Development of an unambiguous and discriminatory multilocus sequence typing scheme for the *Streptococcus zooepidemicus* group

Katy Webb,1 Keith A. Jolley,2 Zoe Mitchell,1 Carl Robinson,1 J. Richard Newton,1 Martin C. J. Maiden2 and Andrew Waller1

1Centre for Preventive Medicine, Animal Health Trust, Lanwades Park, Kentford, Newmarket, Suffolk CB8 7UU, UK
2The Peter Medawar Building for Pathogen Research and Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3SY, UK

The zoonotic pathogen *Streptococcus equi* subsp. *zooepidemicus* (*S. zooepidemicus*) is commonly found harmlessly colonizing the equine nasopharynx. Occasionally, strains can invade host tissues or cross species barriers, and *S. zooepidemicus* is associated with numerous different diseases in a variety of hosts, including inflammatory airway disease and abortion in horses, pneumonia in dogs and meningitis in humans. A biovar of *S. zooepidemicus*, *Streptococcus equi* subsp. *equi*, is the causative agent of strangles, one of the most important infections of horses worldwide. We report here the development of the first multilocus sequence typing (MLST) scheme for *S. zooepidemicus* and its exploitation to define the population genetic structure of these related pathogens. A total of 130 unique sequence types were identified from 277 isolates of diverse geographical and temporal origin. Isolates of *S. equi* shared a recent evolutionary ancestor with isolates of *S. zooepidemicus* that were significantly associated with cases of uterine infection or abortion in horses (*P*<0.001). Isolates of *S. zooepidemicus* from three UK outbreaks of acute fatal haemorrhagic pneumonia in dogs during 1999, 2001 and 2008 were found to be related to isolates from three outbreaks of this disease in the USA during 2005, 1993 and 2006, respectively. Our data provide strong evidence that *S. equi* evolved from an ancestral *S. zooepidemicus* strain and that certain related strains of *S. zooepidemicus* have a greater propensity to infect particular hosts and tissues.

INTRODUCTION

*Streptococcus equi* subsp. *zooepidemicus* (*S. zooepidemicus*) is the most frequently isolated opportunistic pathogen of horses: it is associated with inflammatory airway disease in thoroughbred racehorses (Wood et al., 1993, 2005), uterine infections in mares (Hong et al., 1993; Smith et al., 2003) and ulcerative keratitis (Brooks et al., 1999). It is also associated with disease in a wide range of other animal hosts, including cattle (Sharp et al., 1995), sheep (Las Heras et al., 2002; Stevenson, 1974), pigs (Salasia et al., 2004; Soedarmanto et al., 1996), monkeys (Salasia et al., 2004; Soedarmanto et al., 1996) and humans (Bradley et al., 1991; Downar et al., 2001; Hashikawa et al., 2004). Recently, several cases of acute fatal haemorrhagic pneumonia in dogs have been attributed to *S. zooepidemicus* infection (Chalker et al., 2003; Pesavento et al., 2008). A biovar of *S. zooepidemicus*, *Streptococcus equi* subsp. *equi* (*S. equi*), is the causative agent of equine strangles, one of the most frequently diagnosed and important infectious diseases of horses worldwide. This disease is characterized by pyrexia, followed by profuse nasal discharge and abscessation of the lymph nodes of the head and neck. The swelling of lymph nodes in the head and neck may restrict the airway, and it is this clinical feature that gave the disease its name (Timoney, 1993). However, despite the prevalence of this global endemic disease, *S. equi* infections of other animals are rare, with only a few documented cases in humans (Breiman & Silverblatt, 1986; Elsayed et al., 2003; Popescu et al., 2006) and one ‘strangles-like’ infection in a dog (Ladlow et al., 2006). The variety of distinct infections caused by this group of pathogens suggests that determination of the population genetic structure of the *S. zooepidemicus* group could yield valuable information towards the identification of the genetic basis for the selection of host and site of infection.

**Abbreviations:** ΔlnL, difference in log-likelihood; AHT, Animal Health Trust; DLV, double locus variant; dS/dN, ratios of non-synonymous to synonymous substitutions; ET, electrophoretic type; ML, maximum-likelihood; MLEE, multi-enzyme electrophoresis; MLST, multilocus sequence typing; PFGE, pulsed-field gel electrophoresis; SLV, single locus variant; ST, sequence type.
Strains of *S. zooepidemicus* have previously been discriminated using two different subtyping approaches. Firstly, individual loci or uncharacterized regions of the genome that are highly variable within the bacterial population were exploited. This type of approach is exemplified by pulsed-field gel electrophoresis (PFGE) (Lindmark et al., 1999) and the PCR typing of M-protein hypervariable regions (Walker & Timoney, 1998) or the 16S–23S-RNA intergenic spacer (Chanter et al., 1997). These methods used restriction enzymes or PCR primers to maximize the detected variability within the population. However, the variation that is indexed evolves very rapidly through unknown mechanisms and although useful in short-term epidemiological studies, these methods may be misleading in the study of global epidemiology (Maiden et al., 1998).

In addition, PFGE and PCR subtyping methodologies provide little information regarding the genetic relationships of one bacterial strain with another and the results obtained from these analyses are difficult to compare between different laboratories.

The second type of approach identifies variation that accumulates very slowly in the population that is likely to be selectively neutral. Multilocus enzyme electrophoresis (MLEE) indexes allelic variation by measuring differences in the mobility of housekeeping enzymes on cellulose acetate strips (Selander et al., 1986). Although only a small number of electromorphs corresponding to distinct alleles can be identified within a population, high levels of discrimination are achieved by analysing many loci. The *S. zooepidemicus* MLEE scheme utilizes 10 enzymes with an average number of 2.5 alleles per locus (Jorm et al., 1994). This approach identified 41 electrophoretic types (ETs) of *S. zooepidemicus*, including the ET 12 characteristic of *S. equi*. However, a major problem with MLEE typing methods is that the results obtained from different laboratories remain difficult to compare.

Multilocus sequence typing (MLST) is a development from MLEE, which allows the identification of different alleles directly from the nucleotide sequences of ~400–500 bp internal fragments of usually seven housekeeping genes (Maiden, 2006). This has several advantages over alternative typing methods. Firstly, much more variation can be detected, resulting in many more alleles per locus than can be detected by MLEE (for instance, some of the more established schemes currently have over 400 alleles at most of their loci). Secondly, the majority of the variation detected results in synonymous codon changes and so is not subjected to functional selective pressure. Finally, nucleotide sequence data are fully portable and can easily be compared between different laboratories via electronic databases available on the Internet. These databases are therefore a powerful resource with which to conduct global epidemiological studies (Maiden, 2006). In this report, a MLST scheme using seven housekeeping loci was used to evaluate 253 *S. zooepidemicus* and 24 *S. equi* isolates from horses, donkeys and dogs from several continents and spanning a time period of >30 years.

### METHODS

#### Bacterial isolates

Full details of all of the isolates examined in this study are available on the MLST database (http://pubmlst.org/zooepidemicus/). A total of 46 *S. zooepidemicus* isolates from horses, donkeys and dogs in the UK and USA were kindly provided by Professor John Timoney (University of Kentucky). A further seven isolates of *S. zooepidemicus* from infections of dogs in the UK were kindly provided by Professor Joe Brownlie (Royal Veterinary College). The remaining 200 isolates of *S. zooepidemicus* and 24 *S. equi* isolates from the UK, USA, Canada, Ireland, Spain and Australia were obtained from the Animal Health Trust (AHT) collection. The majority of these isolates had been previously submitted to the AHT diagnostic laboratories from cases of clinical disease in horses or dogs. The *S. zooepidemicus* H70 reference strain was obtained from a healthy thoroughbred racehorse in Newmarket by the AHT and is the focus of the *S. zooepidemicus* genome-sequencing project at the Sanger Institute (http://www.sanger.ac.uk/Projects/S_zooepidemicus/). The *S. equi* 4047 reference strain was isolated from a New Forest pony by the AHT and is the focus of the *S. equi* genome-sequencing project at the Sanger Institute (http://www.sanger.ac.uk/Projects/S_equi/).

#### Preparation of chromosomal DNA

*S. zooepidemicus* and *S. equi* strains were grown on COBA strep select plates (bioMérieux). DNA was then purified from a single colony using GenElute spin columns according to the manufacturer’s instructions (Sigma).

#### MLST

Internal fragments of the carbamate kinase (*arcC*), ribonucleoside-diphosphate reductase (*nrdE*), prolyl-tRNA synthetase (*proS*), signal peptide 1 (*spi*), thymidylate kinase (*tdk*), triphasos-pherase isomerase (*tpi*) and acetyl-CoA acetyltransferase (*yqiL*) genes were amplified by PCR using the following primer pairs: *arcC1* (5′-AGA CAT CTA CGG ACT AAC AC-3′) and *arcC4* (5′-TCT GAA AGG GTT TGG CTA GC-3′); *nrdE1* (5′-TTG TCC TTC AGG TGA CAG ATG-3′) and *nrdE4* (5′-AGA CTA GGC GTC TAA ACC-3′); *proS1* (5′-TTG GAA GAT AAC GAC CAG ATC-3′) and *proS4* (5′-CCT GAT CCT TGA CAT TAA CGG-3′); *spi1* (5′-CTT AGA GCA TTG CGC TAA GC-3′) and *spi4* (5′-TTG CCT GCT ATC TAG GGA AG-3′); *tdk1* (5′-GAA TTC AGG AAA GAC GAT CG-3′) and *tdk4* (5′-TAA TGC TTG CCA GAA ACT GG-3′); *spi1* (5′-GAGT AGT AAG CAA ATT ACC-3′) and *spi4* (5′-AGG AAG GGA AGG AAG CTA TC-3′); and *yqiL1* (5′-CCA CAT GGG ATT AGC AG-3′) and *yqiL4* (5′-TCA ATG CAG TCC TTT GAC C-3′).

The PCRs were performed in volumes of 20 μl using Taq DNA polymerase (Sigma) with 34 cycles of 94 °C for 30 s, 55 °C for 30 s and 72 °C for 1 min. The amplified DNA fragments were purified using a PCR purification kit (Qiagen), and the sequence of each fragment was obtained on both strands using an ABI3100 DNA sequencer with BigDye fluorescent terminators and the following primers, which are internal to those used in the initial PCR: *arcC2* (5′-TGT CGC ACT TGG AGG AAA TG-3′) and *arcC3* (5′-CAC CAC ACC AGA ATC AAC-3′); *nrdE2* (5′-GGC TCT ATT TCT GCC TGG C-3′) and *nrdE3* (5′-GCA TAG GTT GCT CAT GAT GAT G-3′); *proS2* (5′-GAC TTC TTA GCA CAA AGT GC-3′) and *proS3* (5′-AAC GGA GAT GGC TCT TTA GG-3′); *spi2* (5′-CTT ACT TTT TGG ACT CTC ACG-3′) and *spi3* (5′-GAT TTT ATT GAG TGG-3′); *tdk2* (5′-CTT AGA GGA AGC AGA CAA TTA TG-3′) and *tdk3* (5′-GTT CTC ATT GCC ACC AAT TG-3′); and *yqiL2* (5′-AGT ACC GCA AGG TAA AGG TG-3′) and *yqiL3* (5′-GCC TGA CGC TTT TTC ATC TC-3′).

Sequence data were assembled using SeqMan 5.03 (*DNASTAR*), and high-quality double-stranded sequence data were used for further analysis. For each locus, every unique sequence was assigned a distinct allele number, and each sequence type (ST) was defined by a series of seven integers (the allelic profile) corresponding to the alleles at the
seven loci, in the order (alphabetical) arcC, nrdE, proS, spi, tdk, tpi and yqiL. A MLST database containing the sequences of all alleles, the allelic profiles and information about the origins of each isolate is maintained at the University of Oxford and can be found on the S. zooepidemicus pages of the MLST website (http://pubmlst.org/s zooepidemicus/) (Jolley et al., 2004).

Allele sequences for each ST were exported for each locus in turn from the MLST database using the built-in concatenation function. Previously described statistical methods were used to examine the extent of congruence among gene trees inferred from these individual loci (Feil et al., 2001; Holmes et al., 1999). Firstly, the Shimodaira–Hasegawa test was used to determine whether there were significant differences among the gene trees inferred for each gene. This analysis was undertaken by determining the maximum-likelihood (ML) tree for each of the seven genes and then comparing, in turn, the difference in log-likelihood (Δ-ln L) between the topologies generated using sequences from each of the other six genes. To further assess the extent of congruence among the seven ML gene trees, we used the randomization test (Feil et al., 2001; Holmes et al., 1999). In this case the Δ-ln L values of each of the seven ML trees generated using the housekeeping genes were compared to the values calculated for 200 random trees and those calculated from the trees generated for the other housekeeping genes. All of these analyses were carried out using PAUP* version 4.0b10 (Swofford, 2003).

The relatedness of isolates was determined using eBURST (http://eburst.mlst.net) (Feil et al., 2004) and ClonalFrame v1.1 (Didelot & Falush, 2007). The Sequence type Analysis and Recombinational Test (START2) software (Jolley et al., 2001) was used to determine the ratios of non-synonymous to synonymous substitutions (dN/dS) for each locus (Nei & Gojobori, 1986).

Fisher’s exact test was used to test the null hypothesis that there was no statistically significant difference in the proportion of S. zooepidemicus isolates recovered from cases of abortion or respiratory infection that clustered together compared with those isolates that did not cluster. Statistical analysis was conducted using Stata 9.2 software (StataCorp LP), with statistical significance set at P < 0.05.

**RESULTS**

**Development of a MLST scheme for the S. zooepidemicus group**

Chromosomal DNA was obtained from 253 isolates of S. zooepidemicus and 24 isolates of S. equi. Twelve loci were initially chosen as candidates for the MLST scheme. These either had been used successfully for other streptococci (Coffey et al., 2006; Enright & Spratt, 1998; Enright et al., 2001) or were selected with guidance using data from the publicly available Wellcome Trust Sanger Institute S. equi and S. zooepidemicus genome sequencing projects (http://www.sanger.ac.uk). Seven of these loci were then chosen based on the observed sequence diversity in pilot studies with a limited set of nine diverse S. zooepidemicus isolates previously characterized by PCR subtyping (Chanter et al., 1997; Walker & Timoney, 1998) and four diverse S. equi strains (Kelly et al., 2006). Loci selected for this study were evenly distributed across the recently completed S. equi strain 4047 and S. zooepidemicus strain H70 genomes with minimum distances of 30 kb (arcC and nrdE) and 55 kb (arcC and nrdE), respectively (Table 1).

**Population structure**

The collection of 253 S. zooepidemicus and 24 S. equi isolates was assembled with several goals in mind. Firstly, we evaluated many S. zooepidemicus isolates with large temporal and/or spatial distances from their isolation. Secondly, the isolates had been recovered in association with a variety of host tissues and diseases, including respiratory colonization (108 isolates), pneumonia or purulent rhinitis (35 isolates), acute fatal haemorrhagic pneumonia in dogs (18 isolates), non-strangles lymph node abscesses (13 isolates), abortion or uterine infection (48 isolates), wound infection (19 isolates), strangles (24 isolates) and other diseases such as nephritis, keratitis and peritonitis (12 isolates). One of the main aims of this was to determine whether related strains were associated

### Table 1. Characteristics of housekeeping gene loci included in the S. zooepidemicus MLST scheme

<table>
<thead>
<tr>
<th>Locus</th>
<th>Size of sequenced fragment (bp)</th>
<th>No. of alleles</th>
<th>No. of polymorphic nucleotide sites</th>
<th>dN/dS</th>
<th>Chromosomal location (no. of kb from origin)</th>
</tr>
</thead>
<tbody>
<tr>
<td>arcC</td>
<td>437</td>
<td>28</td>
<td>30</td>
<td>0.0160</td>
<td>1586</td>
</tr>
<tr>
<td>nrdE</td>
<td>448</td>
<td>25</td>
<td>49</td>
<td>0.0274</td>
<td>1641</td>
</tr>
<tr>
<td>proS</td>
<td>435</td>
<td>33</td>
<td>51</td>
<td>0.0535</td>
<td>241</td>
</tr>
<tr>
<td>spi</td>
<td>459</td>
<td>28</td>
<td>35</td>
<td>0.0190</td>
<td>1838</td>
</tr>
<tr>
<td>tdk</td>
<td>370</td>
<td>22</td>
<td>24</td>
<td>0.0472</td>
<td>1044</td>
</tr>
<tr>
<td>tpi</td>
<td>424</td>
<td>27</td>
<td>38</td>
<td>0.0793</td>
<td>736</td>
</tr>
<tr>
<td>yqiL</td>
<td>396</td>
<td>42</td>
<td>43</td>
<td>0.0917</td>
<td>601</td>
</tr>
</tbody>
</table>

*S. zooepidemicus* and *S. equi* based on the recently completed S. zooepidemicus H70 genome at http://sanger.ac.uk.

**REFERENCES**

Coffey et al., 2006; Enright & Spratt, 1998; Enright et al., 2001; K. Webb and others

**Table 1. Characteristics of housekeeping gene loci included in the S. zooepidemicus MLST scheme**

Alleles of the seven housekeeping loci can be obtained at http://pubmlst.org/szooepidemicus/.
with a particular disease or host and whether any strains of *S. zooepidemicus* clustered with *S. equi*. Finally, we included several *S. zooepidemicus* and *S. equi* isolates that had been previously subtyped using existing PCR subtyping (Chanter *et al.*, 1997; Walker & Timoney, 1998) and SeM subtyping methods (Kelly *et al.*, 2006), respectively, to provide validation of the new MLST scheme.

The sequences of the seven loci were determined for each of the 277 isolates, and their allelic profiles were assigned based on gene fragments with sequence lengths of between 370 bp (*tdk*) and 459 bp (*spi*). The sequences of all seven loci, and the properties of the 277 isolates studied, are available from the MLST website (http://pubmlst.org/szooe). The number of unique alleles identified for each of the seven housekeeping loci ranged from 22 (*tdk*) to 42 (*yqiL*) (Table 1). The mean number of alleles per locus was 29.3. The most common allele for each locus was *arcC* (57 isolates); *nrdE* (77 isolates); *proS* (46 isolates); *spi* (64 isolates); *tdk* (132 isolates); *tpi* (57 isolates); and *yqiL* (59 isolates) (no isolates were identified that had this allelic profile among the 277 studied). One hundred and thirty allelic profiles were found in this study. Twenty-three of the 24 *S. equi* isolates were found to be ST-179. The remaining *S. equi* isolate was found to be a single locus variant (SLV) of ST-179, ST-151. The most common *S. zooepidemicus* ST was ST-108; all 20 of these isolates came from an AHT study of equine respiratory isolates from 1996 and were included for comparison of MLST with the PCR-typing method. The *dS/dk* ratios for all seven loci were calculated, and all were substantially less than 1 (Table 1).

**Evidence of recombination**

To determine if recombination plays an important role in *S. zooepidemicus* diversity we performed statistical tests of congruence. The Shimodaira–Hasegawa test compares ML trees constructed using data from each locus and tests if they are significantly different from each other. In all cases, topologies were highly significantly different (*P < 0.001*) (Table 2). The randomization test compares these tree topologies and tests whether any congruence between the loci is better than that seen against random trees. In most cases, trees drawn from any two loci are slightly more congruent than random trees. However, on comparing *spi* against other MLST loci, the other loci trees cluster with the random trees (Fig. 1).

**Lineage assignment**

Clustering of STs was investigated using eBURST at the default setting (6 out of 7 matches for group definition). This identified one clonal complex of five STs consisting of ST-1, ST-71, ST-104, ST-108 and ST-166, with ST-71 as the founding member, and one clonal complex of four STs, consisting of ST-114, ST-118, ST-139 and ST-146, with ST-118 as the founding member (Fig. 2). In addition, nine groups of three STs and 12 pairs of STs clustered together. Overall the *S. zooepidemicus* population as represented by eBURST is indicative of a diverse species.

We further analysed the MLST data using ClonalFrame (Didelot & Falush, 2007). A majority-rules consensus tree was generated from 12 independent runs each with 250 000 iterations (100 000 burn-in iterations, thinning interval of 100) (Fig. 3). When tested with ClonalFrame’s built-in tool, full convergence was not achieved between individual runs so the Markov chains were examined further using Tracer (13) (Tracer v1.4, available from http://beast.bio.ed.ac.uk/Tracer). Although the variance between some chains was significantly greater than the variance within

![Fig. 1. Congruence of the *spi* ML tree with trees generated from other MLST loci and 200 random trees.](image-url)

**Table 2. Tests for congruence among *S. zooepidemicus* isolates**

<table>
<thead>
<tr>
<th>Locus</th>
<th>−ln L of ML tree</th>
<th>Δ−ln L of competing ML trees</th>
<th>P</th>
<th>Δ−ln L of random ML trees</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>arcC</em></td>
<td>928.42</td>
<td>808.47–956.60</td>
<td>&lt;0.001</td>
<td>920.30–1116.60</td>
</tr>
<tr>
<td><em>nrdE</em></td>
<td>1098.78</td>
<td>1532.67–2017.40</td>
<td>&lt;0.001</td>
<td>1836.98–2271.61</td>
</tr>
<tr>
<td><em>proS</em></td>
<td>1227.43</td>
<td>1803.59–2138.14</td>
<td>&lt;0.001</td>
<td>2069.47–2337.77</td>
</tr>
<tr>
<td><em>spi</em></td>
<td>1073.43</td>
<td>1274.76–1684.47</td>
<td>&lt;0.001</td>
<td>1326.50–1955.34</td>
</tr>
<tr>
<td><em>tdk</em></td>
<td>714.64</td>
<td>475.26–594.50</td>
<td>&lt;0.001</td>
<td>572.09–640.75</td>
</tr>
<tr>
<td><em>tpi</em></td>
<td>995.46</td>
<td>1221.58–1565.14</td>
<td>&lt;0.001</td>
<td>1419.31–1667.39</td>
</tr>
<tr>
<td><em>yqiL</em></td>
<td>1110.06</td>
<td>1715.71–1980.95</td>
<td>&lt;0.001</td>
<td>1820.33–2125.06</td>
</tr>
</tbody>
</table>
chains, combining all the runs produced a good sample of the posterior and a 50 % consensus tree was considered robust. This analysis was in concordance with eBURST, identifying all 23 eBURST complexes. The analysis of the MLST data by ClonalFrame, however, linked several of the groups identified by eBURST. For example, the two SLV pairs ST-16/140 and ST-134/173 now clustered together. Of particular note was the observation that the S. equi SLV pair ST-151/179 now clustered with the S. zooepidemicus ST-19 and ST-137/149 complexes, ST-6, ST-49, ST-56 and ST-133. Unlike with eBURST, ST-18 did not cluster with ST-134 and ST-173 in the consensus ClonalFrame analysis, although distant linkage was seen in individual runs. ST-18 has 20 nucleotide differences in its proS-1 relative to the proS-15 of ST-173 and ST-134. Nineteen of these differences occur in the 3’ region of the sequenced fragment from nucleotide 200 onwards and a single difference at nucleotide 43. A maximum chi-squared test (performed in START2) between these two alleles identifies a possible recombination event at nucleotide 199, but this is not deemed statistically significant ($P=0.134$).

**Comparison of MLST with other typing methods**

PCR subtyping methods (Chanter et al., 1997; Walker & Timoney, 1998) have previously been conducted on a subset of the S. zooepidemicus isolates reported here (Newton et al., 2008). For isolates of related ST, 39 (57 %) also had concordant 16S–23S-RNA intergenic spacer and M-protein hypervariable region PCR profiles, whereas 30 displayed distinct PCR profiles (http://pubmlst.org/szooepidemicus/). If only the 16S–23S-RNA intergenic spacer PCR was used, 57 isolates (82 %) had concordant related STs, compared with 42 isolates (61 %) if the M-protein hypervariable region PCR was used. Twenty-four S. equi isolates comprising seventeen different SeM subtypes were included in the study. Twenty-three of these were identified as ST-179; the remaining isolate (SeM subtype 6) was found to be a SLV of ST-179, ST-151.

**S. zooepidemicus associated with disease**

A total of 145 isolates of S. zooepidemicus associated with disease in the UK, USA, Spain and Ireland between 1976 and 2007, and 24 S. equi isolates from cases of strangles in the UK, USA, Australia, Canada and Ireland between 1981 and 2006, were included in the study. Some trends in the disease association of clusters of STs were apparent following analysis using the ClonalFrame software (Fig. 3). Of 46 independent isolates in the ST-71 complex, 40 (87 %) were isolated from the respiratory tract ($P<0.001$). This group contained isolates from the UK and USA over a long time period. Other clusters of STs that were also associated with respiratory isolates include 13 of 17 isolates...
(76%) in the ST-8 cluster, 16 of 17 isolates (94%) in the ST-106/164 cluster, all 8 isolates in the ST-103 complex, all 11 isolates in the ST-119 complex and all 5 isolates in the ST-117/148 cluster.

The *S. zooepidemicus* STs that were found to cluster with *S. equi* by ClonalFrame analysis were significantly associated with cases of uterine infection or abortion (*P*, 0.001). Of the 11 *S. zooepidemicus* isolates in this cluster, two ST-49 isolates originated from repeated swabs of the same wound infection. Taking into account the small number of repeated isolates (six) across the *S. zooepidemicus* study population (253), of 10 independent *S. zooepidemicus* isolates in this cluster, 8 (80%) were associated with uterine infection and abortion. This cluster accounted for 18% (8 of 44) of all independent isolates from uterine infection or abortion in this study compared with only 1% (2 of 203) of other independent *S. zooepidemicus* isolates studied.

Isolates from eight independent outbreaks of acute fatal haemorrhagic pneumonia in dogs were investigated in this study. An isolate from an outbreak in the UK during 1999 (ST-123) was a double locus variant (DLV) of an isolate

---

**Fig. 3.** A majority-rules consensus tree generated from 12 independent runs of ClonalFrame, each with 250,000 iterations, and imported into MEGA to display as a radial tree. The ST-151/179 and ST-71 clusters are labelled for clarity. *Isolates from outbreaks of acute fatal haemorrhagic pneumonia in dogs.*
from similar outbreak of disease in the USA during 2005 (ST-129). Four isolates in the ST-10 cluster from an outbreak in the UK during 2001 were found to be related to an isolate from an outbreak in the USA during 1993 (ST-169) by ClonalFrame analysis. Three isolates from a recent outbreak in the USA during 2006 (ST-173) were single locus variants (SLVs) of four ST-18 isolates recovered from greyhounds affected in a UK outbreak during 2008. Of the canine STs identified in this study, an isolate of ST-2 was also recovered from an equine tracheal wash; an isolate of ST-10 was also recovered from an equine wound infection; an isolate of ST-162 was also recovered from a case of purulent rhinitis in a foal; and a SLV of ST-173 and a DLV of ST-18, ST-134, was recovered from a case of equine mastitis.

**DISCUSSION**

MLST is a well-established molecular typing method that has been used successfully for the determination of the population structure of many bacterial species (Coffey et al., 2006; Enright & Spratt, 1998; Enright et al., 2001; Maiden et al., 1998). Here we report the development of an unambiguous and discriminatory MLST scheme for the *S. zooepidemicus* group. The data generated from the *S. zooepidemicus* and *S. equi* isolates in the present study have been used to construct a database on the main MLST website that permits the direct comparison of data generated in other laboratories (http://pubmlst.org/szooepidemicus/). This database is fully accessible and searchable and can be accessed and added to by other researchers. We believe that such a scheme is of particular value for the study of a zoonotic pathogen such as *S. zooepidemicus*, which is associated with a variety of diseases in a large number of different animal hosts.

The statistical tests of congruence identified that the topologies of ML trees constructed using data from each locus were significantly different (*P*<0.001). If each gene tree had the same topology (i.e. phylogenetic congruence), as expected under entirely clonal evolution, then they should not differ significantly in likelihood. The data collected in this study suggested that higher levels of recombination may be occurring at the *spi* locus than at the other six loci examined. This is probably best illustrated by the particularly high prevalence of the *S. equi* *spi* allele 45 in the *S. zooepidemicus* isolates examined in this study (40 of 253 isolates: 16 %), compared with the other six *S. equi* alleles, which had the following prevalences: *arcC* (0/253), *nrDE* (0/253), *proS* (1/253), *tdk* (0/253), *tpi* (8/253: 3 %) and *yqiL* (0/253). One explanation for this observation is the occurrence of recombination with *S. equi* around the *spi* locus. Now that the *S. zooepidemicus* genome sequence has been completed, further studies are warranted to determine if there are coding sequences near the *spi* locus of *S. equi* that may be of value to *S. zooepidemicus*. In the current study there was no statistically significant association of the presence of the *spi* allele 45 and the origin of the isolate.

*S. zooepidemicus* was identified as a diverse species by *eburst* analysis. Only two *eburst* groups, of respectively five STs (based around ST-71) and four STs (based around ST-118), were identified. In addition, nine groups of three STs and 12 pairs of STs were found to cluster together. ClonalFrame analysis identified all of the 23 *eburst* groups; however, the ability of this software to account for the effect of mutation as well as homologous recombination events on the clonal pattern of inheritance generated a majority-rules consensus tree that linked many *eburst* groups, including: ST-16/140 and ST-134/173, and the ST-5 complex and ST-7/48; of particular note was the observation that the *S. equi* SLV pair ST-151/179 now clustered with the *S. zooepidemicus* ST-19 and ST-137/149 complexes, ST-6, ST-49, ST-56 and ST-133, albeit via deep branches on the radial tree. ClonalFrame added several STs to extend many of the *eburst* clusters, including ST-20 with the ST-118 complex, and ST-11, ST-46, ST-96, ST-113 and ST-153 with the ST-8 complex. ClonalFrame did not group ST-18 with the ST134/173 cluster in the consensus tree, although it was distantly linked in trees generated from individual runs. This is probably because, even though ST-18 and ST-173 are SLVs of each other, the pattern of nucleotide differences between the proS alleles of these strains is likely to have arisen through at least two genetic events. Overall, on examination of the sequence differences at each of the seven loci used in the *S. zooepidemicus* MLST scheme, the ClonalFrame analysis presented in Fig. 3 appears to most accurately represent the *S. zooepidemicus* population structure and best accounts for the evolutionary events occurring in this bacterial species.

The clustering of isolates by MLST was in good agreement with those obtained previously with PCR typing methods. As expected, of the two *S. zooepidemicus* PCR typing methods, the 16S–23S-RNA intergenic spacer PCR was noticeably more concordant with related STs than the M-protein hypervariable region PCR scheme. The SeM-typing scheme was found to have a much better ability to differentiate isolates of *S. equi* than MLST. *S. equi* is believed to have evolved relatively recently from an ancestral strain of *S. zooepidemicus* (Jorm et al., 1994) and as MLST schemes identify variation in housekeeping genes that accumulates very slowly in the population it was not surprising to find that the majority of *S. equi* isolates studied were of a single ST. In contrast, the SeM gene is known to be subjected to strong selective pressure (*dS*/ *dS* =5.93) (Waller & Jolley, 2007) and is likely to accumulate variation much more quickly, so different strains of *S. equi* can be differentiated using this method.

One of the key objectives of this study was to identify subtypes of *S. zooepidemicus* that shared a closer evolutionary ancestor with the *S. equi* biovar than the *S. zooepidemicus* H70 genome sequencing strain. ClonalFrame analysis identified nine STs that clustered with *S. equi* as illustrated...
in the majority-rules consensus tree shown in Fig. 3. The determination of the S. zooepidemicus population genetic structure presented here will enable research towards unravelling the key genetic events in the evolution of S. equi to focus on a much smaller number of related S. zooepidemicus subtypes that share a closer evolutionary ancestor. Of particular interest was the observation that rather than S. zooepidemicus STs isolated from cases of non-strangles lymph node abscesses, the S. zooepidemicus STs that most closely clustered with S. equi were significantly more likely to be associated with cases of uterine infection or abortion in horses (P<0.001). These data provide one possible explanation for earlier observations that found no evidence of colonization of the equine respiratory tract by S. equi prior to invasion (Sweeney et al., 2005).

We also found that isolates from six outbreaks of acute fatal haemorrhagic pneumonia in dogs in the UK or USA were genetically related, suggesting that certain subtypes of S. zooepidemicus may be more adept at causing disease in dogs.

These data provide the strongest evidence yet that a reclassification of S. equi subsp. zooepidemicus and S. equi subsp. equi is required to accurately reflect the diversity and origins of this important group of pathogens. A reclassification was previously suggested by a study that utilized MLEE to define the genetic structure of 70 isolates of S. equi and 177 isolates of S. zooepidemicus (Jorm et al., 1994). However, the isolates included in this study were of limited geographical diversity and the MLEE scheme developed had only the capability to differentiate 41 different ETs. Sixty-nine of the 70 S. equi subsp. equi isolates fell into ET 12, which suggested that they were members of a single clone and that S. zooepidemicus may be the archetypal species from which the clone designated subspecies equi was derived. The MLST data presented in this report provide strong evidence in agreement with this earlier MLEE study and suggest that these pathogens should be referred to as Streptococcus zooepidemicus and Streptococcus zooepidemicus subsp. equi.

In conclusion, we have developed the first unambiguous typing method for the determination of the population structure of S. zooepidemicus and linked several related STs to particular disease states. It is hoped that the data presented here can be used as a framework for the identification of virulence genes that form the genetic basis for the selection of host and site of infection.

ACKNOWLEDGEMENTS

We gratefully acknowledge Professor John Timoney (University of Kentucky) and Professor Joe Brownlie (The Royal Veterinary College), who provided isolates for inclusion in this study. We are also grateful to Dr Julian Parkhill and Dr Matt Holden of the Wellcome Trust Sanger Institute, who provided help and expertise in the use of the Artemis Comparison Tool and the S. equi and S. zooepidemicus genome sequencing data. We thank Xavier Didelot (University of Warwick) for advice and guidance in interpreting the ClonalFrame results. The Horse Trust funded the S. equi genome sequencing project and the Horserace Betting Levy Board funded the S. zooepidemicus genome sequencing project. The pilot study to select the most suitable seven loci for the S. zooepidemicus MLST scheme was funded by the Animal Health Trust. This work was supported by a project grant from the Horserace Betting Levy Board and the European Breeders Fund.

REFERENCES


http://mic.sgmjournals.org


Edited by: T. J. Mitchell