nature genetics

The complete genome sequence of *Francisella tularensis*, the causative agent of tularemia

Pär Larsson[1], Petra C F Oyston[2], Patrick Chain[3], May C Chu[4], Melanie Duffield[2], Hans-Henrik Fuxelius[5], Emilio Garcia[3], Greger Hälltorp[5], Daniel Johansson[1], Karen E Isherwood[2], Peter D Karp[6], Eva Larsson[1], Ying Liu[7], Stephen Michell[2], Joann Prior[2], Richard Prior[2], Stephanie Malfatti[3], Anders Sjöstedt[8], Kerstin Svensson[1], Nick Thompson[9], Lisa Vergez[3], Jonathan K Wagg[6], Brendan W Wren[10], Luther E Lindler[7], Siv G E Andersson[5], Mats Forsman[1] & Richard W Titball[2,10]

Francisella tularensis **is one of the most infectious human pathogens known. In the past, both the former Soviet Union and the US had programs to develop weapons containing the bacterium. We report the complete genome sequence of a highly virulent isolate of *F. tularensis* (1,892,819 bp). The sequence uncovers previously uncharacterized genes encoding type IV pili, a surface polysaccharide and iron-acquisition systems. Several virulence-associated genes were located in a putative pathogenicity island, which was duplicated in the genome. More than 10% of the putative coding sequences contained insertion-deletion or substitution mutations and seemed to be deteriorating. The genome is rich in IS elements, including IS630 Tc-1 mariner family transposons, which are not expected in a prokaryote. We used a computational method for predicting metabolic pathways and found an unexpectedly high proportion of disrupted pathways, explaining the fastidious nutritional requirements of the bacterium. The loss of biosynthetic pathways indicates that *F. tularensis* is an obligate host-dependent bacterium in its natural life cycle. Our results have implications for our understanding of how highly virulent human pathogens evolve and will expedite strategies to combat them.**

Francisella tularensis is one of the most infectious pathogens known and is the etiological agent of tularemia, a disease of humans and animals[1]. The vector-borne form of the disease (glandular or ulceroglandular tularemia) is usually contracted from the bite of an arthropod vector that previously fed on an infected animal[1]. Respiratory tularemia is less frequent and is usually contracted during farming activities that generate dust from sites where infected animals have resided. The mortality rate of respiratory tularemia may be as high as 5–30% without antibiotic therapy; even if not fatal, the disease may be severely incapacitating for a period of weeks or even months[1]. One outbreak of respiratory tularemia occurred in Martha's Vineyard in the US, probably triggered by the mechanical disruption of a rabbit carcass during lawn mowing. The aerosols that were generated infected two individuals, underscoring the highly infectious nature of the organism by the airborne route.

The infectious dose of *F. tularensis* in humans by the airborne route is as low as 10 cells[1]. Although the bacterium is nutritionally fastidious, it was developed as a weapon by Japanese Germ Warfare units during the 1930s and 1940s and later by the former Soviet Union and the US[2]. There is concern that bioweapons containing this bacterium still exist elsewhere in the world.

The high level of interest in *F. tularensis* and concerns over possible misuse contrast with the paucity of knowledge on virulence mechanisms. The bacterium infects macrophages[1], and a few virulence determinants have been proposed, including an ill-defined capsule[3] and a 23-kDa protein that seems to have a role in downregulating proinflammatory cytokines[4]. Some other genes required for growth in macrophages have been identified[5–7], but their roles are uncertain. There is currently no licensed vaccine for the prevention of tularemia.

We report here the complete genome sequence and a phylogenetic analysis of a fully virulent human isolate of *F. tularensis* subspecies *tularensis* (strain SCHU S4). In the long term, this study will support work to devise improved countermeasures against tularemia.

RESULTS

General features

The genome of the *F. tularensis* strain SCHU S4 consists of a 1,892,819-bp circular chromosome, with an overall G+C content of 32.9% and 1,804 predicted coding sequences (CDSs; including

[1]Swedish Defence Research Agency, SE-901 82 Umeå, Sweden. [2]Defence Science and Technology Laboratory, Salisbury SP4 0JQ, UK. [3]Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, California 94550, USA. [4]Division of Vector-Borne Infectious Diseases, Centers for Disease Control and Prevention, Fort Collins, Colorado, USA. [5]Department of Molecular Evolution, University of Uppsala, S-752 36 Uppsala, Sweden. [6]Bioinformatics Research Group, SRI International, Menlo Park, California 94025, USA. [7]Walter Reed Army Institute of Research, Silver Spring, Maryland 20910, USA. [8]Department of Clinical Microbiology, Umeå University, SE-901 85 Umeå, Sweden. [9]Welcome Trust Sanger Institute, Cambridge CB10 1SA, UK. [10]Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK. Correspondence should be addressed to R.W.T. (rtitball@dstl.gov.uk).

Published online 9 January 2005; doi:10.1038/ng1499

Table 1 Overall features of the genome of *F. tularensis* strain SCHU S4

Size	1,892,819 bp
G+C content (%)	32.9
CDS	1,804
Coding percentage	79.4
Unique genes	302
Pseudogenes or gene fragments	201
IS elements	74
ISFtu1 (IS630 family)	50
ISFtu2 (IS5 family)	16
ISFtu3 (ISHpaI-IS1016 family)	3
ISFtu4 (IS982 family)	1
ISFtu5 (IS4 family)	1
rRNAs	3 operons
tRNAs	38
Other stable RNAs	7

pseudogenes). The low G+C content is typical of that found in small (0.9–2.0 Mb) bacterial genomes (range 25–40%). The overall features of the genome are given in **Table 1**. The origin of replication (*ori*) was identified with the aid of the strand specific mutation bias (**Fig. 1**) and was flanked by genes also present at this position in other species, such as *dnaA* and *rng*.

In total, 1,281 genes in *F. tularensis* SCHU S4 had homologs ($E < 1 \times 10^{-10}$) in one or more γ-proteobacterial genomes (**Fig. 1**). These were randomly distributed around the genome, with the exception of a duplicated region of 33.9 kb (nucleotides 1,374,371–1,408,281 and 1,767,715–1,801,625), which lacked homologs in 16 other γ-proteobacterial genomes (**Fig. 1**). In *F. tularensis* strain LVS, duplication of one of the genes in this region (*iglC*) has been reported, suggesting that this region is also duplicated in this strain[7]. The origin of the duplicated regions is not clear, because these genes do not show significant sequence homology with any other genes in GenBank. The genes encoding hypothetical proteins in these duplicated regions (**Fig. 2**) have a low G+C content (27.5%). But the G+C content of genes in the *iglABCD* operon and their codon usage are similar to those of other *F. tularensis* genes. In contrast to the genomic islands of other species, there are no flanking insertion elements or tRNA genes on both sides, although both copies are flanked on one side by rRNA operons and on the other an ISFtu1 element. Mutation of some genes within the duplicated regions can be attenuating[5–7]; therefore, we believe that these regions are pathogenicity islands.

We clustered the proteins predicted to be encoded by the SCHU S4 genome into protein families using the TribeMCL method[8] and identified 61 clusters with more than two members. The largest cluster with exclusively hypothetical proteins that we identified contained five members (FTT0025, FTT0267, FTT0602, FTT0918 and FTT0919). A Hidden Markov Model constructed using HMMER[9] failed to identify any distant homologs when searched against the SwissProt or TrEmbl databases. BLAST searches against the NCBI nr or nt databases and the Sanger Centre Pfam protein family database also did not identify any significant hits ($E < 1 \times 10^{-6}$). Therefore, this cluster represents a new protein family. Three of the proteins (FTT0025, FTT0918 and FTT0919) were predicted to contain both signal peptides and coiled-coil domains, whereas a signal peptide only was predicted for FTT0267 and a coiled-coil domain only was predicted for FTT0602. Our analysis for motifs, which might indicate a possible function, and a range of bioinformatics tools did not identify any significant associations. Therefore, the functional importance of these proteins remains to be elucidated experimentally.

Two types of IS elements (ISFtu1 and ISFtu2) were previously identified in *F. tularensis*[10]. Our analysis identified 50 copies of ISFtu1, a transposon that belongs to the IS630 Tc-1 mariner family. Tc-1 mariner elements are generally found in eukaryotes and have been

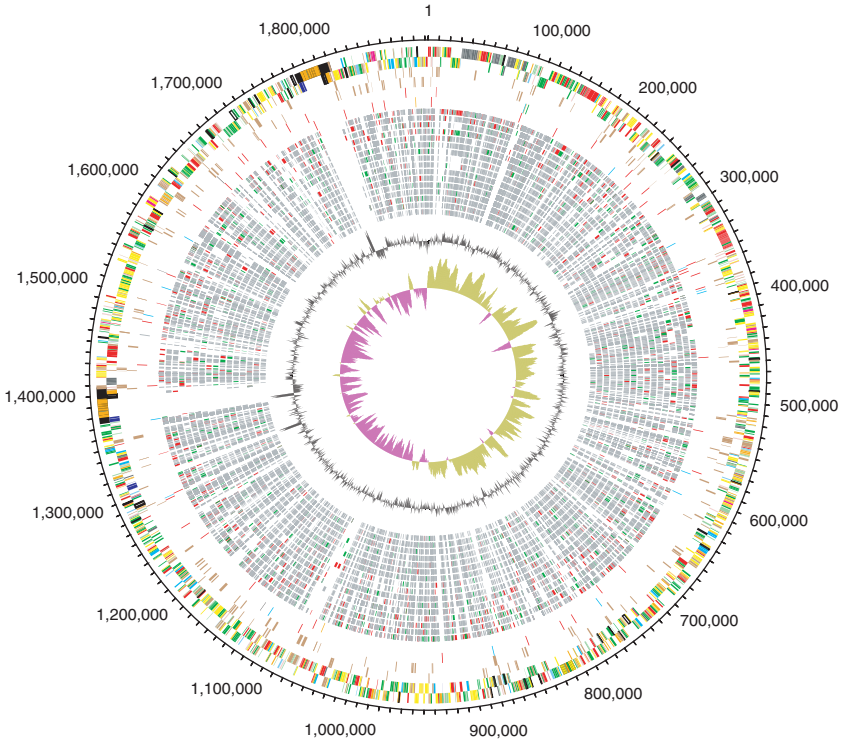

Figure 1 Circular map of the genome of *F. tularensis* strain SCHU S4. The outer scale is marked in base pairs. Circles 1 and 2 (numbering from the outside in) show genes color-coded by function. Circles 3 and 4 show pseudogenes. Circles 5 and 6 show IS elements (ISFtu1, red; ISFtu2, cyan; ISFtu3, orange; ISFtu4, green; ISFtu5, gray; fragments of IS elements, black). The next 16 circles show the locations of genes with matches to *L. pneumophila* (unfinished genome ver. 12 December 2003), *P. aeruginosa*, *V. cholerae*, *C. burnetii* RSA 493, *B. anthracis* Ames, *S. oneidensis*, *E. coli* K12, *H. influenzae*, *P. multocida*, *S. enterica* serovar Typhi, *S. enterica* serovar Typhimurium LT2, *X. axonopodis*, *X. campestris*, *Y. pestis*, *S. flexneri* 2a and *X. fastidiosa*, respectively. Red color marks the top hit, green shows the second best hit and gray shows genes with sequence similarity less than 10^{-10}. The innermost circles show G+C content (%; black) and GC deviation (G–C)/(G+C).

reported in a range of invertebrates such as nematodes and insects. The presence of this element in a bacterium is unusual. *F. tularensis* is often transmitted by infected insect vectors; the IS630 element that we identified may have been acquired originally from an insect. One copy of ISFTu1 is located in the O-antigen cluster, like the IS630 element found in the O-antigen cluster of *Shigella sonnei*. In *S. sonnei*, this element has a key role in the stable expression of form 1 of the O antigen, which is essential for virulence[11]. The IS630 element in the *F. tularensis* O-antigen cluster may have a similar function.

Sixteen copies of ISFtu2, an IS5 family element, were present. The genome also contained three types of IS element previously unreported in *F. tularensis*: ISFtu3 (two complete copies and one fragment), ISFtu4 (one copy) and ISFtu5 (one copy). ISFtu3 has homology to ISHpaI-IS1016 elements, and ISFtu4 and ISFtu5 belong to the IS families IS982 and IS4, respectively. Also present are three IS element fragments, which share homology with ISHpaI-IS1016 elements. These fragments possess terminal inverted repeat sequences not previously reported and therefore represent a new type of IS element.

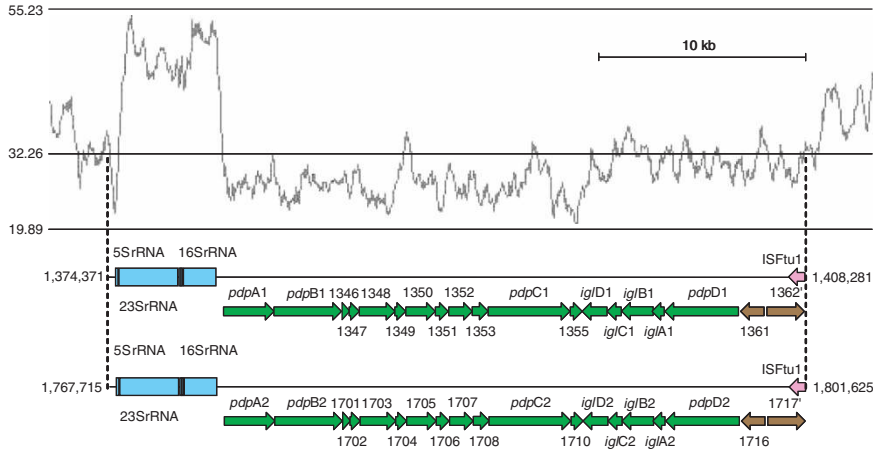

Figure 2 The organization of the duplicated region in *F. tularensis* strain SCHU S4. The leftmost scale shows G+C content. Blue indicates RNA coding regions; green indicates open reading frames encoding hypothetical proteins; brown indicates pseudogenes; and pink indicates IS elements. Open reading frame labels refer to the corresponding annotated gene or FTT number in the genome sequence of SCHU S4.

Most members of the IS630 and Tc-1 mariner family of insertion sequences possess a single open reading frame[12], but translation of the ISFtu1 CDS requires ribosomal frameshifting. In ISFtu1, the first aspartic acid residue of the DDE triad, which is essential for transposition, is generated only after a frameshift[13]. The programmed ribosomal frameshifting motif in ISFtu1 may be used to control the transposition rate of this element.

More than 10% of the CDSs in the SCHU S4 genome are pseudogenes or gene fragments that may have become fixed as the result of a recent evolutionary bottleneck (**Supplementary Table 1** online). The proportion of pseudogenes due to disruption by IS elements (14%) is broadly similar to the proportion in other pathogens such as *Yersinia pestis* (34%), *Leifsonia xyli* (5.5%) and *Bordetella pertussis* (20.6%). Most of the pseudogenes were found among genes uniquely present in *F. tularensis*, hypothetically conserved or encoding proteins involved in transport, DNA metabolism or amino acid biosynthesis (**Fig. 3**). In agreement with this observation, pseudogenes in these categories were also overrepresented in *F. tularensis* compared with other bacteria having more than 50 pseudogenes (**Supplementary Fig. 1** online).

Phylogeny

Francisella is the only genus of the family *Francisellaceae*, which belongs to the γ subclass of proteobacteria. The *Francisellaceae* have no close pathogenic relatives, as inferred from sequence similarity or 16S rRNA phylogenies[14]. Instead, 16S rRNA data suggests that *F. tularensis* is a sister clade with arthropod endosymbionts like *Wolbachia persica*[14] and only distantly related to the human pathogens *Coxiella burnetii* and *Legionella*. This suggestion is supported by a phylogenomic analysis of more than 200 genes with homologs in *F. tularensis* and 15 other γ-proteobacterial genomes, as previously undertaken for a smaller set of γ-proteobacterial genomes[15]. More than 40% of the single gene trees suggest with strong support (>75%) that *F. tularensis* is the most deeply diverging lineage among the 16 γ-proteobacterial species examined here. This is also shown in a tree derived from a concatenated alignment of ten proteins (**Fig. 4**). The *C. burnetii* genome is 1.9 Mbp, with 2,134 predicted CDSs[16]. *Coxiella*, *Legionella* and *Francisella* are γ-proteobacterial pathogens with many lifestyle similarities. But they are not sister clades, and their deep divergences explain the lack of overall gene-order conservation among the three genomes. Therefore, although these pathogens have similar lifestyles and their similar genome sizes seem to reflect this similarity, their positions in the phylogenetic tree suggest that they experienced independent, convergent evolution.

Predicted metabolic pathways and growth requirements

We identified genes encoding 350 enzymes involved in small-molecule metabolism in the annotated genome. We inferred 429 distinct enzymatic reactions to be catalyzed by these enzymes and predicted 155 small-molecule metabolic pathways to be present (**Supplementary Table 2** online). Pathway predictions and information parsed from the annotated genome were output as a pathway-genome database called FrantCyc. Each predicted pathway, P, was assigned a score X/Y/Z: P consists of X reactions; enzymes for Y reactions were identified in the genome; and Z of the Y reactions are used in other predicted

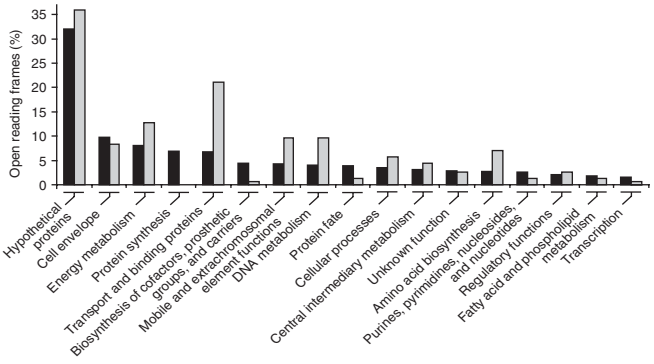

Figure 3 Percentage of total *F. tularensis* strain SCHU S4 CDSs (black) and pseudogenes or fragments (gray) attributed by predicted biological function.

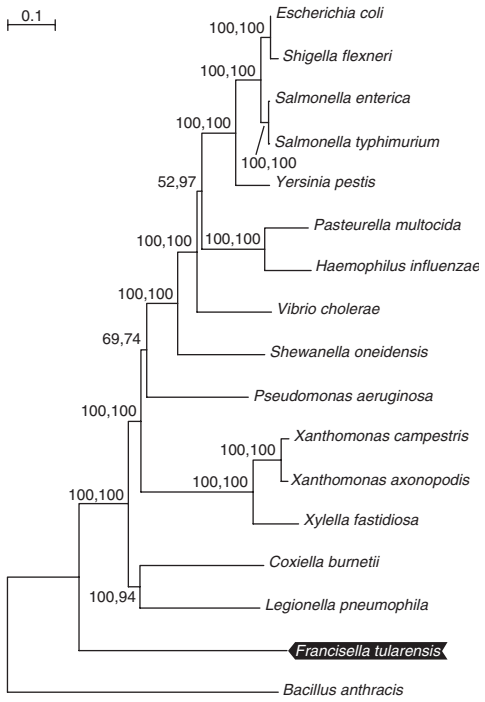

Figure 4 Phylogenetic relationship of 16 γ-proteobacterial species inferred from a concatenated alignment of the proteins encoded by *dnaA*, *ftsA*, *mfd*, *mraY*, *murB*, *murC*, *parC*, *recA*, *recG* and *rpoC*. *B. anthracis* was used as the outgroup. The topology, branch lengths and bootstrap support are according to the reconstruction with the neighbor-joining method. Values at nodes are bootstrap support values for the neighbor-joining and maximum parsimony methods (in that order).

pathways. We inserted all pathways for which Y was nonzero into FrantCyc. In total 1,105 operons were predicted using the Pathway Tools operon predictor and inserted into FrantCyc.

Overall, we identified 390 pathway holes in 137 predicted pathways, corresponding to 54% of the reactions involved in the predicted metabolic pathway network of *F. tularensis*. This percentage is higher than we have observed in other bacteria[17] and is consistent with the proposal that the *F. tularensis* genome is in an advanced state of decay. But we cannot exclude the possibility that because of the relative phylogenetic isolation of this bacterium, some of the pathways holes are filled by divergent orthologs that have not been identified. Pathway holes were input to a program that identifies candidate genes with functions corresponding to each pathway hole[17]. We evaluated each candidate gene manually, and for those deemed sufficiently reliable, we assigned new gene functions to reflect the function postulated by the program (**Supplementary Note** online). Application of this algorithm to a complete genome has not been published previously to our knowledge; in this case, it resulted in the identification of high-probability putative functions for 74 genes whose functions were not identified by classical sequence analysis. We then rescored pathways and removed from the pathway-genome database any pathways for which Y/X was less than 1/3. **Figure 5** shows one part of the entire predicted metabolic map of *F. tularensis*, with the unfilled pathway holes highlighted.

A growth medium consisting of 14 essential compounds (**Table 2**) was developed to support the growth of avirulent *F. tularensis* strain 176 (ref. 18) and was also reported to support growth of SCHU S4 (ref. 19). *F. tularensis* strain SCHU S4 also has a requirement for cysteine[20], which seems to be due to a nonfunctional pathway for sulfate assimilation resulting from a pseudogene (missing start codon) encoding adenylylsulfate kinase (EC 2.7.1.25). It remains to be determined experimentally how many of the other 13 essential compounds[18] are absolutely required for growth. Our analysis indicated that biosynthetic pathways were present for 7 of these 13 compounds. The pathways for sulfate assimilation, threonine biosynthesis, valine biosynthesis and isoleucine biosynthesis seemed to be incomplete (*i.e.*, they contained pathway holes). The available evidence does not indicate whether the enzymes for these pathway holes are truly missing from the genome, thus inactivating the pathway, or whether the enzymes are present but the activity of these pathways is too low to support growth. We found genomic evidence for loss of biosynthetic capabilities for valine, isoleucine and threonine, indicating that these amino acids are required for growth. Specifically, we identified a pseudogene encoding homoserine kinase that mapped to the one step missing from the predicted pathway for threonine biosynthesis and a pseudogene encoding the large subunit of acetolactate synthase that mapped to the one step missing from the biosynthetic pathways for both valine and isoleucine. This loss of biosynthetic capacity may have followed a change of evolutionary niche (such as a move from a free-living organism to one or more specific host cells) that resulted in these amino acids being readily available in the environment. It remains to be determined whether one or more of the other four compounds for which pathways were predicted to be present (*i.e.*, serine, aspartic acid, leucine and proline) is absolutely required for growth. Conversely, the other compounds for which we found little or no evidence of a biosynthetic pathway have probably always been readily available to the organism across a diverse range of evolutionary niches, such that the corresponding pathways were never required by the organism. These compounds are probably required for growth.

We correctly predicted functional biosynthetic pathways for all seven nonessential amino acids (alanine, asparagine, glutamate, glutamine, glycine, phenylalanine and tryptophan; **Supplementary Table 3** online). Because the genomic evidence for these pathways (fraction of enzymes present) was comparable to that observed for the 'false positive' pathways above, we infer that the predominant mechanism of gene-function inactivation rendering biosynthetic pathways inactive or insufficiently active involved relatively small sequence changes, such as one or more point mutations. Biosynthetic pathways for several polyamines, including putrescine and spermine, were also disrupted. This is consistent with the observations that *F. tularensis* is unable to survive under hypotonic conditions[21] and that osmotolerance can be attained by addition of micromolar amounts of putrescine and spermine[21].

Candidate mechanisms of virulence

Little is known about the virulence mechanisms of *F. tularensis*, but growth in macrophages is central to the ability of *F. tularensis* to cause disease. Mutation of the genes *iglA*, *iglC* or *pdpD* in the 33.9-kb duplicated region reduces the ability of *F. tularensis* to survive in amoebae or macrophages and is attenuating[5–7]. These genes, and others in this region, are regulated by the transcriptional regulator MglA[6]. The precise functions of these genes are not known; the gene products do not show sufficient homology with any other genes in GenBank to infer their functions. Therefore, undiscovered mechanisms of virulence are probably encoded in the 33.9-kb pathogenicity island in *F. tularensis*. Within the macrophage, the bacterium can degrade the phagosomal membrane and escape into the cytosol[22]. We identified genes encoding a phospholipase C *acpA* (FTT0221) and a

phospholipase D family protein (FTT0490), which may have a role in this process. FTT1043 encodes a macrophage infectivity potentiator protein previously found to confer virulence for several pathogens, including *Legionella pneumophila*[23]. We also identified a homolog of *mce*, involved in entry of *Mycobacterium tuberculosis* into host cells[24], in the SCHU S4 genome sequence.

When *F. tularensis* is cultured in acidified medium, the pH of the medium increases[25], reportedly owing to the generation of ammonia[18]. The generation of ammonia, and subsequent buffering of the endosomal compartment, may allow pathogens to survive in macrophages[26]. Deaminases such as L-glutaminase, L-asparginase and citrulline ureidase, which could be responsible for ammonia generation, have previously been reported in *F. tularensis*[27]. In addition, citrulline ureidase activity is used to differentiate strains with high virulence (subspecies *tularensis*) from strains with low virulence (subspecies *holarctica*)[20], and low levels of glutaminase activity have been associated with low virulence[27]. We identified several genes in the SCHU S4 genome that could have a role in ammonia production. In addition to genes potentially encoding an L-asparginase (FTT0591) and an L-glutaminase (FTT0195), we also identified an operon predicted to encode a peptidyl-arginine deaminase (FTT0434) and a candidate gene encoding citrulline ureidase (FTT0435). The latter seems to encode a carbon-nitrogen hydrolase family protein and possesses a Pfam (PF00795) motif, indicating that it is an enzyme capable of reducing organic nitrogen compounds and producing ammonia[28].

Type I secretion systems transport substances across the bacterial envelope using transporters containing ATP-binding cassettes. The genome of SCHU S4 is predicted to contain 15 potentially functional ATP-binding cassette systems (H. Garmory, personal communication). We did not identify gene clusters encoding type III, type IV or type V export systems, but we did identify some candidate cell surface–located virulence factors. The presence of pili on the surface of *F. tularensis* has been reported[29], and we identified all currently known genes necessary for type IV pili biosynthesis. The exact role of type IV pili in *Francisella* is not yet known, but in other bacteria, they contribute to virulence. The makeup of the poorly characterized capsule surrounding *F. tularensis* is not known, but we identified a gene cluster (FTT0789–FTT0801) that could encode a polysaccharide additional to the lipopolysaccharide O antigen. We also identified homologs of the genes *capB* (FTT0805) and *capC* (FTT0806) required for capsule biosynthesis in *Bacillus anthracis*. Therefore, the capsule of *F. tularensis* might contain poly-D-glutamic acid.

Virulence and iron acquisition

The ability of the bacterium to acquire iron in the phagosome seems to be crucial for virulence of *F. tularensis*[30], and growth under iron-limited conditions results in changes in the composition of the cell envelope[31]. For many microorganisms, the ferric uptake regulator (Fur) has a key role in modulating iron uptake, and the genome of *F. tularensis* strain SCHU S4 is predicted to encode a Fur protein (FTT0030). We also identified a number of genes that may be regulated by Fur, including *ftnA*, *fumB*, *acnA*, *sodB* and an ortholog of *iraB* (FTT0651), which is associated with iron uptake in *L. pneumophila*[32]. A gene (*frgA*; FTT0029) belonging to a family of hydroxamate-siderophore synthetic genes[33] was located downstream of *fur*, and a putative iron-box was found in the promoter region. Recent papers have described the ability of *F. tularensis* to escape the phagosome[34,35]. In the cytoplasmic environment, iron is highly insoluble, and a TonB-dependent system for complex-bound iron uptake would be expected. Although a low–molecular weight iron-binding compound, growth-initiating

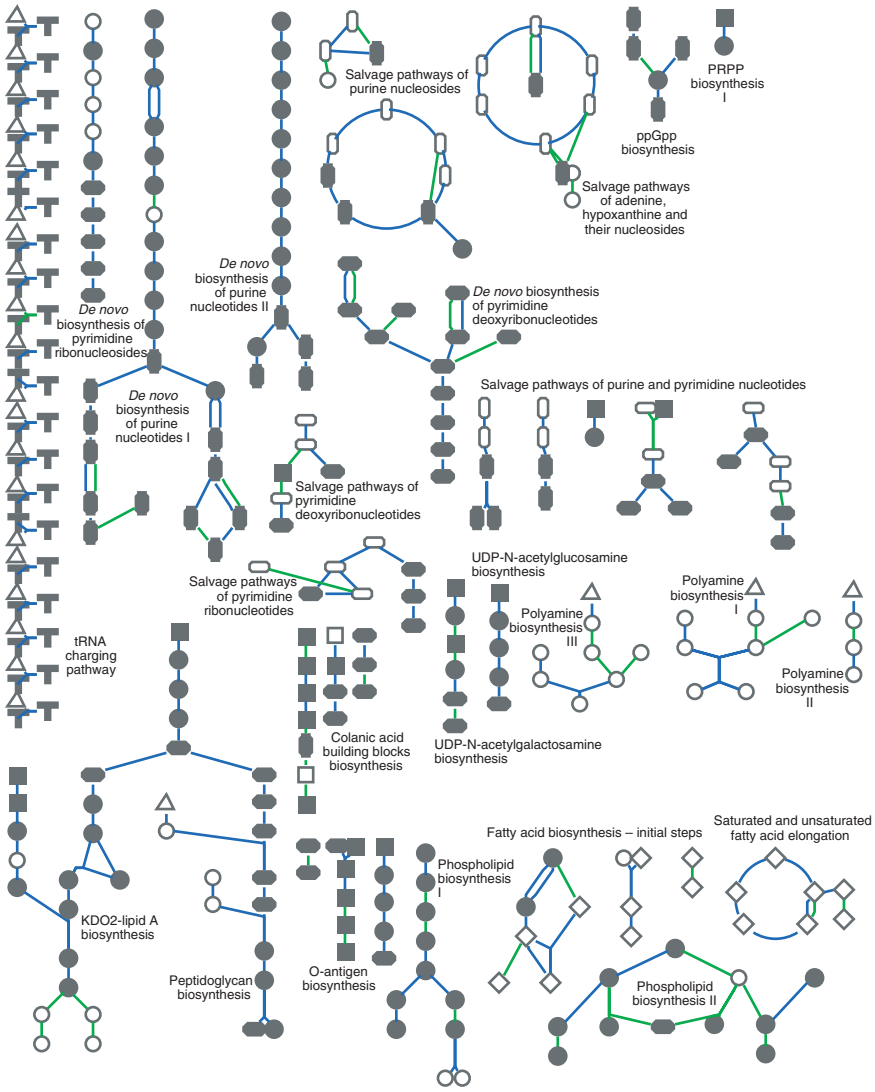

Figure 5 A portion of the FrantCyc cellular overview for *F. tularensis*, showing a region of the predicted metabolic map of the organism. Each line is a single metabolic reaction, and each node is a single metabolite. Green lines indicate reactions that have no enzyme assigned in the genome (pathway holes); blue lines indicate reactions that do have an assigned enzyme. Upward triangle, amino acid; square, carbohydrate; diamond, protein; vertical oval, purine; horizontal oval, pyrimidine; downward triangle, cofactor; T, tRNA; open circle, other; shaded symbol, phosphorylated.

Table 2 A pathway-genome view of the 14 compounds supporting growth of *F. tularensis*

Growth medium requirement	Biosynthetic pathway predicted	Pathway score
Amino acids		
DL-isoleucine	Yes	5/4/1
DL-methionine	No	–
DL-serine	Yes	3/2/2
DL-threonine	Yes	2/1/0
DL-valine	Yes	4/3/2
L-arginine	No	–
L-aspartic acid	Yes	1/1/1
L-cysteine	Yes	5/3/3
L-histidine	No	–
L-leucine	Yes	4/2/0
L-lysine	No	–
L-proline	Yes	4/2/1
L-tyrosine	No	–
Other		
Thiamine HCl	No	–

Evidence was found in the genome for biosynthetic pathways for 8 of these 14 compounds. No evidence was found for pathways for the other 6 compounds and no pathway score was assigned. Absent pathways have probably never been present in this organism, whereas the eight 'false positive' pathways almost certainly reflect the genomic remnants of once-active pathways. The pathway score predicted pathways is denoted X/Y/Z: a pathway P consists of X total reactions; enzymes for Y of the reactions were identified in the genome; and Z of these Y reactions are used in other predicted pathways.

substance, has been reported to be secreted by *F. tularensis*[36], the evidence does not indicate growth-initiating substance to be a hydroxamate siderophore[37]. No genes encoding TonB; outer membrane uptake receptors for ferric siderophore-complexes; or receptors for transferrin, lactoferrin, heme, hemoglobin or hemopexin were found in the genome.

DISCUSSION

Pathogens are frequently thought to evolve by acquiring DNA fragments encoding virulence determinants. But an emerging theme in genome biology is that several pathogens that cause severe disease have evolved by losing genetic information instead. Different genome sequences provide different snapshots of this process of evolution. For example, *Y. pestis* seems to be at the very early stages of evolution and has both lost and acquired genes during this process[38]. *Mycobacterium leprae*[39] and *Rickettsia prowazekii*[40] seem to have evolved solely by gene loss from a progenitor species[39]. The genome sequence of *F. tularensis* SCHU S4 shows extensive inactivation of genes and a duplicated region that is strongly implicated in virulence and may be a pathogenicity island. The origins of the pathogenicity islands are not known, and the function of the genes in this region cannot be inferred on the basis of sequence homology with gene products of known functions. This finding raises the possibility that new mechanisms of virulence operate in *F. tularensis*. SCHU S4 lacks coding potential for several expected features. The ability to import complexed ferric (Fe^{3+}) iron should be important for *Francisella*, because it can escape from the phagosome[22,35], thereby losing its access to soluble iron present in the acidic milieu. But no previously known ferric iron uptake systems were found in the genome sequence.

F. tularensis is considered one of the microorganisms most likely to be used as a biological warfare or bioterrorism agent, but there is a paucity of information on the biochemical makeup of the organism and mechanisms of virulence. In part, this lack of information is a consequence of the difficulties associated with working with highly virulent strains. The complete genome sequence of *F. tularensis* strain SCHU S4 is a key advances in our understanding of this pathogen and will fuel future work to devise defensive countermeasures against this potential biological warfare and bioterrorism agent.

METHODS

F. tularensis subspecies *tularensis* strain SCHU S4 was derived from an isolate from a case of human tularemia in the US[41]. A clonal seedstock of the bacterium has a median lethal dose in the murine model of disease of less than 1 colony-forming unit[42]. We isolated DNA from a culture derived from this seedstock. We constructed plasmid libraries from randomly sheared DNA in pUC18 or pUC19 with insert sizes of 1–2 kb or 2–4 kb, respectively. We also constructed five libraries with insert sizes of 1–4 kb using the TOPO Shotgun subcloning kit (Invitrogen), from nebulized DNA, and one M13 library with insert sizes of 1–2 kb using the double adaptor method[43] from nebulized DNA. We carried out DNA sequencing and assembly as previously described[44]. We produced a total of 32,743 sequence reads, resulting in an overall genomic coverage of ×12.9. For finishing and gap closure, we used PCR, multiplexed combinatorial PCR, single primed PCR and pulse field gel electrophoresis.

Gene prediction was done using Glimmer[45]. We carried out annotation and curation, facilitated by Artemis, as described previously[39] and checked them manually. We identified protein motifs using CONSENSUS and SMART. We used Pathway Tools software to determine metabolic pathways[46], operons[47] and pathway hole fillers[17]. Pathway holes are reactions in a pathway for which no catalyzing enzymes were identified in the genome annotation. The algorithm for identifying candidate genes for each pathway hole involves querying the public protein sequence databases for proteins in other organisms that are known to catalyze the reaction associated with each pathway hole, BLAST searching these sequences against all *F. tularensis* open reading frames and then scoring each matching gene using a Bayesian network that integrates several types of evidence[48]. For example, the Bayesian network will score a given candidate gene higher if multiple query sequences show similarity to it and if the candidate is adjacent to, or in the same direction as, another gene in the same pathway (a direction is a contiguous group of genes transcribed in the same direction). This algorithm differs from previous work[49,50] in that it is completely automated, it applies a reverse BLAST search with increased sensitivity compared with other methods and it computes a probability value for each candidate (validated through cross-validation studies) that allows ranking of the candidates.

URLs. CONSENSUS is available at http://npsa-pbil.ibcp.fr/. SMART is available at http://smart.embl-heidelberg.de/. A version of **Figure 5** that can be explored interactively is available at http://biocyc.org/FRANT/new-image?type=OVERVIEW. A pathway-genome database that describes the *F. tularensis* chromosome; its genes and their predicted operons; the product of each gene; the biochemical reaction(s), if any, catalyzed by each gene product; the substrates of each reaction; and the predicted organization of those reactions into small-molecule metabolic pathways is available at http//biocyc.org/server.html. Also available at this site (http://biocyc.org/FRANT/pathologic-index.html) is a complete listing of all predicted *F. tularensis* pathways and their corresponding evidence scores.

GenBank accession numbers. Genome sequences of *F. tularensis* subspecies *tularensis* strain SCHU S4, AJ749949; *Pseudomonas aeruginosa*, NC_002516; *Vibrio cholerae*, NC_002505 and NC_002506; *C. burnetii* RSA 493, NC_002971; *B. anthracis* Ames, NC_003997; *Shewanella oneidensis*, NC_004347 and NC_004349; *Escherichia coli* K12, NC_000913; *Haemophilus influenzae*, NC_000907; *Pasteurella multocida*, NC_002663; *Salmonella enterica* serovar Typhi, NC_003198, NC_003384 and NC_003385; *Salmonella enterica* serovar Typhimurium LT2, NC_003197 and NC_003277; *Xanthomonas axonopodis*, NC_003919; *Xanthomonas campestris*, NC_003902; *Y. pestis*, NC_003131, NC_003134 and NC_003143; *Shigella flexneri* 2a, NC_004337; and *Xylella fastidiosa*, NC_002488, NC_002489 and NC_002490.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS
We thank the Pathogen Sequencing Unit at the Sanger Institute for advice on the annotation and analysis of this genome sequence. This work was supported by the UK Ministry of Defence, Swedish Ministry of Defence, Defense Advanced Research Projects Agency and US Department of Energy. Work carried out at Lawrence Livermore Laboratory was done under the auspices of the US Department of Energy by the University of California.

COMPETING INTERESTS STATEMENT
The authors declare that they have no competing financial interests.

Received 29 September; accepted 8 December 2004
Published online at http://www.nature.com/naturegenetics/

1. Ellis, J., Oyston, P.C.F., Green, M. & Titball, R.W. Tularemia. *Clin. Microbiol. Rev.* **15**, 631–646 (2002).
2. Dennis, D.T. *et al.* Tularemia as a biological weapon - Medical and public health management. *J. Am. Med. Assoc.* **285**, 2763–2773 (2001).
3. Hood, A.M. Virulence factors of *Francisella tularensis*. *J. Hyg. (Lond.)* **79**, 47–65 (1977).
4. Telepnev, M., Golovliov, I., Grundstrom, T., Tarnvik, A. & Sjostedt, A. *Francisella tularensis* inhibits Toll-like receptor-mediated activation of intracellular signalling and secretion of TNF-alpha and IL-1 from murine macrophages. *Cell. Microbiol.* **5**, 41–51 (2003).
5. Nano, F.E. *et al.* A *Francisella tularensis* pathogenicity island required for intramacrophage growth. *J. Bacteriol.* **186**, 6430–6436 (2004).
6. Lauriano, C.M. *et al.* MglA regulates transcription of virulence factors necessary for *Francisella tularensis* intraamoebae and intramacrophage survival. *Proc. Natl. Acad. Sci. USA* **101**, 4246–4249 (2004).
7. Golovliov, I., Sjostedt, A., Mokrievich, A.N. & Pavlov, V.M. A method for allelic replacement in *Francisella tularensis*. *FEMS Microbiol. Lett.* **222**, 273–280 (2003).
8. Enright, A.J., Kunin, V. & Ouzounis, C.A. Protein families and TRIBES in genome sequence space. *Nucleic Acids Res.* **31**, 4632–4638 (2003).
9. Eddy, S.R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).
10. Thomas, R. *et al.* Discrimination of human pathogenic subspecies of *Francisella tularensis* by using restriction fragment length polymorphism. *J. Clin. Microbiol.* **41**, 50–57 (2003).
11. Houng, H.S. & Venkatesan, M.M. Genetic analysis of *Shigella sonnei* form I antigen: identification of a novel IS630 as an essential element for the form I antigen expression. *Microb. Pathog.* **25**, 165–173 (1998).
12. Mahillon, J. & Chandler, M. Insertion sequences. *Microbiol. Mol. Biol. Rev.* **62**, 725–774 (1998).
13. Doak, T.G., Doerder, F.P., Jahn, C.L. & Herrick, G. A proposed superfamily of transposase genes - transposon-like elements in ciliated protozoa and a common D35E motif. *Proc. Natl. Acad. Sci. USA* **91**, 942–946 (1994).
14. Forsman, M., Sandstrom, G. & Sjostedt, A. 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.* **44**, 38–46 (1994).
15. Canbäck, B., Tamas, I. & Andersson, S.G.E. A phylogenomic study of endosymbiotic bacteria. *Mol. Biol. Evol.* **21**, 1110–1122 (2004).
16. Seshadri, R. *et al.* Complete genome sequence of the Q-fever pathogen *Coxiella burnetii*. *Proc. Natl. Acad. Sci. USA* **100**, 5455–5460 (2003).
17. Green, M.L. & Karp, P. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* **5**, 76 (2004).
18. Traub, A., Mager, J. & Grossowicz, N. Studies on the nutrition of *Pasteurella tularensis*. *J. Bacteriol.* **70**, 60–69 (1955).
19. Nagle, S.C.J., Anderson, R.E. & Gary, N.D. Chemically defined medium for the growth of *Pasteurella tularensis*. *J. Bacteriol.* **79**, 566–571 (1960).
20. Sjostedt, A. Gram-negative aerobic cocci. Family XVII. *Francisellaceae*. in *Bergey's Manual of Systematic Bacteriology* (ed. Brenner, D.J.) 200–210 (Springer, New York, 2004).
21. Mager, J. The stabilizing effect of spermine and related polyamines and bacterial protoplasts. *Biochim. Biophys. Acta* **36**, 529–531 (1959).
22. Clemens, D.L., Lee, B.Y. & Horwitz, M.A. Virulent and avirulent strains of *Francisella tularensis* prevent acidification and maturation of their phagosomes and escape into the cytoplasm in human macrophages. *Infect. Immun.* **72**, 3204–3217 (2004).
23. Cianciotto, N.P., Eisenstein, B.I., Mody, C.H. & Engleberg, N.C. A mutation in the *mip* gene results in an attenuation of *Legionella pneumophila* virulence. *J. Infect. Dis.* **162**, 121–126 (1990).
24. Arruda, S., Bomfim, G., Knights, R., Huima-Byron, T. & Riley, L.W. Cloning of an *M. tuberculosis* DNA fragment associated with entry and survival inside cells. *Science* **261**, 1454–1457 (1993).
25. Chamberlain, R.E. Evaluation of live tularemia vaccine prepared in a chemically defined medium. *Appl. Microbiol.* **13**, 232–235 (1965).
26. Gordon, A.H., Hart, P.D. & Young, M.R. Ammonia inhibits phagosome-lysosome fusion in macrophages. *Nature* **286**, 79–80 (1980).
27. Fleming, D.E. & Foshay, L. Studies on the physiology of virulence of *Pasteurella tularensis*. I. Citrulline ureidase and deamidase activity. *J. Bacteriol.* **70**, 345–349 (1955).
28. Bork, P. & Koonin, E.V. A new family of carbon-nitrogen hydrolases. *Protein Sci.* **3**, 1344–1346 (1994).
29. Gil, H., Benach, J.L. & Thanassi, D.G. Presence of pili on the surface of *Francisella tularensis*. *Infect. Immun.* **72**, 3042–3047 (2004).
30. Fortier, A.H. *et al.* Growth of *Francisella tularensis* LVS in macrophages: the acidic intracellular compartment provides essential iron required for growth. *Infect. Immun.* **63**, 1478–1483 (1995).
31. Bhatnager, N.B., Elkins, K.L. & Fortier, A.H. Heat stress alters the virulence of a rifampin-resistant mutant of *Francisella tularensis* LVS. *Infect. Immun.* **63**, 154–159 (1995).
32. Viswanathan, V.K., Edelstein, P.H., Pope, C.D. & Cianciotto, N.P. The *Legionella pneumophila iraAB* locus is required for iron assimilation, intracellular infection, and virulence. *Infect. Immun.* **68**, 1069–1079 (2000).
33. Hickey, E.K. & Cianciotto, N.P. An iron- and *fur*-repressed *Legionella pneumophila* gene that promotes intracellular infection and encodes a protein with similarity to the *Escherichia coli* aerobactin synthetases. *Infect. Immun.* **65**, 133–143 (1997).
34. Lindgren, H. *et al.* Factors affecting the escape of *Francisella tularensis* from the phagolysosome. *J. Med. Microbiol.* **53**, 953–958 (2004).
35. Golovliov, I., Baranov, V., Krocova, Z., Kovarova, H. & Sjostedt, A. An attenuated strain of the facultative intracellular bacterium *Francisella tularensis* can escape the phagosome of monocytic cells. *Infect. Immun.* **71**, 5940–5950 (2003).
36. Mager, J.A. Factor required for growth initiation of *Pasteurella tularensis*. *Nature* **203**, 898 (1964).
37. Halmann, M. & Mager, J. An endogenously produced substance essential for growth initiation of *Pasteurella tularensis*. *J. Gen. Microbiol.* **49**, 461–468 (1967).
38. Parkhill, J. *et al.* Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **413**, 523–527 (2001).
39. Cole, S.T. *et al.* Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007–1011 (2001).
40. Andersson, S.G.E. *et al.* The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**, 133–140 (1998).
41. Eigelsbach, H.T., Braun, W. & Herring, R. Studies on the variation of *Bacterium tularense*. *J. Bacteriol.* **61**, 557–570 (1951).
42. Russell, P., Eley, S.M., Fulop, M.J., Bell, D.L. & Titball, R.W. The efficacy of ciprofloxacin and doxycycline against experimental tularemia. *J. Antimicrob. Chemother.* **41**, 461–465 (1998).
43. Andersson, B., Wentland, M.A., Ricafrente, J.Y., Liu, W. & Gibbs, R.A. A "double adaptor" method for improved shotgun library construction. *Anal. Biochem.* **236**, 107–113 (1996).
44. Ewing, B. & Green, P. Base-calling of automated sequencer traces using PHRED II. *Genome Res.* **8**, 186–194 (1998).
45. Delcher, A., Harmon, D., Kasif, S., White, O. & Salzberg, S. Improved microbial gene identification with Glimmer. *Nucleic Acid Res.* **27**, 4636–4641 (1999).
46. Karp, P.D., Paley, S. & Romero, P. The Pathway Tools software. *Bioinformatics* **18**, S225–S232 (2002).
47. Romero, P. & Karp, P.D. Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway/genome databases. *Bioinformatics* **20**, 709–717 (2004).
48. Krieger, C.J. *et al.* MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* **32**, D438–D442 (2004).
49. Osterman, A. & Overbeek, R. Missing Genes in metabolic pathways: a comparative genomics approach. *Curr. Opin. Chem. Biol.* **7**, 238–251 (2003).
50. Reed, J.L., Vo, T.D., Schilling, C.H. & Palsson, B.O. An expanded genomescale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol.* **4**, R54 (2003).