Infection, Genetics and Evolution 30 (2015) 186–194

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Development of a Multilocus Sequence Typing scheme for the study of *Anaplasma marginale* population structure over space and time

Eliana C. Guillemi [a,*], Paula Ruybal [a], Verónica Lia [a], Sergio Gonzalez [a], Sergio Lew [b], Patricia Zimmer [c], Ludmila Lopez Arias [a], Jose L. Rodriguez [d], Sonia Y. Rodriguez [d], Roger Frutos [e], Silvina E. Wilkowsky [a], Marisa D. Farber [a]

[a] *Instituto de Biotecnologia, Centro de Investigaciones en Ciencias Veterinarias y Agronómicas, INTA, Buenos Aires, Argentina*
[b] *Instituto de Ingeniería Biomédica, Universidad de Buenos Aires, Buenos Aires, Argentina*
[c] *Estación Experimental Agropecuaria Mercedes, INTA, Corrientes, Argentina*
[d] *CORPOICA, Bogotá, Colombia*
[e] *Cirad, UMR 17, Cirad-Ird, TA-A17/G, Campus International de Baillarguet, Montpellier, France*

ARTICLE INFO

Article history:
Received 8 October 2014
Received in revised form 21 December 2014
Accepted 22 December 2014
Available online 27 December 2014

Keywords:
Anaplasma marginale
MLST
Tick-borne diseases
Population structure
Molecular epidemiology

ABSTRACT

Bovine Anaplasmosis caused by *Anaplasma marginale* is a worldwide disease prevalent in tropical and subtropical regions where *Rhipicephalus microplus* is considered the most significant biological vector. Molecular markers previously applied for *A. marginale* typing are efficient for isolate discrimination but they are not a suitable tool for studying population structure and dynamics. Here we report the development of an MLST scheme based on the study of seven genes: *dnaA*, *ftsZ*, *groEl*, *lipA*, *recA*, *secY* and *sucB*. Five annotated genomes (Saint Maries, Florida, Mississippi, Puerto Rico and Virginia) and 53 bovine blood samples from different world regions were analyzed. High nucleotide diversity and a large proportion of synonymous substitutions, indicative of negative selection resulted from DnaSP 5.00.02 package application. Recombination events were detected in almost all genes, this evidence together with the coexistence of more than one *A. marginale* strain in the same sample might suggest the superinfection phenomena as a potential source of variation. The allelic profile analysis performed through GoeBURST shown two main CC that did not support geography. In addition, the AMOVA test confirmed the occurrence of at least two main genetically divergent groups. The composition of the emergent groups reflected the impact of both historical and environmental traits on *A. marginale* population structure. Finally, a web-based platform "Galaxy MLST-Pipeline" was developed to automate DNA sequence editing and data analysis that together with the Data Base are freely available to users.

The *A. marginale* MLST scheme developed here is a valuable tool with a high discrimination power, besides PCR based strategies are still the better choice for epidemiological intracellular pathogens studies. Finally, the allelic profile describe herein would contribute to uncover the mechanisms in how intracellular pathogens challenge virulence paradigm.

1. Introduction

Bovine Anaplasmosis is a tick borne disease caused by the Gram negative bacterium *Anaplasma marginale*, an obligate intracellular parasite of bovine erythrocytes that causes a moderate to severe hemolitic anemia, jaundice and hemoglobinuria without hemoglobinemia (Kocan et al., 2003). *A. marginale* belongs to the phylum Proteobacteria, alpha Proteobacteria class, order Rickettsiales, Anaplasmataceae family.

Rhipicephalus microplus is considered the most important biological vector for *A. marginale* in tropical and subtropical regions of the world (de la Fuente et al., 2007). Since *R. microplus* eradication, tick transmission of *A. marginale* in the United States is mediated by *Dermacentor andersoni* and *Dermacentor variabilis* (Futse et al., 2003). Additionally, other hematophagous insects and the use of infected blood fomites could cause mechanical transmission (Kocan et al., 2003). The economic losses generated by this disease are not only associated with morbidity and mortality in cattle, but also with a lower weight gain rate, lower milk production, abortions and treatment costs. Among control measures to prevent severe morbidity and mortality due to anaplasmosis, *A. marginale*

* Corresponding author.
E-mail address: guillemi.eliana@inta.gob.ar (E.C. Guillemi).

http://dx.doi.org/10.1016/j.meegid.2014.12.027
1567-1348/© 2014 Elsevier B.V. All rights reserved.

*subsp. central*e as live attenuated vaccine is currently being used in tropical and subtropical countries in Africa Asia, Australia, and Latin America (Bock and De Vos, 2001).

It is relevant to achieve accurate methods for genotyping and characterizing strains as well as for studying the structure and dynamics of *A. marginale* populations in the field. Several methods and markers have been developed to characterize the genetic diversity of *A. marginale*. These are mostly focused on the Major Surface Proteins (MSPs) MSP4 and MSP1a (Almazán et al., 2008; de la Fuente et al., 2001, 2002, 2004, 2007; Lew et al., 2002; Mtshali et al., 2007; Palmer et al., 2001; Ruybal et al., 2009; Vidotto et al., 2006) and are useful for discriminating isolates. However, when the genetic population structure is under study, it is important to choose multiple loci that are selectively neutral. In fact, loci under positive selection may give a distorted view of the genetic population structure and its transmission dynamics, since selection rather than population history may determine the patterns of distribution of alleles within populations for these loci.

Previously published data (Ruybal et al., 2009) analyzing the *msp1a* marker refers to a wide genetic diversity of *A. marginale* population in Argentina. In this study, the authors demonstrated that the genetic population diversity was higher for tick-infested regions than for tick-free areas. Moreover, Estrada-Peña et al. (2009) studied the variability of MSP1a sequence worldwide and reported that this molecular marker is associated to the world ecological regions; therefore, the evolution of *A. marginale* may be linked to ecological traits affecting tick vector performance.

In 1998, Multilocus Sequence Typing (MLST) was first proposed for the characterization of isolates of the human pathogen *Neisseria meningitidis* (Maiden et al., 1998). This tool enables genotypic characterization of isolates and the study of the global dispersion of some new variants of pathogens (Mayer et al., 2002). In addition to these epidemiological studies of medical interest, the data obtained by MLST strategy apply to evolutionary and population studies (Jolley et al., 2000). In fact, it can be employed to estimate the frequency of recombination events and mutations and to investigate evolutionary relationships between organisms belonging to the same genus (Godoy et al., 2003).

We report here the development of the first MLST scheme for *A. marginale* and its application for population structure studies. In this scheme, 7 genes are employed for the discrimination of even very closely related strains. The design of the MLST scheme was assisted by the availability of the complete *A. marginale* genome. We have also developed a bioinformatic pipeline for the automated analysis of raw sequences and further diversity and phylogenetic analysis.

The MLST scheme developed in this work was applied for the study of 58 isolates from different world regions. Taking into account the results previously published by Ruybal et al. (2009) and Estrada-Peña et al. (2009), we hypothesized that geographically related isolates will tend to have a more similar genotype composition compared to the geographically distant isolates.

2. Materials and methods

2.1. Strains and genomic DNA isolation

A total of 58 *A. marginale* strains were analyzed. Five of them came from annotated genomes (Saint Maries, Florida, Mississippi, Puerto Rico and Virginia) and the other 53 were collected from countries in North and South America, Europe, and Africa (Table 1). Field samples were detected as positive for *A. marginale* by microscopic observation of Giemsa-stained blood smears and by PCR amplification of the *msp5* gene. Some of the field samples were from the same geographic region (Argentina provinces), even from the same ranch. Additionally, four of them came from outbreaks (Table 1). The genomic DNA extraction was performed by phenol/chloroform method and a standard ethanol precipitation (Sambrook et al., 1989) from PBS-washed and packed infected erythrocytes.

2.2. Target loci

The search was performed using genes applied for other MLST schemes in related microorganisms (Adakal et al., 2009), which have been previously described in the literature (Baldo et al., 2006; Jacobson et al., 2008; Vitorino et al., 2007) as reference. Those genes were single copy and encoded conserved proteins. Fourteen candidate genes were pre-selected. Primers were designed using Primer 3 program (http://frodo.wi.mit.edu/primer3/) for the first approach and then parameters were adjusted manually by IDT OligoAnalyzer tool from Integrated DNA Technologies (http://www.idtdna.com/analyzer/Applications/OligoAnalyzer/).

Specificity was initially evaluated *in silico* by BLASTn (http://blast.ncbi.nlm.nih.gov/) against *Bos taurus*, *Babesia bovis* and *B. bigemina* database, since DNA from these organisms can be found as contaminant in blood samples from field cattle. Later, specificity was corroborated experimentally using DNA from uninfected *B. taurus* and from *Babesia* species. Two reference strains from *A. marginale* (Mercedes and Salta) were amplified with candidate primers and only those PCR products with high sensitivity and specificity were finally selected.

Amplicons were sequenced from both strands and loci showing a neutral pattern were selected for further analysis. The degree of selection operating on the target genes was determined according to the ratio of mean non-synonymous substitutions per non-synonymous site/mean synonymous substitution per synonymous site (*dN*/*dS* ratio). The *dN*/*dS* ratio was calculated using the START2 program available from http://pubmlst.org/software/analysis/start2/ (Jolley et al., 2001).

Seven genes homogeneously distributed through the genome (Table 2 and Supplementary Fig. 1) were finally chosen: *dnaA* (DnaA chromosomal replication initiation protein; AM430), *ftsz* (cell division protein FtsZ; AM1261), *groEl* (Chaperonin GroEL; AM944), *lipA* (lipoyl synthase; AM820), *recA* (RecA recombination protein; AM085), *secY* (preprotein translocase subunit SecY; AM892) and *sucB* (dihydrolipoamide acetyltransferase component; AM1087).

2.3. PCR amplification and gene sequencing

The primers used to amplify and sequence the seven target genes are listed in Table 2. PCR was performed in a 50 µl reaction mixture containing 0.4 µmol of each primer, 0.2 mM of each deoxyribonucleotide triphosphate (Promega, Madison, WI, USA), 1.25 U of GoTaq DNA polymerase (Promega), 10 µl of 5× PCR buffer and 200 ng of genomic DNA. Amplification was carried out in a thermocycler (Bio-Rad MyCycler Thermal Cycler) with an initial 3 min denaturation at 94 °C, followed by 35 cycles, which consisted of denaturation at 95 °C for 30 s, annealing at 60 °C for 30 s and elongation at 72 °C for 45 s, followed by a final extension step of 72 °C for 10 min. Five microliters of each amplified product were analyzed by electrophoresis in 1% agarose gel stained with ethidium bromide. A molecular size marker (1 Kb Plus DNA Ladder, Invitrogen) was used to determine PCR product size. The remaining 45 µl of the amplified products were purified by precipitation with 11.25 µl of 125 mM EDTA and 135 µl of absolute ethanol, centrifugation at 10,000*g*, precipitation with 70% ethanol and resuspension in pure water. Both strands of the purified amplicons were sequenced on a Big Dye Terminator v3.1 kit from Applied Biosystems and analyzed on an ABI 3130XL genetic analyzer from the

Table 1
A. marginale strains, their origin and their corresponding ST based on MLST scheme.

Isolate name	ST	Origin	Data source
St. Maries	1	USA	GeneBank: NC_004842.2
Florida	2	USA	GeneBank: NC_012026.1
Mississippi	3	USA	GeneBank: NZ_ABOP00000000.1
South Idaho	3	USA	This study
Puerto Rico	4	Puerto Rico	GeneBank: NZ_ABOQ00000000.1
Rosali	5	Argentina, (Salta province)	This study
Salta	5	Argentina, (Salta province)	This study
Virginia	6	USA	GeneBank: NZ_ABOR00000000.1
Mercedes	7	Argentina (Corrientes province)	This study
Virasoro	8	Argentina (Corrientes province)	This study
Oklahoma	9	USA	This study
MIR14	10	Argentina (Corrientes province, Ranch 1)	This study
Quitilipi	11	Argentina (Chaco province)	This study
COB11	12	Argentina (Chaco province, Ranch 2)	This study
COB14	13	Argentina (Chaco province, Ranch 2)	This study
COB4	14	Argentina (Chaco province, Ranch 2)	This study
LA802	15	Argentina (Misiones province, Ranch 3)	This study
LA846	16	Argentina (Misiones province, Ranch 3)	This study
LA862	17	Argentina (Misiones province, Ranch 3)	This study
LF224	18	Argentina (Salta province, Ranch 4)	This study
LF240	19	Argentina (Salta province, Ranch 4)	This study
LF252	20	Argentina (Salta province, Ranch 4)	This study
LH917	21	Argentina (Salta province, Ranch 5)	This study
LM2	22	Argentina (Chaco province, Ranch 6)	This study
LM3	23	Argentina (Chaco province, Ranch 6)	This study
LM7	24	Argentina (Chaco province, Ranch 6)	This study
LM9	25	Argentina (Chaco province, Ranch 6)	This study
LM941	26	Argentina (Entre Rios province, Ranch 7)	This study
LM949	26	Argentina (Entre Rios province, Ranch 7)	This study
LM951	26	Argentina (Entre Rios province, Ranch 7)	This study
Sal8	27	Argentina (Corrientes province, Ranch 8)	This study
SJ165	28	Argentina (Entre Rios province, Ranch 9)	This study
SJ170	29	Argentina (Entre Rios province, Ranch 9)	This study
SJ184	29	Argentina (Entre Rios province, Ranch 9)	This study
SJ175	30	Argentina (Entre Rios province, Ranch 9)	This study
T365a	31	Argentina (Entre Rios province, Ranch 10)	This study
T365b	32	Argentina (Entre Rios province, Ranch 10)	This study
Africa	33	South Africa	This study
Brasil1	34	Brazil	This study
Brasil2.1	35	Brazil	This study
Brasil2.2	36	Brazil	This study
Brasil nuevo	37	Brazil	This study
Italia10	38	Italy	This study
Italia6	39	Italy	This study
Italia7	40	Italy	This study
Italia8	41	Italy	This study
Uruguay	42	Uruguay	This study
Batel1[a]	43	Argentina (Corrientes province, Ranch 11)	This study
Batel2[a]	43	Argentina (Corrientes province, Ranch 11)	This study
Dragones	44	Argentina (Salta province)	This study
Tamaulipas6	45	Mexico	This study
Tamaulipas17	46	Mexico	This study
Tamaulipas19	47	Mexico	This study
Tamaulipas25	48	Mexico	This study
Colombia_Ur	49	Colombia	This study
Colombia_01Te	50	Colombia	This study
Mer_1_May13[a]	51	Argentina (Corrientes province 12)	This study
Mer_2_May13[a]	52	Argentina (Corrientes province 12)	This study

[a] Cattke blood samples from outbreaks in Corrientes province.

same supplier. Sequences were deposited in GenBank under accession numbers KM090857–KM091227.

2.4. Data processing and analysis

In order to overcome manually intensive steps including raw sequence data processing and downstream analysis an automatic pipeline was set up. Sequence data and typing results were stored in MySQL, an open-source relational database management system, whose conceptual scheme is shown in Supplementary Fig. 2.

For the MLST-Pipeline, the string processing was called from dedicated python wrappers designed according to the Galaxy open web-based platform (http://galaxyproject.org/). Briefly, raw traces files from each gene target (forward and reverse chromatogram in ab1 format) were uploaded to the Galaxy web page according to the designed MLST scheme. Base calling was performed using default PHRED parameters (Ewing et al., 1998), except for trim_cut off (0.3). Assembling was carried out from forward and reverse base calling outputs with CAP3 (Huang and Madan, 1999), using the following parameters: *a* 20; *b* 20; *c* 10; *d* 200; *e* 30; *g* 6; *m* 2; *n* −5; *o* 30; *p* 75; *s* 500; *u* 3; *v* 2. Furthermore, sequence trimming

Table 2
Loci used for *A. marginale* MLST, with primer details.

Gene	Product	Accession code[a]	Genome coordinates[a]	Primer sequence	Amplicon size (bp)
dnaA	Chromosomal replication initiation protein	AM430	389.525–390.940 bp	F: 5′GTTCATAAGCGGGAAGGACA 3′ R: 5′CTTGTCTCGGTCTGGCTAGG 3′	512
ftsZ	Cell division protein FtsZ	AM1261	1.119.261–1.120.514 bp	F: 5′CCTGACCACCAATCCGTATC 3′ R: 5′CCCGTATGAAGCACCGTATC 3′	575
groEL	Chaperonin GroEL	AM944	865.932–867.581 bp	F: 5′AGCATAAAGCCCGAGGAACCTT 3′ R: 5′CAGAAGGAAGGACATGCTCGGC3′	699
lipA	Lipoyl synthase	AM820	754.976–755.917 pb (antisense strand)	F: 5′TGTGGATAGGGACGACCTTC 3′ R: 5′AAAGTCATCCTCAGCGTGGT3′	538
recA	Recombinase A	AM085	68.810–69.889 pb	F: 5′GGGCGGTAACTGTGCTTTTA 3′ R: 5′ACGCCCATGTCGACTATCTC 3′	579
secY	Preprotein translocase subunit SecY	AM892	822.069–823.370 bp (antisense strand)	F: 5′TTCACGCTGCTAGCCCTAAT 3′ R: 5′TACGAGGGAAATGCCGTTAC 3′	501
sucB	Dihydrolipoamide acetyltransferase component	AM1087	981.783–983.096 bp (antisense strand)	F: 5′GAGATAGCATCTCCGGTTGC 3′ R: 5′CTCCCCTGGCCTTTTTACTC 3′	808

[a] Using St. Maries strain as reference.

was performed so as to assure proper sequence alignment. During trimming operations, gene-specific primers were defined to detect the sequence start site while the end site was computed according to predetermined sequence lengths. MLST data were analyzed by the standard MLST approach (Maiden, 2006); for each gene, a number was attributed to each allelic variant, and the sequence type (ST) of a strain corresponded to the combination of the allele numbers of the seven genes, using St. Maries strain as the reference strain for ST assignment. To this end, an specific wrapper searched for existent allelic variants in the local MySQL DB. The whole process end up with an HTML report including the isolate haplotype and the set of strain alleles concatenated in FASTA format. Sequences that do not match to preexistent stored variants are reported as new alleles. Every time a new allele come up the database is automatically refreshed, keeping the allele status until upgrading to the ultimate stage through supervision step by the system administrator. The criteria used to define a true SNP in a sequence were two: only nucleotide changes in both forward and reverse sequences were accepted and that the SNP should be biallelic (Krawczak, 1999).

For the MLST-DB, an interface application website was built using Web2py framework (http://www.web2py.com/). The DB could be explored in a friendly way for searching and download the complete set of alleles for all gene targets and the ST numbers and haplotypes for the stored strains. In addition, both single FASTA file or the set of seven sequences could be used for matching against the DB in order to identify existent alleles or haplotypes respectively.

Both the pipeline and the DB are freely available at http://bioinfoinformatica.inta.gov.ar/galaxy/, and http://bioinfoinformatica.inta.gov.ar/mlst/ respectively, under username and password request at ibiotecno.bioinfo@inta.gob.ar.

2.5. DNA polymorphism analysis

Simpsońs index of diversity was used to assess the discriminatory power of the MLST method (Hunter and Gaston, 1988). Confidence intervals were calculated as implemented in the phyloviz on-line tool (http://darwin.phyloviz.net/ComparingPartitions/index.php?link=Tool) as proposed by Carriço et al. (2006) following the method of Grundmann et al. (2001). Genetic similarity and distance matrices were constructed from ClustalX2 alignments using default conditions (Thompson et al., 1997) using BioEdit 7.0.9.0 (Hall, 1999). DNA sequence polymorphism and all subsequent tests were investigated using several functions from the DnaSP 5.00.02 package (Librado and Rozas, 2009). Allele frequencies were calculated according to Nei (1987).

Nucleotide diversity, Pi (π), was also calculated according to Nei (1987), using Jukes and Cantor (1969) correction; π refers to the average number of nucleotide differences per site between two sequences. Theta (Watterson's mutation parameter) was calculated for the whole sequence from *S* (Watterson, 1975). Eta (η) is the total number of mutations and *S* is the number of segregating (polymorphic) sites. *Ka* (the number of non-synonymous substitutions per non-synonymous site) and *Ks* (the number of synonymous substitutions per synonymous site) for any pair of sequences were calculated according to Nei and Gojobori (1986).

Linkage disequilibrium among polymorphic sites within genes was estimated based on the squared allele-frequency correlations (r^2) (Hill and Robertson, 1968). Significant pairwise associations were assessed by Fisher exact tests and corrected for multiple comparisons using Bonferroni procedures. The ZnS statistic (Kelly, 1997), which is the average of r^2 over all pairwise comparisons, was computed to summarize the extent of linkage disequilibrium.

Linkage disequilibrium among genes was assessed by applying the Standardized index of Association (I_{AS}), as implemented in the software LIAN (Haubold and Hudson, 2000).

Intragenic recombination events were investigated by computing the *ZZ* test statistic (Rozas et al., 2001) as implemented in DnaSP5.00.02 and by applying the algorithms included within the RDP 3 package (Martin et al., 2010) available at http://darwin.uvigo.es/rdp/rdp.html.

Tajima's *D* (Tajima, 1989) and MacDonald–Kreitman tests (McDonald and Kreitman, 1991) were used for testing the null hypothesis of neutrality (Kimura, 1983). In the Tajima's *D* test the average number of nucleotide differences between pairs of sequences is compared with the total number of segregating sites (*S*). If the difference between these two measures of variability is larger than what is expected on the standard neutral model, this model is rejected. The MacDonald–Kreitman test explores the fact that mutations in coding regions come in two different categories: nonsynonymous mutations and synonymous mutations, then these mutations are analyzed within and between species (Nielsen et al., 2005). The MacDonald–Kreitman test seeks to determine whether the mutations are fixed in the species. In order to apply this test, sequences from organisms related to *A. marginale* were analyzed. Orthologous genes from two other *Anaplasma* species (*A. centrale* and *A. phagocytophilum*) and three *Ehrlichia* species (*E. ruminantium*, *E. canis* and *E. chaffeensis*) were used. Deviations from neutrality due to demographic events were evaluated using the R_2 (Ramos-Onsins and Rozas, 2002) and *Fs* (Fu, 1997) indices.

2.6. Genetic relationship among isolates

The relatedness between two strains can be inferred by the differences between allelic profiles. The most popular analysis of the allelic profiles in order to infer a hypothetical phylogenetic relationship between sequence types (STs) is that performed by the eBURST algorithm (Feil and Enright, 2004). STs are linked resulting in clonal complexes (CC). A given genotype becomes the founder clone of a CC as a result of a fitness advantage or random genetic drift. The increase in the frequency of this genotype in the population is accompanied by a gradual diversification by mutation and recombination forming a cluster of phylogenetically closely related strains. Those related genotypes differ from the founder in one housekeeping gene, becoming a single locus variant (SLV). STs in a given CC can be linked to another ST differing in two (DLV) or even three genes (TLV) (Francisco et al., 2009). We explored the relationships among the 58 study strains by using GoeBURST that is a globally optimized implementation of the eBURST algorithm available at http://goeburst.phyloviz.net/ (Francisco et al., 2009).

In order to analyse the partitioning of genetic variation within and among groups of isolates, an analysis of molecular variance (AMOVA) (Excoffier et al., 1992) was conducted based on Jukes and Cantor (Jukes and Cantor, 1969) distances using Arlequin 3.11 (Excoffier et al., 2007). Analyses were performed on a concatenated dataset. Statistical significance of each variance component was assessed based upon 999 permutations of the data.

3. Results

3.1. Assignment of haplotypes and STs

As a result of the implementation of the MLST scheme to 58 *A. marginale* isolates, 52 STs were identified and assigned to each allelic profile. From the 52 STs identified, only four were found in more than one isolate: ST 3, in two strains from USA; ST 5, in two strains from northwest Argentina; ST 26, in three strains from northeast Argentina; and ST 29 in two other strains from northeast Argentina. The rest of the STs were unique. The MLST alleles and STs for *A. marginale* isolates are shown in Supplementary Table 1.

Superimposed double nucleotide peaks on the sequence electropherograms were visualized in several samples, indicative of mixed infections. Only the predominant allele present at each locus within each infection was considered. In our approach, base-calling was based on peak height, as it was demonstrated in a previous work that peak height on a pherogram is a product of the true proportions of malaria parasite clones (Anderson et al., 2000; Ford and Schall, 2011).This procedure resulted in unbiased estimation of allele frequencies within a population.

3.2. DNA polymorphism analysis

Simpson's index (Si) was calculated for the whole *A. marginale* MLST scheme and for each locus individually. A high discrimination power (Si = 0.996; CI: 0.991–1) was achieved when applying the seven genes. Studying only six genes (excluding lipA) the Simpsońs index was conserved, and applying only five genes (excluding lipA and groEl) the discrimination power was hardly reduced (Si = 0.995; CI: 0.990–1).

Allele sizes for the genes included in the *A. marginale* MLST scheme varied between 501 bp (*secY*) and 808 bp (*sucB*) (Table 2). The number of distinct alleles per gene (H) was variable and ranged from 5 (*lipA*) to 14 (*dnaA*). Nucleotide diversity in each locus (π) was low and varied from 0.01568 (*dnaA*) to 0.00221 in the least polymorphic gene (*lipA*) (Table 3). When comparing the number of alleles and the number of polymorphic sites (S) a degree of positive correlation was observed (Spearman coefficient: 0.81, $p < 0.05$).

Table 3
Assessment of DNA polymorphism and neutrality tests for the *A. marginale* strains used in this study.

Gene	*H*	*S*	η	Pa	$\eta_{(s)}$	θ	π	Tajima's	*D**	*F**	Rm	ZnS
dnaA	24	40	41	14	26	0.02065	0.01568	−0.80386	−3.91875	−3.30178	5	0.4104
ftsZ	10	10	10	9	1	0.00453	0.00696	1.49222	0.76003	1.18568	2	0.4038
groEl	8	22	22	13	9	0.00966	0.00297	−2.18502	−1.49793	−2.07178	4	0.2642
lipA	5	6	6	6	0	0.00259	0.00221	−0.35609	1.16314	0.79508	1	0.4274
recA	19	12	12	10	2	0.00533	0.00657	0.67026	0.54796	0.35723	4	0.1178
secY	6	6	6	4	2	0.00311	0.00420	0.86830	−0.60065	−0.15347	0	0.3234
sucB	15	27	27	20	7	0.00857	0.00441	−1.5634	−0.32648	−0.92012	3	0.1830

Gene	*Na*	*Ns*	*Ka*	*Ks*	*Ka*/*Ks*
dnaA	329.5	99.5	0.69	119.32	0.0057
ftsZ	357.28	119.72	0	47.5693	0
groEl	378.02	113.98	0	22.3311	0
lipA	380.36	120.64	0	15.5488	0
recA	370.17	115.83	0.3078	45.8357	0.0067
secY	314.69	102.31	2.8992	19.8775	0.1458
sucB	516.52	164.68	2.8235	21.9213	0.1288

H: haplotypes.
S: polymorphic sites.
ŋ: total number of mutations.
Pa: parsimony informative sites.
ŋ(*s*): number of singletons.
θ: Watterson's mutation parameter (calculated from Eta).
π: nucleotide diversity.
Tajima's *D*: Tajima's *D* test of neutrality.
*D**: Fu and Li's *D** neutrality test.
*F**: Fu and Li's *F** neutrality test.
Rm: minimal recombination events.
ZnS: ZnS statistic.
Na: number of non-synonymous substitutions.
Ns: number of synonymous substitutions.
Ka: non-synonymous substitutions frequency.
Ks: synonymous substitutions frequency.

The proportion of nucleotide substitutions that altered the aminoacid sequence (*Ka*) and the proportion of silent changes (*Ks*) were calculated for each gene. The *Ka*/*Ks* ratio varied from 0 (*ftsZ*, *groEl* and *lipA*) to 0.1458 (*secY*), which are very low owing to the quasi absence of non-synonymous mutations (Table 3).

Intragenic recombination events were detected for all genes except for *secY*. In some cases putative parental sequences were found (Supplementary Fig. 3). Among genes statistically significant LD was detected while the ZnS index within genes did not show significant departures from linkage equilibrium.

Neutrality tests were also conducted using Tajima's *D* and McDonald–Kreitman tests. Tajima's *D* test showed no statistical significance for five out of seven genes (*dnaA*, *ftsZ*, *lipA*, *recA* and *secY*), indicating that a neutral model of sequence evolution could not be rejected. On the other hand, groEl and sucB showed negative and significant values for this test suggesting that these genes could be under purifying selection. However, the occurrence of selection was not always supported by McDonald–Kreitman tests, since significance varied depending on the species used as outgroup (Supplementary Table 2). McDonald–Kreitman tests showed no statistical significance for any of the genes when compared to sequences from *Anaplasma phagocytophilum*, whereas *groEl* was the only gene to exhibit an excess of non-synonymous fixations when compared to *Ehrlichia* sequences.

3.3. Genetic relationship among isolates

Applying eBURST algorithm, those strains sharing the same ST are represented in the same node, the size of which is proportional to the number of strains with that particular profile. Each circle represents one ST and the different STs are organized in CC based on their relatedness. To analyse the 58 *A. marginale* strains we applied a TLV criteria. This means that those related genotypes that differ in up to three genes from the founder will be arranged in the same CC.

As a result of this approach we found two CCs (Fig. 1). The largest CC was comprised of STs from diverse LatinoAmerican countries (Argentina, Brazil, Uruguay, Mexico and Colombia) and two STs from Italia. This CC represented 45 strains and it founder was ST 11 which belongs to an Argentinian strain. ST 26 that was represented by three Argentinian strains, was the most common profile and was a single locus variant from the founder; STs 29 and 43 were represented by two strains each (all of them Argentinian). The second CC consisted on strains from USA, ST 3 was the most common profile and was represented by two strains from USA. This CC was founded by the ST 4 originated from Puerto Rico. ST 33 which corresponds to an South African strain remained as a singleton.

The population structure of *A. marginale* was explored at different hierarchical levels using AMOVA, partitioning the population in

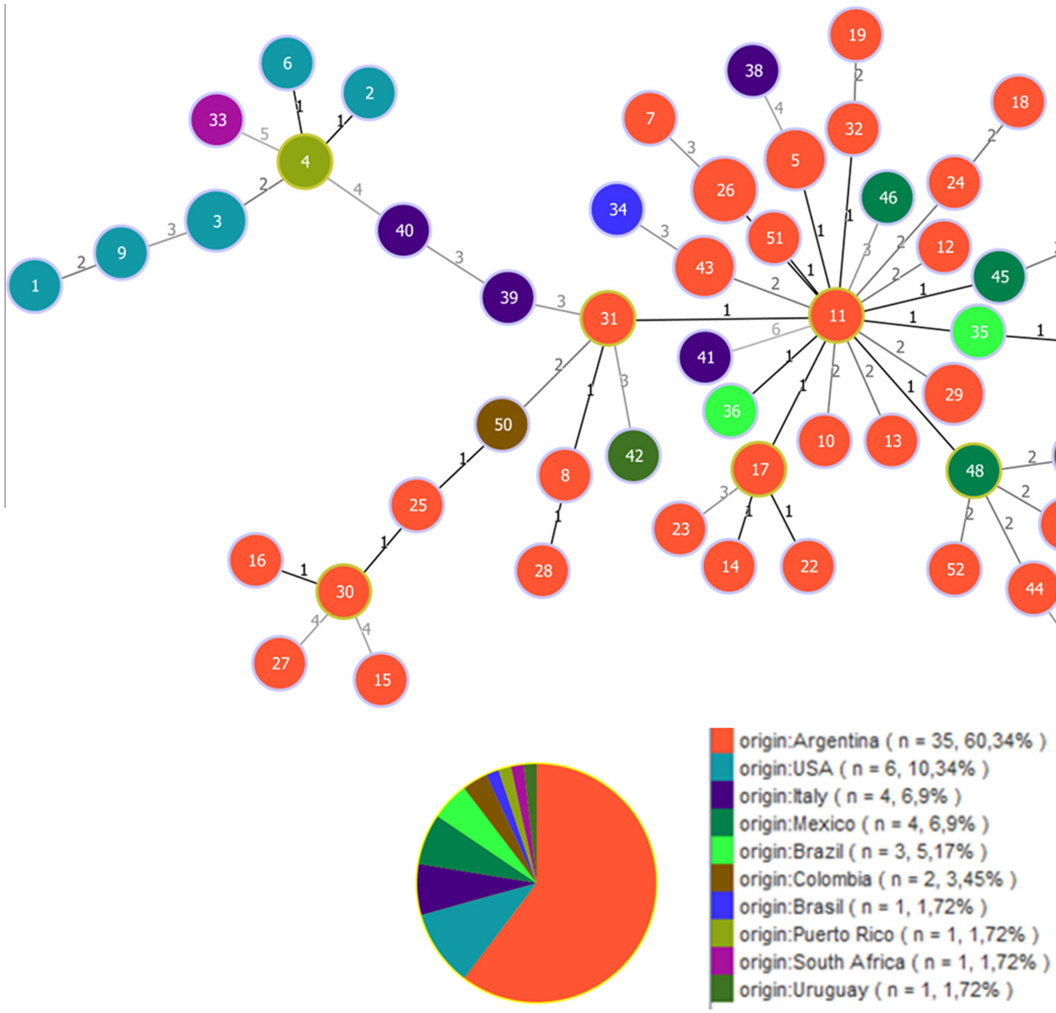
1
9
3
4
6
2
33
40
39
31
50
25
30
16
27
15
8
28
42
41
36
17
23
14
22
10
13
11
34
43
51
26
7
5
38
32
19
46
24
18
12
45
47
35
37
29
48
49
21
52
44
20
🟥 origin:Argentina (n = 35, 60,34%)
origin:USA (n = 6, 10,34%)
origin:Italy (n = 4, 6,9%)
origin:Mexico (n = 4, 6,9%)
origin:Brazil (n = 3, 5,17%)
origin:Colombia (n = 2, 3,45%)
origin:Brasil (n = 1, 1,72%)
origin:Puerto Rico (n = 1, 1,72%)
origin:South Africa (n = 1, 1,72%)
origin:Uruguay (n = 1, 1,72%)

Fig. 1. Graphic representation of the relationship between STs by e-BURST.

Table 4
AMOVA for 7 genes MLST data of *A. marginale* isolates, Groups: North America vs South America vs South Africa vs Italy.

Structure tested	d.f.	Variance component	% of variance	*P*
Among groups	3	9.00033	55.79	0.00000
Among countries within groups	5	0.07647	0.47	0.03812
Within isolates	49	7.05553	43.74	0.00000

Table 5
AMOVA for 7 genes MLST data of *A. marginale* isolates, Groups: Argentina vs Brasil and Uruguay and Colombia and Mexico.

Structure tested	d.f.	Variance component	% of variance	*P*
Among groups	1	0.03899	0.63	NS[a]
Among countries within groups	3	0.66304	10.73	0.03030
Within isolates	45	5.47687	88.64	0.01564

[a] NS, non significant.

four groups defined a *posteriori* of *eBurst* analysis: (a) isolates from USA and Puerto Rico; (b) isolates from Mexico, Colombia, Brazil, Argentina and Uruguay, (c) isolates from Europe (Italy); and (d) one African isolate (South Africa). Within each group, isolates were categorized according to the country of origin, generating a hierarchical structure of countries within groups. AMOVA on all isolates (Table 4) showed that 55.79% of genetic variation was attributable to the four gathered groups. Moreover, non-significant between-group differentiation was detected when comparing Argentinian isolates to those from the rest of the group (Brazil, Colombia, Mexico, Uruguay) (Table 5).

4. Discussion

We describe the development of the first MLST scheme for *A. marginale* based on the allelic polymorphism of seven housekeeping genes. Analysing 58 strains we were able to differentiate isolates with a high discrimination power and also estimate some basic population biology parameters including diversity indices and the impact of homologous recombination.

High nucleotide diversity was identified (Si = 0.9958) with a high number of STs per strain (52 STs in 58 strains) while the *Ka*/*Ks* ratio showed a large proportion of synonymous substitutions, indicative of negative selection. These results are in accordance with what is expected in an obligate intracellular bacterium, as this kind of organism may display genomic stasis after adaptation to intracellular parasitism (Tamas et al., 2002); this condition was postulated for *E. ruminantium* (Adakal et al., 2009).

Recombination events were detected in almost all genes suggesting that this could be a source of population diversity. In some cases parental sequences were found and a recombination pattern of substitution was identified. In addition, homologous recombination recently became evident from MLST results in related organism like *Orientia tsutsugamushi* (Sonthayanon et al., 2010) and *A. phagocytophilum* (Huhn et al., 2014). Furthermore, specific horizontal gene transfer (HGT) events were identified in *R. bellii* and *Orienta* sp. among others members of the order Rickettsiales, highlighting this ancestral mechanism (Georgiades et al., 2011). Even though HGT was not detected in *A. marginale* (Brayton et al., 2005; Georgiades et al., 2011), possible due to its strict adaptation to erythrocytes, that represent a vast niche sheltered from microbial competitors, the superinfection phenomena should be considered as a potential source of variation. The ability of a second strain or even more to establish infection in a host previously infected with other strain (superinfection), it is a common feature of enzootic regions, supported herein and by previous reports (Futse et al., 2008; Ruybal et al., 2009; Ueti et al., 2012).

In the present work we did not find an evident association between geographical regions and genotypes as can be seen in the eBURST algorithm where STs from different locations are grouped in the same clonal complex. The CC that clearly gathered the Puerto Rico and USA strains is founded by the former. Moreover, the AMOVA (Table 4) maximized the largest differences among groups defined *a posteriori* when the samples from USA and Puerto Rico were grouped together. Such a clustering seemed to be unexpected since strains from USA form a genetically distinct clade as compared to a cluster of tropical strains isolated from Mexico, the Caribbean region, and Central and South America when using a gen coding for a surface protein as a genetic marker (de la Fuente et al., 2002). This genetic distance was supported by the fact that tick transmission of *A. marginale* in the United States is mediated by *Dermacentor andersoni* and *Dermacentor variabilis* since *R. microplus* eradication in the 1940s, being the last one the primary vector in tropical regions (Futse et al., 2003). However, historical evidence reinforce the hypothesis of clonality among Puerto Rico an USA isolates revealed by MLST: in 1913 Bishopp first reported cattle infestations with *R. microplus* after collecting the ticks in 1912 from animals in Key West, Florida and he speculated that Cattle Fever Ticks arrived from the Caribbean islands through commerce (Pérez de León et al., 2012). Remarkably, MLST gene fragments belong to the core genome, thus are not influenced by selective pressure (Dark et al., 2009). Therefore, the allelic profile might exposed the true evolutionary history rather than the most recent events of vector co-evolution/host interaction reflected by genes under strong positive selection, as revealed by the alleles encoding the immunodominant outer membrane proteins (Ruybal et al., 2009; Ueti et al., 2012) .

On the other hand, the larger CC grouped together the samples from Latin American countries and two from Italy, where two isolates from Argentina and one from México were the founders (Fig. 1). Furthermore, the AMOVA revealed that most of the variation was found among isolates within countries (88.64%), the variation among countries was low but significant though the partitioning among groups was rejected (Table 5). When alternative *a priori* groupings of samples from Argentina were tested based on geographic criteria, no differentiation was shown between groups (Supplementary Table 3), supporting the South America + Mexico as a whole group. Merging both analyses, the occurrence of at least two main genetically divergent groups of populations became evident. This scenario is equivalent to the one described by Ueti et al. (2012), where in tropical and subtropical regions of highly endemic infection rate, *A. marginale* populations had significantly greater diversity in comparison to temperate regions of low endemicity. Nevertheless, this spatial pattern of infection was revealed quantifying the multiples variants of the immunodominant outer membrane protein, major surface protein 2 (MSP2) after strong selective pressure of mammalian host.

It has been demonstrated that genotypic diversity in *A. marginale* relayed on *msp2* and *msp3* gene alleles (Futse et al., 2008), while maintaining a closed-core genome (Dark et al., 2009). Thus the process of strain variation is influenced by competing selective pressures associated to level of endemicity and population immunity (Ueti et al., 2012) and environmental factors regulating vector prevalence (Estrada-Peña et al., 2009). The genes included in our MLST scheme shape a representative sample of the core genome: *dN*/*dS* values obtained were below 1 for all loci; Tajima's *D* test failed to detect deviations from neutrality in 5 genes. Although *A. marginale* MLST seemed to capture variation due to high endemicity, the inclusion of a marker under positive selection pressure would ensure a better framework for studying infection epidemiology.

The MLST strategy involves the generation and analysis of a large amount of data, therefore, a custom-designed bioinformatic pipeline named "Galaxy MLST-Pipeline" was developed to automate DNA sequence editing and analysis and to significantly reduce the time required for processing data. Another advantage of the "Galaxy MLST-Pipeline" developed here is the use of free software which makes it available to low-income research or health institutions, being fully adaptable to other MLST schemes of any organism.

To date whole genome sequencing usually requires isolation or amplification of the strain, hence PCR based strategies are still the better choice for approaching epidemiological studies of intracellular pathogens. With the rapidly increasing sequencing capacity of Next Generation Sequence Technologies combined with target capturing methods, MLST is becoming a powerful and affordable approach to perform high resolution strain typing of a large sample collection.

In summary, the MLST scheme developed here is a robust, objective and easily adoptable technology. We think that this kind of approach is a valuable tool where to superimpose the data recovered from the analysis of target genes under positive selection pressure, so as to shed light in how intracellular pathogens challenge virulence paradigm.

Acknowledgements

We would like to acknowledge those people who kindly provided us the *A. marginale* isolates used in this study: Dr. Camila Bastos, (Parasitology department, ICB-UFMG, Belo Horizonte, Brazil), Dr. María A. Solari (DILAVE, Uruguay), Dr. Alessandra Torina (Instituto Zooprofiláctico de Sicilia, Palermo, Italia), Dr. Consuelo Almazán (Tamaulipas University, México), Dr. Guy Palmer and Dr. Kelly Brayton (Washington State University, WA, USA); Dr. Nestor Sarmiento (EEA-Mercedes, INTA, Argentina) and Dr. Ignacio E. Echaide (EEA-Rafaela, INTA, Argentina) Thanks are also extended to Dr. Julia Sabio y García for English language editing. Financial support was obtained by ANPCyT-PICT 1634 and INTA-PNBIO 1131043. ECG is a Post-Doctoral, SG and LLA are doctoral fellows from CONICET. MDF, SEW, PR, VL are Career Members of CONICET.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.meegid.2014.12.027.

References

Adakal, H., Meyer, D.F., Carasco-Lacombe, C., Pinarello, V., Allègre, F., Huber, K., Stachurski, F., Morand, S., Martinez, D., Lefrançois, T., Vachiery, N., Frutos, R., 2009. MLST scheme of *Ehrlichia ruminantium*: genomic stasis and recombination in strains from Burkina-Faso. Infect. Genet. Evol. 9, 1320–1328.

Almazán, C., Medrano, C., Ortiz, M., de la Fuente, J., 2008. Genetic diversity of *Anaplasma marginale* strains from an outbreak of bovine anaplasmosis in an endemic area. Vet. Parasitol. 158 (1–2), 103–109.

Anderson, T.J., Haubold, B., Williams, J.T., Estrada-Franco, J.G., Richardson, L., Mollinedo, R., Bockarie, M., Mokili, J., Mharakurwa, S., French, N., Whitworth, J., Velez, I.D., Brockman, A.H., Nosten, F., Ferreira, M.U., Day, K.P., 2000. Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. Mol. Biol. Evol. 17, 1467–1482.

Baldo, L., Dunning Hotopp, J.C., Jolley, K.A., Bordenstein, S.R., Biber, S.A., Choudhury, R.R., Hayashi, C., Maiden, M.C., Tettelin, H., Werren, J.H., 2006. Multilocus sequence typing system for the endosymbiont Wolbachia pipientis. Appl. Environ. Microb. 72 (11), 7098–7110.

Bock, R., De Vos, A., 2001. Immunity following use of Australian tick fever vaccine: a review of the evidence. Aust. Vet. J. 79, 832–839.

Brayton, K.A., Kappmeyer, L.S., Herndon, D.R., Dark, M.J., Tibbals, D.L., Palmer, G.H., McGuire, T.C., Knowles Jr., D.P., 2005. Complete genome sequencing of *Anaplasma marginale* reveals that the surface is skewed to two superfamilies of outer membrane proteins. Proc. Natl. Acad. Sci. U.S.A. 102 (3), 844–849.

Carriço, J.A., Silva-Costa, C., Melo-Cristino, J., Pinto, F.R., de Lencastre, H., Almeida, J.S., Ramirez, M., 2006. Illustration of a common framework for relating multiple typing methods by application to macrolide-resistant *Streptococcus pyogenes*. J. Clin. Microbiol. 44 (7), 2524–2532.

Dark, M.J., Herndon, D.R., Kappmeyer, L.S., Gonzales, M.P., Nordeen, E., Palmer, G.H., Knowles Jr., D.P., Brayton, K.A., 2009. Conservation in the face of diversity: multistrain analysis of an intracellular bacterium. BMC Genom. 11 (10), 16.

de la Fuente, J., Van DenBussche, R.A., Kocan, K.M., 2001. Molecular phylogeny and biogeography of North American isolates of *Anaplasma marginale* (Rickettsiaceae: Ehrlichieae). Vet. Parasitol. 97 (1), 65–76.

de la Fuente, J., Van Den Bussche, R.A., Garcia-Garcia, J.C., Rodríguez, S.D., García, M.A., Guglielmone, A.A., Mangold, A.J., Friche Passos, L.M., Barbosa Ribeiro, M.F., Blouin, E.F., Kocan, K.M., 2002. Phylogeography of New World isolates of *Anaplasma marginale* based on major surface protein sequences. Vet. Microbiol. 88 (3), 275–285.

de La Fuente, J., Passos, L.M., Van Den Bussche, R.A., Ribeiro, M.F., Facury-Filho, E.J., Kocan, K.M., 2004. Genetic diversity and molecular phylogeny of *Anaplasma marginale* isolates from Minas Gerais, Brazil. Vet. Parasitol. 121 (3–4), 307–316.

de la Fuente, J., Ruybal, P., Mtshali, M.S., Naranjo, V., Shuqing, L., Mangold, A.J., Rodríguez, S.D., Jiménez, R., Vicente, J., Moretta, R., Torina, A., Almazán, C., Mbati, P.M., de Echaide, S.T., Farber, M., Rosario-Cruz, R., Gortazar, C., Kocan, K.M., 2007. Analysis of world strains of *Anaplasma marginale* using major surface protein 1a repeat sequences. Vet. Microbiol. 119 (2–4), 382–390.

Estrada-Peña, A., Naranjo, V., Acevedo-Whitehouse, K., Mangold, A.J., Kocan, K.M., de la Fuente, J., 2009. Phylogeographic analysis reveals association of tick-borne pathogen, *Anaplasma marginale*, MSP1a sequences with ecological traits affecting tick vector performance. BMC Biol. 7, 57.

Ewing, B., Hillier, L., Wendl, M.C., Green, P., 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8 (3), 175–185.

Excoffier, L., Laval, G., Schneider, S., 2007. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol. Bioinform Online 23 (1), 47–50.

Excoffier, L., Smouse, P., Quattro, J., 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131, 479–491.

Feil, E.J., Enright, M.C., 2004. Analyses of clonality and the evolution of bacterial pathogens. Curr. Opin. Microbiol. 7, 308–313.

Ford, A.F., Schall, J.J., 2011. Relative clonal proportions over time in mixed-genotype infections of the lizard malaria parasite *Plasmodium mexicanum*. Int. J. Parasitol. 41, 731–738.

Francisco, A.P., Bugalho, M., Ramirez, M., Carriço, J.A., 2009. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. BMC Bioinform. 18 (10), 152.

Fu, Y.X., 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147, 915–925.

Futse, J.E., Brayton, K.A., Dark, M.J., Knowles Jr., D.P., Palmer, G.H., 2008. Superinfection as a driver of genomic diversification in antigenically variant pathogens. Proc. Natl. Acad. Sci. U.S.A. 105 (6), 2123–2127.

Futse, J.E., Ueti, M.W., Knowles Jr., D.P., Palmer, G.H.J., 2003. Transmission of *Anaplasma marginale* by *Boophilus microplus*: retention of vector competence in the absence of vector–pathogen interaction. Clin. Microbiol. 41 (8), 3829–3834.

Georgiades, K., Merhej, V., El Karkouri, K., Raoult, D., Pontarotti, P., 2011. Gene gain and loss events in Rickettsia and Orientia species. Biol. Direct. 8 (6), 6.

Godoy, D., Randle, G., Simpson, A.J., Aanensen, D.M., Pitt, T.L., Kinoshita, R., Spratt, B.G., 2003. Multilocus sequence typing and evolutionary relationships among the causative agents of melioidosis and glanders, *Burkholderia pseudomallei* and *Burkholderia mallei*. J. Clin. Microbiol. 41 (5), 2068–2079.

Grundmann, H., Hori, S., Tanner, G., 2001. Determining confidence intervals when measuring genetic diversity and the discriminatory abilities of typing methods for microorganisms. J. Clin. Microbiol. 39, 4190–4192.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. Nucl. Acid Symp. 41, 95–98.

Haubold, B., Hudson, R.R., 2000. LIAN 3.0: detecting linkage disequilibrium in multilocus data. Linkage analysis. Bioinformatics 16, 847–848.

Hill, W.G., Robertson, A., 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38, 226–231.

Huang, X., Madan, A., 1999. CAP3: a DNA sequence assembly program. Genome Res. 9, 868–877.

Huhn, C., Winter, C., Wolfsperger, T., Wüppenhorst, N., Strašek Smrdel, K., Skuballa, J., Pfäffle, M., Petney, T., Silaghi, C., Dyachenko, V., Pantchev, N., Straubinger, R.K., Schaarschmidt-Kiener, D., Ganter, M., Aardema, M.L., von Loewenich, F.D., 2014. Analysis of the population structure of *Anaplasma phagocytophilum* using multilocus sequence typing. PLoS One 9 (4), e93725.

Hunter, P.R., Gaston, M.A., 1988. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J. Clin. Microbiol. 26 (11), 2465–2466.

Jacobson, M.J., Lin, G., Whittam, T.S., Johnson, E.A., 2008. Phylogenetic analysis of *Clostridium botulinum* type A by multi-locus sequence typing. Microbiology 154, 2408–2415.

Jolley, K.A., Feil, E.J., Chan, M.S., Maiden, M.C., 2001. Sequence type analysis and recombinational tests (START). Bioinformatics 17 (12), 1230–1231.

Jolley, K.A., Kalmusova, J., Feil, E.J., Gupta, S., Musilek, M., Kriz, P., Maiden, M.C., 2000. Carried meningococci in the Czech Republic: a diverse recombining population. J. Clin. Microbiol. 38 (12), 4492–4498.

Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, H.N. (Ed.), Mammalian Protein Metabolism. Academic Press, New York, pp. 21–132.

Kelly, J.K., 1997. A test of neutrality based on interlocus associations. Genetics 146, 1197–1206.

Kimura, M., 1983. Rare variant alleles in the light of the neutral theory. Mol. Biol. Evol. 1, 84–93.

Kocan, K.M., de la Fuente, J., Guglielmone, A., Meléndez, R.D., 2003. Antigens and alternatives for control of *Anaplasma marginale* infection in cattle. Clin. Microbiol. Rev. 16 (4), 698–712.

Krawczak, M., 1999. Informativity assessment for biallelic single nucleotide polymorphisms. Electrophoresis 20 (8), 1676–1681.

Lew, A.E., Bock, R.E., Minchin, C.M., Masaka, S., 2002. A msp1 alpha polymerase chain reaction assay for specific detection and differentiation of *Anaplasma marginale* isolates. Vet. Microbiol. 86 (4), 325–335.

Librado, P., Rozas, J., 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25, 1451–1452.

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M., Spratt, B., 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. U.S.A. 95 (6), 3140–3145.

Maiden, M.C.J., 2006. Multilocus sequence typing of bacteria. Annu. Rev. Microbiol. 60, 561–588.

Martin, D.P., Lemey, P., Lott, M., Moulton, V., Posada, D., Lefeuvre, P., 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26 (19), 2462–2463.

Mayer, L.W., Reeves, M.W., Al-Hamdan, N., Sacchi, C.T., Taha, M.K., Ajello, G.W., Schmink, S.E., Noble, C.A., Tondella, M.L., Whitney, A.M., Al-Mazrou, Y., Al-Jefri, M., Mishkhis, A., Sabban, S., Caugant, D.A., Lingappa, J., Rosenstein, N.E., Popovic, T., 2002. Outbreak of W135 meningococcal disease in 2000: not emergence of a new W135 strain but clonal expansion within the electophoretic type-37 complex. J. Infect. Dis. 185 (11), 1596–1605.

McDonald, J.H., Kreitman, M., 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351 (6328), 652–654.

Mtshali, M.S., de la Fuente, J., Ruybal, P., Kocan, K.M., Vicente, J., Mbati, P.A., Shkap, V., Blouin, E.F., Mohale, N.E., Moloi, T.P., Spickett, A.M., Latif, A.A., 2007. Prevalence and genetic diversity of *Anaplasma marginale* strains in cattle in South Africa. Zoonoses Public Health 54 (1), 23–30.

Nei, M., 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3, 418–426.

Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A., Tanenbaum, D.M., Civello, D., White, T.J., Sninsky, J., Adams, M.D., Cargill, M., 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 3 (6), e170.

Palmer, G.H., Rurangirwa, F.R., McElwain, T.F., 2001. Strain composition of the ehrlichia *Anaplasma marginale* within persistently infected cattle, a mammalian reservoir for tick transmission. J. Clin. Microbiol. 39 (2), 631–635.

Pérez de León, A.A., Teel, P.D., Auclair, A.N., Messenger, M.T., Guerrero, F.D., Schuster, G., Miller, R.J., 2012. Integrated strategy for sustainable cattle fever tick eradication in USA is required to mitigate the impact of global change. Front Physiol. 14 (3), 195.

Ramos-Onsins, S.E., Rozas, J., 2002. Statistical properties of new neutrality tests against population growth. Mol. Biol. Evol. 19, 2092–2100.

Rozas, J., Gullaud, M., Blandin, G., Aguadé, M., 2001. DNA variation at the rp49 gene region of *Drosophila simulans*: evolutionary inferences from an unusual haplotype structure. Genetics 158, 1147–1155.

Ruybal, P., Moretta, R., Perez, A., Petrigh, R., Zimmer, P., Alcaraz, E., Echaide, I., Torioni de Echaide, S., Kocan, K.M., de la Fuente, J., Farber, M., 2009. Genetic diversity of *Anaplasma marginale* in Argentina. Vet. Parasit. 162, 176–180.

Sambrook, J., Fritsch, E.F., Maniatis, T., 1989. Molecular Cloning: A Laboratory Manual, second ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Sonthayanon, P., Peacock, S.J., Chierakul, W., Wuthiekanun, V., Blacksell, S.D., Holden, M.T., Bentley, S.D., Feil, E.J., Day, N.P., 2010. High rates of homologous recombination in the mite endosymbiont and opportunistic human pathogen Orientia tsutsugamushi. PLoS Negl. Trop. Dis. 4 (7).

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585–595.

Tamas, I., Klasson, L., Canbäck, B., Näslund, A.K., Eriksson, A., Wernegreen, J.J., Sandström, J.P., Moran, N.A., Andersson, S.G.E., 2002. 50 million years of genomic stasis in endosymbiotic bacteria. Science 296 (5577), 2376–2379.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25 (24), 4876–4882.

Ueti, M.W., Tan, Y., Broschat, S.L., Castañeda Ortiz, E.J., Camacho-Nuez, M., Mosqueda, J.J., Scoles, G.A., Grimes, M., Brayton, K.A., Palmer, G.H., 2012. Expansion of variant diversity associated with a high prevalence of pathogen strain superinfection under conditions of natural transmission. Infect. Immun. 80 (7), 2354–2360.

Vidotto, M.C., Kano, S.F., Gregori, F., Headley, S.A., Vidotto, O., 2006. Phylogenetic analysis of *Anaplasma marginale* strains from Paraná State, Brazil, using the msp1alpha and msp4 genes. J. Vet. Med. B Infect. Dis. Vet. Public Health 53 (9), 404–411.

Vitorino, L., Chelo, I.M., Bacellar, F., Zé-Zé, L., 2007. Rickettsiae phylogeny: a multigenic approach. Microbiology 153, 160–168.

Watterson, G.A., 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7, 256–276.