Francisella tularensis: persistence, dissemination and source attribution

A theoretical and computational approach

Chinmay Dwibedi

Department of Clinical Microbiology
Laboratory for Molecular Infection Medicine Sweden (MIMS)
Umeå University
Umeå 2019
To Ma, Bapa, Chinki, Roshan and Meenu!
Table of Contents

Abstract .................................................. i
Enkel sammanfattning på svenska ................iii
List of publications .................................. v
Abbreviations ........................................... vi

Introduction .............................................. 1
Microbiology and disease in animals .......... 3
Population genetics and genomics .......... 4
Genome architecture ............................... 4
Population structure ................................. 6
Factors affecting genetic diversity .............. 7
Epidemiology and Ecology ......................... 9
Genomic epidemiology and bacterial source attribution 11

Aims ....................................................... 13

Methods and methodological considerations .... 14
Study population ....................................... 14
Sequencing platform and assembly ............... 15
Multiple sequence alignment .................... 17
Detecting homologous recombination .......... 19
Phylogenetic reconstruction ....................... 20
Estimation of substitution rate .................... 23
Association of genomic diversity with spatial distance 25
Laboratory enrichment of mutations ............ 25

Result and discussion ................................ 27

Conclusion and future work ...................... 35

Acknowledgement .................................... 38
References ............................................. 41

Appendix
Paper I
Paper II
Paper III
Abstract

The bacterium *Francisella tularensis* causing tularemia in humans and other mammals displays little genetic diversity among genomes across temporal and spatial scales. *F. tularensis* infects humans with an extremely low infectious dose and causes natural seasonal tularemia outbreaks. During the Cold War, this bacterium was developed as a biological weapon.

In paper I, we aimed at investigating the genetic diversity of *F. tularensis* over space and time and were especially interested in the influence of spatial dispersal on the genetic diversity. By analyses of single-nucleotide polymorphisms (SNPs) among 205 *F. tularensis* genomes, we found that tularemia had moved from East to West over the European continent by dispersal patterns characterized by multiple long-range dispersal events. Evolutionary rate estimates based on the year of bacterial isolation from 1947 to 2012 indicated non-measurable rates. In outbreak areas with multiple recent outbreaks, however, there was a measurable rate of 0.4 SNPs/genome/year indicating that in areas with more intense disease activity, there is a detectable evolutionary rate. The findings suggest that long-range geographical dispersal events and mostly very low evolutionary rates are important factors contributing to a very low genetic diversity of *F. tularensis* populations.

In paper II, we focused on a geographically restricted area with a history of frequent tularemia outbreaks to study *F. tularensis* persistence. By analyzing *F. tularensis* genomes from 138 individuals infected from 1994 to 2010 in Örebro County in Sweden and performing a long-term laboratory storage experiment, we explored the microbial population concept of a pathogen seed-bank. We found that eight indistinguishable genomes – each of them defined by no SNPs across 1.65 million whole-genome nucleotides – locally persisted over 2–9 years. We found unmeasurable SNP accumulation rates and overlapping
bacterial generations among the outbreak genomes and that \textit{F. tularensis} survived in saline for four years without nutrients. By these findings, and analyses of nucleotide substitution patterns, we suggest that a pathogen seed-bank effect is an important feature of \textit{F. tularensis} ecology influencing genetic diversity.

In paper III, we developed a new concept for source attribution of a \textit{F. tularensis} sample. We aimed to identify genetic variation that is characteristic to laboratory culturing and we used culture amplification to identify genetic variation present at exceedingly low frequencies in a sample. Based on a biological enrichment scheme followed by high-throughput sequencing, we could track genetic variation back to a source sample. These results suggest that the concept has potential for linking a \textit{F. tularensis} sample to its laboratory source sample.

Taken together, the results presented in this thesis provide new understanding of the dissemination patterns and local persistence of tularemia. This is important for the interpretation of molecular epidemiology investigations of the disease. In a wider context, the results demonstrate how spatial dispersal and a microbial seed-bank effect may contribute to the diversity of a disease-causing agent. Finally, we have described a promising concept for source attribution of \textit{F. tularensis} samples.
Enkel sammanfattning på svenska


I artikel II fokuserade vi på ett mindre geografiskt område där det förekommit många utbrott av tularemia. Vi ville studera *F. tularensis* persistens (kvarstannande) från år till år. Vi analyserade *F. tularensis* hela arvsmassa i bakterier som odlats från 138 individer infekterade mellan 1994 och 2010 i Örebro län i Sverige. Vi gjorde också ett långtidsexperiment med laboratorieförvaring av *F. tularensis*. Syftet var att undersöka om bakterierna ungefär som trädfrön kan sparas år till år i en
”mikrobiell fröbank”. Vi fann åtta oskiljbara F. tularensis-arvsmassor – där fanns inga SNPs bland 1,65 miljoner nukleotider – som återkom lokalt under 2-9 år. Vi kunde inte uppmäta någon evolutionshastighet i området och vi fann att bakteriegenerationerna överlappade varandra vid släkttskapsanalys. Vi fann också att F. tularensis överlevde i saltlösning i fyra år utan näringsämnen. Baserat på dessa fynd föreslår vi att det finns en effekt ungefär som en fröbank för F. tularensis där den genetiska mångfalden som kan orsaka infektionsutbrott består av både nya varianter av bakterien och sådana som varit vilande sedan lång tid innan de åter aktiveras och orsakar infektion.


List of publications


III. **C. Dwibedi**, P. Larsson, P. Lindgren, K. Myrtennäs, M. Granberg, E. Larsson, C. Öhrman, A. Sjödin, P. Stenberg, J. Ahlinder, M. Forsman and A. Johansson, Biological amplification of low frequency mutations for source attribution of *Francisella tularensis* (Manuscript)
# Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEAST</td>
<td>Bayesian Evolutionary Analysis Sampling Trees</td>
</tr>
<tr>
<td>Beauti</td>
<td>Bayesian Evolutionary Analysis Utility</td>
</tr>
<tr>
<td>CanSNP</td>
<td>Canonical Single Nucleotide Polymorphism</td>
</tr>
<tr>
<td>ECDC</td>
<td>European Centre for Disease Prevention and Control</td>
</tr>
<tr>
<td>Indel</td>
<td>Insertion and/or deletion</td>
</tr>
<tr>
<td>LRT</td>
<td>Likelihood Ratio Test</td>
</tr>
<tr>
<td>MCC</td>
<td>Maximum Clade credibility</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum Likelihood</td>
</tr>
<tr>
<td>MLVA</td>
<td>Multiple-Locus Variable number tandem repeat Analysis</td>
</tr>
<tr>
<td>MP</td>
<td>Maximum Parsimony</td>
</tr>
<tr>
<td>NJ</td>
<td>Neighbor Joining</td>
</tr>
<tr>
<td>SDI</td>
<td>Spatial Distribution Independence</td>
</tr>
<tr>
<td>UMI</td>
<td>Unique Molecular Identifiers</td>
</tr>
<tr>
<td>VBNC</td>
<td>Viable But Not Cultivable</td>
</tr>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
</tr>
</tbody>
</table>
**Introduction**

Tularemia is a zoonotic infectious disease caused by the facultative intracellular bacterium *Francisella tularensis*. It represents one of the most infectious bacterial pathogens known, with a dose of just ten organisms being sufficient to cause a human infection Saslaw, Eigelsbach et al. (1961). Depending on the pathogen subtype and route of infection, mortality rates of tularemia in humans can vary dramatically. If infected by an *F. tularensis* strain of the subspecies *tularensis* by the respiratory route, without prompt and appropriate antibiotic therapy, the mortality may reach as high as 30 % - 60 % (Dennis, Inglesby et al. 2001). For reasons which include these aforementioned characteristics, *F. tularensis* was developed as a biological weapon during the Cold War by both the United States and the former Soviet Union and is currently considered as one of the most feared potential bio threat organisms (Dennis, Inglesby et al. 2001, Sjöstedt 2007). Aside its potential for nefarious misuse, natural seasonal tularemia outbreaks continue to pose public health risk in the Scandinavian countries, in e.g. Central and Eastern Europe (Tärnvik 2007) and in the South-central and Northeastern United States (https://www.cdc.gov/tularemia/statistics/index.html). In these natural outbreak areas, there is pattern of disease recurrence suggesting there exists an unknown environmental reservoir of *F. tularensis*.

It is yet unclear what constitutes the primary reservoir(s) of *F. tularensis* bacteria in the environment and it has been reported that the bacteria can survive in the environment for long periods of time (Parker, Steinhaus et al. 1951, Pomanskaia 1957, Pavlovsky 1966). *F. tularensis* appear to have an unusually complex life cycle for a pathogen as it can be found in environmental samples including water and mud, in various arthropods that may serve as disease vectors and in a broad range of vertebrates including
mammals (Champion, Zeng et al. 2009, Desvars, Furberg et al. 2015). It is therefore likely that the population dynamics of *F. tularensis* is much influenced by multiple ecological factors. In this work, we have aimed for a synthesis of bacterial genomics, epidemiological investigations, evolutionary theory and ecological concepts to better understand genetic diversity and population dynamics of *F. tularensis*. In paper I we combined genomic and geographic information of *F. tularensis* genomes to define a model of dissemination of the bacteria in Europe. In the paper II we employed a reverse genomic epidemiology approach by combining epidemiology and genomics data of an extensive sample in an endemic region to address local persistence of disease (figure 1).

**Figure 1:** A schematic illustration of the reverse genomic epidemiology approach used in this work in relation to the more common genomic epidemiology approach which is often in use to strengthen epidemiological investigations.
Previous studies have shown that *F. tularensis* has a monomorphic population structure making it difficult to distinguish strains. The minimal genetic variation among strains makes it challenging to use genetic data for tracking disease transmission and to trace the laboratory source of a suspected bioterror release of *F. tularensis*. In paper III, we combined experimental work and evolutionary theory to put forward a concept of biological amplification of low frequency mutations and high throughput sequencing to distinguish lab-cultured strains from wild ones and to link strains to its laboratory source.

**Microbiology and disease in mammals**

In the laboratory *F. tularensis* organisms are observed as faint staining gram negative rod or coccoid shaped bacteria that range in size between 0.2-0.7 by 0.2 µm(Sjöstedt 2015). They are fastidious requiring cysteine rich media for growth and grow best, albeit slowly, at mesophilic temperatures of 35-37°C(Bernard, Tessier et al. 1994), a likely adaptation to a natural ecological cycle with growth within a warm-blooded host. *F. tularensis* is regarded a facultative intracellular bacterium and can readily survive and replicate in a variety of different host cell types. Macrophages are considered the primary mammalian host cells and these cells are often used as laboratory models of infection (Clemens, Lee et al. 2004). Other laboratory models for *F. tularensis* replication include *Drosophila melanogaster*-derived S2 cells (Ozanic, Marecic et al. 2015) and the protozoan species *Acanthamoeba castellani*(Abd, Johansson et al. 2003)

Taxonomically, *F. tularensis* belongs to the class γ-Proteobacteria and is further subdivided into four subspecies, which vary distinctly in geographical distribution and virulence. Two subspecies are of clinical importance; the more virulent *F. tularensis subsp. tularensis* or “type A”, which is exclusively found in North America, and *F. tularensis subsp. holarctica* or...
“type B”, which is widely distributed across the northern hemisphere but comparatively less virulent. Bacteria of both subspecies cause tularemia in humans but subsp. *tularensis* causes a more fulminant disease. The remaining subspecies are *F. tularensis* subsp. *mediasiatica* and *F. tularensis* subsp. *novicida*. Organisms of the subspecies *mediasiatica* have only been isolated in republics of Central Asia and are not known to cause natural infections in human populations (Johansson, Farlow et al. 2004, Kingry and Petersen 2014) The subsp. *novicida* strains are of lower virulence and commonly used as a model organism for the more virulent subspecies of *F. tularensis* (Champion, Zeng et al. 2009).

In mammals infected by tularemia, the primary organs that are affected are lymph nodes, spleen and liver. *F. tularensis* doesn’t release any toxins to target the host tissues. A main pathogenic feature of *F. tularensis* is the ability of to suppress the immune system of the host (Tärnvik 2007). Some *Francisella* species *F. noatunensis*, *F. endociliophora* and *F. adeliensis* are pathogens among fish (Sjodin, Svensson et al. 2012, Vallesi, Sjodin et al. 2018). The isolation of these new species has improved our understanding of the symbiotic lifestyle of several species of the *Francisella* genus and the evolution of pathogenicity in this genus. With the advancement of molecular methods, several francisella like endosymbionts have been detected in in ticks for example, but little is known about their ability to infect humans (Scoles 2004, Barns, Grow et al. 2005, Escudero, Toledo et al. 2008).

**Population genetics and genomics**

**Genome architecture**

The *F. tularensis* genome is about 1.9 million base pairs with a very high AT content (~68%), a characteristic of small size bacteria (Mira, Ochman et al. 2001). It has also been observed that *F. tularensis* is going through a reductive evolution as signified by a
remarkably high number of pseudogenes and dysfunctional metabolic pathways (Larsson, Oyston et al. 2005). There has also been marked increase the IS elements in the *F. tularensis* genomes (figure 2). Such increase in IS elements is also considered a sign of an evolutionary process with loss of functionality of genes. It has been suggested that *F. tularensis* has transitioned from a free-living bacterium to a host-associated pathogen in recent history (Larsson, Elfsmark et al. 2009).

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

**Population structure**

Previous studies have indicated *F. tularensis* is a genetically homogeneous species (Johansson, Farlow et al. 2004, Vogler, Birdsell et al. 2009). The inheritance patterns of single-nucleotide polymorphisms (SNPs) and allele frequencies suggest a strict clonal inheritance pattern with no recombination events detected so far. Bacteria are known to evolve either by accumulating de novo variants and/or by taking up foreign genetic elements at different rates of homologous recombination. Some bacteria such as *F. tularensis*, *M. tuberculosis*, *B. anthracis* and *S. aureus* are at one end of the spectrum, being canonical examples of highly clonal species whereas other bacteria like *H. pylori*, *S. pneumonia* and *N. meningitidis* undergo a highly variable inheritance pattern with clear signs of frequent horizontal transfers (Israel, Salama et al. 2001, Achtman 2008, Hanage 2016).

The lack of diversity based on several molecular methods within *F. tularensis* subsp *holarctica* indicate that the bacteria perhaps evolved recently as a result of genetic bottleneck and went through a global expansion(Vogler, Birdsell et al. 2009). With the advent of genetic and high throughput genomic approaches, our understanding of *F. tularensis* and its natural populations has increased considerably during recent years. Genetic analysis of multiple-locus variable number tandem repeat analysis (MLVA) (Johansson, Farlow et al. 2004) has demonstrated structuring of strains into clades with loose geographical associations(Svensson, Back et al. 2009). However, the limitations of the typing-based techniques coupled with the extremely high monomorphic nature of *Francisella* meant that limited strain level resolution was achieved.
The arrival of extensive genomic datasets has provided further evidence of a strictly clonal population structure within the human pathogenic *F. tularensis* genetic lineages of the subspecies *tularensis, mediasiatica,* and *holarctica*. The subspecies *novicida* however, has appeared different in this respect and behaves like other environmental *Francisella* lineages by demonstrating evidence of more frequent recombination events. The extreme clonal nature of the human pathogenic *F. tularensis* genetic lineages means it is difficult to pick up low-level signal of recombination. The inability to detect or the absence of low level recombination appears therefore to provide a distinctive feature to the human pathogenic *F. tularensis* (Larsson, Elfsmark et al. 2009, Vogler, Birdsell et al. 2009)

**Factors affecting genetic diversity**

Bacterial genomes, influenced by several factors, undergo changes continuously (Robinson 2010). Mutations appear by random during replication and the majority is removed because they are deleterious to the bacteria, i.e. the bacteria are subject to an overall negative selection pressure. Mutations are primarily of three types a) SNPs or point mutations, b) insertion or deletion of nucleotide segments, c) structural variants, i.e. large duplication, deletion, or transposition events. Factors such as the rate of *de novo* accumulation of mutations arising per population doubling, the frequency of population doubling per unit time, the number of mutations arising during a possible non-replicative lifestage, and the probability of their fixation determines the evolutionary change in bacterial populations. Mutations that are beneficial for the bacteria get fixed in the genome whereas deleterious mutations and slightly deleterious mutations are removed over time. The rate at which deleterious mutations are removed in the population depends on the selection (Rocha, Smith et al. 2006). With higher selection pressure in the event of a) high competition, b) unfavorable environmental conditions c) large population or d) external factors like dispersal and spatial variation, the deleterious
mutations may be removed faster. However this phenomenon has been a major debate among the selectionists and neutralists in the 70s and 80s on how the deleterious or slightly deleterious mutations were removed. The selectionists argue that the selection pressure is the primary driver while the neutralists believe neutral mutations (mutations that are neither beneficial nor detrimental to the bacterium) are primarily removed or fixed by genetic drift effects caused by e.g. population bottlenecks (Nei 2005). The debate is since ongoing and there are proponents of both the theories that continue to debate. The role of the effective population size is another debated area. The traditional view is that larger populations are under stricter selection pressure as compared to smaller ones but some recent studies have challenged the paradigm. (Lanfear, Kokko et al. 2014). The human pathogenic clade of *F. tularensis* demonstrates distinctly different levels of purifying selection (evident from levels of non-synonymous mutations, but also expansion of insertion sequence elements) as compared to its environmental counterparts, where the human pathogens appear to have evolved under less stringent selection conditions (Larsson, Elfsmark et al. 2009)

Bacterial populations also get influenced by several other external factors (Jones and Lennon 2010). A spatial effect is one such factor because bacteria tend to adapt to local conditions and this may result in geographically isolated populations. Import of genetic diversity may occur as a result of local or long range dispersal events increasing this diversity. Another factor affecting genetic diversity is that some bacteria tend to pause the replication cycle and enter into a low metabolic activity phase to tide over adverse conditions. In some cases it has been proposed that the bacteria can be maintained in reservoirs like seeds in a seed bank, a theory well studied in plant ecology (Shoemaker and Lennon 2018).

The rate of evolutionary change in *F. tularensis* is so far not established and this makes the interpretation of molecular epidemiology studies of *F. tularensis* challenging. Recent
publications based on whole-genome data of relatively few genomes from interspersed outbreak areas showed strange patterns with either extremely similar strains described as clones or presence of multiple genetic lineages over large outbreak areas (Johansson, Larkeryd et al. 2014, Kilic, Birdsell et al. 2015)

**Epidemiology and ecology**

Tularemia is often described as a disease limited to the northern hemisphere but data that are more recent suggest this is not entirely true. The large majority of cases among humans and animals are still reported from the Euro-Siberian region including Russia, Finland and Sweden, from Turkey, and from the United States. Reports that are more recent, however, suggest that tularemia infections historically have been under reported in other geographies. Studies from China and West-Asia, e.g., have contributed with case reports and studies of *F. tularensis* strains from these areas that have increased the known genetic diversity of *F. tularensis* (Wang, Peng et al. 2014, Zargar, Maurin et al. 2015). The cases reported by surveillance systems set up in countries on the northern Hemisphere typically show irregular outbreak patterns with large variation year to year. This can be observed in the ECDC reports of tularemia in European countries (Figure 3).

A distinguishing characteristic of *F. tularensis* is its ability to infect an exceptionally wide range vertebrate and invertebrate species (>250) (Keim, Johansson et al. 2007). Transmission to mammal hosts is known to occur by arthropod bites, including bites of ticks, mosquitos and flies, inhalation of aerosolized *F. tularensis* organisms, and ingestion of infected water (Johansson, Larkeryd et al. 2014, Kilic, Birdsell et al. 2015). Despite the exceptional infectivity of the etiological agent, no human to human transmission has been established so far. Infection of humans thus aborts the transmission chain and humans appear to be accidental hosts.
Despite extensive knowledge on the hosts, it is not very well known if and how *F. tularensis* survives or maintained in nature outside of its hosts. *F. tularensis* bacteria has been found to survive in the environment at low temperatures for extended periods of time and also enter a viable but not cultivable state (VBNC), from which it however has not yet been possible to resuscitate the organisms (Forsman, Henningson et al. 2000). In the last few decades *F. tularensis* was believed to be maintained within ticks and a variety of rodents although the host range was determined to be very wide, from small mammals, aves to that of the non-human primates. Several amphibians have also been shown to be associated with *F. tularensis*. Increase in rodent populations has been shown be associated with a rise in tularemia cases in outbreaks in Spain (Rodriguez-Pastor, Escudero et al. 2017). Recent studies have also shown correlation of tularemia cases with geographical proximity to natural water bodies (Desvars, Furberg et al. 2015).
**Figure 3:** ECDC surveillance data shows variation over time in outbreak patterns of tularemia across countries in Europe. Here illustrated by contrasting figures from Scandinavia in the years 2015 (upper panel) and 2016 (lower). Source: European Center for Disease Prevention and Control.

**Genomic epidemiology and bacterial source attribution**

Microbial genomics is a powerful tool to perform high-level strain resolution and has been successfully employed to study the transmission patterns of infectious agents both in hospital environments and settings with spread of disease in the community. A recent example is that genomic epidemiology studies were performed to understand spread of Zika and Ebola viruses (Gire, Goba et al. 2014, Faria, Azevedo et al. 2016). The approach has also been used to reconstruct ancient divergence spread patterns of diseases including plague and anthrax among others (Keim, Van Ert et al. 2004, Wagner, Klunk et al. 2014).
The microbial genomics principle has also been employed in microbial forensics for clone tracking and source attribution. When the anthrax spores were sent out in postal letters, post the 9/11 attacks in the US, forensic experts turned to genomics to mitigate the challenge. The incident infamously referred to as “Amerithrax investigation” saw the potential of advancements in genome sequencing techniques to aid in microbial forensic investigations for the first time (Rasko, Worsham et al. 2011). Genomic sequencing has improved since, with higher coverage, reduced error rate making it possible to detect low frequency mutations in the genome.

However, there are limitations in resolving monomorphic bacteria like *F. tularensis, Bacillus anthracis* among others, using available genomic analysis (Salk, Schmitt et al. 2018). The lack of diversity makes it difficult to efficiently distinguish genomes and attribute them based on fixed SNP or indel differences. To address the challenge, novel approaches are required to resolve strains of clonal origin. To identify if strains found at a suspected crime scene were previously extensively laboratory cultured there is a need for distinguishing these strains from wild strains. It is known that laboratory grown bacteria like e.g. *Escherichia coli* acquire mutations in a different way than the wild ones because of different selection pressure in the laboratory conditions (Barrick and Lenski 2013, Dragosits and Mattanovich 2013). Therefore, it might be possible to use the patterns of evolution to track the lab induced mutation as a strategy to identify lab culturing.
Aims

1. Can we use population genomics and genetic diversity to infer dispersal barriers, patterns, range and the model of dissemination of *F. tularensis* at larger geographical scales? (paper I)

2. Can we combine epidemiological data, population genomics and spatial distribution of *F. tularensis* in a geographically restricted endemic region to infer factors that are important to local disease ecology? (paper II)

3. Can we differentiate laboratory cultured *F. tularensis* strains from that of wild ones? Can we link bacterial cultures after several population doublings to a source strain? (paper III)
Methodological considerations

Study population selection

In paper I a total of 205 strains were collected. At the time of study design and analysis this represented all the *F. tularensis* strains of the desired genetic group that were available from the study region (Dempsey, Dobson et al. 2007, Vogler, Birdsell et al. 2009). Some of these strains were received from our collaborators in the form of DNA, because as a tier 1 bioweapon agent there are strict restrictions on the maintenance and sharing of strains among labs. Out of the 205 strains, 67 were whole genome sequenced and the remaining 138 were assigned phylogenetic groups by running the CanSNP assays constructed from the novel SNPs identified from this study and from previously available ones. Most of the strains that were not sequenced were available to us at a later time point. We understand and acknowledge that it would have been ideal to sequence all the strains for a high resolution clone tracking. Availability of more genomes from different time points would also have allowed us to perform mutation rate and molecular clock studies.

In paper II the aim was to perform a deep sampling of all tularemia cases in a geographically restricted endemic region over a period of time to look for outbreak dynamics. Therefore we whole genome sequenced 36% of all *F. tularensis* strains reported in the Örebro county of Sweden between 1994 and 2010. Örebro County has been a previously studied endemic region (Svensson, Back et al. 2009). While some environmental samples from the outbreak area would have been of great interest for this work, no such isolates were available to us because it is extremely difficult to isolate and culture the bacteria from extra host environmental material. Another areas of improvement could have included shotgun sequencing directly on biological samples from the infected individuals to provide further insight into possible within
host diversity or to understand if individuals sometimes were infected with multiple clones.

The study in paper III was designed to mimic the production of large batches of bacteria from a single founder population in different batches and to look for differences within the batches that could reliably differentiate them from wild strains and each other. One obvious drawback to our study is the small scale and limited complexity of the experiment. There was however practical and economic difficulties involved in making a more complex design. The laboratory work is time consuming and additional complexity could not be justified. We do however acknowledge this study would have been improved with an experimental design that included more parallels, multiple strains and multiple slightly different growth conditions. Furthermore, it would have improved tracking of the population changes if we had sequenced additional samples between the 100th and 260th population doublings. With the result at hand, the ultra-deep sequencing of the founder populations seemed inadequate as it was difficult to distinguish signal from noise. Alternative techniques could have been used to confirm preexisting diversity or unique molecular identifiers (UMI) for genome sequencing (Kou, Lam et al. 2016). Therefore, additional large studies are required to optimize the method to a point that it can be employed in forensics. One potential refinement that can be added is to look for the downstream variants in the source samples with target enrichment panels or targeted amplicon sequencing to further validate and strengthen the design.

**Sequencing platforms and assembly**

More than 300 sequences were used in the three projects as part of the PhD thesis. The main goal was to produce high quality sequences and avoid false variants. As *F. tularensis* is a monomorphic bacterium the accuracy of every single variant was
essential to properly infer genome relatedness. The quality of genomic assemblies generally improve by high coverage, long read-lengths, a combination of long and short fragment sizes, and high read quality (Ekblom and Wolf 2014). All sequencing runs were performed on illumina platforms with suitable numbers of samples per lane, as recommended by the manufacturer, with an aim to obtain high quality read coverage. Sequencing on Illumina HiSeq machines resulted in an average read coverage per genome of >600X while the Illumina MiSeq runs produced an average read coverage of >300X.

Two principal routes exist to identify genetic variation among isolates. The first method is to perform de novo-assembly of sequence reads to produce draft genomes (Simpson, Wong et al. 2009). These are subsequently aligned and scanned for variation. The second method is re-sequencing that entails separate mapping of all sequence reads for isolates onto a common reference sequence (Li and Durbin 2009). Variants are subsequently identified using a variant caller, such as GATK, varscan (Warden, Adamson et al. 2014). All variation is hence identified with respect to the common reference. Both strategies have advantages and weaknesses. The de novo-assembly method generally produces good quality consensus sequences. However, misalignments at indels can occur during the multiple alignment steps, resulting anomalous SNPs when the alignment is processed. If the aim is to identify sub-clonal variation, the de novo-assembly method is clearly not suitable as variants between samples are identified from consensus sequences. The re-sequencing strategy is fast and produces good results where the variation as compared to the reference sequence is small. This method is also suitable for identifying sub-clonal variation within a strain. Misaligned read maps may occur and may result in anomalous variant detection. Modern haplotype callers, however, mitigate errors at problematic regions, commonly by performing local de novo-assemblies on sequence reads at problematic loci. Nonetheless, a caveat on re-sequencing is that a missing variant call may not necessarily mean
variant is missing. Missing calls may be due to insufficient sequence coverage at a locus, so keeping track of the local sequencing coverage is important.

De-novo assembly was performed to assemble the reads in to draft genomes. One advantage of producing an assembly is they provide effective identification of insertion and deletion sequences over read only alignment-based methods. This is especially important to make evolutionary interpretation using complex models. De novo assembly method may lead to loss of genomic region and variants when compared with reads only variant calling in some bacterial species but the loss is minimal in clonal bacteria. We obtained assemblies of circa 1.6 million bases (complete genome 1.9 million bases) of *F. tularensis* in paper I and II, with the majority of the loss corresponding to removing multiple copies of IS elements. The assembly selection was done after meticulous testing of all available assemblers based on the following parameters: contigs >=100bp, mean bp size, N50 (bp), largest contig, genome coverage, number of incorrect contigs and open source availability. We found AbySS to have the best overall performance over other assemblers (Velvet, Edena, SSAKE, EULER), consistent with other studies (Lin, Li et al. 2011). We however did not perform assemblies for the sequences in paper III because the objective was not to test evolutionary models.

**Multiple sequence alignment**

Alignments are performed to map the coordinates of the genomes to each other or to a reference genome in a way that positional similarity or dissimilarity can be studied. The multiple alignments are imperative for inferring the evolutionary relatedness of the genome sequences (Felsenstein and Sinauer 2007). It is assumed that nucleotide positions in a multiple alignment of genomes are related in a tree like fashion with ancestors and children. When several hundred sequences are aligned together a balance is maintained between the ideal combination and reasonable time
taken to get there in the form of scoring schemes for gaps, identity and mismatches (Eddy 2004). Therefore alignments are seldom perfect, with the principle of collinearity compromised in fast mutating organisms, quality of the alignments dropping with low homologous relationships and the assumption of rate of mutations as priori. Another potential problem in the absence of a reference sequence is the reduction of the homologous core of the alignment popularly known as the funnel effect.

In paper I and paper II we have used the *F. tularensis* subsp. *holarctica* strain FSC200 as the reference sequence for alignments (Svensson, Sjodin et al. 2012). This genome is well annotated and bioinformaticians in our group both at the Swedish Defense Research Agency [FOI] and Umeå University have much experience working with this strain for several years. We have also at times made use of the SCHUS4 (Twine, Bystrom et al. 2005), another well studied strain, to provide alternate coordinates of canonical SNPs in the papers. The *F. tularensis* subsp. *tularensis* sequence FSC237 clp B (type A) has been used as a reference for the paper III, as it has been widely studied by the Francisella community as a reference for the *F. tularensis* subsp. *tularensis*. In paper III however we did not perform multiple alignments because in this study the primary objective was to identify all genomic variants and their frequencies without making evolutionary relationships.

Multiple alignments were performed in a two-step process to reduce loss of genomic segments to the “funnel effect”, where higher number of genomes translates to lesser conserved regions thereby producing smaller alignments. In the first step we aligned every genome to the well annotated FSC200 reference genome then combining all such pairwise alignments in to one single large multiple alignments. The pairwise alignments were performed using ProgressiveMauve because it aligns best for closely related genomes as compared to other tools like Multi-LAGAN, Shuffle-LAGAN (Darling, Mau et al. 2010). ProgressiveMauve
performance drops a little with genomes that undergo large scale horizontal gene transfers, a process that has not been detected in *F. tularensis holarctica* (Darling, Mau et al. 2004).

Variant calling has been performed both on raw reads and alignments in the different projects. A custom program was written to pull out the variants from the consensus alignments. Overall SNP differences were also measured using MEGA and DNAsp (Librado and Rozas 2009, Tamura, Peterson et al. 2011). SNP calling was performed after the flanking regions around indels in the alignments were trimmed by 10 base pairs to correct for misalignments. It has been reported in previous studies and we observed in our data that regions around the indels are prone to misalignments (Olson, Lund et al. 2015). In paper III, to capture extremely low frequency variants in ultra-deep sequenced genomes, we used VarScan (Koboldt, Zhang et al. 2012). VarScan exhibited both high sensitivity and selectivity in detecting both SNPs and indels (Sandmann, de Graaf et al. 2017). All the diversity analysis in paper 1 has been performed on the Euclidean distance between genomes based on the number of SNP differences among pairwise sequences. This is because there was little diversity among the strains we used, and all the genomes were highly related (clonal) making the Euclidean distance the best measure for such analysis.

**Detecting homologous recombination**

Recombination in bacteria is defined as the ability to exchange genetic material and modify its genome. One way to detect the presence of such regions is to look for identical SNPs in evolutionarily unrelated clades. These incongruent SNPs can be conveniently mapped in network trees. They go against standard evolutionary principles of relatedness and often obliterate signals making evolutionary inferences very difficult.
We did not detect any recombination events within the *F. tularensis* population in either paper I or paper II. The phylogenies reconstructed from the two datasets were clean with no incongruent SNPs across clades. We performed a network analysis using Splits tree program but we did not observe any network, indicating absence of any shared region among the clades (Kloepper and Huson 2008). The highly clonal nature of the bacteria however, makes it difficult to recognize past exchange of genetic regions, if any. It is difficult to distinguish *de novo* SNPs from allelic exchange in the event of extreme homogeneity as observed in *F. tularensis* genomes. However we believe the failure to detect any signal does not strongly rule out the possibility of recombination events and further studies need to be performed to establish the absence of incongruences in *F. tularensis*. Other types of variants like indels, tandem repeats and inversions could provide more insights in to possible events of homologous recombination.

**Phylogenetic reconstruction**

The differences computed among the multiple sequence alignments can be represented in a tree. The sequences in the root of the tree should be an ancestor to other strains to facilitate the interpretation of the tree structure. While the phylogenetic tree itself is an attempt to reconstruct the evolution, it is in reality the combination of tree topology and branch lengths that provides a more detailed picture of the evolutionary processes in past evolutionary history. For example, the star phylogeny with long terminal branches shown in paper I is considered typical for a highly successful clone that spread rapidly out-competing other clones. Another pattern of several clades with shorter terminal branch lengths as in paper II may be an indication of evolution over time, under stronger selection.

Phylogenetic trees in paper I and II in this thesis have been inferred using the distance method Neighbor-Joining (NJ) (Saitou
and Nei 1987) and the simplistic number of differences distance metric, i.e. pairwise distances being defined by the total number of differences between taxa in a sequence alignment. Advantages of this method include that it is fast and permits inference of trees using large numbers of taxa. In combination with the number of differences distance metric it also allows simple interpretation of phylogenies as branch lengths can be translated into exact numbers of accumulated changes. This method works well when the diversity among investigated taxa is small and when there are no homoplastic variants in the data, as has been the case in the studies herein described. The algorithm makes no hard assumptions and use just distance as a measure without any complex statistics. At the same time this algorithm does not include the possibility of multiple events occurring at the same site.

For phylogenetic inference of more divergent taxa, it is generally suggested to employ more complex algorithms. These algorithms are able to employ stochastic models to take in to account multiple events at the same genomic site. Therefore these algorithms not only provide the evolutionary distance but also divergence information. However, these statistically complex phylogenetic algorithms assume that all sites in the alignment evolve independently and that mutations occur at a fixed rate. The maximum likelihood method uses an algorithm that generates the tree with the above mentioned assumptions and another that computes the likelihood score of the tree parameters thus returning the tree with the best likelihood result (Felsenstein 1981). The branch estimation parameter for instance uses the full alignment instead of only pairwise sequences as the Neighbor joining tree does (Price, Dehal et al. 2009). This results in the ML tree raking up longer computing time and higher variance estimates.

Probably, it could also be argued from a philosophical standpoint that, for example, the method of Maximum Parsimony (MP) [Fitch
would be preferential to NJ. The MP method is based on an optimality criterion that explicitly seeks to find the phylogenetic tree that minimizes the number of changes (Kannan and Wheeler 2012). Under the assumption that the simplest solution is the most likely to be true, akin to Occam’s razor, it searches for the simplest solution that fits the data. It relies on counting on the no. of changes in the character states of the internal nodes; therefore the one with the minimum score is the maximum parsimonious phylogeny. However, while MP is philosophically appealing it is also the case that it quickly becomes unfeasible to infer phylogenies using MP as the number of taxa increases (Carmel, Musa-Lempel et al. 2014). This is because MP requires an exhaustive search of all possible trees to ensure that the best solution has been found. The use of MP on larger sets of taxa therefore requires heuristics to reduce the search space. There has also been precedence of the long branches being clustered despite being evolutionarily very different just because the no of mutations are high (Long Branch attraction)(Philippe, Zhou et al. 2005).

We attempted to generate the tree topology of the 163 F. tularensis genomes in the paper II using all the three NJ, ML and MP algorithms. The topology remained the same, a testimony to the absence of any incongruence among the genomes.

Finally, Bayesian trees are also a popular method of tree generation and more computationally intensive than the ML and the MP method, depending on the heuristics. The Bayesian method works on an iterative principle of sampling trees and model parameters according to a priori probability to create a posterior distribution using the Markov Chain Monte Carlo (MCMC) integration. These trees are capable of addressing more difficult problems but require keen knowledge for assessment of the results. We have also made use of the Maximum Clade Credibility Trees, especially from BEAST analysis to infer mutation rate estimates and evolutionary relationships (Bouckaert, Heled et al. 2014). These trees are a summary of the Bayesian phylogenetic
inference (O'Reilly and Donoghue 2018). More information on the assumptions and weakness is detailed in the following section.

**Estimation of substitution rate**

Substitution rate estimates has proved to be the primary parameter to study differences in transmission patterns, persistence, dissemination and related epidemiological analysis in microbes. The rates can only be an estimate of the genomic variation over time because of the randomness associated with the appearance, fixation and loss of mutations. Substitution rates differ from mutation rate in that they do not take in to account all the mutations that occur due to the errors that come up during replication (Ho, Lanfear et al. 2011). Substitution rates rather consider only those mutations that get fixed in the population as a result of selection or genetic drift. Therefore there are certain pre requisites for the calculation of substitution rate a) availability of samples from a reasonably stretched time period; b) phylogenetic tree and c) positive temporal signal in the sample. Errors occurring from a bad alignment or a phylogenetic tree will affect the substitution rate estimates. *F. tularensis* like other clonal bacteria is a difficult candidate owing to very little diversity among strains that were isolated across several years. Therefore here we have employed three major approaches: linear regression, maximum likelihood and Bayesian inference.

Regression is the simplest method that takes in to account the root to tip SNP distance between the strains and the out group. Clade-wise Euclidean distance of each strain was calculated from an out-group and plotted against the respective year it was isolated. The slope thus indicated the rate. This method is generally a good first step to look for a positive temporal signal. The primary weakness of this method is it considers each root to tip distance to be independent and that the strains do not share any evolutionary relationships. In addition, linear regression must also be accompanied with a measure of error estimate such as by
bootstrapping or random sampling of original data. We performed a cluster permutation test where the data was normalized to account for more strains that were identical and isolated in the same year. In the last couple of decades, the ML methods have gained in popularity over the distance methods. In the ML algorithm the tips of the trees are dated while internal nodes are assigned with dates based on the substitution rate parameter. This method offers powerful framework for model testing with the Likelihood ratio test (LRT) that can be used to test which hypothesis fits the data the best, something that is unavailable with the linear regression methods. However the ML method relies heavily on a single tree. In addition to the tree not being the true representative the single-rate dated tip in the ML model could be different from the ML tree generated with the different rates model.

Bayesian evolutionary rate estimates are performed using the MCMC integration that takes into account the tree topology, substitution rate, probability density function and provides the parameter values of the selected model fitting the data. The Bayesian method of mutation rate estimation was performed using BEAST 1 (Drummond, Suchard et al. 2012). The full alignment was used as input with time and location stamps. A relaxed clock method without site rate heterogeneity model was employed to consider different rates in different clades. The population was considered to undergo exponential growth. 10 million iterations were performed with the first 10% as a burn in. The error estimation was provided by the highest posterior density and the central posterior density. The burn in period lets the MCMC algorithm to converge then it samples the values of the selected parameters at a frequency that is directly proportional to the posterior probability density interval. There are also some weaknesses: the Bayesian method currently makes several assumptions including no population subdivision, absence of recombination and selection.
Association of genomic diversity with spatial distance

The relationship of geographic distribution of bacteria with that of the genomic differences in the population can be broadly summed up as phylogeography. Historical events and ecology plays an important role in the distribution of species, subspecies as well as different strains of the same species. Phylogeography was first studied in plant ecology and now is widely employed in the microbial genomics.

Previous studies of *F. tularensis* have shown limited geographical segregation of genotypes even in restricted endemic regions. In paper I *F. tularensis* strains were assigned different phylogenetic groups and the R package named js-marker-clusterer was used to mark the clusters on Google maps and post processed. In paper II Isolates were colored based on their genotype and plotted on a map based on their coordinates as provided by the patients in a questionnaire. In both the studies if multiple isolates were from the same place pie charts were drawn showing the different frequency of each genotype and the size of the chart was directly proportional to the number of isolates.

More advanced Bayesian methods are being employed that can show the population diffusion over time and space but they require a temporal signal from the genomes to process the mutation rate and prepare a Maximum clade credibility (MCC) tree with precise time and geographic coordinates. In our datasets we lacked genome sequences for some strains in paper I and we did not obtain a temporal signal in paper II. We employed the mantel test along with the spatial distribution independence (SDI) test to look for geographic and genomic correlations during the study period.

**Laboratory enrichment of mutations**

Laboratory culturing of strains is routinely performed in labs to maintain and produce stocks for research, study evolution or to
address other growth related questions. It has been reported that lab passaging leads to certain adaptations in the bacteria that are otherwise not present in wild strains (Moxon, Rainey et al. 1994). In some cases prolong lab adaptation leads to loss of virulence, intra cellular survival and impairment of other essential functions that are otherwise not seen in free range pathogens. This process also has been sometimes popularly referred to as domestication (Eydailln, Ryall et al. 2014). Laboratory culturing also provides us with a glimpse of evolutionary processes that the bacterial population undergoes in a limited number of population doublings(Barrick and Lenski 2013), that are otherwise not seen in outbreak or environmental strains.

The experiment was designed such that the primary source strain would be a single clone to ensure no preexisting diversity. The FSC237 clpB mutant was grown on McLeod agar plates and two separate single colonies were selected as primary inoculum for the two arms (A and B) of the experiment. The seed colonies were cultured initially for 30 generations in flasks using Chamberlain’s medium to mimic stock culturing from a pure culture, before dividing the cultures into seven serial passage lines for each arm.

This method is designed as a concept to see how different genetic variants originate, change over time and reach fixation. The method could have been scaled up to include different laboratory conditions, strains of interest and more than just two lineages to make the comparison. The most important limitation to this experiment is probably the lack of biological replicates that would prove the findings are stable between different experiments.
Results and Discussion

- Can we use population genomics and genetic diversity to infer dispersal barriers, patterns, range and the model of dissemination of F. tularensis at larger geographical scales? (paper I)

Phylogenetic reconstruction based on 67 F. tularensis subspecies holarctica genomes from Europe belonging in one of the four major genetic groups of the subspecies resulted in two phylogenetic clades and a single basal strain. The bigger clade B.45 appeared to have undergone a rapid clonal expansion as observed from a tightly clustered star phylogeny with short basal and long terminal branches. The tightly clustered strains were widely distributed all over Western Europe with limited geographical clustering suggesting little dispersal barriers. The second clade B.46 appeared to be geographically restricted around the Alps region and was genetically more diverse indicating a longer evolutionary history. There was no indication of recombination in the dataset, although the genetic homoogeneity of the strains makes such events difficult to distinguish from single nucleotide substitutions. Strains from both the clades were isolated from different hosts including humans, mammals, rodents and arthropods, thus indicating that the reason for few samples of B.46 was not under-sampling of one or several of these hosts. Isolation of related strains from different hosts suggests that diverse hosts get infected from the same disease reservoir(s) and that geographically dispersed reservoirs have limited genomic diversity.

The F. tularensis population in the western parts of the study region was found to have less genetic diversity as compared to the eastern parts, suggesting that F. tularensis in the Alps region represent a longer evolutionary history and may be the likely source of entry of the bacteria into this region. An analysis of
mutation rates based on the year of isolation and SNPs showed that there was no reliable temporal signal using the full dataset of 67 genomes over the 65 year sampling period. A positive mutation rate, however, at 0.4 mutations/genome/year could be determined in the Spanish subset of genomes where the disease recently entered the country and there have been many recent outbreaks. This suggests that *F. tularensis* undergoes variable rates of mutation over longer evolutionary periods. Such variable mutation rates with increased rates during outbreak periods have been previously reported in other bacteria e.g. *Yersinia pestis* (Cui, Yu et al. 2013) While attempting to make inferences based on mutation rates of *F. tularensis*, it appears that variability over time must be considered and that ecological factors including a genetic storage effect as described in paper II must be considered.

The nucleotide substitution patterns of the genomes included in this study as well as in paper II showed high AT mutation bias in an already AT rich genome, a phenomenon that may indicate weak selection forces (Wernegreen 2015). Possibly, a recent host adaptation has conferred strong genetic drift effects because of repeated population bottlenecks in the host population with a relaxed selection as described previously (Larsson, Elfsmark et al. 2009). The star phylogeny is also indirect evidence that we have captured a genetic population an intermittent stage wherein several of the polytomic clades will be lost over time due to stochastic events or because of selection forces that will remove several of the evolutionary most recent genetic variants.

There was no correlation between genetic and geographic distance among closely related strains (<6 SNPs) being found between long distances (150-1500 kms). This indicates that dispersal is rapid with respect to evolution and the complete absence or little barrier for dispersal over long ranges. However, two clades in Spain (B.48, B.52) showed positive correlation between genetic and geographic distances suggesting that there is some spatial structure locally.
The genetic network analysis further indicated a complex mix of long and short jumps with multiple movements between certain locations. The short jumps between adjacent locations formed the majority but the long distance movements were relatively common events. There have been previous reports of *F. tularensis* being transported long distance (Feldman, Enscore et al. 2001, Johansson, Larkeryd et al. 2014), and we suggest that given how easily *F. tularensis* can form aerosols, it could be easily transported by wind, or it could be following animal imports or even birds. Although this study does not prove the exact mechanism of dispersal it suggests that extremely similar strains of *F. tularensis* can be found at both small and large distances and as a result - the phylogenetic clades are not directly related to the geographic distribution. It should be noted that merely the transportation of the bacteria does not guarantee successful establishment. Therefore, we interpret the data as the result of more or less continuous *F. tularensis* dispersals at shorter and longer geographical scales. It appears that there are weak barriers for this dispersal as the correlation between genotype and geography is weak.

In this study we present that the bacteria follows a dissemination model wherein the local diversity is maintained by short range movements, perhaps limited by host or vector range restrictions while the long range imports with little dispersal limitation influences the genetic diversity of the bacterial population in the region. By employing genomic analysis we infer that variable mutation rate, a biology with a resting phase, low selection pressure with genetic drifts and long range dispersal events with limited barriers have shaped the genetic diversity of *F. tularensis* in continental Western Europe.

- *Can we combine epidemiological data, population genomics and spatial distribution of F. tularensis in a geographically restricted tularemia endemic*
region to infer factors that are important to local disease ecology? (paper II)

Phylogenetic reconstruction of the genomes from the outbreaks in Örebro County, Sweden were assigned to six clades within two of the four globally known clades of *F. tularensis* suggesting a lot of the global genetic diversity was present in a small endemic region. The topology of the phylogenetic tree was constant irrespective of the algorithm used (NJ, MP, ML) with no incongruent SNPs. The basal branches were longer while the terminal branches were short suggesting that both the genetic clades had long evolutionary histories outside of the outbreak area while there were also signs of very recent local evolution at the terminal branches. Consistent with the results in Western Europe, identical strains were isolated from animals (hare, field mouse) and humans in Örebro further suggesting the same source of infection for different disease hosts. We searched for and found genomes that were closely related to the outbreak genomes. All these 25 closely related genomes were from Sweden and were within a 2 SNP distance from an outbreak strain. Because all the 25 were related to the B.81 clade only, we interpret this as a sign of one or several past imports of this clade from other areas in Sweden. We did not observe any correlation between genetic clades and geographic localization suggesting local dispersal rate is rapid with respect to the mutation rate. This is consistent with little to no geographic barriers to dispersal locally as proposed in paper I. There was a lack of temporal signal in the substitution rate analysis in the dataset similar to the results in the paper I.

There was however remarkable clonal persistence among the outbreak strains in Örebro. Ninety-three percent of all outbreak strains belonged to one of seven clones, defined by a recent common ancestor and a maximal 6 SNP difference at the genome scale. All these seven clones persisted for more than one outbreak year. 90/138 outbreak genomes belonged to 8 completely identical
clones (0 SNP difference) at the whole genome level despite being isolated from different hosts and across several outbreak years. One such *F. tularensis* genome persisted for a period of 9 years. This suggests the bacteria have a biology that is marked by absence of mutation between outbreak years. In addition, we observed overlap of generations among these clones, meaning basal strains were isolated in later years as compared to terminal strains in the tree. Many of the strains with genomes fulfilling the definition of a persistent clone were placed at the short terminal tips of the tree suggesting they are recent descendants of a parental genome. Given the monomorphic nature of the *F. tularensis* subspecies *holarctica* population as a whole, it seems likely to us that many of the 1 or 2 SNP short tips visible in the tree represent genetic variants that would probably be lost stochastically or by selection in the future. To make an estimate of the total number of clones present in the outbreak area we compared the number of clones against the total no of samples each year. Using rarefaction analysis we found that the resulting curve approached saturation, a finding indicating the presence of finite no of clones in the outbreak area. Taken together these results support a scenario where the main part of the local bacterial population is stored long-term in a reservoir and randomly released in different outbreak years. To account for the closely related genomes found in other part of Sweden, and for the findings of long-range dispersal events in paper I, we need to add occasions if *F. tularensis* imports into this main scenario of local persistence.

There was a very high incidence of GC to AT substitutions in the total genome data. Additionally the SNPs in the terminal branches were more often biased towards AT as compared to the internal branches. The average AT-bias standardized for genome GC content was 3.3 mutations toward AT for every mutation toward CG in internal branches, and 9.3 mutations toward AT for every mutation toward CG in terminal branches, suggesting an evolutionary time effect, perhaps resulting from that purifying
selection counteracting increased AT-content get overwhelmed by random drift during short-term outbreak events.

In an experiment in the laboratory that mimics a potential long-term survival of *F. tularensis*, colonies remained viable for a period of 4 years. This suggests the bacterium has the biological ability to survive in non-nutrient environment for several years perhaps by limiting its metabolic activity to a null stage. The viable colonies were able to grow back on nutrient media indicating the bacteria can shuttle between the two metabolic stages. All the 11 SNPs observed in the surviving colonies at the end of the storage experiment were GC to AT, which is strikingly similar to that of the outbreak pattern. There was however more transversions in the storage experiment as compared to that of the outbreak genomes. Speculatively, this may be typical for mutations that arise in bacteria that survive with little or no replication. A similar pattern was recently reported in *Mycobacterium tuberculosis* surviving in a latent stage while infecting macaques (Ford, Lin et al. 2011).

A synthesis of the results is that there is: a) long term persistence of *F. tularensis* locally, b) overlap of generations over the outbreak years, c) little to variable rate of mutation, and d) a capability of *F. tularensis* to survive in the absence of nutrients for several years. Taken together these findings indicate a disease ecology including a seed bank effect. The seed bank appears to have a finite but diverse mix of ancestral and descendant strains similar to findings in seed-banks of plants. The finding contradicts the existing hypothesis that the bacterium, as a zoonotic pathogen is maintained in a perpetual infectious cycle, in which case we would have observed a positive mutation rate between outbreaks. The phylogenetic tree is a mix of 6 diverse clades yet very tightly clustered strains in each clade with overlapping generations.

- *Can we differentiate lab cultured F. tularensis strains from that of wild ones? Can we link*
**bacterial cultures after several population doublings to a source strain? (paper III)**

In this work serial culture passage experiments of an *F. tularensis* type A strain were performed to analyze changes in mutation frequency. The work is based on results of a pilot experiment with very interesting findings suggesting that the approach used here can detect mutations characteristic to laboratory passages of *F. tularensis* and provide evidence to link a sample to its laboratory source. The results will need to be confirmed in additional investigations but are encouraging. The results support the hypothesis that laboratory culture of a biological attack sample leaves behind a traceable trail of mutations at low frequency that can be biologically amplified to facilitate their detection.

Using the experimental set up described in paper III, we have written a narrative describing three different scenarios for when mutations emerged during the experiment.

In short, we describe the likely timing of mutations and identify those mutations that can be valuable for detecting previous laboratory passages of *F. tularensis*. The most interesting mutations to our purpose of providing data for attributing a sample to a specific source were the one large size duplication, a single large deletion and two SNPs found in multiple parallels of either arm A or B of the experiment but not both. These mutations are indicative of preexisting diversity generated in early culture stages of A or B by the 30th generation that was further amplified as a result of passaging. Such mutations are most informative for the aims of this study, and could potentially act signatures to distinguish a specific source (here exemplified by batch A or B).

We observed strong convergent evolution in capsule genes, which are essential for the survival of the bacteria inside host cells (Raynaud, Meibom et al. 2007, Kim, Sebastian et al. 2010, Sjodin, Svensson et al. 2012). 10/26 SNPs were observed in the wbtA
locus, 5/26 SNPs were observed in wzy loci, genes that are vital for capsule biosynthesis of the bacteria. The presence of 15/26 SNPs in genes essential for O antigen or capsule biosynthesis is consistent with the principle of convergent evolution. The remaining genes with SNPs, i.e., FTT_1236, FTT_1238, maeA, lpcC, fabF and rpoB are reportedly critical for the intracellular survival of F. tularensis. In a comparison with strains lacking extensive laboratory culture passage we found that 4/245 mutations present among 19 other “type A” F. tularensis genomes isolated from infected humans other mammals. These results give further support to the fact that these mutations are a result of in vitro lab adaptations and their presence would make it impossible for the wild strains to survive in intracellular environment of the host.

Several evolutionary processes were observed during the course of the experiment. Selective sweeps by certain SNPs offering fitness advantage were observed by rapid frequency increases in the total population. Clonal interference phenomena were observed (arm A passage 2, arm B passage 1) as well as genetic hitchhiking phenomena (arm A passage 7). The data suggested that low frequency variants could be enriched to detectable signals by passaging up to 60 population doublings but after 100 doublings the variant could be quickly obliterated by the arrival of additional de novo mutations as a result of lab adaptations.

In conclusion, these results supports that it is possible to amplify specific mutations that are present in low frequency within a F. tularensis population culture passaged in a laboratory. Some of these mutations were laboratory adaptations in a set of genes encoding components of the bacterial capsule that could be used to distinguish lab cultured bacteria from that of wild strains. Subsets of the mutations observed were amplified in either A or B arm of the experiment only and could potentially be used as batch signatures to distinguish different culture batches from each other.
Conclusion & future work

In conclusion, the research presented here provides new knowledge as to how *F. tularensis* subspecies *holarctica* spreads to new areas, persists and causes recurrent disease outbreaks. The life-style of this bacterium apparently includes a combined ability to enter new areas over long distances and to persist locally several years without any change in its genome. These new results further suggest an environmental maintenance of this bacterium similar to that of seeds of plants. We think that this seed-bank effect has a major role in tularemia ecology and epidemiological disease patterns. These results mean that persistence is a prominent part of the disease ecology and the bacteria are somehow stored in the environment between outbreaks. With the current results at hand, a scenario emerge with *F. tularensis* subspecies *holarctica* having developed a life-style based on mammal infections as amplifying hosts during shorter time periods followed by longer periods of time of environmental persistence that is not connected to amplification within a host. The findings have important implications for our view of tularemia ecology. The classical view of *F. tularensis* subspecies *holarctica* maintenance by circulation between arthropods and mammalian hosts should be revisited based on these findings.

Besides specific findings regarding *F. tularensis* subspecies *holarctica*, some of the findings, especially the genetic storage effect between disease outbreaks, may be worth investigating in a wider microbial context. We suggest to consider this effect on genetic diversity patterns in other outbreaks of disease caused by bacteria with survival properties outside infected mammalian hosts including e.g., *Clostridium difficile*, *Listeria monocytogenes* and *Legionella* spp.

We have also addressed basal research questions regarding factors that determine genetic diversity within microbial populations. We
found that in the case of *F. tularensis* spatial dispersal is complex and includes both long-range dispersal events and local persistence patterns over long periods. With this knowledge, it is easier to understand the complex patterns of genetic diversity reported in previous work over the years from various investigations of tularemia outbreaks. It is apparent that the genetic diversity of *F. tularensis*, and probably many other microbes, to a large extent is affected by the rate and mode of geographical microbial dispersal as well as the microbial ability to survive over long periods locally with little mutation rate. The inherent low mutation rate identified in both paper I and II suggests this is a general characteristic of *F. tularensis* populations in nature. This low rate contrast to the clearly measurable rate in the laboratory experiments described in this thesis. We noted 11 SNPs in a four-year experiment with *F. tularensis* resting without nutrient and we noted 26 SNPs in an experiment with laboratory culture passages suggesting that an inherent low mutation rate due to an unusually stringent DNA synthesis machinery is not the cause of a monomorphic population structure in nature. It seems more likely to us that the low rate in nature largely is determined by population factors and ecological factors that may include a resting phase of the bacterium with decreased rate of evolution. To investigate this further, it would be feasible to mimic a resting phase in the laboratory, perform genetic analyses over longer periods, and test surviving *F. tularensis* cells for their virulence in mammals.

The pattern of mutations identified in the persistence laboratory experiment and in the outbreaks further supported that bacteria do not maintain an active life cycle in environmental reservoirs. It would therefore be interesting to look for more insights as to how *F. tularensis* shuts itself on and off, especially with respect to the environmental or genetic triggers that are in play.

Finally, this thesis includes work aimed at distinguishing a laboratory cultured *F. tularensis* sample from that of the wild
ones. This work like the other parts of the thesis is aiming at making a connection between evolutionary/population genetic theory and applied science. The results of paper III suggest that it might be possible to amplify certain laboratory induced low frequency mutations in an *F. tularensis* population that would otherwise be a fitness disadvantage for *F. tularensis* in the wild. The mere presence of such mutational signature in essential genes for intracellular survival could distinguish a laboratory-cultured strain from that of the wild ones. We also could use the mutational profile to differentiate culture batches. Together these two findings pave way for more research to establish a novel microbial forensics approach for bacteria. What lacks is more repeats of the experiments to verify that the observed patterns are consistent and more optimization and calibration of experimental conditions. One important factor is the preferred number of generations for amplification and another is to find a more sensitive technique to confirm the presence of the low frequency variants in the source sample.

One area of information that has eluded us and that we would continue to work on is an appropriate molecular clock analysis of *F. tularensis*. We did not find any temporal signal for the mutation rate in our datasets which is a pre requisite for such analyses. The gradual increase in the availability of more genomes, a wider study involving strains from multiple geographies over large time scales should make it easier to perform these studies in the future.
I would first like to acknowledge the two fine institutes: Umeå University (UmU) and the Swedish Defense Research Agency (FOI), where I performed the PhD studies. To all the coworkers that made these workplaces fun to work.

I would start with thanking my two excellent supervisors Anders Johansson and Pär Larsson for several things: mentoring, supporting and for sharing your knowledge in this area of research. Thanks for the training to think critically, to let me work independently, to back my decisions and most importantly for teaching me the good practices. For your patience, especially in this last year, when I have harassed you both to the point, I feared you wouldn’t recognize me any longer. Thanks for being critical of my work yet friendly, I have not once hesitated to discuss my concerns and issues. I’ll continue to build upon everything that I have learnt from this PhD and hopefully make you both proud.

To Mats Forsman and Kerstin Myrtennäs, who for most of my time at FOI, acted as my co-supervisors and were always ready to hear my doubts, rants and help me with the pettiest of queries without judging. Thank you Kerstin, for sharing the office with me and for being the first one to proof read all my work over the years. Mats for being an optimist and you have been a huge motivator, especially after every single time I failed. I would miss your way of greeting me with “everything under control?” followed by long scientific chats. Your simplicity is inspiring. Am very certain you both are a bit relaxed, now that am Phinally Done with this thesis.

The entire bioinformatics unit (both past and present) at FOI has been a remarkable team to work with: I got to learn a lot from all of you. Andreas, Caroline, Per: For your suggestions and critical comments on the bioinformatics questions, thanks for all the helpful discussions at group meetings, at the fika and lunch breaks. Petter, Jon, thanks for helping with the stats and for being extremely down to earth despite being the most wanted colleagues in the building. Adrian takes the cake for having to bear the bulk of my rant and for dealing with all my tantrums. You probably left for London only to get rid of me, nice try mate!
Moa, Emil, David, Sämuel: Thank you for not making me feel alone in a building full of legends, I could be myself with you guys and thanks for all the help in and out of the work place. Special thanks to the experimental group: Johanna, Eva, Jonas, Malin, Mona, Edvin and Elin for the support and for helping me every time I needed your inputs. As a dry lab PhD I can only say: you all made my life very easy. A big shout to the IT team at FOI: My apologies friends, sorry for being such a pain and thank you for all the help with the servers and access. Also to the rest of the colleagues at FOI, that I didn't work with directly but they often took time out for quick chats and always had smiles to share. Thank you everyone, for being very nice despite me forcing you to speak in English all the time. I promise to learn Swedish!

To our fine collaborators from multiple institutes: for sharing the data and expertise without which this research wouldn't have been possible. To Dave and Dawn from NAU: for your continuous inputs on all the projects. To Patrik: for working closely with us on multiple statistical projects. To my PhD examiner Anders Sjostedt: for always being appreciative of every tiny achievement, for sharing your expertise on the subject on many occasions and for being extremely supportive over these years. To Edward Feil: from the University of Bath for having me at his lab and reviewing my PhD work. To Therese: A colleague turned friend, for all the fun discussions on staph epi and for all the tips about the Swedes.

Big thanks to all the PI's, administrators, PhDs and postdoc friends at the departments of Climi, MolBio and MiMS within the Medical faculty for creating a vibrant international work space. Geetanjali, my best buddy, for being there during all the highs and lows, am so happy I met you, Swarupa: the sole Odia connection in Umeå, for helping me even before my arrival and for introducing me to Nil. Chaitanya: probably spent most of my free time with you in all these years, thank you for being who you are! Arun, Naresh, Koushik, Anubhav: For all the dinners, travels and discussions till late nights. Anandi, my first friend in Umeå and corridor mate, thanks for all the game evenings, dinners and parties. Steffi & Mahesh: for all the fun times, singing and dancing. To Gautam, Brijesh & Lidiya, Ganesh & Divya, Radha & Ivgen, Damini, Tanu & Christoph, Valli & Alags, Zahra, Mridula, Lalitha, Saba, Agi & Tamas, Edvin &
Maria, Tushar, Shrikant, Karthik, Ala, Anne-Laure, Roshani, Akbar, Kerstin & Thomas, Goni: thanks for all the support over the years and for the fun times outside of work. Life would have been very boring without you guys!

To my friends from my undergraduates days: Ashwani, Gourav, Neha and Saket for cheering me all the way! Your phone calls and chats would more than make up for the dull days. To all the members of the badminton group: Laxmi & Sanat, Rajesh, Amit, Jay, Shashank and others. Thank you for some amazing game evenings and after parties! Also for all the help, care and motivation during the injuries! To the satsang group: for the bonhomie and calm, amidst all the chaos in my PhD life.

Rupali & Rishi: for being such warm and wonderful individuals and for taking care of me over these years like family, am lucky to have met you. Jyoti & Tony, for opening your house and hearts, for all the celebrations, I really didn’t miss India much, thanks to you guys. Anu & Nitin: You guys are so full of life and I always have been happy for no reason after every single meet up with you! Seema & Gireesh for all the advice on career and life in general, you have been my go to people for all practical questions. Tripti & Pravin: You have been with me in this journey, from my very first visit to Sweden in 2009, thanks for all the fun times in Stockholm in all these years. Finally, Stina and Ylva for being amazingly kind first ladies! Thank you!

To the Dwibedi and Dash family: it has not been easy living so far from you all and for missing out on several family moments, I’ll try to make up for them now onward. To my grandparents, uncles and aunts: for their blessings and unconditional love. Kalia mamu: for inspiring me to be in science and Kaku: for always checking on my progress. Ma, Bapa: thank you for everything, for all the love and motivation, for supporting all of my decisions and firmly standing behind me, all the time. I know this thesis means way more to you than me. Chinki & Roshan: for bringing so much happiness in to our lives, for taking my back and for always making yourself available whenever we needed you guys. Meenu: for being extremely understanding and for your unwavering belief in me. I couldn’t have asked for a better life partner! Thank you!
References


