An entropy-optimized multilocus approach for characterizing the strains of *Anaplasma phagocytophilum* infecting horses in the Czech Republic

Petr Zeman¹ and Petr Jahn²

¹Medical Laboratories, Konevova 205, 130 00 Prague 3, Czech Republic
²Equine Clinic, University of Veterinary and Pharmaceutical Sciences, Palackeho 1-3, 612 42 Brno, Czech Republic

*Anaplasma phagocytophilum* is a tick-borne rickettsial pathogen that has measurable genetic heterogeneity across its geographical range and reservoir spectrum. In the present study, publicly available sequences of the genes that have prevailingly been used for typing *A. phagocytophilum* were analysed to identify the segments giving the highest resolution with respect to the predictability of host and geographical provenances of the strains. Selected partial sequences of 16S rRNA, *groL*, *msp4* and *ank* genes were then employed in a tentative multilocus typing scheme used to characterize the strains causing equine granulocytic anaplasmosis (EGA). We were able to both identify alleles characteristic for equine strains of *A. phagocytophilum* and distinguish two unique genetic variants infecting horses in the Czech Republic. This resolution far exceeded the discriminatory potential of any of the four sequenced genes when used singly. The two novel *A. phagocytophilum* variants appeared to be phylogenetically closer to the strains reported as causing human disease in Slovenia than to strains thus far isolated from other European EGA cases. A decline in the quality of recently deposited *A. phagocytophilum* sequences was also demonstrated.

**INTRODUCTION**

*Anaplasma phagocytophilum* is an emerging rickettsial pathogen. This species comprises forms previously referred to as *Ehrlichia phagocytophilae*, *Ehrlichia equi* and HGE (human granulocytic ehrlichiosis) agent (Dumler et al., 2001) associated with diseases of ruminants, horses and humans, respectively, as well as strains identified in a variety of animal hosts, including vector ticks (mainly *Ixodes* spp.) throughout the northern hemisphere. The high degree of biological and clinical diversity within this taxon is attributed to the existence of genetic variants, adapted to specific reservoir hosts and/or vector ticks in different geographical areas. Molecular studies have shown that some variants are implicated in human and animal disease, and some variants are implicated in animal disease only (Massung et al., 2002, 2005); however, to date no standard genotyping method is available.

The most common approaches to the molecular characterization of *A. phagocytophilum* employ the sequencing of genes encoding the small-subunit rRNA (16S rRNA) or a heat-shock protein (*groL*); however, the discriminatory capacity of these gene sequences is insufficient for typing the wide variety of circulating strains. Attempts to compare sequences of the highly diversified *ank* gene, encoding a 190 kDa secreted protein, are hindered by great strain-to-strain differences, demanding the usage of multiple specific primers for both the amplification and sequencing of a particular strain (Massung et al., 2000; Von Loewenich et al., 2003). Highly discriminatory targets are presented by genes encoding the outer membrane surface proteins. The *msp4* gene was shown to provide information that can be used to differentiate those strains obtained from horses, dogs and humans from those obtained from ruminants (De la Fuente et al., 2005; Bown et al., 2007). The highest sequence diversity was shown for *msp2/p44*. This gene exists in multiple paralogues (Hotopp et al., 2006) that differ widely from each other in a central hypervariable
region. Comparative studies have shown that strains of similar geographical and host origin tend to share more similar sets of these paralogues (Casey et al., 2004; Lin et al., 2004; Ohashi et al., 2005; Barbet et al., 2006; Masuzawa et al., 2008). Due to a reliance on cloning and the lack of any capability to distinguish co-infecting strains, the usage of msp2/p44 is impractical, however, and inappropriate.

Finer genetic resolution can be achieved by combining multiple polymorphic loci. Recently, a multilocus variable-number tandem-repeat scheme was proposed for the genetic fingerprinting of European strains of A. phagocytophilum (Bown et al., 2007). The multi-gene polymorphism approach was also shown to be satisfactorily selective for distinguishing among diverse strains infecting dogs in the western USA (Shukla et al., 2007).

With A. phagocytophilum, the utility of a consensual multilocus typing system is hindered by the disparateness of those targets preferred for sequencing in different laboratories. In this study, the nucleotide sequences of those genes primarily used for typing A. phagocytophilum were analysed to delimit segments with the highest potential for resolution, with respect to the predictability of strain provenances. Selected partial sequences of 16S rRNA, groL, msp4 and ank genes were then used to tentatively characterize the strains of A. phagocytophilum infecting horses in the Czech Republic.

**METHODS**

**Sequences.** A. phagocytophilum sequences were retrieved from GenBank using the Entrez search engine. Only sequences with a specified host and geographical origin were included in this study.

**Gene selection.** The 16S rRNA, groL, msp4 and ank genes were selected, based on their good representation in GenBank, as well as their distant location within the A. phagocytophilum chromosome, precluding joint recombinational replacement (Hotopp et al., 2006). Of the two closely spaced genes, msp4 and gltA, the former was given preference, as it outnumbers gltA in GenBank.

**Optimization of target sequences.** The selected sequences were aligned using CLUSTAL W (Thompson et al., 1994); partial sequences were aligned separately with a full-length sequence, and global alignments were then assembled and corrected by hand using BioEdit (Hall, 1999). Within each alignment block, sequences derived from the same host species were grouped together. Some host species were represented by single sequences; more than one sequence was available for others. Whenever possible, the latter sequences were further categorized by geographical origin. This gave rise to the provenance categories.

The Shannon information uncertainty (entropy) value was used as the measure of residue variability at each aligned position (Schneider & Stephens, 1990):

$$H(\text{residue}) = - \sum_i P_i \log_2 P_i$$

where $P_i$ gives the frequency at which the $i$-th residue is found at an alignment position. The amount of useful information at a given position was rated as the difference of two entropies, $R=H(\text{residue})-H(\text{background})$, where $H(\text{background})$ was considered as corresponding to the mean sequencing error rate simulated as 1% random nucleotide substitution by the Monte Carlo method. The contribution of the $j$-th sequence to overall entropy at a given position was estimated as the absolute deviation, $D=|H(\text{residue})-H_\ast(\text{residue})|$, where $H_\ast(\text{residue})$ gives the entropy of the remaining alignment block after leaving out the $j$-th sequence. The joint information uncertainty coefficient was used to estimate quantitatively the association between the residue and provenance variables at each aligned position:

$$U = 2 \left[ \frac{H(\text{residue}) + H(\text{host}) - H(\text{residue, host})}{H(\text{residue}) + H(\text{host})} \right]$$

where $H(\text{residue})$ and $H(\text{host})$ are the marginal entropies of the respective residue and provenance variables, and the joint entropy term, $H(\text{residue, host})$, measures entropy of the residue–provenance pairing:

$$H(\text{residue, host}) = - \sum_i \sum_j P_{ij} \log_2 P_{ij}$$

where $P_{ij}$ is the frequency at which the $i$-th residue is met in the $j$-th provenance category (Press et al., 1992). The significance of $U$ estimates was assessed by 100 random permutations of sequences between the provenance categories. The genotyping potential of an alignment position was considered high when the distribution of residues, measured by entropy, exceeded the background ‘noise’; and jointly when $U$ indicated a significant correlation of the residue distribution with the provenance distribution. Gene segments of approximately 500 bp were then located to cover the highest density of the information-rich positions; these segments were flanked with primers designed accordingly. Secondary structure prediction of the target sequences were carried out with RNAstructure 4.5 (Mathews et al., 2004) and PredictProtein (Rost et al., 2004).

**Clinical samples.** A. phagocytophilum DNA was isolated using a DNeasy tissue kit (Qiagen) from buffy coat, prepared from jugular blood drawn from nine horses hospitalized during 2002–2005 for equine granulocytic anaplasmosis (EGA) at the Equine Clinic, University of Veterinary and Pharmaceutical Sciences. EGA diagnosis was confirmed by nested PCR (Massung et al., 1998), and the selected fragments of the 16S rRNA, groL, msp4 and ank genes were amplified by PCR (Table 1). Amplons were cleaned with the aid of agarose gel electrophoresis and a QIAquick PCR purification kit (Qiagen), and then sequenced in both directions using BigDye 1.1 and 3.1 chemistries for amplicons of <600 and ≥600 bp, respectively, following the manufacturer’s standard protocol (Applied Biosystems). The reactions were analysed with an ABI Prism 3100 Avant genetic analyser (Applied Biosystems); sequences obtained for both strands were combined using BioEdit, and compared with both the other EGA cases and with the sequences published in GenBank.

**RESULTS AND DISCUSSION**

**Quality of deposited sequences**

For assessing the quality of the A. phagocytophilum sequence data, 16S rRNA sequences were chosen as a representative set; in consideration of the fact that 16S rRNA is the most conserved and prevalent gene in public repositories. Fig. 1 shows a plot of mean absolute deviation of entropy, as a function of the distance to the end of the sequences. The diagram can be interpreted in terms of
mean error probabilities along the length of the sequences. Paradoxically, the sequences published more recently exhibit, on average, poorer quality than those published earlier. Based upon calibrating simulations (data not shown), the sequencing error rate for the historical and recent subsets was estimated to vary by about 1 and 5%, respectively, within the most error-prone terminal region, and to decrease to about 0.05 and 0.5%, respectively, more centrally. Accordingly, a mean background error rate of 1% was assumed throughout this study. This corresponds to the accuracy of single-pass sequencing (Clayton et al., 1995).

Contaminating error noise in the public sequence databases has been a long recognized problem (Clark & Whittam 1992). Several attempts to estimate the overall error rate in these databases produced quite divergent numbers, in the range of 0.037–3.5% (Wesche et al., 2004). While data originating from large-scale sequencing projects are thought to be relatively error free (in the order of 1 error, or less, in 10,000 bases), the accuracy of individual sequence submissions is less certain (Clayton et al., 1995; Richterich, 1998; Hill et al., 2000; Wesche et al., 2004; Ashelford et al., 2005). Although the error rate in this study might be somewhat overestimated, its increasing trend should be an alert, and evokes the concerns expressed recently on the threatened quality of the public repositories (Ashelford et al., 2005).

16S rRNA gene

A total of 111 16S rRNA sequences, sorted into 34 provenance categories, was included in the analysis; the results are summarized in Fig. 2(a, b). Owing to an insufficient variance-to-background contrast, only a few nucleotide positions can be considered as reliable genotyping markers. Scrutiny of the data revealed that most of those positions corresponded to apparent sequencing artefacts anyway; an exception being a short variable region between nts 75 and 85. A prediction of secondary 16S rRNA structure suggests that the latter variable positions are confined to an ascending strand of a short helix, and that substitution of a base paired nucleotide in the ascending strand does not require substitution in the

Table 1. Primers used in this study for amplification and sequencing of the hypervariable regions of the A. phagocytophilum genes indicated

<table>
<thead>
<tr>
<th>Target</th>
<th>Primer name</th>
<th>Sequence (5’→3’)</th>
<th>PCR annealing temperature [°C]</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>16S rRNA</td>
<td>593II</td>
<td>GTTTGATCCTGGGTCAGGDAACG</td>
<td>65.0</td>
<td>Olsson-Engvall et al. (1996)</td>
</tr>
<tr>
<td></td>
<td>ERB2</td>
<td>CCTCCGGGACTCTAGTCTGGC</td>
<td></td>
<td>Olsson-Engvall et al. (1996)</td>
</tr>
<tr>
<td>groL</td>
<td>GRO686F</td>
<td>GGTTGAGTTGATGTCGCCGGA</td>
<td>50.0</td>
<td>This study</td>
</tr>
<tr>
<td></td>
<td>GRO1185R</td>
<td>AGCGTTCTTTGTCATAATCGAGAT</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>msp4</td>
<td>MSP4F</td>
<td>TGGCTGTAGGGAGCCTATCTGC</td>
<td>59.4</td>
<td>This study</td>
</tr>
<tr>
<td></td>
<td>MSP4R</td>
<td>ACGTCGTAAGGCCGGAACGAA</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>ank outer</td>
<td>ANK_extF</td>
<td>TTACGTACCTGGCTCATC</td>
<td>52.0</td>
<td>This study</td>
</tr>
<tr>
<td></td>
<td>ANK_extR</td>
<td>CTITGAGGAGCTCTATGTTG</td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>ank inner</td>
<td>ANK_intF</td>
<td>GATGTGCGATAGCTATCA</td>
<td>58.0</td>
<td>This study</td>
</tr>
<tr>
<td></td>
<td>ANK_intR</td>
<td>TCAGTGTGGCATATAATCGGCTC</td>
<td></td>
<td>This study</td>
</tr>
</tbody>
</table>

Fig. 1. A comparison of error rates between two subsets of A. phagocytophilum 16S rRNA sequences: (a) deposited before July 2005 (n=55), and (b) deposited latterly (n=55); the diagrams show mean absolute deviation of entropy, as a function of the distance to the proximate end of sequences. Note that given the datasets parity greater dispersion of a graph means worse data quality.
corresponded to the amino acid change of serine to alanine in *A. phagocytophilum* strains from European roe deer and some rodents.

**msp4 gene**

A total of 43 *msp4* sequences, representing 10 provenance categories, was included in the analysis; the results are shown in Supplementary Fig. S2 (available with the online journal). Information-rich positions were shown to cluster in several regions throughout this gene. The highest genotyping potential in terms of discriminatory capacity can be attributed to the segment between nts ~150 and 500, which hypothetically encodes the outer portion of the MSP4 protein, including 6 amino acid positions showing significant provenance associations (Supplementary Fig. S2 available with the online journal). The outer portion of the protein potentially takes part in the interaction of the bacteria with host cells, and so may carry a host signature.

**ank gene**

A total of 35 *ank* sequences, falling into 9 provenance categories, was analysed; the results are shown in Supplementary Fig. S3 (available with the online journal). The region with most information in terms of discriminatory capacity was apparently situated between nts ~1500 and 3000, and included a prominent variable-length region encoding EPIYA motifs (Ijdo *et al.*, 2007) in the terminal part of this sequence. A representative segment of this highly heterogeneous region was identified spanning from the nt 2596 position to beyond the variable-length region (Supplementary Fig. S3 available with the online journal). This configuration allows for variable-length amplicons, believed to facilitate diagnosis of strain mixtures.

**Genotyping of horse strains**

Primers flanking the gene fragments, picked to facilitate the characterization of the *A. phagocytophilum* strains, were designed and applied, as detailed in Table 1. To simplify the classification scheme, information redundancy in the sets of polymorphous positions within these fragments was minimized by sorting-out subsets of such positions that exhibited the highest *U* level and jointly the least mutual correlation. The resulting profile of markers is summarized in Table 2. All nine strains detected in the Czech horses shared the same 16S rRNA, groL and *msp4* sequence homology, while they differed in the *ank* sequence. Nonetheless, the sequences conformed to the classification scheme shown in Table 2 and all exhibited markers of the central European subtype of EGA agents. Specifically, the 16S rRNA and groL sequences were 100% identical to the sequences of equine strains of *A. phagocytophilum* reported from Switzerland, Germany and Sweden (GenBank accession nos AF057707, AF482761, AY527213 and U96735, AF482760, AY529490, respectively). The *msp4* sequence was 99.8% identical to a sequence previously

![Fig. 2. Residue polymorphism analysis of 111 *A. phagocytophilum* 16S rRNA sequences: (a) nucleotide entropy at the positions indicated compared with simulated 1% random substitutions (shown in grey); the coverage (i.e. no. of aligned sequences per position) is shown by dashes; (b) magnitude of nucleotide-provenance association measured as *U* at the positions indicated compared to 100 random permutations interval (shown in grey), significant links are highlighted by + symbols (at the top of the panel), and the chosen typing segment is indicated beneath the baseline; (c) predicted secondary structure of the main variable region – the four variable bases are indicated in bold (relative coordinates refer to the 16S rRNA gene of HZ strain, GenBank accession no. NC_007797).](image-url)
reported from a German horse (GenBank accession no. AY706390), and showed only a single silent mutation with respect to that German strain, having a T instead of a C at nt position 510. By contrast, it was only 98.6 % identical to the msp4 sequence variant shared by Italian donkeys and an American horse (GenBank accession nos AY702925, AY530196), which differed in 7 nt positions. At the protein level, however, all these sequences were 100 % identical within the compared frame. Two ank variants were distinguished; they diverged in the variable-length portion of the sequence, while they were 100 % homologous in the constant-length portion, both at the nucleotide and protein levels. Hereafter, they are referred to as ank-variant 1 (shorter) and ank-variant 2 (longer); these variants were detected in three and six EGA cases, respectively (Fig. 3). Both were most closely related to the strains reportedly causing human disease in Slovenia: at the respective nucleotide and protein levels, the ank-variant 1 showed 88.5 and 91.5 % similarity to the GenBank sequence AF100886; and the ank-variant 2 even showed a 95.9 % similarity and 100 % identity to the GenBank sequence AF100887. Of the horse strains available in GenBank, both detected variants were most closely related to a Swedish strain (GenBank accession no. AY529487), showing 88.3 and 91.0 % similarity to the ank-variant 1, and 88.1 and 91.7 % similarity to the ank-variant 2, at the nucleotide and protein levels, respectively.

Previous single-gene studies indicated that equine strains are closely related (if not identical) to those inducing disease in humans and dogs (Johansson et al., 1995; Madigan et al., 1996; Pusterla et al., 1998; Chae et al., 2000); they cannot be resolved well, even with the multilocus genotyping system we used. This observation corresponds well to their biological properties displayed in cross-infection experiments: the horses were found to be highly susceptible to those strains pathogenic to humans (Madigan et al., 1995; Chang et al., 1998; Pusterla et al., 1999a; Kim et al., 2002), but seroconverted only (showing no signs of infection) when challenged with a bovine strain (Pusterla et al., 1999b). In turn, equine strains were capable of inducing clinical disease in dogs, cats and non-human primates (Lewis et al., 1975); while in sheep and cattle some other tested equine strains induced only subclinical or no disease (Stuen et al., 1998; Pusterla et al., 1999b, 2001). Despite the comparable biological and clinical properties, equine strains of A. phagocytophilum circulating in North America and Europe differ genetically, and are not homogeneous, even within the two continents (Table 2). In Europe, there are indications of two distinct clades, differentiable by groL and msp4 sequences; one distributed throughout central and northern Europe, and the other found in the Mediterranean region (the latter clade clustering more accurately with the North American strains) (Alberti et al., 2005).

Multilocus genotyping methods have shown potentialities for capturing a view of the clonal structure of microbial populations and for enabling the identification of clones that are particularly associated with disease (Maiden et al., 1998; Urwin & Maiden 2003; Perez-Losada et al., 2006). The classification scheme used in this study relies on the sequence collection that is now available in public repositories. However, the current dataset on the genetic diversity of A. phagocytophilum is limited, being predominantly focused on strains isolated from humans and

Table 2. Molecular markers of the strains of A. phagocytophilum associated with Equus spp.

Markers were defined by contracting the information of multiple polymorphous positions within the gene fragments indicated. A profile characteristic of the strains detected in the Czech horses is shown in bold (relative coordinates refer to HZ strain, GenBank accession no. NC_007797).

<table>
<thead>
<tr>
<th>Gene/protein</th>
<th>Coordinate</th>
<th>Marker</th>
<th>Provenance</th>
<th>GenBank accession no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>16S rRNA</td>
<td>76, 77, 80, 84</td>
<td>A, A, A</td>
<td>Horse, Europe</td>
<td>AF057707</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A, A, A/G, A</td>
<td>Horse, America</td>
<td>AF172165/AF172166</td>
</tr>
<tr>
<td>groL</td>
<td>729, 933, 969, 972, 996</td>
<td>A, C, G, A, T</td>
<td>Horse, southern Europe and America</td>
<td>U96727</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A, C, A, C</td>
<td>Horse, north central Europe</td>
<td>AF482760</td>
</tr>
<tr>
<td>Msp4</td>
<td>63, 90, 95, 96, 142, 155</td>
<td>A, A, H, D, V, T</td>
<td>Horse &amp; donkey, Europe &amp; America</td>
<td>AY530196</td>
</tr>
<tr>
<td>Ank</td>
<td>874, 878, 922, 935</td>
<td>A, L, V, P</td>
<td>Horse, Europe</td>
<td>AF482759</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A, F, A, A</td>
<td>Horse, America</td>
<td>AF100882</td>
</tr>
</tbody>
</table>

Fig. 3. A. phagocytophilum ank-variants differentiated by variable-length PCR products: lanes 1–9, DNA isolates from EGA cases; lane 10, EQ-1 strain (isolated from case number 2 and passaged in splenectomized mice); lane 11, control strain NTN-1 (North American isolate, courtesy of Dr S. Telford, Cummings School of Veterinary Medicine, Tufts University, Grafton, MA, USA).
domesticated animals in North America and Europe. An inclusion of sequences from additional host sources and/or geographical areas, once available, could lead to a finer and more comprehensive classification scheme. The subset of tick-derived sequences (which was abandoned in this analysis because of the lack of host data) documents that the pool of genetic variants of *A. phagocytophilum* circulating worldwide in nature is broader. For example, peculiar *ank* variants were detected in *Ixodes ricinus* ticks in Germany; however, the animal reservoirs remain unknown (Von Loewenich et al., 2003). Further studies are needed to gather a more complete picture of the geographical and host-related diversity of *A. phagocytophilum* populations.

Overall, a multilocus typing system employing suitable segments of several housekeeping and structural genes, and also possibly some non-coding regions, may present a valuable tool for studying *A. phagocytophilum* epidemiology. To that end, a consensus over target genome segments, as well as a surveillance of consistent sequence quality, will be needed so that results from one study can be accurately compared with others.

**ACKNOWLEDGEMENTS**

This study was supported by grant MSM 6215712403 provided by the Ministry of Education of the Czech Republic.

**REFERENCES**


of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A 95, 3140–3145.


