UPTEC X 05 047
SEP 2005

ISSN 1401-2138

DIMITRI F. GUALA

Identification and screening of *Francisella tularensis* indel markers

Master's degree project

UPPSALA UNIVERSITET

Molecular Biotechnology Programme

Uppsala University School of Engineering

UPTEC X 05 047	**Date of issue 2005-09**

Author

Dimitri F. Guala

Title (English)

Identification and screening of *Francisella tularensis* indel markers

Abstract

In silico search for and identification of new type of markers for phylogenetic relation studies of *Francisella tularensis*, the causative agent of tularemia, and screening of twenty-four representative strains of the pathogen, allowed for correct identification of *F. tularensis* subspecies and groups. The insertion/deletion (indel) markers were able to confirm earlier findings of phylogenetic relationship between the subspecies of *F. tularensis* and position *F. tularensis* ssp. *mediasiatica* in the phylogenetic tree of *F. tularensis*. The data also supports previous suspicions that the Japanese strains of *F. tularensis* ssp. *holarctica* should be considered a separate subspecies. Some of the markers also suggest possible horizontal gene transfer between the subspecies of *F. tularensis*. In conclusion, indel markers appear to have suitable discriminatory capabilities for reliable subspecial identification and correct phylogenetic inference within *Francisellaceae*.

Keywords

Francisella tularensis, tularemia, markers, indel, phylogeny, subspecies, screening.

Supervisors

Pär Larsson, Kerstin Svensson
Swedish Research Defence Agency (FOI), NBC-defence

Scientific reviewer

Mats Forsman
Swedish Research Defence Agency (FOI), NBC-defence

Language	Security
English	
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **30**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 555217

Identifiering och systematisk undersökning av *Francisella tularensis* insertions- och deletionsmarkörer

Dimitri F. Guala

Sammanfattning

Francisella tularensis är en mycket infektiös patogen som har funnits i biovapenprogrammen hos flera av medlemmarna i G8 sedan slutet av andra världskriget. I dagsläget anses *F. tularensis,* i likhet med mjältbrandsbakterien, *Bacillus anthracis*, utgöra ett potentiellt bioterrorhot. Alla av *F. tularensis* underarter är dock inte lika farliga för människan och ett snabbt sätt att identifiera de harmlösa utbrotten från de livshotande samt ett sätt att härleda de ansvariga bakteriestammarnas ursprung, är högst önskvärt. De nuvarande metoderna för identifiering och klassificering av *F. tularensis* underarter är otillräckliga. Dessutom skulle detta fastställande av släktskapsförhållanden bidra till den allmänna förståelsen av *F. tularensis* evolution och därmed underlätta forskningen kring ett framtida vaccin.

I denna studie har en ny metod, baserad på genetiska markörer i form av tidigare oanvända likheter och skillnader på DNA nivå, tagits fram. Dessa markörer, bestående av insertioner och deletioner av DNA sekvenser i arvsmassan, identifierades med hjälp av en rad bioinformatiska sökverktyg tillämpade på DNA sekvenser från fyra representativa *F. tularensis* bakteriestammar. Därefter undersöktes och klassificerades ett antal av *F. tularensis* bakteriestammar med avseende på de hittade markörerna. Släktskapsindelningen bekräftade tidigare resultat, med undantagen att en tidigare misstänkt ny underart, *F. tularensis* ssp. *japonica* kunde urskiljas samt en tidigare identifierad men icke-inplacerad underart, *F. tularensis* ssp. *mediasiatica* kunde positioneras i släktskapsträdet. Unika markörer för de olika grupperingarna kunde också fastställas.

Denna studie har således resulterad i att nya genetiska markörer för korrekt identifiering av *F. tularensis* underarter, har påvisats samt att dessa markörers användbarhet vid bestämmandet av ett pålitligt släktskapsförhållande inom *Francisellaceae,* har bekräftats. Denna nya typ av genetiska markörer har potentialen att bli ett standardverktyg för fastställandet av släktskapsförhållanden inom och mellan bakterieunderarter och på så sätt bidra till ökad förståelse för bakteriell evolution.

Examensarbete 20p, Molekylär bioteknikprogrammet

Uppsala Universitet september 2005

TABLE OF CONTENTS

1 Introduction

1.1 Background

Francisella tularensis is one of the most contagious pathogens known. It is a small intracellular, Gram-negative bacterium associated with disease in a wide range of animal species [1] and transmittable through contact with infected animals, arthropod vector bites, ingestion of contaminated food or water and inhalation [2]. The resulting disease is called tularemia and ranges in severity from a self-resolving- to a severely incapacitating- or even fatal form. The respiratory type of tularemia is the most ruthless and is associated with a mortality rate of 5-30% [3] if not treated with antibiotics in time. The extremely low infectious dose (10 cells [2]) is one of the reasons for *F. tularensis* place among the top six on CDC's (Center for Disease Control and Prevention http://www.bt.cdc.gov/agent/agentlist.asp) category A-list of potential bioterrorist agents, together with anthrax, botulism and others [4].

Another reason for tularemia's place on the category A-list is its history as a biological weapon (BW) present in the BW-programs of United States, Japan and former Soviet Union [5] and the absence of a widely available vaccine. According to a calculation made by World Health Organization (WHO) in 1970 a release of 50 kg of virulent *F. tularensis* over a densely populated area of 5 million inhabitants would result in 250,000 incapacitating casualties including 19,000 deaths [6].

1.2 Properties and previous work

Based on the sequence data from the small subunit RNA, *F. tularensis* has been classified to the γ-subgroup of proteobacteria [7] (Fig. 1). Currently, four subspecies of *F. tularensis* are recognized: *tularensis* (type A), *holarctica* (type B), *mediasiatica* and *novicida* [8]. These are spread across the whole northern hemisphere, each one associated to a specific geographical region. *F. tularensis* subspecies (ssp.) *tularensis* is responsible for the severe respiratory form of tularemia in humans and is mostly present in North America. Type B or *F. tularensis* ssp. *holarctica* and the other two subspecies are less life threatening for humans and are dominant in different parts of Eurasia. Despite the difference in virulence and geographical adaptation of the subspecies, their sequence similarity is extremely high [10]. This poses a considerable problem when assessing the danger of a strain by identifying the subspecies it belongs to and makes it difficult to diagnose the acute respiratory infection with type A tularemia

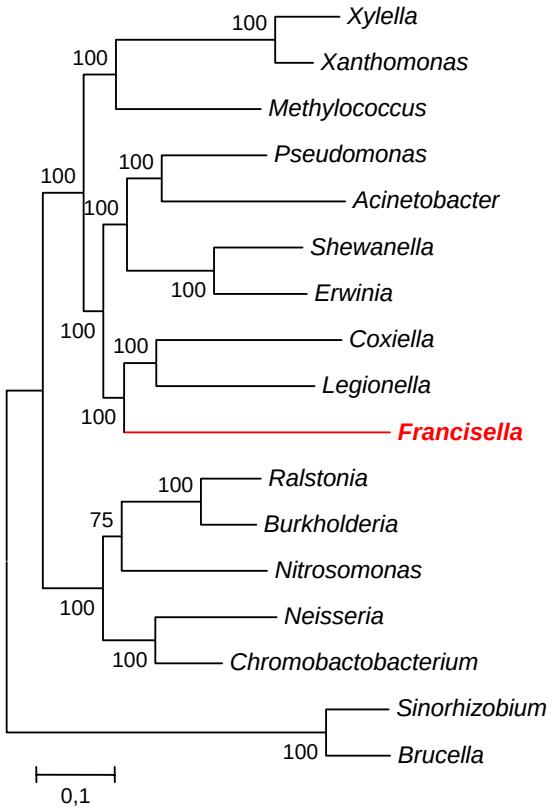

Figure 1. *Phylogeny of proteobacteria (courtesy of Pär Larsson, FOI).*

that could turn out to be fatal in as little as 3 days [11]. Presently available DNA-based typing methods including Repetitive element PCR (Rep-PCR) [9], arbitrary primed PCR [9], DNA Microarray [12] etc. all support the subspecies classification but have problems identifying and grouping strains. This is due to the high sequence similarity among the strains of *F. tularensis*. Multiple-Locus VNTR Analysis (MLVA) [13] has much higher sensitivity and is proposed to have the ability to discriminate between individual strains [13]. However the extremely high resolution obtained by this method hints a possibility of homoplasy induced by the randomness of VNTRs. This could result in different clones of the same strain considered as different strains. A method based on the identification of large size regions of difference (large deletions) in a whole-genome microarray study of *F. tularensis* [12] is much more conservative than MLVA and supports the current notion of four-subspecial division of *F. tularensis* (Fig. 2), but suggests that the Japanese strains of *F. tularensis* ssp. *holarctica* may be considered a separate subspecies. The method is, however, unable to resolve the branching order between *F. tularensis* ssp. *tularensis* and *mediasiatica*. Furthermore, characterisation of plastic regions flanked by direct repeats in the genome of *F. tularensis* strains was also unsuccessful in resolving the branching of *F. tularensis* ssp. *tularensis* and *mediasiatica* [14]. This poses a need for studies aiming at finding additional markers that, in a reliable manner, are able to discriminate between different subspecies and groups of strains of *F. tularensis*

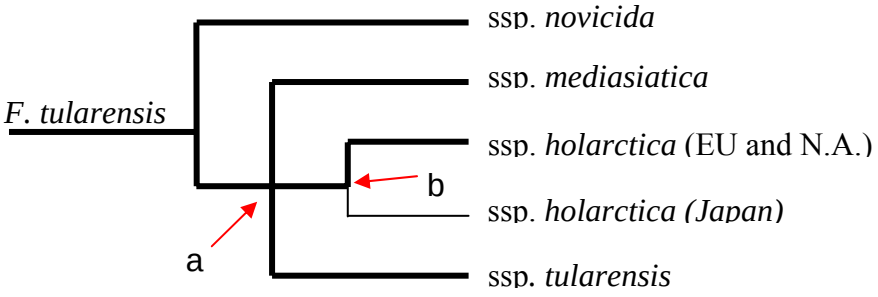

Figure 2. *Phylogeny of* F. tularensis *based on the identification of regions of difference [12, 14]. a) unresolved branching order between ssp.* tularensis *and* mediasiatica*. b) potentially new subspecies consisting of* holarctica *strains from Japan*

1.3 Indels

A method, based on shared conserved insertions and deletions (indels), has been proposed to address the issue of resolving phylogenetic relationship among groups and subgroups of similar bacterial strains [15]. The rationale for this approach and its ability to deduce groups of closely related bacterial strains and determine the phylogenetic relationship among them is that an indel of the same size, sequence and present at the same position in all members of one or more groups of bacteria has the simplest and most parsimonious explanation of being introduced only once in a common ancestor of this group. Thus, the presence or absence of an indel can be used to divide a collection of bacterial strains into distinct groups. A set of such informative indels can be used to distinguish several bacterial species and sub-species and perhaps even subgroups. Evolutionary events dividing the species can be determined, because all the species emerging from the ancestral clone where the indel was introduced will contain that indel signature, whereas all the species that existed prior to this event or which did not evolve from this ancestor will lack the signature [15].

1.4 Goals

This study aimed to find indel markers that demonstrate variation between different subspecies of *F. tularensis* and to investigate whether these markers allow a reliable identification and reflect phylogenetic relationship among the identified groups.

2 Theory

2.1 Alignment

A high sequence similarity usually means close structural and/or functional relationship, which is often an important issue in biology. A way to compare and measure nucleotide- or amino acid sequence similarity is to align the sequences of interest to each other. An alignment over the entire sequence is called global and is performed using the Needleman-Wunch algorithm [16]. When only parts of the sequences are compared at a time, the alignment is called local and the Smith-Waterman algorithm [17] is used. The type of the problem usually determines which of the alignment methods is more suitable.

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	2	2	2	2	2
A	0	1	2	2	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3	3
C	0	1	2	2	3	3	3	4	4	4	4	4
G	0	1	2	2	3	3	3	4	4	5	5	5
A	0	1	2	3	3	3	3	4	5	5	5	**6**

Figure 3. *A scoring matrix representing the alignment of 2 nucleotide sequences.*

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	2	2	2	2
A	0	1	2	2	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3	3
C	0	1	2	2	3	3	3	4	4	4	4	4
G	0	1	2	2	3	3	3	4	4	5	5	5
A	0	1	2	3	3	3	3	4	5	5	5	**6**

Figure 4. *One possible path through a scoring matrix.*

2.1.1 Dynamic programming and alignment parameters

The two algorithms used are similar in nature. First the sequences are arranged designating rows and columns of a matrix. Then a score is calculated for every position of the matrix (Fig. 3), based on three possible events: alignment of the amino acids/nucleotides in the two sequences, inserting a gap in one or in the other sequence. Finally, the path through the matrix yielding the highest score (Fig. 4), calculated using dynamic programming [18], determines the optimal alignment. In the global alignment a path through the whole matrix is calculated while in the local alignment, short stretches are considered separately to yield best local motifs.

The three events forming the basis for the calculation of scores for the alignment matrix are based on parameters defined by the user. The penalties for opening and extending a gap (gap penalties) are used when score for a gap insertion is calculated. The scores for matching of certain nucleotides/amino acids are extracted from a substitution matrix (Fig. 5), that is constructed based on the chemical and physical properties of the nucleotides/amino acids.

	A	T	C	G
A	5	-4	-4	-4
T	-4	5	-4	-4
C	-4	-4	5	-4
G	-4	-4	-4	5

Figure 5. *BLAST nucleotide substitution matrix. A match gives a score of 5 while a mismatch a score of -4.*

2.1.2 BLAST

The Basic Local Alignment Search Tool [18] is an approximation of the Smith-Waterman algorithm that finds the locally optimal alignment of an input sequence (query) and a database. The database is searched for a "word" of predetermined length (a "hit") with some minimum threshold parameter and then BLAST tries to extend the hit until the score falls below the maximum score yet attained. A High-Scoring Segment Pair (HSP) is the query and a sequence from the database yielding a score of the path through the alignment matrix that is above a certain threshold. The expected number of HSP's (length N, scores of at least S) is called the expectation value (E) and is calculated using the following equation: $E = \frac{N}{2^S}$ [19].

Table 1. *IUB code*

N	A, C, G, T
V	G, A, C
D	G, A, T
B	G. T. C
H	A, T, C
W	A, T
M	A, C
R	A, G
K	G, T
S	G, C
Y	C, T

2.1.3 Multiple Alignment

The generalization of dynamical programming for alignment of multiple sequences has not been very successful since the matrix space n^m (length of sequence = n, number of sequences = m) increases exponentially ($O(n^m)$) with the number of sequences. Instead, an approach where an initial pair-wise alignment of all the pairs of sequences is used to make a distance matrix for construction of a guide tree [20]. The guide tree showing phylogenetic relationship is used to align the most closely related sequences and then gradually add more distantly related ones to the alignment, resulting in a consensus sequence with IUB (International Union of Biochemistry) coded bases (Table 1), where several possible bases can be represented by one letter.

2.2 TP-PCR

Polymerase Chain Reaction is used to study genomic markers. There are variations of PCR suitable for different amplification goals but their general structure is the same (Fig. 6). The most suitable PCR technique in this case was the tailed primer method (TP-PCR) (my acronym).

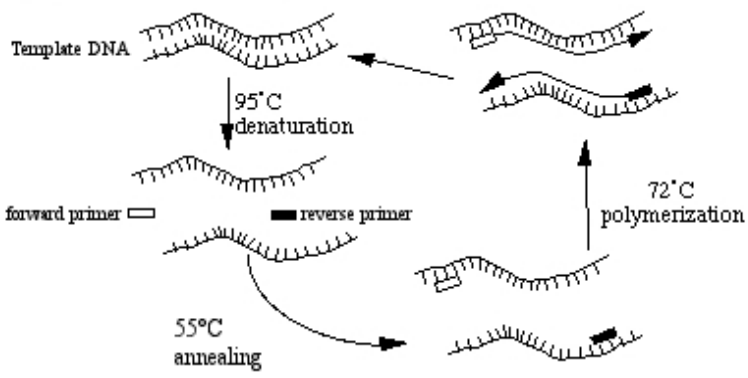

Figure 6. *PCR. The steps in a generic PCR involve an initial melting of the double strand (denaturing) followed by annealing of the two primers and their elongation by a polymerase. The cycle is repeated to obtain sufficient amount of DN2.*

To enable detection of PCR fragments in an automated sequencer, chromophore labelling of the PCR products is required. Most often, and most conveniently, one primer in each pair used for the PCR is labeled. However, the labelling costs become exceedingly high if many markers are explored. Therefore, in this work a TP-PCR [21] method with a few modifications was used, offering a much less expensive way. In TP-PCR, each forward or reverse primer, of every primer pair, is tailed at the 5'-end using the sequence of a standard primer (i.e. universal M13 primer). A labeled M13 primer with a lower melting temperature than the specific primers (forward and reverse) is also added to the PCR reaction (Fig. 7). The initial amplification is performed using the annealing temperature of the unlabeled primers, but subsequently during the

amplification process, the annealing temperature is lowered and fluorescently labeled primers become incorporated. In this way, only one labeled primer is necessary for labelling PCR fragments of all markers.

2.3 Phylogeny

2.3.1 "Model-free" phylogenetic reconstruction - Maximum Parsimony

The logic behind the reconstruction method maximum parsimony (MP) simply states that the (optimal) tree requiring the minimum number of changes to explain the observed data is likely to be correct, analogously justified by the *Occam's razor* precept that the simplest solution usually is the best. MP does not make any assumptions about the process of changes, so it is easily applied when the data at hand is not easily modeled [22]. However, it requires that the similarity among taxa is high. Under certain conditions, MP is known to consistently produce misleading results. Long branches are, for instance, inclined to attract when the evolutionary rate differs among the lineages, a phenomenon dubbed 'long-branch attraction' [23].

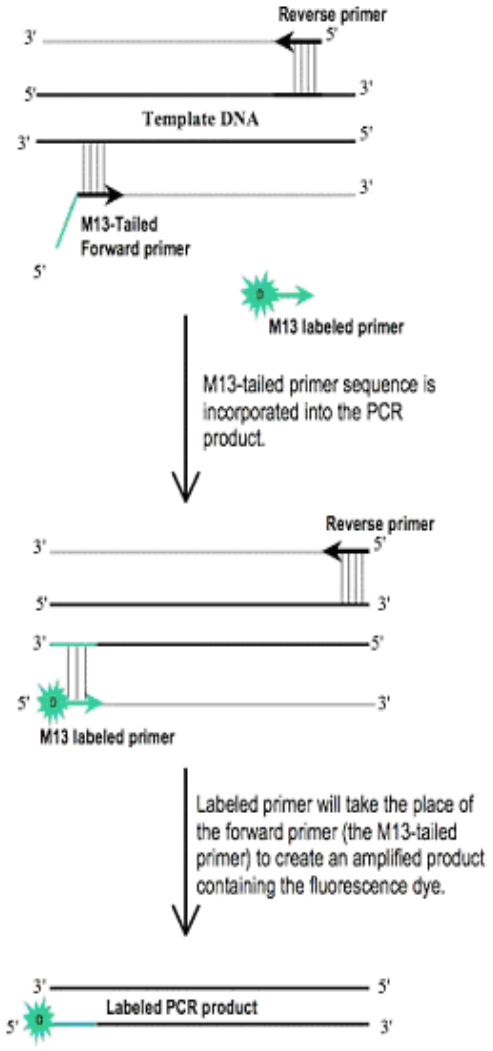

Figure 7. *TP-PCR is a 2-step amplification technique that utilizes one primer with an M13 tail and a labeled*

2.3.2 Model-based phylogenetic reconstruction

Except for MP, all methods for phylogenetic reconstruction are based on explicit models of molecular evolution. The complexity ranges from simple methods to account for multiple base or amino acid substitutions, transition/transversion ratios to more complex models that accommodate heterogeneous rate and composition equilibrium frequencies and co-variation among sites. Unfortunately, as parameter-rich models better can fit the data, they can also result in over-fitting and give incorrect results. Thus, complex models should be used cautiously and, when possible, the validity should be tested using statistical criteria.

2.3.2.1 Distance methods

All distance methods are based on a transformation of the data into pair-wise distances for all taxa prior to the inference of phylogenetic trees. This transformation, in turn, is based on the chosen evolutionary model. One of the drawbacks of the distance methods is that some of the information is lost when the observations are transformed into pair-wise distances. Examples of methods include Unweighted Pair-Group Method Using Arithmetic averages (UPGMA), Neighbor-Joining (NJ) and Minimal Evolution (ME). UPGMA [24] and NJ [25] are algorithmic methods where the tree is directly produced by branch clustering of the algorithms. UPGMA gives rooted trees and relies on the assumption that data is ultrametric (giving equal total branch lengths to the root). However if the data is not ultrametric due to absence of a molecular clock, the constructed tree will be wrong. Conversely, trees produced

by NJ are unrooted and not ultrametric. For ME [26], the optimality criterion dictates that the best tree has the lowest total sum of branch lengths.

2.3.2.2 Maximum Likelihood (ML) and Bayesian Inference (BI)

ML differs from MP in that it is an explicit model, stating how the character state changes and including branch lengths of the trees [27]. In its general form, ML requires independence of characters because their likelihood is calculated and multiplied to give the total likelihood of the model. Usually, other assumptions are also made about the data, such as the constant rate of change over time and that the conditional probability is the same for all characters and is also constant over time. These assumptions are unfortunately not always valid but ML can still serve a useful tool where other fail. The likelihood of a model (and the tree) is the conditional probability of getting the data given the model and the tree. Bayesian inference [28] turns this around to produce the probability of the model and the tree, given the data, using an iterative combined search-and-evaluate algorithm (Metropolis-Hastings coupled Markov chain Monte Carlo) that ultimately converges on a topology if correct prior probabilities have been chosen. Use of Bayesian methods is still a new and active research topic in phylogenetics. However, Bayesian phylogenetics appears to have clear advantages being fast, robust to the choice of prior probability parameters, and that it provides built-in support values for the analysis.

2.3.3 Searching tree-space

For phylogenetic reconstruction methods that depend on searching tree-space (MP, ML, ME), the search process is the hardest part of the analysis, since scanning all trees is an NP-complete problem. The number of possible trees quickly reaches extreme figures; 20 taxa give $2.2 * 10^{20}$ possible (unrooted) trees, 100 taxa give $4.5 * 10^{190}$ combinations. Exhaustive search methods can be used for a small number of taxa but for larger datasets the computational time will be prohibitively large and heuristic methods are used. The heuristic methods rely on building one or a few likely initial trees from the original data and then performing a few changes (branch swapping etc) investigate other related trees, keeping the best. The disadvantage here is that the search method can get stuck in a local maxima and never find the globally best tree.

2.4 Multivariate data analysis

Multivariate data analysis (MLVDA) is a useful tool for understanding a dataset containing many different measurements, -variables and -properties of a process or a system [29]. These multivariate data contain much more information than univariate data but they are also much harder to understand without the proper tools. PCA and PLS are two projection methods used in MLVDA, to represent the difficult-to-grasp amounts of data. The disadvantage of these methods is that some data is inevitably lost when only a few principal components are selected. However the information content of the lost data is usually noise.

2.4.1 Principal component analysis

Principle Component Analysis (PCA) is a way to represent the information from the initial data matrix, in an easily overviewed form. The initial multidimensional space formed by the quantified variables is projected onto a few descriptive model dimensions, denoted principal components [29]. Each observation (strain, species etc.) can then be displayed graphically and analyzed to show correlation structures in the reduced dimensional space, given by the principal components, to reveal relationships between observations and variables, and among

the observations and variables themselves to uncover groupings and trends in the data.

2.4.2 Partial least square projections to latent structures

Partial least square projections to latent structures (PLS) is a method for simultaneous projections of both the X and Y spaces on low dimensional hyper planes, uncovering correlations in reduced variable space between sets of dependent and predictor variables X <--> Y for observations [29]. The advantage of PLS is that it easily handles many, noisy, collinear and even incomplete variables in both data sets. Its precision also intuitively increases with the increasing number of relevant variables collected.

3 Materials and Methods

Scripting tools were constructed in Perl and BioPerl (http://bio.perl.org/) and combined with existing software for analysis of bio-molecular sequences to analyze four available genome sequences of *F. tularensis* (A, B, C and D). The found similarities and differences were later verified and screened in a subset of strains of *F. tularensis*, to obtain a set of multivariate data. The data was finally analyzed using clustering- and phylogenetic inference programs.

3.1 In silico search for indel markers

The strategy for finding potential indel markers was to extract the relevant information from a multiple alignment of the sequenced genomes. Since multiple alignment algorithms become computationally expensive when aligning long genome sized sequences, an approach where an initial pair-wise alignment followed by clustering of the found hits was performed, prior to passing the found subsequences to a multiple alignment algorithm. This general approach and the created scripting tools can be applied on any number of genomes for which a sequence is available, if only the parameters described in the following subsections are tuned correctly. With a few modifications, the algorithm (Fig. 8) can also be used to search for other markers than indels. In this case four available genome sequences for *F. tularensis* were used: "A" a ssp. *novicida* genome, "B" a ssp. *tularensis* genome and "C" & "D" two different ssp. *holarctica* genomes. The sizes of the sought indels were constricted to an interval of 5 to 200 bp.

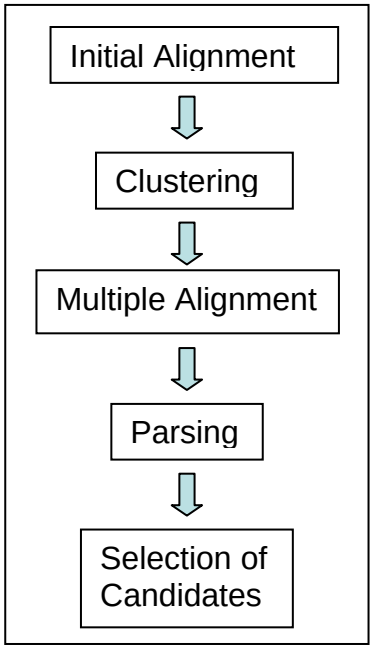

Figure 8. *Search algorithm*

Before beginning the initial alignment one of the sequences was selected to be the query and a database of the remaining sequences was established and formatted. To reduce redundancy in the results from the initial alignment, a program called *Crossmatch* (http://www.phrap.org/) was used on the query sequence. This step eliminated three of the most abundant IS-elements (ISFtu1, 2 and 3) [3] that pollute the genome of *F. tularensis*.

3.1.1 Initial Alignment

The standalone BLAST (version 2.2.6) for UNIX was used to align the query sequence against the database of the other three sequences, in a pair-wise manner. Most of the BLAST default parameters were used, but some crucial alterations (Table 2) were made to improve the finding of potential indel markers in the specified size range. Since both the query and the target sequences were composed of nucleotides the *blastn* search routine was used. The expectation value was set to 0.0001, because search hits with higher expectation value are unlikely to reflect true sequence identity in a comparison of such closely related genomes. The value controlling the region of the path graph explored by BLAST during its gapped extension process (X drop-off value) and the gap extension penalty were adjusted to find and extend gaps of the desired size. The low complexity filter, which masks significant hits with no apparent biological interest (i.e. hits against common acidic-, basic- or proline-rich regions), was turned off because any significant hit is biologically interesting since it can represent a potential indel marker.

Table 2. *Specific BLAST parameters*

	Value
Search routine	blastn
Alignment view	Pair-wise
Expectation value (e)	0.0001
Gap extension penalty (E)	1
X drop-off value (Z)	50, 200, 500
Filter for low complexity (F)	off

A script utilizing BioPerl was written to extract the hits from BLAST output. The hits were sorted according to their position in the query genome and stored in separate FASTA format files: one with all the query subsequences and another with all the hits. To enable a new BLAST search, a database was made from the file containing all the hits.

3.1.2 Multiple Alignments

To perform a multiple alignment of the subsequences returned by BLAST, clustering and removal of redundancy, resulting from repeats and multiple copies of certain regions, was needed. The clustering algorithm grouped the sequences on the basis of their starting positions and lengths, picking the longest hit subsequence (one from each genome) for every specific starting position of a query sequence. The resulting clusters were passed to ClustalW for multiple alignment.

3.1.3 Parsing

The output files from ClustalW were converted to FASTA format using the *clustalign2fasta* tool from SEALS package [30]. Each resulting FASTA file, corresponding to one multiply aligned cluster of subsequences, was searched for indels by a parsing tool. This search was performed by comparing the value (A, C, G, T or '-') at every position in the subsequences from the different genomes. Every time a deletion was encountered, in any of the clustered subsequences, its position was stored for further

Table 3. *Classification of* indels. *"X" means the sequence is present, while "-" means the sequence is absent.*

class	***Genome A***	***Genome B***	***Genome C***	***Genome D***
1	X	-	-	-
2	-	X	-	-
3	X	X	-	-
4	-	-	X	-
5	X	-	X	-
6	-	X	X	-
7	X	X	X	-
8	-	-	-	X
9	X	-	-	X
10	-	X	-	X
11	X	X	-	X
12	-	-	X	X
13	X	-	X	X
14	-	X	X	X

analysis. Later only the indels of the desired size were saved, along with their size and the class they belonged to (Table 3).

A number of indel markers were chosen to be tested and screened on a subset of relevant strains of *F. tularensis*.

3.2 DNA isolates of F. tularensis

The study included 24 geographically distinct isolates of *F. tularensis* from Sweden, Russia, Spain, Central Asian Republics, Japan and North America (Table 4).

Table 4. *The collection of strains, used in this study (courtesy of Swedish Defence Research Agency).*

ID	Strain name	Subspecies	Source	Origin	Year
FSC 012	425 F4G	*holarctica*	Tick, Montana,	USA	1941
FSC 017	S-2	*japonica*	Human	Japan	1926
FSC 021	Tsuchiya	*japonica*	Human	Japan	1958
FSC 022	Ebina	*japonica*	Human	Japan	1950
FSC 035	B423A	*holarctica*	Beaver, Montana	USA	1976
FSC 040	U112	*novicida*	Water, Utah	USA	1950
FSC 041	Vavenby	*tularensis*	Tick, British Columbia	Canada	1935
FSC 046	Fox Downs	*tularensis*	Human, Ohio	USA	1940
FSC 054	Nevada 14	*tularensis*	Rabbit, Nevada	USA	1954
FSC 147	543	*mediasiatica*	Miday gerbil	Central Asia	1965
FSC 148	240.84	*mediasiatica*	Ticks	Central Asia	1982
FSC 149	120	*mediasiatica*	Hare	Central Asia	1965
FSC 155	LVS	*holarctica*	Unknown	Russia	1949
FSC 171	77II	*holarctica*	Human	Sweden	1995
FSC 230	CCUG2112	*tularensis*	Human, Utah	USA	1920
FSC 237	SCHU S4	*tularensis*	Human, Ohio	USA	1941
FSC 257	503/840	*holarctica*	Tick, Moscow region	Russia	1949
FSC 398	2-5081	*holarctica*	Human, Örebro	Sweden	2003
FSC 412	2-5787	*holarctica*	Human, Örebro	Sweden	2003
FSC 429	2-5786	*holarctica*	Human, Örebro	Sweden	2003
FSC 454	FNSp1	*novicida*	Human	Spain	2003
FSC 519	32-23	*holarctica*	Human, Örebro	Sweden	2004
FSC 595	F58	*novicida*	Human	Brazil-UK	2004
FSC 604	O-363	*tularensis*	Fowl	USA	1959

The *F. tularensis* strains were obtained from the *Francisella* Strain Collection (FSC) at the Swedish Defence Research Agency. All strains have been characterized by specific agglutination, specific PCR amplification of the *F. tularensis* specific *lpnA* gene and biochemical analyses. The *F. tularensis* cultures were inoculated directly from frozen seed stocks and grown for 48 hours on modified Thayer-Martin agar plates at 37° C and 5% ambient CO_2, harvested by scraping, and resuspended in saline (0.85% NaCl) at a concentration of 10^9 CFU mL^{-1}. Bacterial thermolysates were produced by incubating the resulting suspensions at 65ºC for 2 hours. Sterility was verified by lack of growth after inoculation of 10 µl of lysates on modified Thayer-Martin agar plates [31], incubated for 10 days at 37ºC in 5% ambient CO_2 concentration. Lysates were diluted 1:1 with pH_2O and DNA was extracted with phenol; adding an equal volume of phenol vortexing, centrifuging the sample (13000rpm for 5') and collecting the upper phase. The extraction procedure was repeated three times and followed by a precipitation with 2.5 volumes of 99.5% EtOH and 10% of 3M NaAc in -20ºC. The precipitate was centrifuged (13000 rpm, 15'), washed with 70% EtOH, dried and solved in sterilized H_2O.

3.3 Optimization

Before proceeding with the actual screening of the selected strains for the found indel markers, an optimization of the TP-PCR protocol was performed.

3.3.1 Oligonucleotide primers

Four markers were randomly selected and flanking oligonucleotide primers for their optimization were constructed using JEMBOSS' built-in tool *eprimer3* [32]. The forward primer for each primer pair was synthesized (MWG-Biotech AG, Ebersberg, Germany) with an additional 19-bp M13 tail (5'-GTAAAACGACGGCCAGT-3') added to the 5'end. Both the forward- and the reversed primers were designed with a predicted annealing temperature of 55-57ºC. An M13 primer, with an annealing temperature of 51ºC, labeled with IRD700 chromophore, was used for detection of the amplified fragments (Proligo Primers & Probes™, Hamburg).

3.3.2 Amplification

The indel marker loci were amplified using PCR. DNA templates were from strains for the genomes B and C (from the *in silico* search). The variable parameters from the optimization were: annealing temperature, number of cycles and concentration of $MgCl_2$/-primers/-DNA/-buffer. The negative control contained water instead of DNA and the positive controls contained a directly labeled primer pair for amplification of Ft-M10 [13].

3.4 Screening

3.4.1 Primer design

A Perl script was written to supply the information necessary for primer design using *eprimer3* tool from JEMBOSS package [32]. The script uses a genome sequence file in FASTA format and a coordinate file containing the indel coordinates. It then finds the subsequences, containing the indels and their surroundings and aligns them using BLAST against the other available genome sequences. The BLAST outputs are finally multiply-aligned with ClustalW and a consensus sequence for each indel is created using the cons tool from the JEMBOSS package. The consensus sequence is a sequence where IUB code (Table1) is used for all the mixed base sites. The consensus sequence is necessary to design primers that work for all of the tested genomes.

The forward primer for each primer pair was synthesized with an additional 19-bp M13 tail added to the 5' end of the oligonucleotide. The M13 primers were labeled with D2-PA or D3-PA or D4-PA dyes, at the 5'end (Proligo Primers & Probes™, Hamburg, Germany).

3.4.2 Refined primer selection

All the primer pairs chosen for screening were primarily tested on the four strains used in the *in silico* discovery of the indel markers. The ones that worked were used on the chosen subset of *F. tularensis* strains.

3.4.3 Amplification

The reactions were performed in 96-welled micro titer plates. Each PCR-reaction contained

0.15 mM dNTP, 0.6 U DyNAzymeII DNA polymerase (F-501L Finnzymes, Espoo, Finland), 1×PCR buffer for DyNAzyme DNA polymerase (F-511), 2 µl of template DNA (20 ng/µl), 0.3 pM forward primer, 0.8 pM reversed primer and 0.8 pM labeled M13 primer. Filtered sterile water was added to a final volume of 25 µl. The thermal cycling (Table 5) was performed in MyCycler™ (BioRad, Hercules, CA).

Table 5. *Thermal cycling protocol for amplification of indels from* F. tularensis.

	temp /ºC	*time*	*cycles*
Initial denature	95	2 min	
denature1	95	30 s	
annealing1	56	30 s	15
extension1	72	45 s	
denature2	95	30 s	
annealing2	51	30 s	20
extension2	72	45 s	
final extension	72	7min	
finalize	4	∞	
total time		01:10:15	

3.4.4 Detection

After the amplification, 3 differently labeled PCR reactions, 2 µl from each, were pooled and diluted fifteen-fold. 1 µl of every diluted sample was added to 40 µl of Sample Loading Buffer (SLS) buffer, containing DNA Size Standard-600 (Beckman Coulter Inc., Fullerton CA), and sealed with 1 drop of Mineral oil (Sigma-Aldrich Sweden AB, Stockholm, Sweden). Finally the samples were analyzed in the Genetic Analysis System CEQ™ 8800 (Beckman Coulter Inc., Fullerton CA).

3.5 Analysis

The widely used Simpson's diversity index (D) [33] was calculated for each indel marker ($D = 1 - \sum_i \left(\frac{n_i}{N}\right)^2$, where n_i is the number of strains belonging to a specific indel signature, N is the total number of strains and n_i/N can be viewed as the allele frequency). Phylogenetic inference, including tree construction was made in Bionumerics software (Applied Maths BVBA, Sint-Martens-Latem, Belgium) and Phylip [34]. Neighbor-Joining (NJ) and Single Linkage (SL) analysis methods from Bionumerics were used for the construction of phylogenetic trees for indel markers and for a mixture of indel- and VNTR markers. The *Pars* method from Phylip was used for the construction of a parsimonious tree, based on Maximum Parsimony (MP) and a bootstrap of 1000, for indel markers. The measure of similarity used in the two distance methods, was Pearson Correlation. SIMCA (Umetrics AB, Umeå, Sweden) software was used to perform PCA and PLS-DA (PLS-discriminant analysis) on the dataset of indel markers and on the mixture of indel markers and the seven VNTR markers [13] available for the investigated taxa. The clusters, discovered in the PCA analysis, served as the "dummy" variables in the PLS-DA analysis. The discriminatory quality and identification capability of the found indel markers was discovered per se and supported by the values of importance (VIP) based on the contribution of a marker to the final clustering in the MLVDA.

4 Results

4.1 Found indel markers

The *in silico* search for indel markers was performed using all four sequences (one per experiment) as query sequences in the initial alignment to ensure the best possible coverage of the existing indels and to test the robustness of the method. The effect of having different values for the X drop-off in the alignment was different number of clusters (Table 6). It also resulted in a different number of indel markers found (Table 6). Despite the increasing number of the initially found indel markers, only a part of them appeared to be valid. The false positive markers originated mostly form repetitive sequences abundant in the genome of *F. tularensis*. Only a few novel, valid indel markers were found changing the X drop-off value and using a different sequence as the query sequence (Table 6).

Table 6. *Number of clusters and* indels *found using different values of X-drop-off. Within parenthesis is the value of valid not previously seen* indels.

	clusters			*indels*		
Genome\Z	50	200	500	50	200	500
A	680	465	420	243(128)	722(5)	833(1)
C	669	449	n/a	254(16)	859(12)	n/a
B	677	n/a	n/a	268(3)	n/a	n/a
D	653	n/a	n/a	241(1)	n/a	n/a

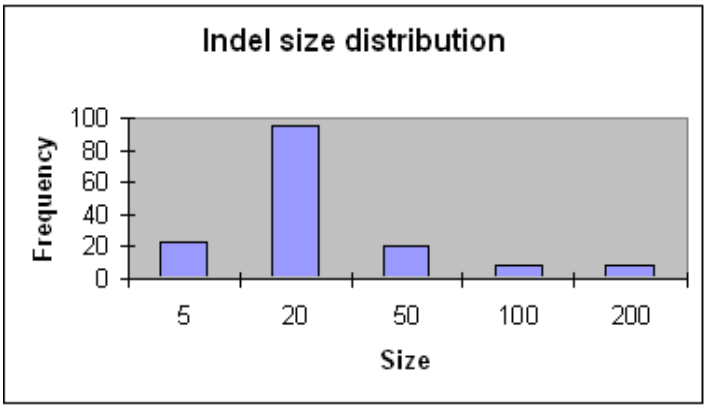

Figure 9. *Distribution of indels*

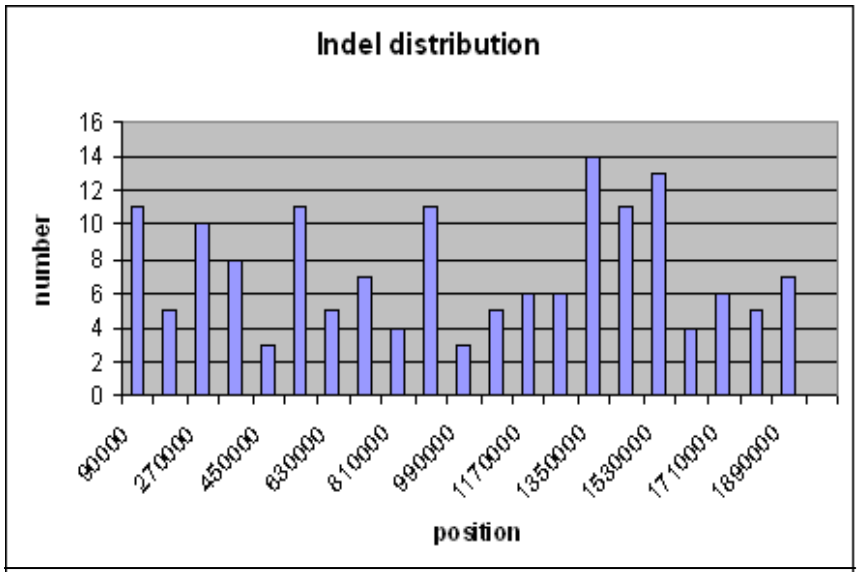

Figure 10. *Indel distribution in the genome of* F. tularensis

After verification 155 markers remained. The position of indel markers was mapped in the Genome A (*F. tularensis* ssp. *novicida*) since it is believed to be the oldest of the subspecies studied and thereby has the greatest probability to contain the largest number of the found markers. The majority of the found indel markers were 20 bp large (Fig. 9) and randomly distributed (Fig. 10) across the genome.

Classification of the found indels revealed that none of the found markers belonged to classes 5, 6, 9 and 10 and the class containing the largest amount of indel markers was class 3.

4.2 Selected indels

Since the scope of the project did not include an exhaustive search for indel markers, only a subset of the originally found 155 markers was chosen for screening. The subset contained 65 indel markers (App. A) including all the markers from the smaller classes and a selection of markers from the other classes (Table 7).

Table 7. *The number of found indels in each class.*

class	*Genome A*	*Genome B*	*Genome C*	*Genome D*	Number
1	X	-	-	-	5
2	-	X	-	-	5
3	X	X	-	-	20
4	-	-	X	-	2
7	X	X	X	-	5
8	-	-	-	X	7
11	X	X	-	X	5
12	-	-	X	X	7
13	X	-	X	X	4
14	-	X	X	X	5

4.3 Informative markers

Forty-seven functional and informative indel markers could be amplified in the screening of the previously mentioned *F. tularensis* strains (App. C). Their sizes ranged from 5 bp (Ft-ID22, Ft-ID50) to 182 bp (Ft-ID23) and corresponded well with the, *in silico*, predicted values. Naturally, only two alleles were predicted for each marker in the *in silico* survey (insertion/no insertion or deletion/no deletion), however four of the markers (Ft-ID1, Ft-ID12, Ft-ID51 and Ft-ID57) showed 3 alleles. The diversity index (D) was calculated for every marker (App. A) and ranged from 0.10 (Ft-ID32) to 0.61 (Ft-ID51) with an average value of approximately 0.39 and a median of 0.47 (Fig. 11).´

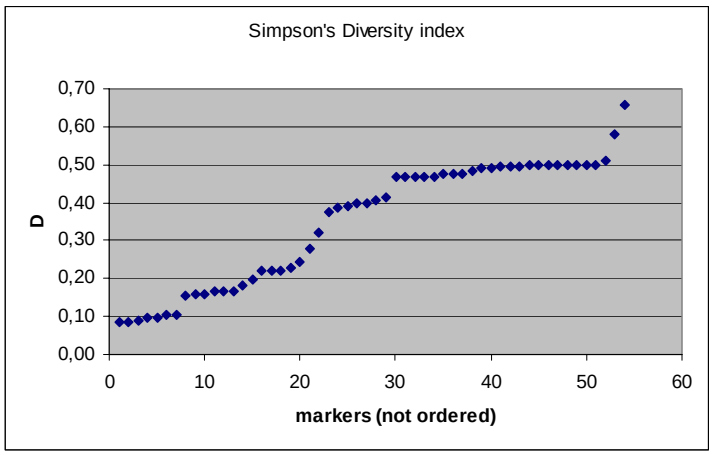

Figure 11. *Simpson's diversity index (D)*

4.4 Genetic diversity

4.4.1 Grouping

Phylogenetic inference analysis on the found indel markers using NJ and Single Linkage (SL is a variation of UPGMA method) with Pearson correlation suggested presence of 7 clades (Fig. 12). The three *novicida* strains (FSC 040, 454 and 595) clustered together, forming a group with the greatest genetic diversity (long branches within the group). The ssp. *tularensis* strains clustered into 2 distinct clades (A.I: FSC 041, 046, 237 and A.II: FSC 054, 230, 604). The strains from Central Asia (FSC 147, 148, 149) clustered together in the ssp. *mediasiatica* group. The Japanese strains formed a distinct ssp. *holarctica* group. The two N. American ssp. *holarctica* strains (FSC 035, 012) also formed a distinct clade. The remaining ssp. *holarctica* strains (FSC 155, 171, 257, 398, 412, 429 and 519) formed a group where FSC

519 and 171 appeared to deviate slightly from the others. Similar grouping was observed when both indel and VNTR markers were used for phylogenetic inference, with the exception of the European ssp. holarctica group, the Örebro strains FSC 398, 412, 429 now appeared to form a slightly distinct clade (no VNTR data on FSC 171 was available) (Fig. 13). Since only a few (7 out of 25) VNTR markers were available for the investigated strains, a phylogenetic inference based only on VNTR markers was omitted. The bootstrapped MP from Phylip, with one altered, questionable character value for FSC230, a member of A.II, supports the grouping found by SL method using indels (Fig. 14). Without this alteration the separation of A.I and A.II groups was not as obvious (data not shown). The bootstrap values are very high for all the groupings except for branching-of of ssp. *mediasiatica* and the split between A.I and A.II. However the branch-of of the N. American ssp. *holarctica* group is highly supported by bootstrap and branch length.

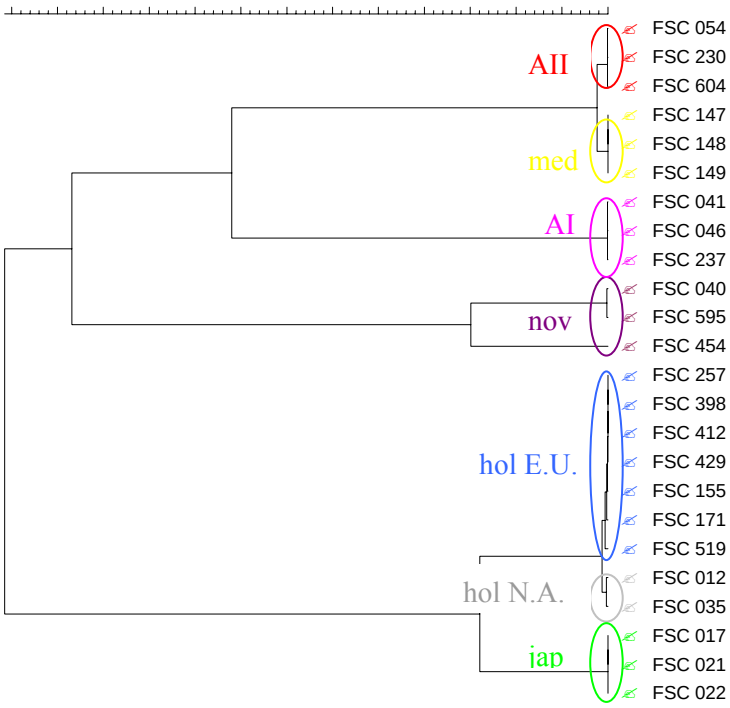

Figure 12. *Single Linkage analysis of* indel *markers. Identifiable groups:* novicida *(purple: FSC 040, 595, 454),* tularensis *A.I (pink: FSC 046, 237, 041),* tularensis *A.II (red: FSC 230, 604, 054),* mediasiatica *(yellow: FSC 147, 148, 149),* japonica *(green: FSC 017, 021, 022),* holarctica *(grey: FSC012, 035; blue: FSC155, 257, 519, 412,429,398).*

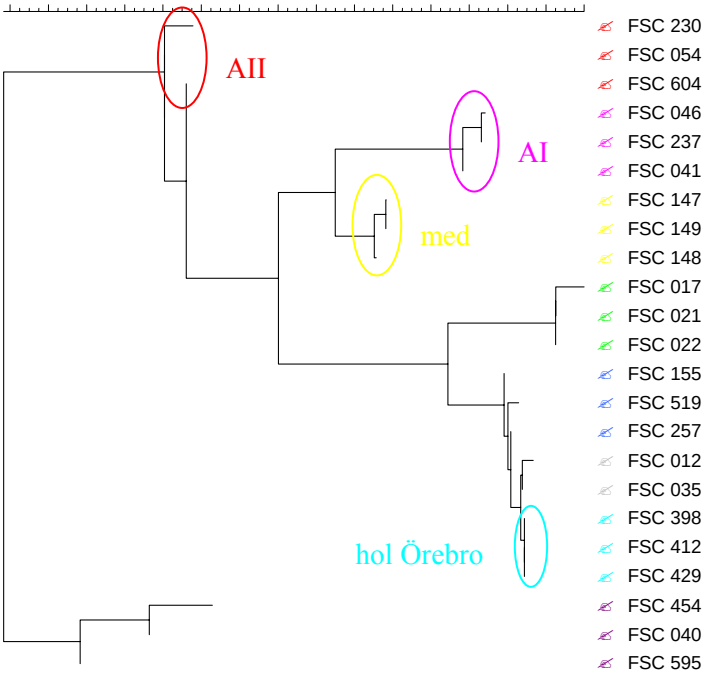

Figure 13. *Neighbour Joining analysis of* indel- *and VNTR data. New identifiable group:* holarctica *"Örebro" (turquoise: FSC 412,429,398*

4.4.2 Phylogenetic relationship

The phylogenetic analysis, using the found indel markers, showed a closer relationship between A.II and *mediasiatica* groups, than between A.II and A.I when SL method was used (Fig. 12), however both the NJ and MP analysis showed a closer relationship between the two ssp. *tularensis* clades A.I and A.II

(Fig. 14, 15). These three groups: A.I, A.II and *mediasiatica* also showed a closer relationship to each other than to the remaining ssp. *holarctica* strains. Of the ssp. *holarctica* groups the Japanese clade seems to have diverged the most, while a closer relationship was seen between the groups N. American and the European strains. When VNTR markers were added to the calculations a slightly different phylogenetic relationship was observed. The NJ method suggested that A.I and *mediasiatica* were more closely related than A.I and A.II (Fig. 13), while SL suggested once again that ssp. *mediasiatica* and A.II were more closely related to each other than A.I and A.II (Fig. 16). The clustering of ssp. *holarctica* strains from Örebro was confirmed in the SL analysis; however the N. American strains from ssp. *holarctica* did no longer form a distinct clade. The NJ method supported the previously found grouping but showed closer relationship between the N. American ssp. *holarctica* group and the previously mentioned European group of ssp. *holarctica* (Fig. 13).

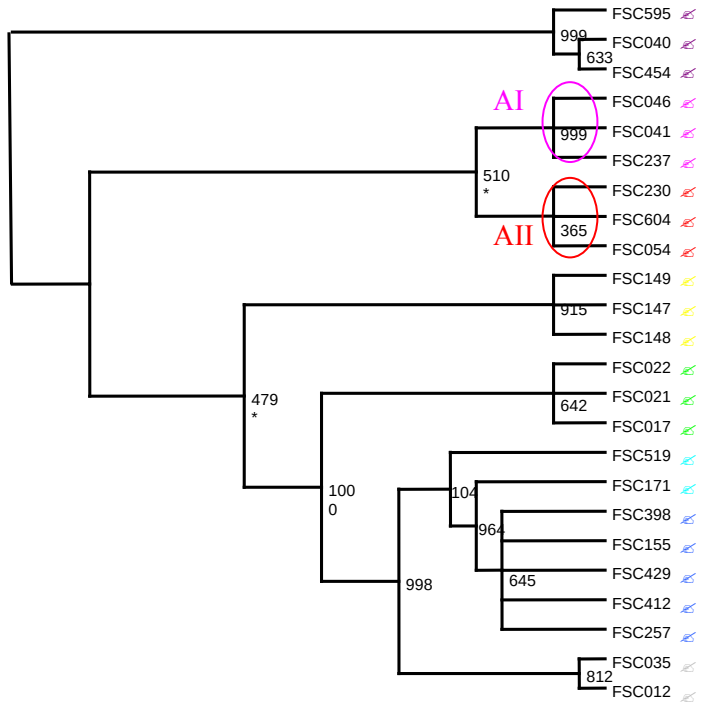

Figure 14. *Bootstrapped MP of indel markers. Bootstrap values are represented where branches divide. The important group separations are supported by bootstrap values (although two are relatively low*). A.II appears to be more closely related to A.I than to* mediasiatica *group.*

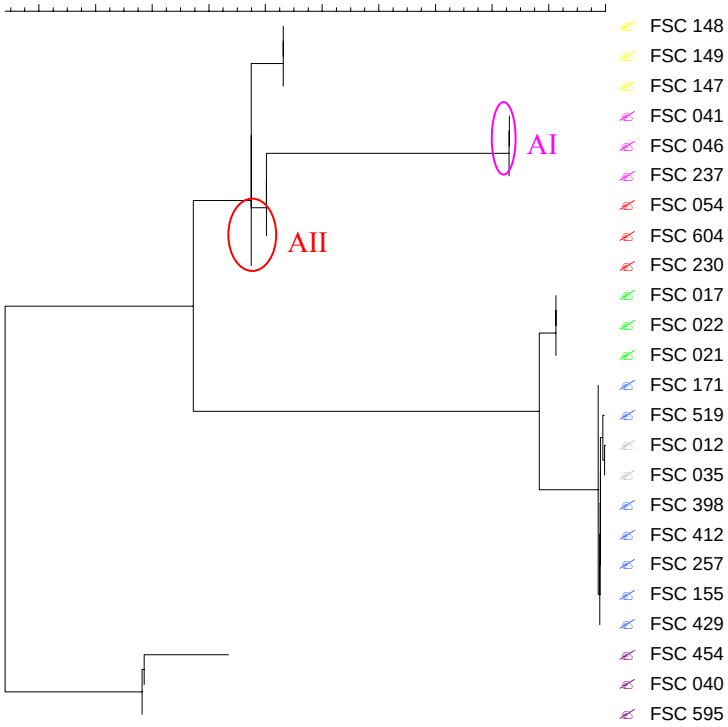

Figure 15. *NJ analysis of* indel *markers.*

4.5 MLVDA

Both PCA and PLS-DA performed on indel and on indel + VNTR data could successfully identify the seven groups found in the phylogenetic inference. Clustering trends of these groups, corresponding to the tree structures, were also discovered. The first two principal components, in the PCA and PLS-DA analysis, were sufficient to cluster all the groups and show

previously seen relationship patterns among them, with the exception of the division between N. American ssp. *holarctica* strains and one strain (FSC 519) from the European ssp. holarctica clade (Fig. 17). To make that separation a third principal component was needed (Fig. 18). The closer relationship between A.II ssp. *mediasiatica*, previously suggested in NJ of indel data and SL of indel + VNTR data (Fig. 12 and 16), was supported both by PCA and PLS-DA. Addition of VNTR data was also here able to distinguish a previously mentioned Örebro subgroup among the European ssp. *holarctica* strains (Fig. 19).

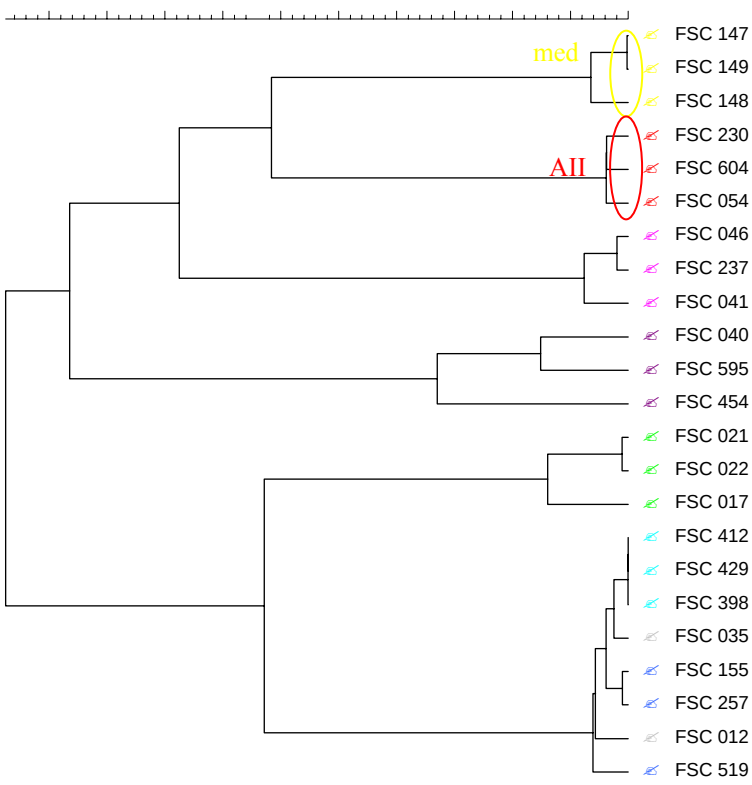

Figure 16. *Single Linkage analysis of* indel- *and VNTR data. Groups:* novicida *(purple: FSC040, 595, 454),* tularensis *A.I (pink: FSC046, 237, 041),* tularensis *A.II (red: FSC230, 604, 054),* mediasiatica *(yellow: FSC147, 148, 149),* japonica *(green: FSC017, 021, 022),* holarctica *(turquoise: FSC412, 429, 398; blue: FSC257, 155, 519; grey: FSC012, 035).*

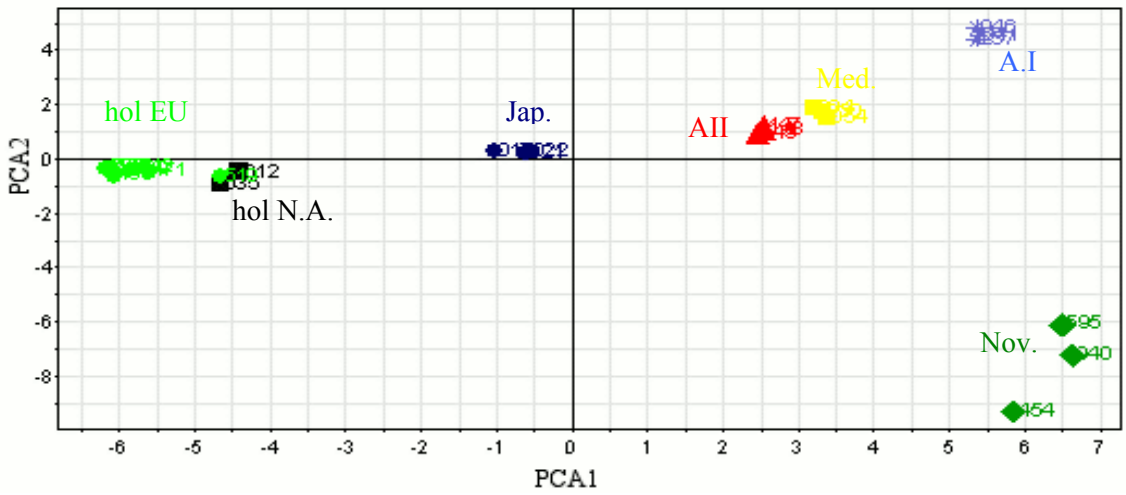

Figure 17. *PCA of indel markers. Groups:* novicida *(green: FSC040, 454, 595),* tularensis *A.I (blue stars: FSC041, 046, 237),* mediasiatica *(yellow: FSC147, 148, 149),* tularensis *A.II (red: FSC054, 230, 604),* japonica *(blue circles: FSC017, 021, 022), N. American* holarctica *(black: FSC012, 035) + FSC019 (light green), other* holarctica *(light green: FSC155, 171, 257, 398, 412, 429).*

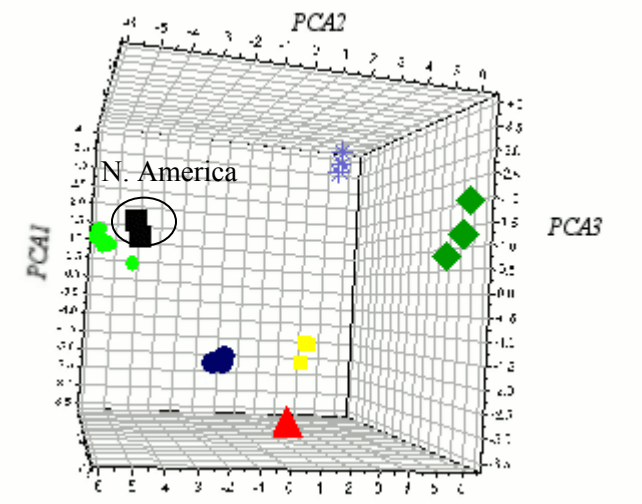

Figure18. *PCA of indel markers using 3 principal components. N. American ssp.* holarctica *strains (encircled) appear to be separated from the rest.*

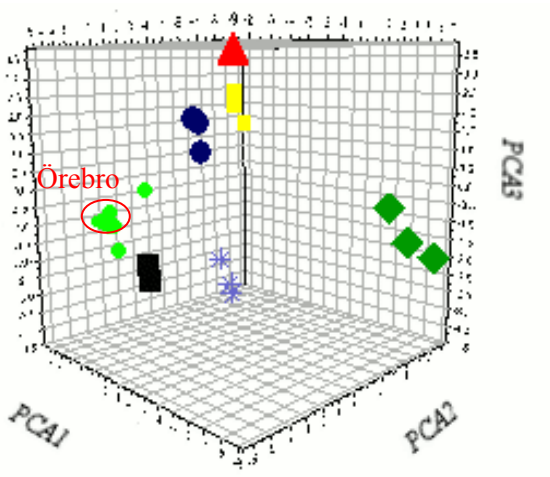

Figure 19. *PCA of indel and VNTR data. Grouping of 3 of the strains from* holarctica *(light green: FSC398, 412, 429) can be seen.*

4.6 Identification markers

A number of specific markers, able to distinguish the previously mentioned subspecies and groups of *F. tularensis* were found (Table 8). An even greater subset of the found markers can be used for identification if combinations are permitted to identify the right subspecies (Table 8). To distinguish the group A.II from all others a combination of any marker specific for A.I with Ft-ID60 specific for *F. tularensis* ssp. *tularensis* can be used. A combination of any marker specific for *F. tularensis* ssp. *holarctica* excl. *japonica* with Ft-ID31 can be used to identify the group consisting of *holarctica* strains from Europe. Remaining groups can be identified using the discovered unique markers (Table 8).

Table 8. *Identification markers. The markers uniquely presented in one group are shadowed.*

group	*unique markers*
novicida	1,2,3,65
tularensis	60
A.I	6,7,8,57,58,59
A.II	60 & A.I
holarctica - japonica	11,19,30,50,52,53,54
holarctica + japonica	12,14,15,18,24,28,
holarctica *(EX.) - FSC519*	47,48,49
holarctica *(E.U.)*	31 & *holarctica-japonica*
holarctica *(NA)*	44
mediasiatica	12
japonica	57

The VIPs for different markers confirm the results above in cases when the markers discriminate between many strains, however the markers identifying a smaller number of strains appear at the end of the VIP list (App. B). That notion is also supported by the results from PLS-DA analysis with only the markers appearing at the beginning of the list, because it cannot separate the N. American ssp. *holarctica* strains from the European ones (data not shown).

Interesting and unexpected results were obtained for several of the investigated markers. Marker Ft-ID21 show similar values for ssp. *novicida*, *tularensis* A.I and *mediasiatica* while *tularensis* A.II and holarctica have another value. Marker Ft-ID31 showed similar values in ssp. *novicida* and the distant N. American ssp. *holarctica* –strains (App. C).

5 Discussion

In silico indel search

The *in silico* search resulted in 155 apparently informative indel markers subdivided in 14 different classes, based on the presence/absence of the indel sequence in the four available genome sequences. The classes 5, 6, 9 and 10 contained no indels. It is a reasonable result since these classes represent phylogenetically contradictory marker classes, based on previously found sub-special division of *F. tularensis* [14]. A marker from any of these empty classes would represent an indel showing similarity in distinct subspecies at the same time as it shows dissimilarity within its own subspecies. For that to happen, a deletion or an insertion of a certain size and sequence has to take place at the same position in the genome of two distinct subspecies. This type of orthology is highly improbable when it comes to indels that are not VNTRs or have another repetitive character.

The fact that the number of markers distinguishing the two ssp. *holarctica* genomes (C & D) is so low is also an expected result since they both belonged to the same subspecies and have therefore diverged much later from each other than from the other genomes in the search, acquiring fewer differences.

The robustness of the search algorithm and the correct selection of ssp. *novicida* genome (A) as the primary query sequence in the BLAST search was demonstrated by the diminishing findings of new valid indels when BLAST parameters and the query sequence were changed.

The problem with the large number of false positives in the *in silico* search can probably be solved by eliminating more or all of the repetitive sequences. Instead of only eliminating the three most common of the large repetitive sequences, discovered in the annotation of SCHUS4 sequence [3], one could find all the large repetitive sequences for each genome using BLAST and eliminate them using the *Crossmatch* tool in all the analyzed sequences.

The algorithm used and implemented in the *in silico* search of this study would also be more powerful if more genome sequences were used, providing even more genetic information about the relationships between the subspecies. Therefore this approach will gain further strength and accuracy as more *F. tularensis* genomes are sequenced and annotated, a process well under way.

Screening

The TP-PCR method, used in indel screening, proved to be a very reliable and fast tool for performing large scale amplifications of the kind necessary for this project. It will most certainly be used when the remaining indels will be investigated.

Because of the nature of this study (non-exhaustive) all available strains and found indels could not be screened and a selection had to be made. The strains were selected based on prior assumptions about their relationship (subspeciation) and their need to be classified as a member of one or another sub-specie or group of *F. tularensis*. There are currently more than 600 different *F. tularensis* strains available at the Swedish Defence Research Agency, where the 24 analyzed strains were obtained. This means that a better selection of strains, to answer some of this project's questions, is possible and to obtain definitive answers about phylogeny a lot more strains need to be screened. When it comes to the selection of a subset of markers

for the screening, it was mostly guided by the selection of strains and the questions posed by this project. Many markers were initially selected from class 3 (Table 7) because the location of *F. tularensis mediasiatica* on the phylogenetic tree was sought after and there were suspicions that it might be somewhere between *F. tularensis* ssp. *holarctica* and *tularensis*.

Since the laboratory at Swedish Defence Research Agency lacked strain D, whose sequence was used in the *in silico* search (Genome D), the indels found exclusively in it could not be properly verified. The problem was however partly solved by using a closely related strain (determined by MLVA).

A difficult and crucial part of the screening process was the design of oligonucleotide primers for the amplification of the marker regions. It could also be the reason for why some of the initially selected (65) markers failed to work. During the optimization of the amplification procedure it was noted that the annealing temperature (T_m) had a large effect on amplification efficiency. That could be explained by the fact that the estimated T_m for the primers was not necessarily very accurate. In fact, different calculation programs gave different T_m values for the primers. There was also no way of knowing how much the tail-sequence would affect the ability of a primer to anneal to its target, but since most of the primers worked it is safe to assume that its presence did not seriously affect the binding of the molecular primers.

Clustering and phylogeny

The found indel markers confirmed previous results about the phylogenetic relationship among subspecies of *F. tularensis* and allowed a further development of phylogeny, positioning one of the previously unpositioned subspecies and identifying several clades within subspecies. The average- and the median values for the diversity index (D) suggest a general stability of the indel system. The fact that most of the indel markers have a diversity that is close to 0.5 means that indel markers probably lack the sufficient variability to distinguish individual bacterial clones from one another but should be perfectly suited for a higher level of discrimination. This reasoning is well supported by indel markers' ability to correctly identify the subspecies of *F. tularensis* as well as groups of strains within these subspecies. This is shown both in the clustering analysis (PCA and PLS-DA) and methods for phylogenetic inference (MP, NJ and SL). The overall structure with 7 clades (both sub-special and on a lower level) is well supported in all the methods when only indels are used (Fig. 12, 14, 15 and 16). The addition of partial (7 of 25 available markers) VNTR data allows separation of a new subgroup when analyzed with NJ and PCA but prevents clustering of another previously established group (N. American ssp. *holarctica*). The VNTR markers have previously been shown to have an extremely high discriminatory capability (diversity index up to 0.95) [13], which can explain the extra subgroup found when the partial VNTR data was used. The loss of one of the previously classified groups could have been caused by problems in the NJ method (since it was the only method to loose that cluster) or the incompleteness of the VNTR data, due to both the limited number of strains used in this study and because only the data from seven out of twenty-five VNTRs were available. Random variations in the number of repeats in VNTR analysis might also have interfered with the correct clustering.

All the phylogenetic inference methods and the projection methods used in this study support the fact that the Japanese strains seem to form intermediate clade separating the ssp. *holarctica* strains from the rest. The suspicion about the Japanese strains forming a separate subspecies has previously been presented [14] and this study supports it as well. However, to

be classified a subspecies more strains need to be screened and included in the analysis.

The main difference in results between the model based method MP and the distance methods NJ and SL in their phylogenetic inference, when only indel data was considered, lies in the relationship between *F. tularensis* ssp. *tularensis* and *mediasiatica*. The most probable result is that *tularensis* A.II is more closely related to *tularensis* A.I rather than to *mediasiatica* and is supported by the MP analysis. The explanation for why the distance and projection methods fail to come to the same conclusion lies in their nature to consider both similarities and differences (distances) between different taxa, while MP only looks at the similarities. In this case the complete distances may be somewhat misleading since an indel in one group of taxa can, in distance methods, bring the groups lacking this indel closer together when in reality a lack of an indel does not infer closer phylogenetic relationship, since that indel can have happened after the split (App. D, Ft-ID 6, 7, 8, 58, 59). A model based method like MP will however not cluster the taxa lacking an indel closer together because it only infers greater phylogenetic relationship if two groups/taxa share a common feature.

Identification

A subset of the found indel markers possesses the ability to, individually and/or in combination, discriminate and identify the found phylogenetic groups of *F. tularensis* present in this study (Table 8). The only groups that needed a combination of two indel markers to be identified were: ssp. *tularensis* A.II and the European ssp. *holarctica*. The other groups could be identified using individual markers. The *in silico* search predicted several individual indel markers for the European and N. American ssp. *holarctica* strains but only one of these predicted indels has been verified. This can be a consequence of the previously mentioned fact that strain of Genome D used in the *in silico* prediction, was not available for screening. Another source of error could have been the fact that only two N. American *F. tularensis* ssp. *holarctica* strains were used and if one of them was an intermediate strain it could have interfered with the determination of which markers should be chosen for identification.

The fact that not all the indel markers supported the correctly established phylogenetic relationship can be a sign of genetic mechanisms such as recombination and horizontal gene transfer at work in *F. tularensis* genome or a trace of two similar indel events in different groups of strains. Markers Ft-ID21 and Ft-ID31 (App. C) show similar values for distantly related groups of strains, so either these indels two separate indel events or they are evidence of horizontal gene transfer (a recombination event), between *F. tularensis* ssp. *tularensis* A.II and *holarctica* in case of Ft-ID21, and the N. American *F. tularensis* ssp. *holarctica* and any of *F. tularensis* ssp. *novicida* or *tularensis* in case of Ft-ID31. It is impossible to determine which of the suggested events is more likely to have taken place since there are no studies showing one to be more probable than the other.

Some markers (Ft-ID 45, 47, 48, 49) are interesting in a geographic sense, since they show similarity in species that have their natural habitat on different continents, but it could also be that the indels observed in these strains had appeared before the strains split and moved to distinct continents.

A demonstration of the potential of indel markers and the *in silico* approach used in this study was, that despite only four different sequences used in the initial search, all seven groups presented in the screened subset of strains could be identified.

Future developments

This study supports the idea that indel markers can be used both as discriminating markers for identification of groups of strains and subspecies, and for determination of stable and reliable phylogenetic relationships. A natural step towards the complete achievement of these goals is to screen more strains of *F. tularensis* and conclude the investigation of the rest of the found indel markers. More information about the nature and stability of the indels can be obtained by investigating the surrounding sequences; determining if the region is coding for a protein, look for any repetitive sequences in the vicinity of the indel and investigate if the region is subjected to frequent recombination, etc.. When the necessary subset of indel markers is established and verified, an even faster screening and amplification might be achieved by converting the used PCR methodology into RT-PCR protocols. A combination of the results from the use of this indel method with the results from some of the existing typing methods like MLVA and SNP analysis or/and a combination of the different types of markers i.e. indels and VNTRs has the potential to achieve an even greater typing coverage and reliability and will hopefully present the standard future tool for strain identification and fast and reliable phylogenetic inference.

Conclusion

The *in silico* search for a new type of phylogenetic identification markers, using the currently available genome sequences of *F. tularensis*, and a screening of twenty-four representative strains of that pathogen, enabled the finding of a system for identification of subspecies and groups of *F. tularensis* using phylogenetically informative indel markers. Using the newly found type of multi-locus identification markers the study was able to confirm earlier findings of phylogenetic relationship between the subspecies of *F. tularensis* and extend these to include other previously undefined phylogenetic relationships. The study suggests a place for *F. tularensis* ssp. *mediasiatica* in the phylogenetic tree of *F. tularensis* and supports previous implications, that the strains of *F. tularensis* ssp. *holarctica* from Japan should be considered a separate subspecies. Some markers showed possible presence of horizontal gene transfer between the subspecies of *F. tularensis* and demonstrated interesting geographical relationships. The future extension of this study, to more strains of *F. tularensis* and more indel markers, has a potential to result in a standard tool for fast and reliable identification of bacterial strains and determination of phylogenetic relationship within bacterial species.

6 Acknowledgments

Pär Larsson: for support accompanied by invaluable discussions in all aspects of the project.
Kerstin Svensson: for motivation and support, and for careful revision of the report.
Mats Forsman: for revision of the report and inspiring discussions throughout the project.
Per Wikström: for great help and interesting suggestions during the analysis.
Ann-Christine Andersson, **Linda Karlsson** and **Malin Granberg:** for help and support during the screening process.
Mona Byström and **Stina Bäckman**: for providing the screening material.
Lotta Avesson and **Greta Hultqvist**: for revision of the report and help during its presentation.
Johan Tegman: for IT-support.

7 References

1. **Hopla C. E., Hopla A. K., (**1994). Tularemia. *In: Beran GW, Steele JH, eds. Handbook of Zoonoses.* 2nd ed. Boca Raton, Fla: CRC Press; 113-126.

2. **Forsman M., Johansson A**., (2005). Tularemia. *In: Encyclopedia of Bioterrorism Defence*. (Eds.) Richard F. Pitch and Raymond A. Zilinskas. John Wiley & Sons, Hoboken NJ. July. 2005.

3. **Larsson P., Oyston P. C., Chain P., Chu M. C., Duffield M., Fuxelius H. H., Garcia E., Halltorp G., Johansson D., Isherwood K. E., Karp P. D., Larsson E., Liu Y., Michell S., Prior J., Prior R., Malfatti S., Sjostedt A., Svensson K., Thompson N., Vergez L., Wagg J. K., Wren B. W., Lindler L. E., Andersson S. G., Forsman M., Titball R. W.,** (2005). The complete genomic sequence of *Francisella tularensis*, the causative agent of tularemia. *Nature Genetics*, 37(2):153-159.

4. **Rotz L. D., Khan A. S., Lillibridge S. R., Ostroff S. M., Hughes J. M.,** (2002). Public health assessment of potential biological terrorism agents. *Emerg. Infect. Dis.* 8(2):225-230.

5. **Davis CJ.,** (1999). Nuclear blindness: An overview of the biological weapons programs of the former Soviet Union and Iraq. *Emerg. Infect. Dis*. 5(4):509-512.

6. **WHO**. (1970) Health Aspects of Chemical and Biological Weapons. World Health Organization, Geneva. 1970:105-107.

7. **Forsman M., Sandström G., Sjöstedt A.,** (1994). Analysis of 16S ribosomal DNA sequences of Francisella strains and utilization for determination of the phylogeny of the genus and for identification of strains by PCR. *Int. J. Syst. Bacteriol.* 44(1):38-46.

8. **Sjöstedt A. B.,** (2005). *Francisella*. In D. J. Brenner, N. R. Krieg, J. T. Staley, and G. M. Garrity (ed.), The Proteobacteria, part B. *Bergey's manual of systematic bacteriology,* 2nd ed., vol. 2. Springer-Verlag, New York, N.Y. p.200–210.

9**. Johansson A., Ibrahim A., Göransson I., Eriksson U., Gurycova D., Clarridge III J. E., Sjöstedt A.,** (2000). Evaluation of PCR-based methods for discrimination of *Francisella* species and subspecies and development of specific PCR that distinguishes the two major subspecies of *Francisella tularensis*. *J. Clin. Microbiol.* 38:4180-4185.

10. **Titball R. W, Johansson A., Forsman M.,** (2003). Will the enigma of Francisella tularensis' virulence soon be solved? *Trends. Microbiol.* 11(3):118-123.

11. **Dennis D. T., Inglesby T. V., Henderson D. A., Bartlett J. G., Ascher M. S., Eitzen E., Fine A. D., Friedlander A. M., Hauer J., Layton M., Lillibridge S. R., McDade J. E., Osterholm M. T., O'Toole T., Parker G., Perl T. M., Russell P. K., Tonat K.; Working Group on Civilian Biodefense.** (2002). Tularemia as a biological weapon: medical and public health management. *JAMA*, 23-30; 287(4):452-453.

12. **Broekhuijsen M., Larsson P., Johansson A., Byström M., Eriksson U., Larsson E., Prior R. G., Sjöstedt A., Titball R. W., Forsman M.**, (2003). Genome-wide DNA microarray analysis of *Francisella tularensis* strains demonstrates extensive genetic conservation within the species but identifies regions that are unique to the highly virulent *F. tularensis* ssp. *tularensis*. *J. Clin. Microbiol.* 41:2924-2931.

13. **Johansson A., Farlow J., Larsson P., Dukerich M., Chambers E., Bystrom M., Fox J., Chu M., Forsman M., Sjostedt A., Keim P.,** (2004). Worldwide Genetic Relationships among *Francisella tularensis* Isolates Determined by Multiple-Locus Variable-Number Tandem Repeat Analysis. *J. Bacteriol.* 186(17):5808-5818.

14. **Svensson K, Larsson P, Johansson D, Bystrom M, Forsman M, Johansson A.,** (2005). Evolution of subspecies of Francisella tularensis. *J. Bacteriol.* 187(11):3903-3908.

15. **Gupta R. S., Griffiths E.,** (2002). Critical Issues in Bacterial Phylogeny. *Theoret. Pop. Biol.* 61:423-434.

16. **Needleman S. B., Wunsch C. D.,** (1970). Needleman-Wunsch Algorithm for Sequence Similarity Searches. *J. Mol. Biol.* 48:443-453.

17. **Smith T. F., Waterman, M. S.**, (1981). Identification of common molecular subsequences. *J. Mol. Biol.* 147(1):195-197.

18. **Altschul S. F., Gish W., Miller W., Myers E. W., Lipman D. J.,** (1990). Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.

19. **Altschul S. F., Madden T. L., Schaeffer A. A., Zhang J., Zheng Z., Miller W., Lipman D. J.,** (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.

20. **Higgins D., Thompson J., Gibson T., Thompson J. D., Higgins D. G., Gibson T. J.,** (1994). CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673-4680.

21. **Oetting W. S., Lee H. K., Flanders D. J., Wiesner G. L., Sellers T. A., King R. A.,** (1995). Linkage analysis with multiplexed short tandem repeat polymorphisms using infrared fluorescence and M13 tailed primers. *Genomics* 30(3):450-458.

22**. Eck R. V., Dayhoff., M. O.,** (1966). *Atlas of Protein Sequence and Structure 1966*. National Biomedical Research Foundation. Silver Spring. Maryland.

23. **Hendy M. D., Penny. D.,** (1989). A framework for the quantitative study of evolutionary trees. *Systematic Zoology* 38:297-309.

24. **Nei M., Tajima F., Tateno Y.,** (1983), Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. *J. Mol. Evol.* 19(2):153-170.

25. **Sneath P. H., Sokal R.R.,** (1962). Numerical taxonomy. *Nature*, 193:855-860.

26. **Saitou N., Nei M.,** (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Mol. Biol. Evol.*, 4:406-425

27. **Rzhetsky A., Nei M**., (1992). A simple method for estimating and testing Minimum-Evolution trees. *Mol. Biol. Evol.,* 9:945–967.

28. **Yang Z.,** (1993). Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.,* 10:1396-1401.

29. **Mau B., and Newton M., (**1997). Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *JCGS,* 6:122-131.

30. **Eriksson L., Johansson E., Kettaneh-Wold N., Wold S.,** (1999). Introduction to Multi- and Megavariate Data Analysis using Projection Methods (PCA & PLS). *Umetrics AB*. Umeå.

31. **Walker, D.R., and Koonin, E.V.,** (1997). SEALS: A System for Easy Analysis of Lots of Sequences. *Intelli. Syst. Mol. Biol.,* 5:333-339.

32. **Sandstrom G, Tarnvik A, Wolf-Watz H, Lofgren S.,** (1984). Antigen from Francisella tularensis: nonidentity between determinants participating in cell-mediated and humoral reactions. *Infect. Immun*., 45(1):101-106.

33. **Rice P. Longden I. and Bleasby A**., (2000). **EMBOSS**: The European Molecular Biology Open Software Suite. *Trends Gen.,* 16:276-277.

34. **Simpson E. H.**, (1949). Measurement of diversity. *Nature,* 163:688

35. **Felsenstein, J.,** (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics,* 5:164-166.

8 Appendices

Appendix A – Indel markers

(Only the 47 successfully screened markers are shown). The four positions of the coding column represent the four available genome sequences (A, B, C and D). The "X" denotes the presence of the sequence for that particular indel, while "-", denotes its absence. The diversity index is based on the twenty-four analyzed strains.

marker	*size*	*class*	*coding*	*Diversity*
Ft_ID1	47	1	X,-,-,-	0,41
Ft_ID2	9	1	X,-,-,-	0,16
Ft_ID3	178	1	X,-,-,-	0,16
Ft_ID4	14	1	X,-,-,-	0,18
Ft_ID5	6	1	X,-,-,-	0,15
Ft_ID6	15	2	-,X,-,-	0,22
Ft_ID7	6	2	-,X,-,-	0,22
Ft_ID8	10	2	-,X,-,-	0,23
Ft_ID11	24	3	X,X,-,-	0,47
Ft_ID12	6	3	X,X,-,-	0,39
Ft_ID13	55	3	X,X,-,-	0,40
Ft_ID14	13	3	X,X,-,-	0,50
Ft_ID15	8	3	X,X,-,-	0,50
Ft_ID16	102	3	X,X,-,-	0,50
Ft_ID17	27	3	X,X,-,-	0,48
Ft_ID18	65	3	X,X,-,-	0,50
Ft_ID19	8	3	X,X,-,-	0,47
Ft_ID21	9	3	X,X,-,-	0,47
Ft_ID22	7	3	X,X,-,-	0,50
Ft_ID23	180	3	X,X,-,-	0,50
Ft_ID24	21	3	X,X,-,-	0,50
Ft_ID26	50	3	X,X,-,-	0,49
Ft_ID27	10	3	X,X,-,-	0,49
Ft_ID28	34	3	X,X,-,-	0,50
Ft_ID29	5	3	X,X,-,-	0,50
Ft_ID30	63	3	X,X,-,-	0,50
Ft_ID31	19	4	-,-,X,-	0,50
Ft_ID39	8	14	_,X,X,X	0,15
Ft_ID44	6	8	-,-,-,X	0,17
Ft_ID45	10	11	X,X,-,X	0,51
Ft_ID47	5	11	X,X,-,X	0,38
Ft_ID48	5	11	X,X,-,X	0,39
Ft_ID49	10	11	X,X,-,X	0,40
Ft_ID50	23	12	-,-,X,X	0,48
Ft_ID51	5	12	-,-,X,X	0,66
Ft_ID52	7	12	-,-,X,X	0,47
Ft_ID53	6	12	-,-,X,X	0,47
Ft_ID54	6	12	-,-,X,X	0,48
Ft_ID56	5	12	-,-,X,X	0,58
Ft_ID57	91	13	X,-,X,X	0,48
Ft_ID58	27	13	X,-,X,X	0,24
Ft_ID59	118	13	X,-,X,X	0,17
Ft_ID60	7	13	X,-,X,X	0,42
Ft_ID62	10	14	-,X,X,X	0,17
Ft_ID63	8	14	-,X,X,X	0,20
Ft_ID64	36	14	-,X,X,X	0,28
Ft_ID65	48	14	-,X,X,X	0,22

Appendix B – VIP list from PCA analysis of indel marker

ID	PC1	PC2	PC3	Σ
Ft-ID17	1.34339	0.989934	0.980856	3.31418
Ft-ID2	0.588076	1.42761	1.28716	3.302846
Ft-ID11	1.32476	0.984209	0.977833	3.286802
Ft-ID30	1.32531	0.98152	0.973132	3.279962
Ft-ID53	1.26921	0.94714	1.0559	3.27225
Ft-ID50	1.29507	0.952031	1.01022	3.257321
Ft-ID52	1.32301	0.974723	0.958422	3.256155
Ft-ID54	1.27651	0.934014	1.02098	3.231504
Ft-ID65	0.69225	1.34596	1.17637	3.21458
Ft-ID3	0.687054	1.3504	1.17638	3.213834
Ft-ID4	0.638273	1.33281	1.2313	3.202383
Ft-ID63	0.561948	1.38999	1.24966	3.201598
Ft-ID62	0.544189	1.40231	1.24532	3.191819
Ft-ID19	1.30111	0.965157	0.924589	3.190856
Ft-ID1	0.644918	1.33783	1.20394	3.186688
Ft-ID64	0.722585	1.296	1.13	3.148585
Ft-ID29	1.23941	0.911603	0.9829	3.133913
Ft-ID45	1.04962	1.11804	0.924232	3.091892
Ft-ID5	0.829019	1.21375	1.03789	3.080659
Ft-ID23	1.31484	0.964075	0.797098	3.076013
Ft-ID18	1.30932	0.956311	0.789066	3.054697
Ft-ID28	1.30233	0.95218	0.787572	3.042082
Ft-ID15	1.3044	0.950972	0.782441	3.037813
Ft-ID14	1.29939	0.952266	0.784035	3.035691
Ft-ID24	1.29343	0.949964	0.7813	3.024694
Ft-ID16	1.28434	0.938085	0.772999	2.995424
Ft-ID58	0.608145	1.00156	1.383	2.992705
Ft-ID59	0.593685	0.993146	1.40102	2.987851
Ft-ID7	0.679695	0.94437	1.35494	2.979005
Ft-ID51	1.24668	0.928707	0.782789	2.958176
Ft-ID6	0.572519	0.96089	1.41994	2.953349
Ft-ID60	0.790673	1.07269	1.05594	2.919303
Ft-ID8	0.531847	1.01248	1.35354	2.897867
Ft-ID13	1.15922	0.925064	0.777028	2.861312
Ft-ID27	1.13472	0.93283	0.792463	2.860013
Ft-ID57	0.581854	1.00954	1.23423	2.825624
Ft-ID48	1.11843	0.866034	0.728134	2.712598
Ft-ID49	1.09048	0.799982	0.728481	2.618943
Ft-ID47	1.09926	0.810981	0.705748	2.615989
Ft-ID56	1.10467	0.809706	0.677546	2.591922
Ft-ID31	0.909578	0.663285	0.980716	2.553579
Ft-ID21	1.04222	0.76593	0.72125	2.5294
Ft-ID26	0.953607	0.692474	0.665076	2.311157
Ft-ID12	0.517812	0.376529	1.21408	2.108421
Ft-ID44	0.670187	0.493321	0.725617	1.889125
Ft-ID32	0.719824	0.526603	0.482551	1.728978
Ft-ID39	0.043703	0.902342	0.750756	1.696801
Ft-ID22	0.137078	0.735164	0.64144	1.513682

Second to fourth column represent each marker's clustering importance value for each of the three principal components. The last column contains the sum of the three. The markers in orange are identifying big clusters. The ones marked in blue identify intermediate clusters and the green ones identify a small number of strains from the rest.

Appendix C – The indel data coded as discrete characters

Vertically – markers. Horizontally – strains

FSC	40	454	595	237	41	46	54	230	604	147	148	149	17	21	22	155	171	398	412	429	519	257	35	12
Ft-ID1	2	2	2	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
Ft-ID2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ft-ID3	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ft-ID4	2	2	2	1	1	1	1	1	1	n/a	n/a	n/a	1	1	1	1	1	1	1	1	1	1	1	1
Ft-ID5	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ft-ID6	1	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ft-ID7	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ft-ID8	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ft-ID11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
Ft-ID12	2	2	2	2	2	2	2	2	2	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
Ft-ID13	1	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Ft-ID14	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Ft-ID15	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Ft-ID16	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Ft-ID17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
Ft-ID18	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Ft-ID19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
Ft-ID21	2	1	2	2	2	2	0	0	0	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0
Ft-ID22	2	0	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1
Ft-ID23	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Ft-ID24	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Ft-ID26	2	2	2	2	2	2	2	2	2	0	0	0	2	2	2	1	1	1	1	1	1	1	1	1
Ft-ID27	1	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Ft-ID28	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Ft-ID29	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
Ft-ID30	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
Ft-ID31	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
Ft-ID39	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Ft-ID44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Ft-ID45	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	0	1	0	1	1
Ft-ID47	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	1	0	1	1
Ft-ID48	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	1	0	1	1
Ft-ID49	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	1	0	1	1
Ft-ID50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
Ft-ID51	1	0	1	0	0	0	1	0	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2
Ft-ID52	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
Ft-ID53	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
Ft-ID54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
Ft-ID56	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	2	0
Ft-ID57	2	2	2	0	0	0	2	2	2	2	2	2	1	1	1	2	2	2	2	2	2	2	2	2
Ft-ID58	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ft-ID59	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ft-ID60	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ft-ID62	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ft-ID63	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ft-ID64	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ft-ID65	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1