three notable instances where segments from different $CR_{holarctica}$ in LVS are positioned adjacent to one another in the SCHU S4 genome. Specifically, segments from CR1 and CR16 are adjacent in the SCHU S4 genome as are segments from CR13 and CR15, and CR4 and CR10, and all three events are illustrated in sFig. 4. The juxtapositioning of these segments in subsp. *tularensis* suggests that their organization in the *tularensis* subspecies was the ancestral state while their organization in *holarctica* is a derived state. As shown in sFig. 4c, this notion is further supported by the finding that both CR13 and CR15 contain genes that are involved in glycerol fermentation and it seems likely that their ancestral condition would have been functionally clustered.

Genes affected by rearrangements. Further bioinformatics analysis of the junction of the juxtaposed CR4-CR10 segment revealed the complete deletion of a gene of unknown function (FTT1308c) from subsp. *holartica*. In addition nine genes were found which are disrupted as a consequence of the rearrangements or translocation of genome segments in the CR. The intact versions of these genes in the SCHU S4 genome encode proteins with significant similarity to oligopeptide transporters (*oppD* and *oppF*), a ribosome modification gene (*rimK*), an acetyltransferase (FTT0177c), and genes of unknown

Deleted: With regard to content of
Deleted: CR, it should be noted that nearly all of the genes within the
Deleted: ,
Deleted: indeed
Deleted: , albeit at unique positions.
Deleted: juxtaposed
Formatted: Font: Italic
Formatted: No underline
Formatted: Font: Italic
Deleted: LVS as compared with its intact presence in the SCHU S4 genome.
Formatted: No underline
Deleted: ,
Deleted: our analysis has shown that there are

function (corresponding to FTT0898c, FTT1122, FTT0921, and FTT1311). SFig. 4c illustrates the region of CR13 in SCHU S4 containing the intact *oppD* and *oppF* genes which are truncated in the LVS genome. The *aceF* gene, encoding the E2 of pyruvate dehydrogenase lies near the junction of CR3 (along with *aceE* and *lpd*) and carries a 300 base in-frame deletion. A fine-structure genome comparison of the entire $CR_{holarctica}3$ mapped onto SCHU S4 is shown in (Fig. __ or sFig. __). The deletion corresponds to loss of a repeated biotin-binding repeat region, leaving *holarctica* strains with two biotin-binding domains while the *tularensis* strains contain three. Whether the deletion occurred during the translocation event and whether it affects function of the pyruvate dehydrogenase complex is not clear. The AceF orthologues from several pathogenic species, including *Vibrio*, *Yersinia*, *Shigella* and *E. coli*, do carry three domains. It has, however, been shown that deletion of two of the three domains of AceF in *E. coli* has little affect on function [25]. Whether the additional binding domain influences efficiency of the reaction and whether it could contribute to virulence will require further experimentation. It is worth noting that the *aceF* truncation was also identified by comparative genome hybridization studies [12, 18, 19], but the translocation event corresponding to CR3 was not detected, underscoring the importance of using multiple approaches for comparative genome analyses.

In addition to direct disruption as a consequence of translocation, genes near the junctions of the translocation events could also be subject to control by unique regulatory machinery. In this light, it is interesting to note that some of the genes within the CR and near the junctions could have functions related to physiology and virulence of *F. tularensis*. Two different genes encoding pilin subunits (*pilE* homologues) of a type IV pilus, are present within the CR2 and CR10, and fine-structure genome comparisons of these $CR_{holarctica}$ mapped onto SCHU S4 are shown in sFig. __ and sFig. __, respectively. The gene encoding the pilin subunit *pilE5* (FTT0230c) [26] is embedded within CR2. Another member of the *pilE* family is also present in CR10 and the region upstream is disrupted by IS elements. These IS elements are also associated with disruption and duplication of the FTT1311gene in LVS as compared to a single, intact copy in SCHU S4.

Formatted: No underline

Deleted: al

Deleted: al

Deleted: , which also appear to have

Formatted: No underline

Deleted: ed and duplicated

Because glycerol fermentation has classically served as a biochemical marker distinguishing the *tularensis* and *holartica* subspecies, we also scanned the CR for genes associated with glycerol metabolism. Several genes encoding enzymes associated with glycerol fermentation were found within CR11, CR13, and CR15. Fine-structure genome comparisons of each of these $CR_{holarctica}$ mapped onto SCHU S4 are shown in (Fig. __ or sFig. __ and/or sFig. __). There is no difference in the content of genes within CR11, CR13, and CR15 between the two subspecies, however it is possible that their unique organization contributes to the different glycerol fermentation phenotypes.

Formatted: Font: Italic

Formatted: Font: Italic

Deleted: Metabolic genes besides *aceF* were also found within the CR; and in particular,

Deleted: s

Deleted: are

Deleted: the

Deleted: al

Deleted: The inability to ferment glycerol is a hallmark of the *holarctica* subspecies, so it will be interesting it will be interesting to determine if altered expression context could be a simple explanation.

Discussion:

Whole genome sequencing has provided an outstanding resource for comparative genome studies, allowing high-resolution snapshots of the genetic diversity found within a given species. One of the drawbacks of comparative genome sequencing, however, is that only limited numbers of strains or taxa can reasonably be compared, making it difficult to distinguish between strain-level genomic differences and true lineage or population-specific sets of genes. Comparative genome hybridization using DNA microarrays circumvents this problem to some degree by providing a single platform for comparison of multiple strains or taxa. On the other hand, the array approach is limited to assessing the diversity in genetic content that is represented on the array. Here, we have shown that PESM can help circumvent the strain bias and the representation problems associated with whole genome sequencing and DNA microarrays. PESM was originally developed as a means to identify unique genomic islands [20]. In our application, we scaled PESM for comparative genome studies. Given at least one reference genome sequence, PESM provides an economical means to identify candidate regions of genomic difference, and these regions can be further examined in larger strain sets by nested CG-PCR. The PESM library used in our study carried modest sized fragments of the genome (averaging 14 Kb), such that coverage could be obtained with a reasonable amount of sequencing without severely limiting the ability to measure physical size. PESM libraries, however, can be made using different insert sizes in different types of vectors, such

Deleted: To overcome the limitations of whole genome sequencing and comparative genome hybridization,

Deleted: have combined the resolution of comparative sequencing by PESM with the power of multiple strain comparison.

that coverage per clone can be increased with larger segments while resolution can be increased by sequencing a larger number of segments from small-fragment libraries.

In addition to economy, the PESM approach allows any strain to be used as a source of the library, thereby allowing the user to choose the best taxonomic unit as a reference. This is particularly important when multiple subpopulations of a species may display unique characteristics that are of interest. Indeed, although we have used PESM in a binary comparison (*holarctica* compared to *tularensis*), it is possible to scaffold multiple libraries into the same PESM pipeline using only a single reference genome upon which to scaffold the data.

Genome diversity in *F. tularensis*. Populations of highly virulent bacteria display a wide spectrum with respect to genetic diversity. On the one hand, populations of species such as *Bacillus anthracis* display very little diversity and only a limited number of clones appear to be spread worldwide [27]. These clones can only be differentiated by examining variation in tandem repeats, which are some of the most rapidly evolving loci in the genome [28]. With the availability of multiple genome sequences, single nucleotide polymorphisms will soon complement or displace the MLVA-based approaches. At the other end of the spectrum are subpopulations of *E. coli* O157:H7 which, despite the presence of highly clonal signatures in their genomic backbone, display substantial genomic diversity, even being detectable by a relatively low-resolution method as Pulsed Field Gel Electrophoresis [29-31].

Based on the data described in our study, we believe that *Francisella tularensis* may represent an intriguing model of genome evolution. Previous studies of genetic diversity in *F. tularensis* detected only limited diversity [11, 32]. The four *F. tularensis* subspecies are known to share 98% identity in their 16S rRNA, show very similar biochemical profiles, and have quite similar antigenic compositions [11, 33]. Only very high resolution methods can provide any phylogenetic signal that reasonably correlates with biochemical and virulence characteristics.

Deleted: represents an

Despite the apparently limited degree of genetic diversity, the *F. tularensis* subspecies display quite distinct geographic distribution and virulence characteristics. Thus, it was initially believed that

although limited, the diversity in genomic content would parallel phylogeographic and epidemiologic characteristics and provide clues to the genetic basis for these traits. With the exception of differences in numbers of *pilE*-like loci [26] and the loss of the *pdpD* locus [21], no other obvious candidate virulence genes [6, 22] emerged from comparative genome hybridization studies [12, 18, 19], and >99% of the genetic material present in the more virulent subspecies *tularensis* can also be found in the *holarctica* genome.

In the present study, we now show that despite the high degree of genetic conservation, the genome organization of the *tularensis* and *holarctica* subspecies is vastly different. At least 17 substantial genomic events have occurred during divergence of these two subspecies and have been preserved among multiple strains of each population. The events correspond to extensive translocations and rearrangements, many if not all of which were mediated by movement of IS elements. As shown in Fig. 3, the IS elements are plentiful in the SCHU S4 genome, with 50 different copies of ISftu1 and 16 copies of ISftu2 being distributed around the genome [22]. Given the large number of these elements, it is therefore not surprising to find them at or near the junctions of all 17 CR. Certainly, IS elements were also found abutting subspecies-specific regions of genomic difference (RD) observed in comparative genome hybridization studies [18, 19], and our data here further confirm that IS elements are the primary means through which this genome diversifies.

While the degree of diversity in organization between the genomes of subsp. *tularensis* and subsp. *holarctica* is remarkable, perhaps equally remarkable is the degree to which the unique structure is preserved across temporally and spatially distinct taxa of *holarctica* strains. This observation leads to several interesting possible hypotheses. First, it is possible that population growth is very minimal such that little diversity has had time to accrue. However, because *Francisella* is free-living and is also capable of infecting many different mammalian hosts, slow population turnover in the environment would seem to be an unlikely explanation. A second explanation is that IS elements move only at a very low frequency, thus generating diversity only on a very slow timescale. In this instance, the divergence would have been quite ancestral given the degree of diversity that has accrued. With the number of ISftu1 and

ISftu2 elements in the genome, this explanation is unsatisfying. Moreover, we detected significant diversity among the $CR_{holarctica}$ regions within the subsp. *tularensis* strains, suggesting that the ISftu are indeed functional. Lastly, and more likely, it is also possible that the extant populations of *holarctica* are quite homogenous because they share very recent common ancestry. This hypothesis would imply that the populations have recently been through periodic selection or they arose from recent emergence, expansion, and geographic spread of a successful clone. If true, the clonal expansion hypothesis would also beg the question of why the recently emerged *holartica* population can be found in Eurasia whereas the *tularensis* populations seem to be confined to North America.

The *holarctica* subspecies is likely a derived state. Given the high degree of virulence that is displayed by the *tularensis* subspecies, it has been speculated that it represents the ancestral state while the less virulent subspecies are derived states [12] that are more adept at infecting hosts without killing. In support of this hypothesis, the *novicida* subsp. can be found in water, implying that it may survive more effectively in the free living state than the more highly virulent *tularensis* subspecies. Evolutionary analysis of VNTR loci also suggest that *tularensis* is likely more similar to the common ancestor [4, 16, 17]. With respect to genome organization, our data also support this hypothesis, showing that organization of different $CR_{holartica}$ appear to be a derived state, arising by dissociation of genomic units through translocation events occurring in an immediate ancestor of the *holartica* populations. At least three genomic segments were found to be single contigs in the *tularensis* genome but are dispersed into six different CR in the *holarctica* subsp. Moreover, some genes at the junctions of these events are disrupted or even deleted in *holarctica* whereas the respective genes are present with no remnants of gene fragments being present at the junctions of *tularensis*. Furthermore, the translocation events in at least two different CR have separated functionally associated genes that are putatively involved with glycerol fermentation in *holarctica*, again their scattering being consistent with a derived condition. We also note that three additional CR in tularensis (CR3-CR9, CR4-CR8, and CR9-CR11) are adjacent, but not contiguous whereas they are highly dispersed in the *holarctica* subspecies. Therefore, several lines of

Deleted: ly
Deleted: more
Deleted:
Formatted: Font: Italic
Formatted: Font: Italic
Deleted:
Deleted: In the midst of limited genetic diversity, the simplest explanation for the observed population structure of the *F. tularensis* subspecies is that they are essentially clonal populations and shared a common ancestor. Given the
Deleted: one would speculate that
Deleted: e now show that w
Formatted: Subscript
Deleted: *holarctica*
Deleted: s
Formatted: Font: Italic
Formatted: Font: Italic
Formatted: No underline
Formatted: No underline
Deleted: the disruption of the apparent glycerol fermentation operon through translocation
Formatted: No underline
Formatted: No underline
Deleted: is likely the
Formatted: No underline

evidence are beginning to mount in favor of the hypothesis that *F. tularensis* subsp. *tularensis* is likely more similar to the ancestral state while populations of the *holartica* subsp. are derived states. If this is true, then analysis of genomic content and organization between the different subspecies should provide insights not only into virulence characteristics, but also into selective pressures that have led to emergence and geographical spread of the *holartica* populations.

Deleted: evidence is beginning to mount

Deleted: the other

Formatted: Font: Italic

Formatted: Font: Italic

Deleted: additional candidate loci that are intact and required for virulence in subsp. *tularensis* but which are altered in the other subspecies.

Two US strains may represent a unique taxanomic unit within Francisella. Although our search was primarily focused on identifying population-specific regions of genomic difference, the genome organization observed in the subsp. *tularensis* strains WY00W4114 and WY-WSLVL02 is very intriguing. Their pattern of genome organization is clearly distinct from the *tularensis* and *holarctica* populations, sharing CR "alleles" at some loci with *tularensis* strains, CR "alleles" at other loci with *holarctica* strains, and unique alleles at still other loci. The fact that some CR alleles are shared with both subspecies suggests three different possibilities. First, the strains could have descended from a recombination event in which DNA was acquired from one subspecies and recombined into another. Given that the distribution of the CRholartica "alleles" present in these strains is highly dispersed, multiple such recombination events would have to be invoked to account for the genomic organization. A second explanation is that the CRholartica in these strains are homoplasious, arising through independent events within a separate lineage of descent from the common ancestor. This explanation is also unsatisfying in that multiple such events must be invoked, each of which occurred similarly in at least two different lineages. Lastly, it is possible that these strains are derived from an evolutionary intermediate, and thus only some of the $CR_{holartica}$ have occurred and are preserved. This explanation invokes the fewest number of events and requires that the events only occur in a single lineage. Thus, we favor the evolutionary intermediate explanation. Regardless of the explanation for the events, the unique combination of CR in these strains implies that populations represented by these strain should have distinct taxonomic status. It will be interesting to conduct further analyses on additional isolates to determine if the evolutionary intermediate model holds up.

Formatted: Subscript

Deleted: The diversity

Deleted: is such that we propose they represent a new taxonomic unit ([illegible] you wanna name it? Andy, is appropriate to this paper, beg steal *F. tularensis* subsp. *novicida/tularensis*? due to the observations from DO PCR and the ambiguousness of the glycerol fermentation (+/-) and PFGE, both factors in your last article which you concluded were closest to subsp. *tularensis*, though not with overwhelming conclusivity, the RD1 results are subsp. *mid-grade*, however... but *novicida/tularensis* would fit the virulence observation better). ¶

Acknowledgements. We thank John McGraw and Mark Chrustowski for coordination of the AFIP Genomics Laboratory and assistance with DNA sequence analysis. We also thank Steve Francesconi, Bob Burgess, and Wendall Thomas for cultivation of strains and preparation of DNAs for the *Francesella* panel. This work was supported by R-21 AI1234567 to A.K.B. from the National Institutes of Health.

Deleted: (Probably)

Deleted: Sequencing assistance also

Deleted:

Deleted: ¶
We thank XXXX for YYYY and ZZZZ. This work was supported by

References:

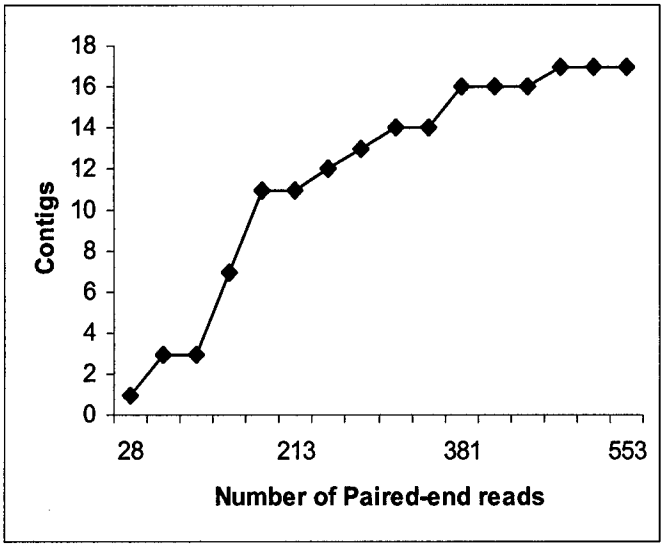
1. McCoy, G.W. and C.W. Chapin., *Further observations on a plague-like disease of rodents with a preliminary note on the causative agent, Bacterium tularense.* J. Infect. Dis, 1912. **10**: p. 61-72.
2. Sjöstedt, A., *Family XVII, Francisellaceae. Genus I, Franciesella*, in *Bergey's Manual of Systematic Bacteriology*, G.M. Garrity, Editor. 2003, Springer-Verlag: New York, N.Y. p. 111-113.
3. Whipp, M.J., et al., *Characterization of a novicida-like subspecies of Francisella tularensis isolated in Australia.* J Med Microbiol, 2003. **52**(Pt 9): p. 839-42.
4. Johansson, A., et al., *Worldwide genetic relationships among Francisella tularensis isolates determined by multiple-locus variable-number tandem repeat analysis.* J Bacteriol, 2004. **186**(17): p. 5808-18.
5. Hopla, C.E. and A.K. Hopla, *Tularemia.* 2nd ed. Handbook of Zoonoses, ed. G.W. Beran. 1994, Boca Raton, FL: CRC Press. 113-126.
6. Oyston, P.C., A. Sjostedt, and R.W. Titball, *Tularaemia: bioterrorism defence renews interest in Francisella tularensis.* Nat Rev Microbiol, 2004. **2**(12): p. 967-78.
7. Petersen, J.M. and M.E. Schriefer, *Tularemia: emergence/re-emergence.* Vet Res, 2005. **36**(3): p. 455-467.
8. Jellison, W.L., *Tularemia in North America*, University of Montana, Missoula, Montana. 276 p.
9. Morner, T., *The ecology of tularaemia.* Rev Sci Tech, 1992. **11**(4): p. 1123-30.
10. Abd, H., et al., *Survival and growth of Francisella tularensis in Acanthamoeba castellanii.* Appl Environ Microbiol, 2003. **69**(1): p. 600-6.
11. Chu, M.C. and R.S. Weyant, *Francisella and Brucella*, in *Manual of Clinical Microbiology, 8th Edition*, P.R. Murray, et al., Editors. 2003, ASM Press: Washington, D.C. p. 20.
12. Svensson, K., et al., *Evolution of subspecies of Francisella tularensis.* J Bacteriol, 2005. **187**(11): p. 3903-8.
13. Sandstrom, G., et al., *Characterization and classification of strains of Francisella tularensis isolated in the central Asian focus of the Soviet Union and in Japan.* J Clin Microbiol, 1992. **30**(1): p. 172-5.

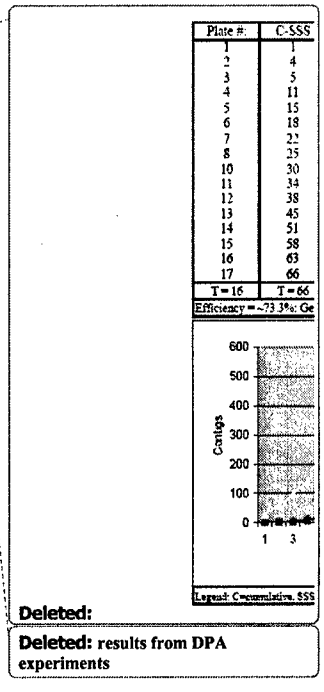
14. Garcia Del Blanco, N., et al., *Genotyping of Francisella tularensis strains by pulsed-field gel electrophoresis, amplified fragment length polymorphism fingerprinting, and 16S rRNA gene sequencing.* J Clin Microbiol, 2002. **40**(8): p. 2964-72.
15. Thomas, R., et al., *Discrimination of human pathogenic subspecies of Francisella tularensis by using restriction fragment length polymorphism.* J Clin Microbiol, 2003. **41**(1): p. 50-7.
16. Farlow, J., et al., *Francisella tularensis strain typing using multiple-locus, variable-number tandem repeat analysis.* J Clin Microbiol, 2001. **39**(9): p. 3186-92.
17. Farlow, J., et al., *Francisella tularensis in the United States.* Emerg Infect Dis, 2005. **11**(12): p. 1835-1841.
18. Broekhuijsen, M., et al., *Genome-wide DNA microarray analysis of Francisella tularensis strains demonstrates extensive genetic conservation within the species but identifies regions that are unique to the highly virulent F. tularensis subsp. tularensis.* J Clin Microbiol, 2003. **41**(7): p. 2924-31.
19. Samrakandi, M.M., et al., *Genome diversity among regional populations of Francisella tularensis subspecies tularensis and Francisella tularensis subspecies holarctica isolated from the US.* FEMS Microbiol Lett, 2004. **237**(1): p. 9-17.
20. Bumbaugh, A.C., et al. *Genomic variation in Shigella dysenteriae type I using a new approach called paired end sequencing mapping.* in *American Society for Microbiology 103rd General Meeting.* 2003. Washington, DC: ASM Press.
21. Nano, F.E., et al., *A Francisella tularensis pathogenicity island required for intramacrophage growth.* J Bacteriol, 2004. **186**(19): p. 6430-6.
22. Larsson, P., et al., *The complete genome sequence of Francisella tularensis, the causative agent of tularemia.* Nat Genet, 2005. **37**(2): p. 153-9.
23. Wilson, K., *Preparation of Genomic DNA from Bacteria*, in *Current Protocols in Molecular Biology*, F.M. Ausubel, et al., Editors. 1997, John Wiley & Sons. p. 2.4.1-2.4.5.
24. Olsufjev, N.G. and I.S. Meshcheryakova, *Infraspecific taxonomy of tularemia agent Francisella tularensis McCoy et Chapin.* J Hyg Epidemiol Microbiol Immunol, 1982. **26**(3): p. 291-9.
25. Cronan, J.E., Jr., *Interchangeable enzyme modules. Functional replacement of the essential linker of the biotinylated subunit of acetyl-CoA carboxylase with a linker from the lipoylated subunit of pyruvate dehydrogenase.* J Biol Chem, 2002. **277**(25): p. 22520-7.
26. Gil, H., J.L. Benach, and D.G. Thanassi, *Presence of pili on the surface of Francisella tularensis.* Infect Immun, 2004. **72**(5): p. 3042-7.
27. Keim, P., et al., *Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within Bacillus anthracis.* J Bacteriol, 2000. **182**(10): p. 2928-36.
28. Keim, P., et al., *Anthrax molecular epidemiology and forensics: using the appropriate marker for different evolutionary scales.* Infect Genet Evol, 2004. **4**(3): p. 205-13.
29. Arbeit, R.D., et al., *Resolution of recent evolutionary divergence among Escherichia coli from related lineages: the application of pulsed field electrophoresis to molecular epidemiology.* J Infect Dis, 1990. **161**(2): p. 230-5.
30. Olive, D.M. and P. Bean, *Principles and applications of methods for DNA-based typing of microbial organisms.* J Clin Microbiol, 1999. **37**(6): p. 1661-9.

31. Scott, L., et al., *The characterisation of E. coli O157:H7 isolates from cattle faeces and feedlot environment using PFGE.* Vet Microbiol, 2006.
32. Titball, R.W., A. Johansson, and M. Forsman, *Will the enigma of Francisella tularensis virulence soon be solved?* Trends Microbiol, 2003. **11**(3): p. 118-23.
33. Forsman, M., G. Sandstrom, and A. Sjostedt, *Analysis of 16S ribosomal DNA sequences of Francisella strains and utilization for determination of the phylogeny of the genus and for identification of strains by PCR.* Int J Syst Bacteriol, 1994. **44**(1): p. 38-46.

Figures:

Figure 1. Cumulative count of CR contigs as a function of paired-end reads.





Deleted:

Deleted: results from DPA experiments