Genetic Genealogy and Epidemiology of *Francisella*

Kerstin Svensson

Department of Clinical Microbiology
Infectious Diseases
Umeå University
Umeå 2009
Of old was the age
when Ymir lived;
Sea nor cool waves
nor sand there were;
Earth had not been
nor heaven above,
But a yawning gap,
and grass nowhere.

Völuspá (The Wise
Woman's Prophecy),
Story of the creation
in the Norse Mythology
**ABBREVIATIONS**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>bp</td>
<td>base pair</td>
</tr>
<tr>
<td>CDC</td>
<td>Centers of Disease Control and Prevention</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>ELISA</td>
<td>enzyme-linked immunosorbent assay</td>
</tr>
<tr>
<td>GIS</td>
<td>geographic information systems</td>
</tr>
<tr>
<td>INDEL</td>
<td>insertion/deletion mutation</td>
</tr>
<tr>
<td>LVS</td>
<td>live vaccine strain</td>
</tr>
<tr>
<td>MIC</td>
<td>minimal inhibitory concentration</td>
</tr>
<tr>
<td>MLST</td>
<td>multi-locus sequence typing</td>
</tr>
<tr>
<td>MLVA</td>
<td>multi-locus variable number of tandem repeats analysis</td>
</tr>
<tr>
<td>NIH</td>
<td>National Institutes of Health</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>PFGE</td>
<td>pulsed-field gel electrophoresis</td>
</tr>
<tr>
<td>RD</td>
<td>region of difference</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td>US</td>
<td>United States</td>
</tr>
<tr>
<td>USSR</td>
<td>Union of Soviet Socialist Republics</td>
</tr>
<tr>
<td>VNTR</td>
<td>variable number of tandem repeats</td>
</tr>
</tbody>
</table>
CONTENTS

SAMMANNFATTNING (Summary of the thesis in Swedish) ................................................................. ii
ABSTRACT ........................................................................................................................................ iv
ORIGINAL SCIENTIFIC PAPERS ................................................................................................ v
INTRODUCTION .............................................................................................................................. 1
  Aims .................................................................................................................................................. 2
  Background ..................................................................................................................................... 3
    DNA – the primary source of genetic information ................................................................. 3
    Francisella tularensis – the causative agent of tularemia (rabbit fever) ....................... 4
    On the origin of species – what is a bacterial species like *F. tularensis*? .............. 9
    Landscape epidemiology of tularemia - the importance of heritage and environment in the dispersal of *F. tularensis* populations over time .................. 13
    Genetic genealogy – the use of genetic methods to trace the origin, and establish the identity of *Francisella* isolates ................................................................. 17
    Genetic methods for identifying *Francisella* in clinical practice ................................. 21
RESULTS AND DISCUSSION ........................................................................................................... 23
  Insertion/deletion mutations (INDELs) flanked by repeats are useful for typing isolates and batches of *Francisella* (Paper I) ......................................................... 23
  Insertion/deletion mutations (INDELs) not flanked by repeats are robust genetic markers for typing isolates of *Francisella* (Paper II) ........................................... 26
  Combining single-nucleotide polymorphisms (SNPs) and insertion/deletion mutations (INDELs) not flanked by repeats in a hierarchical real-time PCR array gives rapid and robust classification of isolates of *Francisella* (Paper III) .......... 29
  A combined epidemiological and genetic approach, using patient interviews and high-resolution typing of *F. tularensis* isolates, reveals a landscape epidemiology of tularemia (Paper IV) ................................................................. 31
CONCLUSIONS .............................................................................................................................. 34
ERKÄNNANDEN / ACKNOWLEDGMENTS ................................................................................ 36
REFERENCES ................................................................................................................................. 39


I den första vetenskapliga artikeln (I) så identifierade och analyserade vi insertions/deletionsmutationer (INDELs) som uppkommit via en mekanism där närhillgående repetitiv sekvans på bakteriekromosomen har rekombinerat och mellanliggande DNA sekvans har klipt bort. Vi hittade åtta nya sådana regioner där det fanns en skillnad mellan olika bakteriestammar. Genom att analysera dessa regioner med INDELs, kallade Regions of Difference (RD) i artikeln, tillsammans med single nucleotide polymorphisms (SNPs) dvs enstaka nukleotidskillnader så kunde vi beskriva ett evolutionärt scenario där vi kunde se i vilken tidsföljd olika varianter av *Francisella* uppkommit. Vi fann att *Francisella novicida* var äldst och förgrenade sig tidigast i släktträdet medan *F. tularensis* subsp. *holarctica* var yngst. Vi fann också att alla analyserade stammar som var attenuerade och alltså hade nedsatt sjukdomsalstrande förmåga, hade DNA klipt bort i två regioner som vi kallade RD18 och RD19. Fyndet pekade på att bortklippning av DNA förmedlad av rekombinerings närliggande repetitiv sekvans kunde vara en mekanism som gjorde att stammar av *Francisella* tappade sin sjukdomsfrekventlade förmåga. Detta har också bekräftats senare genom laboratorieförsök där sjukdomsalstrande stammar har attenuerats när DNA klipt bort från dessa regioner. Sjukdomsfremkallande förmåga har också kunnat återställas i den attenuerade vaccinstammen LVS när den kompletterats med bortklipt DNA i dessa regioner.
I följande artikel (II) så utvecklade vi en metod för kombinerad analys av INDELs som saknade omgärdande repetitiv sekvans samt regioner med korta repetitiva sekvenser som följer omedelbart på varandra (VNTR). De förstnämnda gav stabilitet till förgreningarna i släktträdet medan de sistnämnda gav hög upplösning längst ute i grenarna. Fördelen med vår metod var att både INDELs och VNTRs kunde analyseras samtidigt med samma analysinstrument. Tidigare har dessa två typer av mutationer analyserats var för sig, vilket tar längre tid och blir dyrare.


Slutligen i artikel (IV) visade vi en ny strategi för att förbättra epidemiologiska undersökningar av tularemi. Genom att kombinera geografiska data över smittplatser angivna av tularemipatienternas själva och högupplösande genetisk karaktärisering av F. tularensis subsp. holarctica bakterier isolerade från patienterna ifråga kunde vi fördjupa förståelsen av sjuksdomens spridning. Vi fann att den geografiska utbredningen av specifika genetiska undergrupper av bakterien var väldigt begränsad under utbrott i de två studerade orterna Ljusdal och Örebro. Smittplatserna för några genotyper var begränsade till områden om ca två km², vilket tyder på att det finns en tydlig lokal landskapsepidemiologi för tularemi med åtskilda punktsmittkällor.

Sammanfattningsvis så bidrar resultaten från den här avhandlingen till en skildring av släktskapsförhållandena mellan Francisella-stammar på både global och lokal utbrottsnivå.
ABSTRACT

This thesis is about analyzing genetic differences among isolates of *Francisella tularensis* – the tularemia-causing bacterium. To elucidate how these bacterial isolates are related, and their geographical and genetic origins, I have developed typing assays for *Francisella* and used them to study the epidemiology of tularemia.

Tularemia is an infectious disease of humans and other mammals found throughout the Northern Hemisphere. The severity of the disease depends on the type of *F. tularensis* causing the infection. In Sweden, as in other countries of Europe and Eurasia, tularemia is caused by *F. tularensis* subsp. *holarctica*, while other varieties of the bacterium occur in Middle Asia and North America. It is important to identify a tularemia infection promptly in order to initiate the correct antibiotic treatment. A rapid identification of the causative *F. tularensis* variety gives additional clinical information. In recent years, several genomes of various *Francisella* strains have been sequenced, and in this thesis, I have utilized these genomes to identify genetic markers.

In studies reported in the first paper (I) appended to the thesis, we identified and analyzed insertion/deletion mutations (INDELs) inferred to have resulted from a sequence repeat-mediated excision mechanism. We found eight new Regions of Difference (RDs) among *Francisella* strains. Using RDs together with single nucleotide polymorphisms (SNPs), we were able to predict an evolutionary scenario for *F. tularensis* in which *Francisella novicida* was the oldest variety while *F. tularensis* subsp. *holarctica* was the youngest. We also found that all virulence-attenuated isolates analyzed had deletions at two specific genetic regions - denoted RD18 and RD19 – suggesting that repeat-mediated excision is a mechanism of attenuation in *F. tularensis*.

In subsequent studies (presented in paper II), we developed a combined analysis of INDELs lacking flanking repeats and variable number of tandem repeats (VNTRs). Both markers could be assayed using the same analytical equipment. The inclusion of INDELs provided increased phylogenetic robustness compared with the use of VNTRs alone, while still maintaining a high level of genetic resolution.

In analyses described in the next paper (III), we selected INDELs from paper (II) and discovered novel SNPs by DNA comparisons of multiple *Francisella* strains. Thirty-four phylogenetically informative genetic markers were included in a hierarchical real-time PCR array for rapid and robust characterization of *Francisella*. We successfully used the assay to genotype 14 *F. tularensis* isolates from tularemia patients and DNA in six clinical ulcer specimens.

Finally, in paper (IV) we demonstrated a strategy to enhance epidemiological investigations of tularemia by combining GIS-mapping of disease-transmission place collected from patient interviews, with high-resolution genotyping of *F. tularensis* subsp. *holarctica* isolates recovered from tularemia patients. We found the geographic distributions of specific *F. tularensis* subsp. *holarctica* sub-populations to be highly localized during outbreaks (infections by some genotypes being restricted to areas as small as 2 km²), indicative of a landscape epidemiology of tularemia with distinct point sources of infection.

In conclusion, the results acquired during the studies underlying this thesis contribute to our understanding of the genetic genealogy of tularemia at both global and local outbreak scales.
This thesis is based on the following papers and manuscripts, which will be referred to in the text by the corresponding Roman numerals.

I. **Svensson K**, Larsson P, Johansson D, Byström M, Forsman M, and Johansson A.
   Evolution of subspecies of *Francisella tularensis*.
   Journal of Bacteriology, 2005; 187:3903-8

II. Larsson P, **Svensson K**, Karlsson L, Guala D, Granberg M, Forsman M, and Johansson A.
    Canonical insertion-deletion markers for rapid DNA-based typing of *Francisella tularensis*.

III. **Svensson K**, Granberg M, Neubauerova V, Forsman M, and Johansson A.
    A real-time PCR array for hierarchical identification of environmental and human pathogenic *Francisella* isolates.
    Manuscript.

    A high-resolution landscape epidemiology of tularemia exposed by genetic analysis of the causative agent isolated from infected humans.
    Submitted manuscript.

Reprints were made with permission of the publishers.
INTRODUCTION

DNA is the code that uniquely identifies one individual person, as well as one bacterial clone. It is the template for the proteins which build the organism. Differences in the DNA template often give rise to differences in the external form, function or behavior of each individual. These differences, or mutations, vary in composition and length, and arise from different causes and at different rates. By simultaneously analyzing many mutations, the unique identity of each bacterial isolate or individual person can be established, and phylogenetic (genealogical) trees can be constructed to illustrate the relationships between them.

In contrast to humans that reproduce sexually, a bacterium propagates by cloning itself into multiple copies. These copies are not always identical. Sometimes mutations arise that are inherited by the daughter clones, and if the mutations are favorable for the bacteria it will, after some generations, out-compete other clones lacking the mutation. A bacterium can also evolve by acquisition of DNA that previously evolved elsewhere, or by loss of DNA. These acquisitions and losses can also be inherited and give rise to differences between isolates. The degree of clonality and tendency to exchange or lose DNA differs among bacterial species and subspecies.

I have studied Francisella tularensis, the causative agent of the zoonotic disease tularemia (commonly known as rabbit fever, or “harpest” in Swedish), which is a highly clonal intracellular bacterium. Tularemia is a severe disease that can be found throughout the Northern hemisphere. It is endemic in Sweden with an incidence which varies between zero cases and hundreds of patients annually. It is usually transmitted to humans via bites of infected mosquitoes.

The evolution of F. tularensis isolates and the relationships between them had not been very well characterized until recent advances facilitated by the availability of cheap and powerful methods for DNA sequencing and analysis. I participated in the sequencing of the first published genome of Francisella tularensis in 2005, and from the analyses of that isolate and other isolates sequenced later, a large set of genetic markers have been identified. It is these genetic markers that I have examined in the studies this thesis is based upon (I-IV), in attempts to elucidate the evolution of Francisella, and to trace the genetic and geographical origin of isolates.

The series of papers starts with reports of the discovery and examination of two categories of insertion/deletion mutations (INDELs): those flanked by repeats (I), and those not flanked by repeats (II), useful for typing Francisella isolates. In the third study (III), INDELs from (II) and single-nucleotide polymorphisms (SNPs) were applied in a real-time PCR array for rapid and robust hierarchical identification and typing of Francisella isolates. Finally, paper (IV) describes an approach combining genotyping (including three types of genetic markers: INDELs, SNPs, and variable-number of tandem repeats, VNTRs), with geographical data about the presumed place of tularemia acquisition, as stated by the patients, that was used to trace the fine-scale geographical sites of transmission and causative genotypes during outbreaks.

In this thesis, I present the background to DNA research related to tularemia and the implications derived from epidemiological investigations, and put my findings into a broader perspective.
Aims

The aims of the work this thesis is based upon were:

1. To develop genetic typing methods for *Francisella* in order: i) to investigate evolutionary relationships among isolates, ii) to establish isolate identity, and iii) to study the epidemiology of human tularemia, including the geographical origins and distribution of *Francisella* genotypes.

2. To develop genetic assays for identifying and typing *Francisella* isolates that can be used not only in research laboratories but also in clinical routine laboratories or field laboratories.
Background

DNA – the primary source of genetic information

DNA (deoxyribonucleic acid) is a molecule that contains the genetic instructions used in the development of living organisms and viruses (Figure 1). As early as 1869, Friedrich Miescher isolated nucleic acids, which he called nuclein, and in 1884 Oscar Hertwig et al. suggested that nuclein plays a role in heredity (5). However, the first evidence that DNA was the primary source of genetic information was presented by Oswald Avery, Colin M. MacLeod, and Maclyn McCarty in 1944 (6). In their experiments, they isolated the active substance, DNA, from one type of *Pneumococcus* bacteria and introduced it into another type, which transformed the receiving bacteria into the type from which the DNA had been taken. This finding encouraged scientists to determine the structure of DNA. In the April issue of Nature in 1953, three papers were published presenting evidence, of varying strength, that DNA has an alpha-helical double stranded structure (7-9). In 1962, the Nobel Prize in Medicine was awarded to three of the authors of two of these papers, namely Francis Watson, James Crick, and Maurice Wilkins (10). The authors of the third paper, Rosalind Franklin and Raymond Gosling, contributed to the discovery of the double helix by obtaining X-ray diffraction patterns of crystalline DNA. Franklin gave correct details about the shape and size of the double helix. The only important missing piece that she could not discern from her data was how the bases paired inside the helix, and thus the secret of heredity itself. This was solved by Watson and Crick. What Franklin never knew was that her X-ray pictures were presented to Watson and Crick without her knowledge. After Franklin’s death in 1958 when she was only 37 years old, Watson and Crick made clear in public lectures that they could not have discovered the structure of DNA without her work. However, Rosalind Franklin could not be cited for her essential role in the discovery of the physical basis of genetic heredity, since the Nobel Prize is not awarded posthumously. In 1958, five years after the discovery of the DNA double helix structure, Crick presented the central dogma of genetic information transfer (Figure 2) (11), and in 1961 the genetic code was described (12, 13).

![Figure 1. The DNA molecule. The sequence of nucleotides (A,T,G,C) determines individual hereditary characteristics. Courtesy: National Human Genome Research Institute.](image)

![Figure 2. The central dogma of genetic information transfer.](image)
In 1837, the Japanese physician Homma Soken described a disease with a febrile illness and swollen lymphatic nodes among people who had eaten infected rabbit meat (14). These symptoms usually characterize tularemia. The first definitive report on tularemia dates from 1911 when George Walter McCoy described a plaque-like disease among ground squirrels in the area of Tulare Lake in California, US (tulare is an Aztec word for the tule reed, a marsh plant commonly found in that area (Figure 3) (15). In 1912, the bacterium was isolated and named Bacterium tularense (Figure 4) after the location where it was first reported (16). The discovery was followed by experiments conducted during subsequent decades by researchers such as Edward Francis and Hachiro Ohara, to determine the properties of the bacteria and the disease. By the mid-20th Century, the common taxonomic name had changed to Pasteurella tularensis. In 1959, the bacterium was renamed Francisella tularensis in honor of the American scientist Edward Francis, who linked several clinical symptoms to the bacterial infection, and assigned the names “rabbit fever” and “deer-fly fever” to the disease (17). These nicknames for the disease are somewhat misleading, however, since F. tularensis is associated with an extremely wide range of hosts and arthropod vectors. A recent review listed 304 species as being susceptible to Francisella (18). Scientists have further discovered that it is a highly virulent intracellular bacterium which can be fatal to humans. The potential for F. tularensis to infect and kill humans, has attracted military interest in several countries and the possibility of using F. tularensis as a potent biological warfare agent has been examined. Unethical experiments performed on humans in Manchuria between 1932-45 to investigate the effects of exposure to F. tularensis, have been described (19). Offensive biological warfare programs existed for example in the US, the USSR and Japan (20). No intentional uses of F. tularensis in war have been described, but natural outbreaks of tularemia due to environmental disruption and breakdown of sanitation during wars have been documented. For example, during the 2nd World War and the German invasion of Russia, a large outbreak affecting approximately 100,000 people was recorded in the regions of Rostov, Stalingrad and Woroschilowgrad (21-23). Abandoned agricultural areas probably permitted a rapid increase in the rodent population, which spread tularemia to humans. More recently, in postwar Kosovo, where

*Figure 3. F. tularensis is named after Tulare Lake in California where it first was reported. Tulare is an Aztec word for the Bulrush (tule reed). Photo: Kerstin Svensson. Tule reed in Seattle.*
tularemia had not previously been described, an epizootic among rodent populations was considered to have caused the spread of tularemia to humans between 1999 and 2002 (24). In 1972, the Biological and Toxin Weapon Convention was signed by 22 countries, and later entered into force on 26 March 1975 (25). Its signatories currently include more than 160 states that have agreed to prohibit the development, production, and stockpiling of biological and toxin weapons, including *Francisella*. US President Ford signed the convention in 1975, but stockpiles had already been destroyed by 1973 (20).

*F. tularensis* is on the CDC top-six list of Category A agents (26), and is a notifiable disease in many countries, including the US and Sweden, i.e. it is assessed as harmful to the community, since hundreds of people can be infected within a short period of time. In the US, tularemia was removed from the list of notifiable diseases in 1995, but it was included again in 2000 (27). The sending of deadly anthrax letters to government official and media personalities in 2001 resulted in renewed interest and increased research effort to find efficient vaccines against tularemia and other potential biological warfare agents (28).

**Figure 4.** Scanning electron micrograph of a murine macrophage infected with *Francisella tularensis*, the agent of tularemia. Image courtesy of Elizabeth R. Fischer, Rocky Mountain Laboratories/NIAID/NIH USA.
Clinical manifestations

The onset of the disease is acute. The infectious dose can be as low as ten bacteria, and the incubation period spans from a few hours to weeks, with a mean of three to six days (29). A patient suffers from headache, chills, dizziness, muscle pains, and loss of appetite. Inflammation often proceeds with fever, enlargement of lymph nodes, and gland suppuration. The disease may become life threatening if infection is caused by the most virulent form of *F. tularensis*. Humans can acquire tularemia through a number of routes each of which lead to different clinical forms:

i) Bites by infected vectors such as mosquitoes and ticks, or after direct contact with infected animals, leads mainly to the ulceroglandular form of tularemia (Figure 5). At the point of entry, an ulcer develops, followed by enlargement of one or several adjacent lymph nodes. Often it is the enlargement of a lymph node that patients seek medical attention, and not the primary ulcer, which can not always be detected. When no point of entry can be identified, but inflammation of lymph nodes occurs, the clinical form is referred to as glandular. A special form of ulceroglandular form of tularemia is the oculoglandular, in which the eye becomes inflamed (conjunctivitis) and the regional lymph node in front of the ear is enlarged (Figure 6).

ii) Direct inhalation of the bacteria leads to the respiratory form of tularemia. High fever and non-specific symptoms are associated with this form. Infiltrations of fluid into the lungs and hilar enlargements can be detected by examination of chest X-rays (Figure 7). Other X-ray findings include cavitations and nodular lesions, which can be mistaken for tumors or tuberculosis. Before the introduction of antibiotics in the 1940s, the mortality of respiratory tularemia was 5% to 30% in the US (29). Today, fatal cases are rare but still reported (30).

iii) Ingestion of contaminated food or water, leads to the oropharyngeal form of tularemia, in which the primary lesion is located inside the mouth and throat and accompanied by swelling of cervical lymph nodes and the tonsils.

In older literature, another form of tularemia is described as typhoidal, and characterized by high fever without other obvious focal signs. This form is thus difficult to diagnose. It is commonly accepted that the term typhoidal tularemia should be avoided since it is a heterogeneous form of disease. It may result from different routes of infection and its severity is very variable. All forms of tularemia may progress to a septic stage, where bacteria spread to the blood-stream.

It is generally believed that recovery from tularemia results in life-long immunity, at least against severe disease.
**Pathogenesis**

*F. tularensis* is a gram-negative intracellular bacterium with a coccoid shape (Figure 4). When entering the body of a host, it escapes the inflammatory response by avoiding the respiratory burst (31) and having a less pyrogenic lipopolysaccharide (LPS) than many other bacteria (32). It seems to lack exotoxins and induces cell death through apoptosis (“cell suicide”) (33). There is a strong response to the bacteria from the innate system, which is followed by activation of both the CD$_{4^+}$ and the CD$_{8^+}$ T-cells for clearance of the bacteria and development of immunological memory. It has been shown that the cell-mediated response is more important in immunity to *F. tularensis* than the humoral response (34). Much work on the interaction between bacteria and host has been done in murine models, using the Live Vaccine Strain (LVS) of *F. tularensis* or on *F. novicida*, which rarely causes disease in humans. The results from these types of study may therefore be of limited clinical relevance to cases of human infection.
**Diagnosis**

Diagnosis of human tularemia is usually based on sero-diagnostic tests for antibodies, by polymerase chain reaction (PCR) assays, and/or by culture. Specific antibody response in patient serum is detectable seven to fourteen days after the onset of the disease (35). The agglutination test detects mainly IgM antibodies, which reach their highest titer levels after four to eight weeks and then slowly decline (36, 37). Enzyme-linked immunosorbent assays (ELISA) are used for the detection of IgM, IgG, and IgA antibodies, and generally indicate a positive result slightly earlier during the disease than the agglutination test (35, 38). Both methods are reliable, with high sensitivity and specificity, but are unable to discriminate between infections caused by different types of *F. tularensis*. Culturing bacteria may require seven days, but this is often avoided since it poses a considerable risk of laboratory-acquired infections, and requires BSL-3 laboratory conditions (39-41). Therefore, to shorten the time to diagnosis, PCR assays targeting 16S DNA or specific genes encoding outer membrane proteins such as *fopA* and *lpnA* have been used to detect *Francisella* (42-44). In recent years, several real-time PCR assays have been developed which appear superior to conventional PCR in both sensitivity and rapidity (45-50) (Paper III). The conventional and real-time PCR methods are applicable mainly to samples taken from primary lesions in patients with the ulceroglandular or oropharyngeal form of tularemia. Both types of PCR methods can be used to discriminate between types of *Francisella* on different taxonomic levels.

**Treatment**

Early institution of appropriate antibiotics for treating tularemia is important in order to avoid complications, and can be life-saving in severe cases. In North America, severe tularemia has mainly been treated with aminoglycosides such as streptomycin or gentamycin (51). In Europe, tularemia has traditionally been treated with tetracyclines, especially doxycycline. However, tetracyclines are bacteriostatic drugs and treatment failure and relapses may occur, especially if the course of treatment is not continued for at least 14 days (52). In recent years, bacteriocidal fluoroquinolones, in particular ciprofloxacin, have been used to cure tularemia. Fluoroquinolones exhibit both very low minimum inhibitory concentration (MIC) values for *F. tularensis*, and high intracellular concentrations. Therefore, ciprofloxacin is now recommended as the first choice therapeutic, at least for milder forms of tularemia (53, 54). *F. tularensis* is naturally resistant to beta-lactam antibiotics (55). Treatments with penicillins, cephalosporines, monobactams and carbapenems are thus ineffective. Some *F. tularensis* subsp. *holarctica* isolates from Europe and Russia are resistant to the macrolide antibiotics such as erythromycin and azithromycin (56, 57).

A Live Vaccine Strain (LVS) is used for immunizing military personnel and laboratory staff, but is not approved for public use by the Federal Drug Administration (FDA) in the US, due to uncertainties in the cause of attenuation and stability. Intense research efforts are focused on the development of a new and well-defined vaccine (58).
On the origin of species – what is a bacterial species like *F. tularensis*?

This year (2009) marks the 200th birthday of the British naturalist Charles Darwin, and the 150th anniversary of his world-shattering publication “On the Origin of Species by Means of Natural Selection” in 1859 (59). The publication described a controversial theory of how species evolved into their present forms, which made, and still makes, many people furious because it contradicts the Biblical view of creation. Darwin adopted the Tree of Life metaphor (Figure 8) to explain the relationships between different species and it has since become a unifying concept for understanding the history of life. The root of the tree was the Common Ancestor from which the earliest life-form emerged, followed by branches, each representing a new species diverging from the old, driven by changes in the environment.

But what is a “species”? It is surprisingly difficult to define the word “species” in a way that can be applied to all organisms. As scientific methods improve and new variants of organisms are found, there is a never-ending debate among biologists about how to define “species” and how to identify actual species. Darwin viewed the notion of “species” as a temporarily useful system for naming groups of interacting individuals. This was a shift in the view of species from the previous morphological species concept. Prior to Darwin, naturalists viewed species as ideal types, bearing all the traits general to the species. They tried to find, describe and name organisms, and categorize them as species. The Swedish botanist Carolus Linnaeus (Carl von Linné) (1707-1778) created the first standardized classification system published in 1735 in his “Systema Naturae” (60). He divided known organisms into seven major levels of taxonomic groupings: Kingdom, Phylum or Division, Class, Order, Family, Genus, and Species, and devised the two-part naming system of binomial nomenclature (e.g. *Francisella tularensis*) that we still use today. In this first version the bacterial kingdom was not included. Although the first observation of living bacteria dates from 1683 by Antony Van Leeuwenhoek (61), it was not until later when the German, Ferdinand Cohn first made a systematic attempt to classify them in 1872 (62, 63) and Robert Koch first isolated a disease-causing microbe, *Bacillus anthracis*, in 1876 (64), that the bacterial kingdom was recognized and added to Linnaeus’ taxonomic system.

A contemporary of Darwin was the Augustinian priest and scientist Gregor Mendel. With his study of the inheritance of certain traits in pea plants, genetics made
its entrance in the debate on the species definition problem. Although Gregor Mendel’s paper on genetics was published in 1866 (65), its significance was not immediately recognized (66). Indeed, it was not until the 1940s that the concepts of Darwinian natural selection and Mendelian inheritance were combined to form the biological species concept, which connects the units of evolution (genetics) and the mechanism of evolution (natural selection). In 1942, Ernst Mayr defined species as “groups of actually or potentially interbreeding natural populations, which are reproductively isolated from other such groups”, which is still the most commonly used species concept among the general public (67). The phylogenetic species concept was introduced in the 1960s and is today commonly used in science, and defines species as “the smallest set of organisms that share an ancestor and can be distinguished from other such sets” (68). The discipline strives to discover the evolutionary inter-relationships among all living organisms, which is often visualized in the form of the Tree of Life.

But could this definition also hold for bacterial species? In 1987, Wayne et al. defined bacterial species as an entity that “includes strains with approximately 70% or greater DNA-DNA relatedness and with 5°C or less $\Delta T_m$”, where $\Delta T_m$ is the difference in DNA melting temperatures according to DNA-DNA hybridization analyses, and this definition has since been widely used by bacteriologists (69). However, this definition was formulated just at the beginning of the genomic era. Since the first genomes were sequenced in the 1990s, large numbers of both culturable and unculturable novel microorganisms have been detected and sequenced, partially or fully. Although this new wealth of genomic data has enabled the discovery of new genetic markers for phylogenetic tree construction, it has also presented new challenges for theoretical phylogenetic analysis.

Several branches of the Tree of Life, including that of Bacteria, are proving difficult to resolve. Darwin assumed that descent was exclusively “vertical”, with organisms passing traits down to their offspring. In analogy, it has been assumed for a long time that all bacteria are clonal i.e. that they divide themselves into identical copies. However, it has become evident that many bacteria routinely swap genetic material “horizontally” with other bacterial species (70). In principle, a bacterium can evolve alone or in combination, by the acquisition of genes that previously evolved elsewhere, by the loss of genetic material, or by mutations of genes that already exist. This means that the model of the Tree of Life does not apply to all bacteria. For some bacteria there seems to have been no Common Ancestor: rather there seems to be a DNA pool of genetic material moving freely between even distantly-related species i.e. a Web of Life, as some authors have suggested, rather than a Tree of Life (71, 72). The current genetic makeup of the bacteria, the access to DNA, and the environmental pressure and selection of certain mutations, decides which evolutionary path a bacterium takes. This has led to the widely used definition of bacterial species by Wayne et al. (69) being questioned. At issue are: (i) whether bacteria naturally form cohesive genotypic or phenotypic clusters, (ii) how to recognize such clusters, and (iii)
when they might warrant the status of “species” (73). Other linked questions are: what ecological, genetic and evolutionary processes are responsible for grouping, if and when it does occur? It may seem strange that only approximately 5,000 bacterial species have been formally named, in contrast to at least a million animal species that have been described, despite the early evolution of microbial life, which occurred about three billion years before the evolution of plants and animals (74). This great disparity in species numbers further indicates that the modern theory of evolution requires a re-definition of bacterial species. The issue of defining bacterial species is certainly of importance. Further, the ability to discriminate between isolates is essential in population genetics and microbial epidemiology. It is the basis that allows researchers to classify isolates of microorganisms and taxonomy provides an appropriate nomenclature for the different groups and is used for identifying the so-called species. However, despite great efforts being made to refine the systematization of the bacterial kingdom, no definition has yet been generally accepted: see reviews (75-78).

How then, is Francisella currently defined, and what type of bacterium is it? Based on DNA-DNA hybridization, there are currently three recognized species within the genus Francisella: F. tularensis, F. philomiragia, and F. novicida (79) (Figure 9). The position of F. novicida as a species is debated and is in Bergey’s manual from 2005 re-classified as a subspecies of F. tularensis (80). Analysis of 16S rDNA has placed the genus of Francisella in the γ-proteobacteria class, order of thiotrichales, and family of Francisellacae, where Francisella is, so far, the only accepted genus (42). The species F. novicida and F. philomiragia are often associated with water and are pathogenic to humans only in cases of immunosuppression. Some fish pathogens were recently described as close relatives to F. philomiragia (81-83). Several tick endosymbionts are also closely related to Francisella based on analysis of 16S rDNA, and have been placed in the Francisellacae family (84, 85). Due to a lack of genomic or hybridization data, their exact taxonomic positions remain unclear. Other water associated bacteria also show relationships to Francisella based on analysis of 16S rDNA, such as the protozoan endosymbiont Caedibacter taeniospiralis, the fish pathogen Piscirickettsia salmonis, and Fangia hongkongensis, which was isolated from coastal seawater of Hong Kong (86, 87). The species of F. tularensis consists currently of

![Figure 9. A neighbor-joining tree illustrating that the slowly evolving 16S rRNA gene provides little resolution among F. novicida and F. tularensis strains and that F. philomiragia is one of several distinct groups of environmental Francisella bacteria. Bootstrap values from 1,000 resamplings are indicated at the branch points (from Paper III)
three subspecies: *tularensis* (also known as type A), *holarctica* (also known as type B), and *mediasiatica*. The subspecies have been further divided into subpopulations in several studies (88, 89) (Figure 10). To date, 22 *Francisella* genomes have been completely or partially sequenced, and there are several ongoing genome projects (90).

Recent genome comparisons indicate that the *Francisella* species differ in the ways in which they have evolved: the human pathogenic *F. tularensis* has taken the evolutionary path leading to a highly clonal bacteria with little or no exchange of genetic material with other bacteria, while *F. novicida* and *F. philomiragia* exhibit more frequent signs of DNA exchange (3). This means that *F. tularensis* fits very well into the traditional Tree of Life model, and that genetic mutations are mainly inherited vertically (89, 91): a property that allows for genetic genealogy testing of this species and which is fundamental to my work presented in this thesis.

![Figure 10](image-url)

**Figure 10.** Whole genome phylogeny among 17 *Francisella* strains based on 1,104,129 aligned nucleotide positions. Panel (A) depicts relationships among major clades within the *Francisella* genus and panel (B) relationships within the species *F. tularensis*. The evolutionary tree was inferred using the Neighbor-Joining method. Bootstrap support values (500 replicates) are shown next to branches. Scale bars indicate the number of base substitutions per site. Adapted from (3).
Landscape epidemiology of tularemia - the importance of heritage and environment in the dispersal of *F. tularensis* populations over time

Landscape epidemiology is the study of spatial and temporal variation in disease risk or incidence. Pathogens infecting mammals may be dispersed by many different modes from an infected to an uninfected host, including direct contact between infected and uninfected individuals, indirect contact by inhalation or ingestion of bacteria excreted by an infected host, and transmission via arthropod vectors. In most cases, the probability of transmission declines with distance from an infected host. As a consequence, factors affecting the spatial positions of pathogens, hosts and vectors, and their probability of close encounter, are fundamentally important to disease dynamics. The concept of landscape epidemiology was developed by the Russian parasitologist Evgeny N. Pavlovsky and first published in 1939 (2, 92). It is based on three observations: first, diseases are not evenly distributed geographically, but occur in foci (which he called niđi or nidus in singular); second, this spatial variation arises from underlying variation in abiotic and/or biotic conditions that support the pathogen and its vectors and reservoirs; and third, if those conditions can be delimited on maps, then both contemporaneous risk and future change in risk should be predictable. Examples of such conditions are climate (temperature, precipitation), geological structure (rock, soil), microclimate (nutrient supply, predation pressure), land use, and vector and host distributions.

The pathogenic agent circulates among vectors and hosts, independently of humans, and humans become victims when they become exposed to the vectors or hosts by entering an area which is a focus of the disease. By learning how to recognize these living and physical characteristics of a nidus, it is possible to identify niđi in new and unsurveyed areas. This can be achieved by recognizing characteristics of the locality under survey that suggest the likelihood of diseases being present or absent. The term niđus has been sometimes used to indicate a more expanded territory, sometimes called a nidal center, where a disease was prevalent due to one or more transmission patterns.

The nidality concept was applied in Siberia to tick-borne Russian spring-summer encephalitis, where ticks and/or rodents could be controlled by chemical control and environmental changes to discourage development of ticks and rodents near villages in disease niđi. One procedure was to create clean areas around camps and villages, through burning dead wood and debris, thus creating open areas where rodent burrows and nests could be eliminated. Another method of disease control was to locate new camps and villages away from disease niđi.

Tularemia is another zoonotic disease to which Pavlovsky’s concept has been applied. Tularemia has a locally patchy distribution and exhibits recurrent epidemic outbreaks in geographically restricted natural foci. In endemic areas, outbreaks occur at irregular and unpredictable intervals. Eight nidal centers have been described in the world (Figure 11) (2). Two of the centers are in North America, three in Europe, and
three in Asia. The epidemiology of the disease varies in the different nidal centers (Table 1).

Large programs for the active control of tularemia outbreaks were developed in the former Soviet Union in the 1960s and 1970s, aiming at mitigating the effects of tularemia, by covering grain stocks, poisoning rodents, and burning hay (23, 93). After 1992, when the USSR was dissolved, these programs were splintered and, due to limited resources, scientists were no longer able to acquire accurate information on tularemia incidence and frequency in many of the countries’ natural disease foci (93).

The increased availability of geographical and spatial environmental data, and the development of new and more powerful Geographic Information Systems (GIS) tools, have stimulated renewed interest in spatial epidemiology. Conditions included in the old concept of Pavlovsky can now be combined with other conditions affecting spatial patterns of disease, such as the immune status of hosts and reservoirs, and the genetic genealogy of the disease-causing agent. *F. tularensis* is particularly suitable to study in the latter manner, since it is genetically monomorphic and highly clonal, which should facilitate the establishment of kinships (89, 91).

Tularemia is reported only from locations in the northern part of the world. It is enzootic in Canada (94), Mexico (95), and in all states of the US except Hawaii. In

---

**Table 1.** Enzootic cycles in the tularemia nidal centers.

<table>
<thead>
<tr>
<th>Nidal center</th>
<th>Enzootic cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Eastern North America</td>
<td>Hard tick – cottontail rabbit</td>
</tr>
<tr>
<td>2. Western North America</td>
<td>Hard tick – hare (jackrabbits)</td>
</tr>
<tr>
<td></td>
<td>Deer fly - hare</td>
</tr>
<tr>
<td></td>
<td>Water – muskrat/beaver</td>
</tr>
<tr>
<td>3. Scandinavia</td>
<td>Mosquito – lemming</td>
</tr>
<tr>
<td>4. Western Central Europe</td>
<td>Not mentioned in (2)</td>
</tr>
<tr>
<td>5. Eastern Central Europe</td>
<td>Not mentioned in (2)</td>
</tr>
<tr>
<td>7. Siberian</td>
<td>Soft ticks, horse flies - mice (after World War II)</td>
</tr>
<tr>
<td>8. Japan</td>
<td>Not mentioned in (2)</td>
</tr>
</tbody>
</table>
Europe, outbreaks among animals and/or humans have been reported from most countries except the British Isles, Iceland, and Malta. Russia and other former USSR republics have recurrent outbreaks, and in Asia, tularemia has been reported from Turkey (96), Iran (97), China (98), Japan (99, 100), South Korea (101), and Mongolia (102). In central Europe and Asia, tularemia is reported to be transmitted to humans mainly through contact with lagomorphs (hares and rabbits), but also through contaminated food and water, and via ticks. In Sweden, Finland and parts of Russia, mosquitoes have been reported to be the main vector for disease transmission (103).

However, the to date largest outbreak in Sweden occurred during the winter of 1967 to 1968 in the county of Jämtland, affected 676 patients, most of whom exhibited the respiratory or typhoidal form of tularemia, and was caused by inhalation of hay dust contaminated with bacteria excreted by voles (104, 105). After this outbreak, tularemia, mostly of the ulceroglandular form, generally appeared only in long-term endemic areas of northern and central Sweden, until the late 1990s when tularemia (re-)emerged in southern parts of Sweden. In 2003, 698 patients were diagnosed with tularemia in Sweden, almost half of them in (re-)emerging areas (106).

Differences in the geographical distributions of subpopulations have been found: *F. tularensis* subsp. *holarctica* is found throughout the northern hemisphere, while *F. tularensis* subsp. *tularensis* is found in North America, and *F. tularensis* subsp. *mediasiatica* has been isolated in Central Asia. The genetic subpopulations A.I and A.II of subsp. *tularensis* show distinct geographical distributions, where A.I is predominantly found in eastern US and A.II in western US (107, 108), and the subpopulation B.V of subsp. *holarctica* is restricted to Japan (88). Although the tools previously used to characterize the disease agent were not as detailed as those available today, there are obvious similarities between these areas and the tularemia nidal centers of Pavlovsky (Figure 11). However, the nidal center designated number two is also based on *F. tularensis* subsp. *holarctica* (type B) data (109).

The bacterium has the ability to infect and colonize a wide range of hosts. More than 300 species have been shown to be susceptible to *Francisella*, from large mammals and vertebrates to invertebrates, arthropods, and amoebas (18). Small mammals such as rodents seem to be more sensitive to tularemia than larger animals. Differences in the virulence and enzootic cycle of genetic subpopulations of *F. tularensis* subsp. *tularensis* (type A) have been reported. Type A is associated with domestic cats, cottontail rabbits and jackrabbits, while *F. tularensis* subsp. *holarctica* (type B) is associated with rodents such as beavers and voles (108, 109). Host-species differences have also been observed among type A isolates. The desert cottontail *Sylvilagus audubonii* and the mountain cottontail *S. nuttallii*, are associated with a subpopulation denoted A2, whereas the eastern cottontail *S. floridanus* is associated with another subpopulation denoted A1 (107, 108).

Recent molecular epidemiological studies of *F. tularensis* isolates submitted to the Centers of Disease Control and Prevention (CDC) in the US between 1964 and 2004, have identified differences among different genetic groups in their association to
human mortality rates (110). Using PFGE for strain discrimination and comparison with clinical records, Kugeler et al. found that a subpopulation of *F. tularensis* subsp. *tularensis* A1, named A1b, caused most deaths. Twelve of 49 (24%) examined human infections with A1b were fatal, compared with 2 of 55 (4%) infections with A1a. The subpopulation B caused mortality in 7% of infected humans while A2 strains caused no human mortality.

To improve and enhance operational surveillance and control programs of natural outbreaks of vector-borne tularemia, it is necessary to be able to predict spatial as well as temporal patterns of disease incidence, i.e. to predict habitat development, and where and when vector production and tularemia transmission risks will be greatest. There have been several pertinent recent studies of tularemia. For example, a GIS-based predictive spatial model for tularemia in nine US states has been developed using county-based tularemia incidence from 1990 to 2003, climate data (temperature, precipitation, relative humidity), vegetation data, and land cover classifications (111). In addition, the spatial-temporal pattern of tularemia from 1994 to 2001 has been studied in a smaller geographic area of the Czech Republic (130 x 90 km) (112), and the information obtained has been used to predict areas with high tularemia transmission risk across the whole country (113). These studies show how research that includes modern landscape epidemiological analysis of tularemia can improve predictive models, even though they might not include temporal changes in environmental factors as parameters. Also, the geographical and genetic resolution in these studies was low: the geographical data regarding human cases was based on county data or patients’ residential addresses, which do not accurately reflect the actual places where they acquired tularemia. Furthermore, information about the genetic subpopulations of *F. tularensis* was not taken into account, although they may differ in transmission routes and pathogenicity to humans (108).

An attempt to combine high-resolution genotyping of bacterial isolates from patients, with more fine-scale geographical data, including the actual place of tularemia acquisition in local outbreaks in Sweden, is presented in this thesis (Paper IV), providing knowledge that could contribute to the understanding of the landscape epidemiology of tularemia in Sweden.

In the US, there is a large surveillance network called BioWatch, which continuously monitors airborne *Francisella* in more than 30 cities at locations where many people gather, to provide warnings to the government and public health community of any potential bio-terror event (114). The system has positively detected tularemia in samples on several occasions, but high-resolution genetic methods have classified the positive samples as naturally occurring, non-pathogenic bacteria (115-117). These incidents underline the importance of knowledge about natural occurrence and types of *Francisella* in the area under survey and awareness of ongoing tularemia epizootics, in order to avoid spreading unfounded fear in the community and undermining public trust in warning systems.
Genetic genealogy – the use of genetic methods to trace the origin, and establish the identity of *Francisella* isolates

Genetic genealogy is the term used for the application of genetics to traditional genealogy in humans (118), but is here used for tracking a disease agent (119). Genetic genealogy involves the use of genetic DNA tests to determine the level of genetic relationship between individuals. For bacteria and viruses, it can be synonymous with phylogenetic analysis, but additionally involves tracking from where or whom the infectious agent originated, i.e. the epidemiological tracking of disease transmission paths.

Traditional bacteriology for identification of bacterial isolates has involved use of phenotypic methods such as serotyping, immuno-blotting, testing for the ability to grow on specific media, and assessment of the tested organism’s capacity to perform various fermentation reactions. These methods are routinely used in clinical laboratories to identify an infectious agent, but they do not discriminate isolates very well, if at all. For some purposes it is desirable to have a higher resolution than species or subspecies level. This may be applicable in an epidemiological investigation of communicable diseases, where the identification of specific bacterial isolates in a transmission chain of an infectious disease may be warranted. Similarly, investigating the emergence and spread of bacteria with antibiotic resistance requires methods for tracking individual bacterial isolates, see review (120). In a forensic investigation, a complete picture of the genetic makeup of the bacterial population, is critical to provide statistically valid evidence. Ideally, to establish the identity of a bacterium definitively, its entire genome sequence would be obtained. However, this level of information cannot often be easily achieved by current methods, although many powerful methods for sequencing have been developed in recent years, see reviews (121, 122). Instead, methods for discriminating isolates based on measuring genomic variability are used, often referred to as “typing methods”. In recent decades, various genetic typing methods have been developed for discrimination of isolates, which have facilitated the identification of isolates and establishment of intra-relationships. The best genetic method to use is dependent on the typing purpose and organism. Some methods that have been applied to *Francisella* are summarized below.

Resolution of DNA fragments, obtained by digesting the bacterial genome using rare-cutting restriction enzymes, by pulse-field gel electrophoresis (PFGE) has been used for typing, and in epidemiological investigations of many bacteria, including *Salmonella enteritidis*, Extended Spectrum of Beta Lactamase (ESBL)-producing *Escherichia coli*, and methicillin-resistant *Staphylococcus aureus* (MRSA). In this approach the genome is cut into large pieces, which are separated by size in a gel subjected to an electric field, providing a “fingerprint” of the bacterium. The discriminatory power is high for many bacteria. The method has been standardized and is widely used in laboratory networks in the US, Europe and Japan (PulseNet) (110, 123, 124). It has also been proposed for use in public health laboratories for
typing *Francisella* (125), but its utility for this is doubtful, since PFGE has significant disadvantages. Notably, it has reproducibility problems, and it is labor-intensive, time-consuming, has limited typing resolution, bio-safety level (BSL) 3 facilities are needed, and there is a risk of laboratory-acquired infections.

Polymerase chain reaction (PCR) methods (126) are generally fast methods that do not require culturing bacteria prior to analysis. PCR methods are based on the amplification of a piece of DNA defined by primers that are approximately 18 to 25 nucleotides long and match the target sequence on both sides of the region of interest. Several PCR methods have been developed that target specific defined genetic loci and have been used to characterize *Francisella* isolates, including a 30 bp INDEL mutation denoted Ft-M19 (88), used to distinguish *F. tularensis* subsp. *tularensis* from *F. tularensis* subsp. *holarctica*, and an insertion sequence denoted ISFtu2, used for characterizing *F. tularensis* as subsp. *tularensis* or subsp. *holarctica* (127).

Other PCR methods that have been used for typing *Francisella* include arbitrary primed PCR (AP-PCR), random amplified polymorphic DNA (RAPD) fingerprinting, and repetitive DNA PCR analysis (Rep-PCR) (128, 129).

In multi-locus sequence typing (MLST), DNA fragments amplified by PCR from a variable number of genes (generally seven) are sequenced and analyzed with respect to nucleotide polymorphisms (130). For each of these genes, the different sequences are assigned as distinct alleles designated with a digit and, for each isolate, the alleles at the loci define an allelic profile or sequence type (ST). Each isolate is thus characterized by a series of integers which correspond to the alleles at the loci, and the discriminatory power can be altered by selecting different numbers of loci. Several public databases for sharing MLST data have been established, e.g., pubmlst.org and mlst.net. MLST using seven housekeeping genes of *F. tularensis* has been shown to distinguish the subspecies, but not to separate individual isolates (129, 131) (Paper I).

For *F. tularensis*, multi locus variable number of tandem repeat (VNTR) analysis (MLVA) is the only typing method so far that has enabled discrimination of isolates (88, 107). MLVA measures the variations in copy number of short repetitive sequences located in tandem at several loci in the genome. Some of these loci are highly variable and are considered to be the fastest mutating parts of the DNA. However, the high rates at which MLVA loci mutate may cause homoplasy effects, i.e. isolates may share mutational changes for reasons other than common ancestry, with associated risks of incorrect isolate affiliation.

In work on *Yersinia pestis*, *Bacillus anthracis* and *Francisella tularensis*, the issue of homoplasy effects of VNTRs has been addressed by whole genome single-nucleotide polymorphisms (SNPs) analysis, for defining genetic lineages (89, 132-134). Compared to tandem repeats, SNPs exhibit slow mutation rates, making them valuable for phylogenetic analysis (135), as long as the discovery bias is taken into consideration (136). Since all three mentioned species are genetically monomorphic and their population structure is highly clonal, the mutations are mainly inherited by vertical
descent, and the direction of evolution can thus easily be followed. Once an accurate phylogenetic tree structure has been defined using the SNPs from the whole genome SNP analysis, canonical SNPs, i.e. SNPs that serve as evolutionary guides, can be selected that define each branch in the phylogeny, and can be included in a typing assay to explore the evolutionary structure of the species.

For both phylogenetic robustness and discrimination of isolates, a two-step procedure has been suggested, including assays of canonical SNPs (providing robustness) and MLVA (providing high-resolution) (89, 133). A limitation of the procedure is that it involves two assays, thus increasing time and cost. Therefore, instead of using SNPs, we developed an assay based on MLVA and INDELs (Paper II) (137). INDELs have canonical properties in Francisella, and similar mutation rates and typing resolution to SNPs. A practical advantage is that one assay (Fragment Analysis) can be used for both INDEL analysis and MLVA.

However, since Fragment Analysis can only be performed in specialized laboratories, we also developed a robust real-time PCR assay based on both canonical SNPs and INDELs for robust and rapid analysis of the hierarchical population structure of Francisella, that could also be used in routine clinical laboratories, in which real-time PCR methods are now widely used (Paper III). Real-time PCR assays are rapid and flexible diagnostic tools, as exemplified in Paper III, in that phylogenetic and geographical resolution can be altered by adding or removing genetic markers, which cannot generally be done when using conventional PCR methods. Real-time PCR assays also appear superior to conventional PCR in sensitivity, and enable many PCR-reactions to be run in parallel. Several real-time PCR assays have been developed for Francisella, that have not taken the canonical properties of markers into account and often target specific genes, leading to a similar typing resolution to that achieved by conventional PCR, but which are more rapid (45-50).

Microarray-based comparative genomic hybridization (CGH) is a technique that can be used to analyze intraspecies genetic diversity at the whole-genome level. In this method, the genome of one bacterial isolate usually serves as a reference to which other isolates are compared. The reference genome provides a template for generating DNA sequence probes that are attached to a microarray, then DNA from other isolates is extracted and hybridized to the probes on the microarray. In theory, only DNA probes that match between the reference and test isolate will give a signal. The method is powerful since it measures the genomic diversity but is labor-intensive and has some technical limitations mostly related to problems with the detection of differential hybridization signals. A CGH microarray including 27 isolates representing F. novicida and the three subspecies of F. tularensis revealed eight larger Regions of Difference (RDs) (138), that were further examined in Paper I in this thesis (131).

Genome sequencing is the ultimate identification method for an isolate since it gives the complete and unique DNA sequence. However, not all bacteria and viruses can be cultured, and the widely used Sanger sequencing method (139, 140) is very time-consuming, labor-intensive and costly. However, in recent years, new techniques
for DNA sequencing have been developed and changed this situation dramatically. Five years ago, one genome could cost 10,000 US dollars to sequence, while today the cost has fallen to approximately 2,000 to 4,000 US dollars per genome of the size of *Francisella*, and the cost continues to decrease. The first genome of *Francisella* was published in 2005 (141), and to date 22 genomes have been sequenced, completely or partially, and posted to the US National Institute of Health (NIH) genetic sequence database GenBank. There are also several other ongoing genome projects. The increased numbers of genome sequences will further facilitate the typing of new isolates and improving characterization of the phylogenetic structure and geographic distribution of *Francisella*. 
Genetic methods for identifying *Francisella* in clinical practice

Rapid and reliable detection of *F. tularensis* in clinical specimens has a high diagnostic value. *F. tularensis* causes a severe disease and appropriate treatment with antibiotics is needed instantly to alleviate the severity of the disease. The bacterium is not part of the normal human flora, and detection of the bacteria in clinical specimens is therefore always of significant value. *F. tularensis* is considered to be a potential biological warfare agent and a laboratory hazard, hence bio-safety level 3 laboratory procedures should be used, i.e. immunized staff, protective clothing, safety cabinets with laminar flow, negative pressure ventilation with high efficiency particulate filters, and an autoclave for sterilizing waste prior to disposal (41). Culturing the bacteria is thus often avoided. Diagnosis of tularemia by sero-diagnostic techniques takes weeks and culturing may require an incubation period of seven days. Rapid diagnosis using non-viable bacteria by detecting its DNA thus seems to be an attractive alternative or complement.

However, there are problems with PCR-based assays that should be taken into consideration when introducing them in clinical diagnostics. Notably, high sensitivity may cause problems with false-positive results, due to contamination of the sample with DNA from previous runs or the surroundings, and false-negative results may occur, due to a low sensitivity of the assay or presence of PCR inhibitors in the sample. A positive control amplifying a human DNA target should therefore be used to detect PCR inhibition, if present.

Validation of the typing method should be performed, including assessment of typeability, reproducibility, stability, and discriminatory power (142). The assay should be able to assign the same type to an isolate tested in independent assays, be able to determine the relationships between isolates, and the test collection of isolates used in the evaluation of the assay should reflect as much diversity as possible within the species.

The choice of typing method depends on the aim of the test and organism. The discriminatory power of each method may vary greatly between bacterial species with limited sequence differences, such as *F. tularensis*, and species with extensive sequence differences and genome rearrangements, such as *Campylobacter jejuni* and *Helicobacter pylori* (143, 144). For rapid diagnosis to verify the presence or absence of *Francisella* in a clinical sample, a PCR assay targeting one or a few genes, e.g. *lpnA*, *fopA* and 16S rDNA, could be adequate. In epidemiological investigations where an additional aim is to prevent further infections by tracing the chain of transmission and disease transmission sources, knowledge about the bacterial population structure may be more important. In the case of tularemia, which has sources in natural foci, high-resolution typing of clinical specimens in combination with information about the actual tularemia-transmission places gathered from patient interviews, as exemplified in Paper IV, may increase the understanding of spatial and temporal patterns of human risk of exposure to infection. Knowledge gained through landscape
epidemiology studies could be used to target educational campaigns to people with increased risk of acquiring tularemia, so that they might protect themselves against vectors in certain identified areas during local outbreaks of vector-borne tularemia. In a forensic investigation, where statistically valid evidence is sought, a complete understanding of the bacterial population structure is critical, as exemplified with the anthrax letters sent in 2001, in which typing of canonical INDELs and SNPs in different batches of *B. anthracis* revealed the laboratory where the strains used in the bio-terror attack originated (145).
RESULTS AND DISCUSSION

I. Insertion/deletion mutations (INDELs) flanked by repeats are useful for typing isolates and batches of Francisella (Paper I)

*F. tularensis* isolates from global origins are highly genetically similar. *F. tularensis* is characterized by a clonal population structure and low recombination frequency, which means that changes that arise in the DNA sequence of a bacterium will, to a high degree, be inherited vertically (89). *F. tularensis* is also an intracellular bacterium that seems to have evolved from a more free-living bacterium (146). There appear to have been successive gene losses within *F. tularensis*, which is a characteristic of bacteria that have changed their niche (147, 148), and are, as in this case, in the process of adapting to life within a host cell. Certain pathways, e.g. some amino acid synthesis pathways, become unnecessary for the bacteria to maintain, since the product can instead be obtained from the environment within the host cell. Assuming that *F. tularensis* subspecies have evolved from a common ancestor, identification of genetic differences among isolates might provide information both on the evolution of subspecies, and on functional issues.

The study presented in *Paper I* is based on a multiple-genome microarray study by Broekhuijsen et al. (138), who found that all INDELs within the identified Regions of Difference (RD1-8), resulted from repeat-mediated deletions (exemplified in Fig. 1 in *Paper I*). RD1-7 was flanked by short repeats, while RD8 was flanked by insertion sequence elements, which are longer repetitive sequences commonly associated with recombination events. At the time of this study, the first published *Francisella* genome denoted SCHU S4 had nearly been completed at my institute.

To examine if the presence of spatially close short repeat motifs is a general feature of deletion events within *F. tularensis*, we scanned the unfinished genome sequence of SCHU S4 for repetitive sequences. More than 70 regions were evaluated by PCR in a subset of five *Francisella* isolates representing *F. novicida*, the three subspecies of *F. tularensis*, and *F. tularensis* subsp. *holarctica* isolated in Japan. Regions that exhibited size polymorphism on agarose gels were further analyzed in 45 isolates, and PCR products were sequenced to determine the junctions of the INDEL fragment.

To evaluate the phylogeny derived from the INDELs, we performed a parallel phylogenetic analysis based on single-amino acid polymorphisms within four proteins, and single-nucleotide polymorphisms\(^1\) (SNPs) within seven housekeeping genes, in an MLST-like approach.

In the search for INDELs flanked by repeats, we found 17 Regions of Difference in 45 *F. tularensis* isolates. Nine RDs fulfilled the strict criteria we set up for a

---

\(^1\) Single nucleotide polymorphism (SNP) is synonymous to single-nucleotide variation (SNV) in *Paper I*. 

23
unidirectional deletion event, i.e. that the deletion event was irreversible and inherited vertically. The pattern of deletions suggested that the branching of *F. tularensis* subsp. *tularensis* and *mediasiatica* preceded the branching of *F. tularensis* subsp. *holarctica* from North America and Europe. The Japanese isolates of *F. tularensis* subsp. *holarctica* were found to be intermediates between these extremes. The evolutionary place of *F. novicida* in the tree could not be pin-pointed, due to failure of PCR for three RDs. A complete lack of deletion events and high numbers of SNPs within seven sequenced genes were detected in *F. novicida*, suggesting that *F. novicida* and *F. tularensis* are separate lineages. The proposed evolutionary scenario is graphically illustrated in Figure 2 in *Paper I*.

The position of *F. novicida* was confirmed by the protein phylogeny, which included representatives of *Agrobacterium tumefaciens* (α-proteobacteria) and *Yersinia pestis*, which belong to a different branch of the γ-proteobacteria group from *Francisella*. The tree was completely congruent with the proposed scenario inferred by INDELS (Figure 3 in *Paper I*).

The general order in which the subspecies and isolates have presumably diverged could be outlined by the SNP phylogeny, which had greater resolution than the protein phylogeny, and was rooted in *F. novicida*. We could see here the split of *F. tularensis* subsp. *tularensis* into two branches A.I and A.II, which has been recognized previously by others (88, 107, 125), and a split of *F. tularensis* subsp. *holarctica* into subgroups, confirming the intermediate position of Japanese *holarctica* isolates in the INDEL phylogeny. The phylogenetic tree indicated a common ancestry of *F. tularensis* subsp. *tularensis* and *mediasiatica*, which is in agreement with the INDEL phylogeny and other published analyses (138).

Based on the pattern of deletions and SNPs among isolates, no evidence of recombination across subspecies was detected in the study, implying that the INDELS and SNPs are probably inherited vertically. However, later studies have suggested that *F. novicida* shows signs of recombination to a much higher degree than *F. tularensis*, possibly due to a more free-living style or the occupation of a distinctly separate niche by *F. novicida* (3). In this study (*Paper I*), we only had one isolate of *F. novicida* to examine. Thus, analysis of recombination frequencies within *F. novicida* was not performed.

There are also some functional issues to be discussed regarding the genes affected by repeat-mediated deletions. We found two RDs (RD18 and RD19), also independently identified by Samrakandi et al. (149), that discriminated isolates of *F. tularensis* subsp. *holarctica* and *tularensis*. Furthermore, these two deletions were later found to be the only two large differences between LVS and a virulent counterpart FSC200 (>99.92 % sequence similarity) (150). The INDEL within RD18 affected two genes of unknown function, belonging to a novel protein family, unique to *Francisella*. The INDEL was found in laboratory strains of FSC043, which is an attenuated strain of SCHU S4, and in the live vaccine strain LVS. In a subsequent study, we found that a targeted mutation of one gene (*FTT0918*) in the RD18 region of
the highly virulent isolate SCHU S4, caused attenuation of virulence in mice (151). The biological functions of the two affected genes in the RD18 region are unclear, but another member of the protein family, fslE (FTT0025 in SCHU S4), is suggested to be a siderophore receptor (152, 153), which is a protein involved in iron acquisition from the host cell. The INDEL within RD19 affected three genes of type IV pili, which is a known virulence factor for many bacteria and needed for dispersal (154). Deletion of these genes has occurred in laboratory strains of LVS. Combined restoration of the genes in the RD18 and RD19 regions in LVS by complementation resulted in a fully virulent phenotype (personal communication, Emelie Salomonsson and Kerstin Kuoppa, submitted manuscript), indicating that these two mutations contribute significantly to the observed attenuation of LVS.

LVS gives protective immunity against tularemia, but is currently not accepted for public use partly due to uncertainties regarding the cause of the attenuation, and thus a risk of reversal of the attenuation. If these deletions can be shown in laboratory experiments to be irreversible, restrictions on the use of LVS as a vaccine could be more easily lifted.

While the majority of observed deletions might be regarded as canonical INDELs that have evolved over wide time-scales, RD18 and RD19 seem to represent unidirectional genetic events that have occurred in the laboratory during passage or storage of F. tularensis strains. This is supported by the findings of a mix of two distinct cell populations that exhibit these RDs in laboratory stocks of strains FSC074 (RD19) and FSC158 (RD18). The finding of unique deletion characters among bacterial colonies derived from a single bacterial seed stock suggests that regions flanked by repeats are prone to mutations in the F. tularensis genome.

According to reports in newspapers, similar observations were recently made in the investigation of the anthrax letters that were sent to government officials in the US in 2001 (145). The powder in each letter contained a mixed population of B. anthracis Ames that, when grown on plates, yielded some colonies that were different in texture, color, and size. Twelve such colonies were fully sequenced, and from analysis of the genomes, four INDELs were selected to represent a genetic signature of the attack anthrax. Subsequently, a repository of more than 1,000 samples of the Ames strain that the FBI had collected from laboratories in the United States, Sweden, the United Kingdom and Canada were screened for these characteristic INDELs. Only eight samples had all four mutations, and all eight could be traced to the flask of spores under the main suspect’s charge. Since the observations of INDEL-mutations in laboratory strains of B. anthracis have not yet been published in any scientific journal, interpretations will remain speculative. Judging from public news reports, however, it seems that the mechanisms we observed in F. tularensis are quite similar to the phenomenon that was used by FBI scientists to track the origin of the anthrax attacks.
II. Insertion/deletion mutations (INDELs) not flanked by repeats are robust genetic markers for typing isolates of *Francisella* (Paper II)

In *Paper I* we found that DNA sequence repeats located close to each other on the bacterial chromosome could recombine and give rise to deletions that appeared to be fixed through long-time evolution. We also observed spontaneous deletions during laboratory handling when culturing in artificial media (RD18 and RD19). The latter finding suggests that there is a risk that repeat-mediated deletions may exhibit homoplasy, i.e. that the same deletion event can be generated independently in different bacterial clones, and thus not necessarily be inherited from a common ancestor. This phenomenon can disrupt and confound accurate phylogenetic placement of isolates and prevent correct taxonomic designation. To achieve a robust phylogeny and genetic typing of *Francisella* isolates, INDELs generated from the recombination of two repeats may not therefore be the optimal choice.

For *F. tularensis*, the typing resolution of INDELs is at the level of subspecies. To distinguish isolates, multi-locus variable number of tandem repeat (VNTR) analysis (MLVA) is currently the method of choice. During the last decade, MLVA typing schemes have been developed for several bacteria (155-158). VNTRs are a type of INDELs formed by a mechanism involving slippage of the DNA polymerase during transcription, and are rapidly evolving. VNTR markers can therefore, by chance, have the same number of repeats in isolates that are distantly related. Therefore, MLVA may, like analysis based on INDELs generated from the recombination of two repeats, cause homoplasy effects. It would thus be ideal if MLVA, with its fine-resolution, could be combined with a method that analyses genetic markers evolving at a low rate and which do not show homoplasy, in order to give a phylogenetic analysis or a genetic typing scheme a robust “backbone”.

Single nucleotide polymorphisms (SNPs) are genetic markers that evolve more slowly, and could therefore be combined with MLVA. Multi-locus sequence typing (MLST) schemes are based on the detection of SNPs and have been developed for many bacteria. A SNP can exhibit four states (A, C, G, or T), and can thus cause homoplasy, albeit at a low rate, by random reversion of the SNP. Another disadvantage is that MLVA and MLST require separate assays, thereby increasing the complexity, time and cost of the use of a combined analysis.

In this study, we therefore searched for robust INDELs, which could complement the VNTR markers and could be used in the same assay. We had noticed from our previous study (*Paper I*) that INDELs flanked by short repeats seemed to give no homoplasy among the studied isolates, while longer repeats could give spontaneous deletions, suggesting that separate molecular mechanisms are involved. We reasoned that INDELs that arose by a mechanism other than repeat-mediated excision would give less homoplasy and provide the desired stability to complement MLVA markers in a phylogenetic tree construction.
To identify such INDELs, we used five *Francisella* genomes, representing each subspecies of *F. tularensis* identified at that time. The genomes were automatically aligned and scanned for regions with size polymorphisms. By this method, we identified 280 INDELs: 164 with no or short repeats (≤ 5 nt), 51 with long repeats (>5 nt), and 65 VNTRs. We found a bimodal distribution of INDELs according to repeat size (Figure 2 in **Paper II**), indicating the presence of two separate underlying unimodal distributions. This can be interpreted as an indication that separate molecular mechanisms are responsible for their mutation: one dependent on the presence of repeats, and another that is independent of such repeats. The degree of homoplasy in the data was examined and compared to the overall phylogenetic tree. We estimated the frequency of homoplasy to be lower for the INDEL loci with no or short repeats than for INDEL and VNTR loci, since we found four incongruent loci (not fitting the overall tree) among the latter, while none were found among the others. Collectively, this suggests that INDELs with no or short repeats can function as robust phylogenetic markers in the genotyping of *F. tularensis* isolates.

To evaluate the potential of INDELs with no or short repeats as genotyping markers, we selected 38 INDELs that encompassed the complete set of genotypes identified by the previously described genome alignment operation, and that were physically evenly distributed over the genome of SCHU S4. We then amplified these INDELs together with a set of 25 previously published VNTR markers used in MLVA by PCR, and determined the sizes of the INDELs and the numbers of repeats in 23 *F. tularensis* isolates by the Fragment Analysis method using an automated sequencing machine – a procedure commonly used for MLVA. The isolates were selected to be representative of all previously identified and accepted major genetic groups. The analysis of the resulting genotypes demonstrated that INDELs can provide a phylogenetic structure similar to the structure provided by SNPs, with the advantage that only one laboratory method is needed. Bootstrap analysis showed that the topology inferred using MLVA data was weakly supported at most nodes, while the statistical support increased significantly when MLVA data were used in combination with INDEL data. In addition, combined analysis provided higher resolution for isolates of *F. tularensis* subsp. *holarctica* than any of the individual datasets.

In the study presented in **Paper II**, we amplified each genetic marker individually and used multiple markers bearing the same phylogenetic information, thus the typing resolution was not increased. Future studies should consider multiplexing the PCR and also reducing the number of specific markers for each genotype, to reduce the costs of genotyping and redundancy in the data. The continuously increasing amount of genome sequences also opens up the possibility to include markers that would identify novel genotypes.

The presented approach represents an improvement to the currently widely used MLVA method by combining analysis of the rapidly evolving VNTR mutations with slowly evolving but phylogenetically robust INDEL mutations. It also offers an obvious
advantage over pulsed field gel electrophoresis (PFGE), which is another commonly used genetic typing method for bacteria. PFGE is time-consuming and there is a risk of laboratory-acquired infections. Further, it has reproducibility problems and provides limited typing resolution. None of these issues is a problem in the proposed MLVA-INDEL method.

In conclusion, the developed assay using two types of marker provided better estimates of genetic relatedness among isolates, as well as higher resolution. The inclusion of canonical INDELs was essential to identify correct relationships among isolates at deeper phylogenetic levels.
Combining single-nucleotide polymorphisms (SNPs) and insertion/deletion mutations (INDELs) not flanked by repeats in a hierarchical real-time PCR array gives rapid and robust classification of isolates of Francisella (Paper III)

In the previous Paper (II) a robust genotyping assay using Fragment Analysis on a sequencing machine was presented. This type of equipment is not available in many public health laboratories. Therefore, a rapid and flexible real-time PCR assay with the same robustness as described in Paper II was developed. The assay was designed to be run by a standard 96-well real-time PCR machine, equipment commonly used in modern routine laboratories. The work was based on the observation that mutations in the two clinically relevant subspecies F. tularensis subsp. tularensis (type A) and F. tularensis subsp. holarctica (type B) are mainly inherited vertically.

In contrast to previously developed conventional and real-time PCR assays, which in general target fewer genes with a predetermined typing resolution, we aimed to develop genetic targets representative for each branch in a hierarchical phylogeny of Francisella, including markers to discriminate environmental and weakly pathogenic bacteria from human pathogenic bacteria. Thus, the typing resolution can be altered depending on the purpose of the typing. The geographical resolution can also be altered, since the subspecies and some subclades differ in their geographical distribution.

To find markers with canonical properties, i.e. markers that can serve as evolutionary guides, we compared available Francisella genomes and other DNA sequences published in the US National Institute of Health genetic database GenBank using BLAST and sequence alignment tools. In addition, the literature was scanned for candidate markers previously shown to discriminate subclades of Francisella. Forty-nine single nucleotide polymorphisms (SNPs) and 15 insertion/deletion mutations (INDELs) were tested for typeability and canonical properties in a set of 62 isolates representing maximal genetic and geographical diversity found in our Francisella Strain Collection (FSC). Twenty-six SNPs and four INDELs were discarded due to amplification failure and categorization problems. Two SNPs and one INDEL showed homoplasy and were therefore also excluded.

The final set of markers consisted of 23 SNPs and 11 INDELs with canonical properties. These were arrayed in a hierarchical phylogenetic structure in one plate, to facilitate the categorization of tested isolates. The limit of detection was 100 pg, based on the amount of DNA required to amplify all 34 primer pairs successfully, which was higher than limit amounts for other published real-time PCR assays targeting fewer genes. In a blind test, the final 96-well plate assay correctly categorized six isolates that had been part of the development set of isolates. We could further categorize five isolates from global origins, and isolates from 14 patients in Sweden that acquired
tularemia in 2008 (Figure 12). The categorization of DNA from six clinical ulcer specimens from the same patient group was incomplete due to scarcity of DNA. Nevertheless, four analyzed markers allowed the fine-tuned categorization of these specimens. The assay showed good reproducibility and the failure of classification was low, indicating that the assay was technically robust. Even though failure of amplification occurred for 0.3 to one marker per plate, samples could almost always be classified by reading the results from the other markers included in the plate, which provided the hierarchical context.

An advantage with our hierarchical array is the flexibility to add and remove markers depending on the purpose of the typing, and the desired typing resolution and geographical resolution. An easy-to-use and rapid real-time PCR assay that identifies the exact genotype of Francisella could be of value in several settings. Rapid on-site confirmation of tularemia could be of value in the primary health care system in Scandinavia (the area corresponding to Pavlovsky’s nidal centre number three (Figure 11), since tularemia is endemic in many areas of the Scandinavian countries. This would allow rapid institution of correct treatment. In addition, the results could be matched with previously recorded local genotyping results to detect any unusual occurrences of tularemia that should alert health authorities to initiate an outbreak investigation. As we show in Paper IV, genotyping combined with GIS-mapping of the actual disease-transmission places, revealed a landscape epidemiology of tularemia at two outbreak areas in Sweden. Similarly, the presented assay can be used to refine epidemiological investigations in other areas, by exploring the geographical distribution of disease-causing genotypes, and pin-pointing areas with an increased risk of acquiring tularemia. In a bio-terror scenario, with a stated threat to use F. tularensis, rapid detection of attack strains would be important and, perhaps most importantly, an ability to identify weakly pathogenic Francisella would reduce fear in the community and the unnecessary use of antibiotics.

In summary, the developed real-time PCR array for hierarchical identification and typing of Francisella is flexible in that typing resolution and geographical resolution can be altered depending on the purpose, which cannot be done with other conventional PCRs and real-time PCRs presented to date. The markers identified herein can easily be transferred to other types of technical platforms such as SNaPshot (Single Nucleotide Primer Extension) Assay and the TaqMan SNP Genotyping Assay (Applied Biosystems), which may provide higher sensitivity.
A combined epidemiological and genetic approach, using patient interviews and high-resolution typing of *F. tularensis* isolates, reveals a landscape epidemiology of tularemia (Paper IV)

In the final study (presented in Paper IV) we took genotyping a step further into epidemiology and clinical reality by tracing the landscape origin of strains isolated from tularemia patients.

Traditionally, outbreak investigations of transmissible diseases aim at identifying the disease reservoirs and transmission sources to mitigate an ongoing outbreak and/or prevent future outbreaks. Investigations of tularemia are hampered by a lack of knowledge about sources of pathogenic *Francisella* and the factors that govern the often irregular patterns of outbreaks. Geographical mapping of human tularemia cases on a large geographical scale such as Country or Continental scales has been performed previously but such mapping provides no clues regarding the local geographical sites or areas where *F. tularensis* has been transmitted (or, therefore, where people living in tularemia outbreak areas should be more careful, or possibly protect themselves against tularemia).

We performed an epidemiological investigation of two tularemia outbreaks that occurred between 1995 and 2005 in Sweden and affected 441 humans. In the Municipality of Ljusdal, tularemia has been endemic for several decades, while in the County of Örebro, the disease has recently re-emerged after being virtually absent since early reports in the 1930s. In the investigation, we combined geographical data about the actual places of disease transmission obtained by patient interviews, with genetic analysis of 136 isolates of the etiological agent. We used an 18-marker genotyping system including three types of markers (10 INDELs, six VNTRs, and two SNPs) developed for the purpose, see Paper II, Paper III, and (88).

In Ljusdal, the isolates were genetically homogenous with 49 out of 56 isolates showing an identical genotype, while in Örebro we found two distantly related genetic groups of isolates; group 1 was heterogeneous and composed of 13 genotypes while group 2 was homogeneous and composed of only two genotypes. We noticed that a genetic subgroup (denoted 1d) of group 1 isolates was genetically separated from all other isolates in Örebro by a SNP. This SNP mutation (at marker Ft-SNP1) was also found in 53/56 isolates from Ljusdal (genetic subgroup 1e). Another SNP (Ft-SNP2) separated the Örebro subgroup 1d isolates from those in subgroup 1e in Ljusdal, indicating that these subgroups have evolved separately for quite a long evolutionary time, since SNPs mutate at low rates compared with VNTRs. The Ft-SNP2 mutation was exclusively found among isolates from Ljusdal in the worldwide collection of isolates that was investigated in Paper III (Ft-SNP2 is denoted marker B.22 in Paper III). A mutation at SNP1 (denoted B.21 in Paper III) was found in one isolate from Oulu, Finland, and another from water in the Odessa region, Ukraine in the worldwide
collection. The findings indicate that mutation at SNP2 is local to the Ljusdal 1e subpopulation, which is consistent with the fact that the genome of isolate FSC200 originating from Ljusdal was used for the discovery of Ft-SNP2.

The genetic differences were mirrored in the local geography. Spatial associations were found between subpopulations of *F. tularensis* and the places of disease transmission. We found small areas with high frequencies of tularemia, indicating hitherto unidentified point sources of infection. In both Ljusdal and Örebro, the disease clusters were associated with recreational areas beside water, and genetic subpopulations were present throughout the tularemia season and persisted over years. Genetic groups 1 and 2 in Örebro showed distinct disease occurrence centers and genetic group 1d was the only isolate of group 1 that occurred along the whole stretch of the river Lillån. In Ljusdal, patterns were similar: the three single isolates of genetic subgroups 1b, 1c and 3, respectively, were located in places far from the disease occurrence center of subgroup 1e and from each other, indicating separate disease reservoirs. Interestingly, the single isolate of genetic group 3 was probably acquired at a zoo with wildlife indigenous to Nordic countries, indicating possible introduction into the area by human activities.

There are several scenarios that could explain the distribution of genotypes and clustered occurrence of tularemia, including the following.

Under the first scenario, high genetic diversity and limited spatial distribution of outbreak isolates, as exemplified by genetic group 1 isolates in Örebro, are consequences of a recent *F. tularensis* population expansion. Changes in the climate and/or the local environment, such as the restoration of a wetland area in the disease occurrence center of group 1, may have promoted increased bacterial replication, and diversity may have been generated when the bacteria replicated and the daughter bacteria were dispersed from the original foci by vectors and hosts into new areas favorable for them. The observed distribution patterns of genetic groups mirror such a dispersal process: genetic group 1a has a very limited distribution so far, while 1d is dispersed over a larger area and is predominantly found along the river Lillån. As an alternative explanation for high genetic diversity, it may have already been present in the environment before the increased replication. Small geographically separated populations may have existed in the environment, and random genetic drift may have acted as the main evolutionary force for variation. The observed genotype pattern may then mirror the diversity that has been present in the area at least since the first reports of tularemia in the 1930s.

Under another scenario, *F. tularensis* may not have been present in the environment until just before the outbreak, but were recently introduced into the area by birds or other hosts. In support of this scenario, several genotypes found in Örebro have also been found in other parts of Sweden or in other countries. The observed bacterial diversity may thus mirror the diversity among the introduced populations and their distributions. Many birds stay at the newly restored wetland area, which is artificially flooded every year to make the area attractive to bird life.
In all scenarios, genetic groups 2 and 1e in Ljusdal are present in areas where the bacteria have persisted for a long period of time and selective sweeps in large populations and/or genetic drift in small populations have reduced the original diversity.

We consider that the first scenario, postulating that environmental changes favored a boost in replication of *F. tularensis* populations residing in the environment, is the most likely explanation. However, other explanations may also be possible. We do not yet have the tools available to observe the real diversity of *Francisella* in nature, but fortunately, genetic tools and sampling methods are continuing to improve dramatically.

In conclusion, high-resolution genotyping combined with GIS-mapping of places of disease transmission provided valuable indications of likely places of disease-transmission and the causal genotypes and their distribution areas.
CONCLUSIONS

The highly clonal and genetic monomorphic properties of *F. tularensis* greatly facilitated studies of the genetic genealogy of tularemia, i.e. use of genetic signatures and phylogenetic analyses to establish kinship among *Francisella* isolates and their origins.

Insertion/deletion mutations (INDELs) found in the *F. tularensis* genomes have often resulted from recombination of two spatially close repetitive nucleotide sequences, followed by excision of the DNA between them, leaving a hybrid sequence of the two repeats in the genome of the daughter clone (Paper I).

Analyses of such unidirectional deletion events, and multi-locus sequence typing (MLST) analysis, suggested a phylogeny in which *F. novicida* branched earliest and *F. tularensis* subsp. *holarctica* was, evolutionarily, the youngest of the subspecies of *F. tularensis*. The Japanese *F. tularensis* subsp. *holarctica* branch diverged earlier than other *holarctica* isolates (Paper I).

Repeat-mediated excision of DNA is an important natural attenuating mechanism in *F. tularensis*, since two identified INDELs (RD18 and RD19) identified in Paper I, were later found to attenuate virulent isolates, and to explain the attenuation of the Live Vaccine Strain (LVS).

DNA repeats located close to each other can recombine and give rise to deletions that appear to be fixed through long-term evolution, but may also occur during laboratory passage of *F. tularensis*. This means that the same deletion events observed in different isolates are not necessarily inherited from a common ancestor. This is exemplified by the INDELs RD18 and RD19 that were detected in several laboratory strains lacking a recent common ancestor (Paper I).

Small INDELs in the *F. tularensis* genomes that appear to have been formed by a mechanism other than repeat-mediated excision of DNA, were found to be robust canonical markers for evolutionary analyses. An analysis of such INDELs combined with MLVA provided both phylogenetic robustness and high typing resolution (Paper II).

A real-time PCR assay, including analysis of canonical single nucleotide polymorphisms (SNPs) and INDELs, proved flexible since typing resolution and geographical resolution could be easily altered depending on the purpose of the assay. The results from the developed assay suggest that rapid hierarchical identification and typing of *Francisella* can be performed on clinical specimens and/or isolates from patients in routine laboratories where a real-time PCR machine is available. Genetic markers for discrimination of environmental and/or low pathogenic isolates from human pathogenic isolates were included in the assay (Paper III).
GIS mapping of the places where tularemia was contracted, as stated by the patients, helped to identify local natural foci of tularemia, where there was an increased risk of tularemia-transmission to humans, which may be used for effective infection control in the studied areas. Coupled with analysis of canonical INDELs, canonical SNPs, and VNTRs, the location and geographical distribution of causative genotypes over time could be studied, which can be used to refine epidemiological investigations of tularemia, to understand where the bacteria persist in nature, and to understand bacterial dispersal patterns. Genome sequencing efforts can contribute to the design of genotyping schemes tailored to a specific outbreak investigation (Paper IV).
ERKÄNNANDEN / ACKNOWLEDGMENTS

Under den här vetenskapliga resan så har jag blivit skyldig ett stort TACK till en rad människor som har hjälpt, uppmuntrat och stöttat mig, eller bara varit tålmodig med mig. Det gäller bland andra:

- **Handledarna Anders Johansson och Mats Forsman.** Måste säga att jag som doktorand haft en enorm tur att ha begåvats med två sådana kunniga handledare, som med stort intresse och engagemang har hjälpt mig genom denna process och haft tid och tålmod med mig, trots att herrarna ofta har ett mer än välspäckat eget schema.

- **Malin Granberg och Linda Karlsson** som genom åren hjälpt mig MYCKET med de laborativa delarna. Jag skulle förstås inte ha kunnat disputera utan en fantastiska hjälp.

- **Mona Byström** – alltid glad och uppmuntrande! Tack för hjälp med laborativt arbete och medmänsklighet i svåra stunder.

- **Pär Larsson.** Här flödar tankarna fritt i en stor kokande kittel av inspiration, kunskap, känslor, kreativitet och engagemang. Jag är tacksam för att med jämna mellanrum få ta del av denna brygd och för att du står ut med mina ibland avkylande kommentarer.

- **Daniel E, Andreas S, Per W, Johanna T, Eva L, Anki A, Ulla E, Stina B, Peter M, Anna H, Maria M** (Skyddscentrum), och **Tina B** (numera Länsstyrelsen). Ingen konst att doktorera när man är omgiven av ett sådant expertstöd.

- **Kerstin K och Anna M** som redan vid min första visit på FOI sommaren 1999 tog hand om mig och vägledde mig i labbet. Tack även **Anna-Lena F, Emelie S, Åke F, Solveig L, Yngve Ö**, och **Laila N.**

- Tack till infektionsläkarna **Erik Bäck** och **Henrik Eliasson** i Örebro och **Lennart Berglund** i Ljusdal för gott samarbete; ”Myggjagarna” **Jan Lundström** och **Martina Schäfer** i Uppsala; Umeå Universitetskollegorna **Anders Sjöstedt, Carl, Svenja, Helena, Patrik** m. fl., extra-tack till **Clas Ahlm** för att du ställde upp som examinerator; Tack till antibiotikaresistensjägarna **Stina Englund** och **Maria Finn** vid Statens Veterinärmedicinska Anstalt; Thanks to the DSTL-colleagues in the United Kingdom, especially **Helen Diaper** and **Steve Mitchell**, for good collaboration in works regarding genomics and microarrays. Thanks **Vera Neubauerova** in the Czech Republic for collaborative work on real-time PCR methods.

- **Tina Sehlstedt** - min gamla studiekamrat från Teknisk Biologi. Tänk att du skulle bli min chef! Det var otippat när vi kämpade med kemiska formler och metaboliska vägar hos Stålhult i Obbola.

- Sekreterarna som får det byråkratiska administrativa att fungera så bra som det gör: **Ondina Plasencia** (FOI) och **Gunborg Eriksson** (Umeå Universitet).
Jag vill också tacka anhöriga och vänner, som får mig att tänka på annat än mitt arbete, bland andra:

- **Katarina Persson** som jag ”träffade” på skjutbanan i Sävar. Tack för att du är min vän! Snart dags att öva upp skyttet igen inför hösten. Ska vi träna lite mer på björnfiguren i är kanske? Älgar är ju ingen match längre, eller hur?

- **FDGJ dvs före detta grannj.....** i Signhildsback alias Gräddhyllan i Tavelsjö: Saknar er! En underbar plats att bo på och en trevlig tid i mitt liv. Vi får ordna en altanfest med springkubb i Rödåsel i sommar. Tyvärr har vi ingen badtunna, men det är väl kanske tur det...

- **Anna Kirk**. Ungdomsvänn som för alltid är min vän!

- **FOI skytteförening.** Tack för att jag fick förtroendet att vara ordförande för denna härliga församling under tre års tid. Det var mycket givande och ”kul”. En eloge speciellt till **Lars H** och **Jan B** som med sin erfarenhet och intresse för skytte hjälpte mig att hålla skutan uppe.

- **Jaktlaget i Granliden.** Vilket härligt gäng! Kamrerer och livsfilosofer blandat med brandmän och timerbils/skotar/väghyvelsförare. Alltid rå men hjältlig jargong! Fick dock tyst på det eviga tuggandet om Bomskjutar-Kerstin och Spårspaps-passet till slut när jag lade ner en niotaggare för två år sedan. Och Långman, **Edvard Lindström**, förstås: denne vitale 96-åring som slutade att jaga för några år sedan och började att basta med oss istället efter att ha haft uppehåll med bastandet sedan 1947...

- **Tavelsjö Hembygdsförening** som bidragit till att jag känner fler döda än levande i bygden jag flyttade till för tre år sedan. Tack för förtroendet att driva projektet med att dokumentera gamla kyrkogården i Tavelsjö i vintras och att skapa en hemsida för det. Det var en otroligt spännande och rolig utmaning! Tack **Carina Oscarsson** för ditt ivriga letande i kyrkböcker, härliga inställning till livet och alla trevliga pratstunder över en kaffekopp. Förhoppningsvis ökar andelen levande bekanta i bygden med tiden.
- **Släkt- och hembygdsforskarna.** Denna hobby som fått mig till att leta igen släktingar i Kanadas vildmark och bland Norges fjordar och sedermera till att i mitt yrke fastställa släktskap och leta ursprung för harpestbakterier. Många är ni som jag har haft kontakt med genom åren, ni vet vilka ni är.

- Mina gudbarn **Ester** och **Erik Huss** i Boxholm som jag träffar allt för sällan, men som finns i mitt hjärta.

- Utan all uppmuntran och stöd från **min familj** hade jag förstås inte kommit till denna punkt i mitt liv. **Mamma Gudrun**, du finns hos mig i mina tankar när jag nu tar min doktorsexamen. Vet att du skulle ha varit stolt över mig! **Pappa Allan, Margareta**, storebror **Jan** och **Åsa** och taxen **Frode**, storebror **Anders** och **Karin** och lilla **Bente**: Då ä sô gò`rgò`htt ö våhht ahtt do finns ö å`hltti ställå öhpp nör jë behö´ vå ehck. Jå hâ`ppås ahtt mo få rå`kes ö râmse lite mër no nör dä hånn ä få`låht, sô jë ihht glömå å` vüss n tô`lô süm fölk. (Tack **Jon Olof Olofsson** för korrekturläsning).

- Svärfar **Anders** och svärmor **Marianne** och övriga i **Lindström- och Lövgren-klanerna**: Känner mig alltid välkommen i ert sällskap. Ni är som en extrafamilj för mig!

- Mest av allt vill jag förstås tacka **min ÄLSKADE sambo Mats**. Tänk att en så’n härlig man som Du fanns därute i sajberryymden och väntade på lilla mig! Tack för att du gav mig tid att genomföra detta arbete. Hoppas på ett fortsatt lyckligt liv tillsammans med dig (och **Tarja**-vovven förstås)! Puss och kram!

- Tack också till övriga kollegor, fd kollegor, samarbetspartners, Tännäsingar och Svanabyare, vänner och släktingar och andra som jag råkat glömma att nämna vid namn här.

Arbetet har finansierats av Försvarsdepartementet, FOI projekt A4854 B-analys, samt av Utrikesdepartementet, FOI projekt A4952 Verifikationsmetodik B, grundkompetens.

Tack till FOI och avdelningscheferna Åke Sellström och Anders Norqvist för uppmuntran och tillåtelse att genomföra doktorandstudierna.

This work has been funded by the Swedish Ministry of Defence, FOI project no. A4854 and the Swedish Ministry of Foreign Affairs, FOI project no. A4952.
REFERENCES

1. Darwin CR. First Notebook on Transmutation of Species 1837.


<table>
<thead>
<tr>
<th>Reference</th>
<th>Genetic Genealogy and Epidemiology of <em>Francisella</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>60.</td>
<td>von Linné C. Systema naturae: sive regna tria naturae systematice proposita per classes, ordines, genera, &amp; species. 1 ed; 1735.</td>
</tr>
<tr>
<td>65.</td>
<td>Mendel JG. Versuche über Pflanzenhybriden Verhandlungen des naturforschenden Vereines in Brünn; 1866.</td>
</tr>
<tr>
<td>72.</td>
<td>Tow D. Future Darwinism - The Web of Life AustraliaTO.</td>
</tr>
</tbody>
</table>
| 79.       | Hollis DG, Weaver RE, Steigerwalt AG, Wenger JD, Moss CW, Brenner DJ. *Francisella philomiragia* comb. nov. (formerly *Yersinia philomiragia*) and


117. CIDRAP News 2005: Tularemia agent found in DC air, but no cases seen. 2005 [cited 2009 26/04]; Available from: http://www.cidrap.umn.edu/


