North American and European porcine reproductive and respiratory syndrome viruses differ in non-structural protein coding regions

R. Allende,1, 2 T. L. Lewis,2 Z. Lu,2 D. L. Rock,2 G. F. Kutish,2 A. Ali,1 A. R. Doster1 and F. A. Osorio1

1 Department of Veterinary and Biomedical Sciences, University of Nebraska–Lincoln, Lincoln, NE 68583-0905, USA
2 Plum Island Animal Disease Center, USDA/ARS, Greenport, NY 11944, USA

Although North American and European serotypes of porcine reproductive and respiratory syndrome virus (PRRSV) are recognized, only the genome of the European Lelystad strain (LV) has been sequenced completely. Here, the genome of the pathogenic North American PRRSV isolate 16244B has been sequenced and compared with LV. The genomic organization of 16244B was the same as LV but with only 63.4 % nucleotide identity. The 189 nucleotide 5′ non-coding region (NCR) of 16244B was distinct from the LV NCR, with good conservation (83 %) only over a 43 base region immediately upstream of open reading frame (ORF) 1a. Major differences were found in the region encoding the non-structural part of the ORF1a polyprotein, which shared only 47 % amino acid identity over 2503 residues of the six non-structural proteins (Nsps) encoded. Nsp2, thought to have a species-specific function, showed the greatest divergence, sharing only 32 % amino acid identity with LV and containing 120 additional amino acids in the central region. Nsps encoded by the 5′-proximal and central regions of ORF1b had from 66 to 75 % amino acid identity; however, the carboxy-terminal protein CP4 was distinct (42 % identity). The ORF1a–1b frameshift region of 16244B had 98 % nucleotide identity with LV. Consistent with previous reports for North American isolates, the six structural proteins encoded were 58 to 79 % identical to LV proteins. The 3′ NCR (150 nucleotides) was 76 % identical between isolates. These genomic differences confirm the presence of distinct North American and European PRRSV genotypes.

Introduction

Porcine reproductive and respiratory syndrome virus (PRRSV), the causative agent of PRRS, is a small enveloped virus (50–65 nm) with a single-stranded RNA genome of positive polarity. PRRSV belongs to the newly created family Arteriviridae, order Nidovirales, together with lactate dehydrogenase-elevating virus (LDV) of mice, equine arteritis virus (EAV) and simian haemorrhagic fever virus (SHFV) (Cavanagh, 1997).

PRRS was first reported in 1987 in North America (Keffaber, 1989) and a few years later, Wensvoort et al. (1991) isolated the causative virus in Europe. The disease is characterized by reproductive failure in sows, preweaning mortality and respiratory tract illness that can have severe consequences, especially in piglets (Snijder & Meulenberg, 1998; Rossow, 1998).

Studies with polyclonal antiseria have revealed antigenic differences between North American and European PRRSV isolates, and among North American isolates (Wensvoort et al., 1992). These antigenic differences were confirmed by studies with monoclonal antibodies directed against virus structural proteins. Although these isolates share some common antigenic epitopes, they represent distinct serotypes (Nelson et al., 1993; Drew et al., 1995; Dea et al., 1996).

The genome of the European PRRSV Lelystad strain (LV), the only PRRSV that has been sequenced completely (Meulenberg et al., 1993a), is 15 kb long and contains seven open reading frames (ORFs). ORF1a and ORF1b are located immediately downstream of the 5′ non-coding region (NCR). A ribosomal translational frameshift mechanism encodes a
potentially large polyprotein whose cleavage products are involved in virus transcription and replication. ORFs 2–7 encode the structural proteins of the virion and are located at the 3′ end of the genome.

Many North American isolates have been sequenced partially at the 3′ end (3 kb) of the genome (ORFs 2–7) (Meng et al., 1995a, b; Murtaugh et al., 1995; Kapur et al., 1996; Yang et al., 1998) and the sequences are at least 90% identical. However, there is much greater variation in the structural proteins (58–79% amino acid identity) between North American isolates and LV (Meng et al., 1995a; Murtaugh et al., 1995). Also, analysis of a 600 bp segment in the polymerase-coding region of ORF1b of some North American isolates (Gilbert et al., 1997) indicates that distinct variability exists between the two types. Based on these data, it has been suggested that North American isolates and LV represent different genotypes, as previously reported (Meng et al., 1995a; Suarez et al., 1996).

Here we have sequenced and analysed the entire genome of a North American PRRSV isolate, 16244B, and have compared it to the European PRRSV LV strain. Our data confirm that North American and European PRRSV isolates represent distinct virus genotypes that exhibit marked differences in non-structural proteins (Nsps) and NCRs.

Methods

- **Virus, cells and RNA isolation.** PRRSV strain 16244B (2/18/97 Nebraska) was isolated from a 7-day-old piglet at a farm in Nemaha County, NE, USA, where sows experienced severe reproductive failure from late January to early February, 1997. The diagnosis of PRRS at this farm was confirmed by virus isolation, immunohistochemistry and histopathology.

  MARC-145 cells, a subclone of the monkey kidney cell line MA-104 (Kim et al., 1993), were obtained from J. Kwang (USDA Meat Animal Research Center, Clay Center, NE, USA) and propagated in Dulbecco’s modified Eagle’s medium supplemented with 10% foetal bovine serum. The growth medium was removed 24 h after cell confluence and 1 ml serum-supplemented with 5% foetal bovine serum was added and the cells were incubated at 37°C for 72 h. When approximately 80% of the cells exhibited cytopathic effect, the virus suspension was frozen, thawed and used to infect the cells with guanidinium thiocyanate followed by ultracentrifugation through a 5.7 M CsCl cushion (Sambrook et al., 1989).

  Total RNA from infected cells with guanidinium thiocyanate followed by ultracentrifugation through a 5.7 M CsCl cushion (Sambrook et al., 1989). Total RNA from mock-infected cells was also extracted by the same procedure. RNA size was analysed by electrophoresis in a 0.8% neutral agarose gel.

- **RT–PCR and PCR.** Viral and mock cDNA were constructed by RT–PCR by using random hexanucleotide primers and SuperScript II reverse transcriptase (Gibco BRL) following the supplier’s instructions. The resulting cDNA was used as template in the PCR. The PCR was performed with PRRSV-specific primers designed from the sequence of the 3′ end (3 kb) of the PRRSV VR-2332 genome (Murtaugh et al., 1995). Primers were also designed from conserved domains in other arteriviruses (Godeny et al., 1993). Both ‘regular’ PCR with Taq DNA polymerase (Promega) and long-distance PCR with Advantage cDNA polymerase mix (Clontech) were performed according to the supplier’s instructions.

  The oligonucleotides used for PCR, together with their sequences and positions in the genome of PRRSV 16244B, were:

  - TL-13R (5′ TCGCCCTAATTGAACTGGTAC 3′; 15345–15366);
  - TL-13F (5′ TCTTTTGGCCATCTTGGTGGC 3′; 13753–13774);
  - TL-12R (5′ CGGTCAAGCTATTCTCCAACATCTTA 3′; 13785–13811);
  - TL-12F (5′ CCTGTTTTCGCCAAAGAGCTTAC 3′; 12603–12680);
  - TL-9cF (5′ ATGCTGAAATGACCGACCA 3′; 12503–12522);
  - TL-9cF (5′ GGCATTGTTTGGCTCCAC 3′; 12178–12198);
  - TL-Pol-F (5′ CATTGAAAAAGCGTTTAAT 3′; 8781–8800);
  - TL-Pol-R (5′ AGGGGTCAACTCTTAACT 3′; 9181–9201);
  - RA-1031R (5′ GAGAGCCAGCACCTGGC 3′; 993–1009);
  - RA-900F (5′ TGAATTTGAAACGTGTACCCCAGGGAGT 3′; 815–842).

- **cDNA library construction.** A cDNA library was constructed by using the SuperScript plasmid system (Gibco BRL) with oligo(dT) as the primer, according to the manufacturer’s instructions. The resulting double-stranded cDNA was ligated into the Nol–Sal site of linearized pSPORT1 vector (Gibco BRL) and then used to transform E. coli Max Efficiency DH5α competent cells (Gibco BRL). Colony filters were hybridized with 32P-labelled cDNA probes synthesized by PCR from PRRSV 16244B.

- **DNA sequencing and computer analysis.** PCR amplification products and clones were sequenced with T7 and Sp6 forward and reverse primers (Sambrook et al., 1989), as well as virus-specific primers, by the deoxyxynucleotide chain-termination method (Sanger et al., 1977). Sequencing reactions were prepared with the Dye Terminator Cycle Sequencing kit (Perkin Elmer). Sequences were determined with an ABI PRISM 377 automated DNA sequencer. The ABI sequence software version 3.0 was used for lane tracking and trace extraction. The nucleotide bases from the chromatogram traces were determined with Phred (version 0.961028), which also produced a ‘quality file’ of the predicted probability of error at each position. The sequences were assembled with Phrap (version 0.900731) by using the quality files and default settings to produce a consensus sequence with some subsequent manual editing with the Consed sequence editor (version 4.1). The assembled sequence was also checked for possible frameshift errors with the NAP and AAT computer programs (Huang & Zhang, 1996; Huang et al., 1997), which align DNA sequence by codons to a target protein sequence, i.e. ORFs 1–7 from other arteriviruses.

  Sequences were compared by using the computer programs FASTA (Pearson, 1990), CLUSTAL W 1.74 (Thompson et al., 1994) and SIM2 (Chao et al., 1997). The putative pseudo-knot region between ORFs 1a and 1b was examined by using the computer programs MFOLD (Devereux et al., 1984) and RAGA (Notredame et al., 1997). Predicted proteins sequences were analysed with the GCG programs (Devereux et al., 1984) and SAPS (Brendel et al., 1992) and compared with sequences in the existing genetic databases with the programs FASTA and BLAST2 (Altschul et al., 1997). Protein multiple sequence alignments were made with the MSA (Lipman et al., 1989) and GCC Package programs with the Dayhoff PAM-250 symbol comparison table and a cut-off value of 0.5 for peptide comparison. The programs MEMSAT (Jones et al., 1994) and TopPred II (Claros & von Heijne, 1994) were used for transmembrane prediction.

- **Genome assembly.** The PRRSV 16244B RNA genomic sequence was assembled as follows. After isolation of total RNA from virus- and mock-infected MARC-145 cells and reverse transcription, the cDNA was used as a template for PCR with the primers described previously. Four overlapping PCR products spanned the region from the 3′ NCR to the RNA-dependent RNA polymerase (RdRp) gene. Long-distance PCR products, made by using the primer RA-900F, spanned from the 5′
PRRSV inter-serotype sequence variation

terminus of ORF1a towards the RdRp region. Sequences obtained from six independent cDNA clones covered parts of ORF1a and PCR products were used to fill the remaining gaps in ORF1a.

The cDNA of the 5' and 3' terminal sequences of the genomic RNA were synthesized by using the primers RA-1031R and TL-13F and the rapid amplification of cDNA ends (RACE) with the Marathon cDNA amplification kit (Clontech), following the manufacturer’s protocol. PCR fragments were cloned into linearized pCR 2.1-TOPO vector by using the TOPO TA cloning kit (Invitrogen) for further sequencing.

At least three independent clones from each transformation reaction were sequenced to obtain the sequence of the 5' terminus and the 5' NCR to a point approximately 700 nucleotides from the start codon of ORF1a. The 5'-terminal nucleotide was where the RACE adapter joined the viral RNA; however, most of the cDNA clones from the screened cDNA library stopped 10–17 bases downstream of the start of the RACE products. The 3' end of the viral genome including the poly(A) tail was sequenced in a similar manner.

Results and Discussion

Genome sequence analysis

PCR products from 16244B genomic RNA and clones from a cDNA library were sequenced in 545 reactions and assembled into one contiguous sequence of 15428 nucleotides. The consensus sequence had at least fourfold redundancy at each nucleotide position in both directions, with an average tenfold redundancy. The length of the 16244B RNA genome was similar to the 15101 nucleotide genome of the European LV strain (Meulenberg et al., 1993a). The mean G+C content of the genome of the 16244B isolate was 53%, which is similar to other arteriviruses (51–54%).

The organization of the 16244B genome was the same as that of LV but the two sequences had only 63–4% nucleotide identity (Table 1). Two large overlapping ORFs, ORF1a and ORF1b (11–9 kb in length), were located immediately downstream of the 5' NCR and were separated by a ribosomal frameshift junction. Six small ORFs were identified downstream of ORF1b, followed by a 3' NCR and terminating with a poly(A) tail.

5' NCR

The 5' NCR region, located upstream of the 16244B ORF1a start codon, was 189 nucleotides in length (Table 1) and 99% identical to the 5' NCR of ATCC VR-2332, another North American isolate (Fig. 1(a)). This is 32 bases shorter than,

Table 1. Comparison of the PRRSV 16244B and LV genomes

<table>
<thead>
<tr>
<th>Genome segment</th>
<th>Length (nucleotides)</th>
<th>Sequence identity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>16244B</td>
<td>LV</td>
</tr>
<tr>
<td>Total, excluding poly(A)</td>
<td>15411</td>
<td>15088</td>
</tr>
<tr>
<td>5' NCR</td>
<td>189</td>
<td>221</td>
</tr>
<tr>
<td>ORF1a</td>
<td>7509</td>
<td>7188</td>
</tr>
<tr>
<td>'Slippery' region</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Pseudo-knot</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>ORF1b</td>
<td>4389</td>
<td>4389</td>
</tr>
<tr>
<td>ORFs 2–7</td>
<td>3187</td>
<td>3185</td>
</tr>
<tr>
<td>3' NCR</td>
<td>150</td>
<td>114</td>
</tr>
</tbody>
</table>
Table 2. Comparison of predicted ORF1a and ORF1b polyprotein cleavage products from PRRSV isolates 16244B and LV

Information for LV was taken from Meulenberg et al. (1993a).

<table>
<thead>
<tr>
<th>ORF</th>
<th>Cleavage product</th>
<th>Genomic position (nucleotides)</th>
<th>Polyprotein position (amino acids)</th>
<th>Length (amino acids)</th>
<th>Amino acid sequence identity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>Nsp1</td>
<td>190–1338</td>
<td>1–383</td>
<td>383</td>
<td>385</td>
</tr>
<tr>
<td></td>
<td>Nsp2</td>
<td>1339–4278</td>
<td>384–1363</td>
<td>980</td>
<td>861</td>
</tr>
<tr>
<td></td>
<td>Nsp3</td>
<td>4279–5616</td>
<td>1364–1809</td>
<td>204</td>
<td>203</td>
</tr>
<tr>
<td></td>
<td>Nsp4</td>
<td>5617–6228</td>
<td>1810–2013</td>
<td>204</td>
<td>203</td>
</tr>
<tr>
<td></td>
<td>Nsp5</td>
<td>6229–7563</td>
<td>2014–2458</td>
<td>445</td>
<td>455</td>
</tr>
<tr>
<td>1b</td>
<td>RdRp</td>
<td>7680–9617</td>
<td>1–646</td>
<td>646</td>
<td>645</td>
</tr>
<tr>
<td></td>
<td>CP2</td>
<td>9618–10940</td>
<td>647–1087</td>
<td>441</td>
<td>442</td>
</tr>
<tr>
<td></td>
<td>CP3</td>
<td>10941–11609</td>
<td>1088–1310</td>
<td>223</td>
<td>224</td>
</tr>
<tr>
<td></td>
<td>CP4</td>
<td>11610–12071</td>
<td>1311–1463</td>
<td>153</td>
<td>152</td>
</tr>
</tbody>
</table>

and only 56% identical to, the 5’ NCR of LV (Meulenberg et al., 1998). The first 43 nucleotides upstream of the ORF1a start codon of PRRSV 16244B were 100% identical to VR-2332 but only 83% identical to LV. However, the remainder of the 5’ NCR was less than 40% identical to LV, with 21 nucleotides deleted in three gaps.

Two possible hairpin-loop structures were present in the leader sequence adjacent to the ORF1a start codon of both PRRSV strains (data not shown). These loops resembled the structures postulated by Shieh et al. (1987) to be the termination site for leader RNA synthesis in the coronavirus murine hepatitis virus.

The nucleotide sequence around the 16244B ORF1a start codon contained the bases 5’ UAACCAUG 3’ (Fig. 1a). This is identical to the ORF1a start of LV. A similar consensus sequence, 5’ UAACC 3’, is present in other PRRSV subgenomic RNAs in the leader–body junction region (de Vries et al., 1990; Chen et al., 1993; Meulenberg et al., 1993b; Zeng et al., 1995; Meng et al., 1996) and at the 3’ end of the leader sequence in other arteriviruses (Godeny et al., 1998; Meng et al., 1996; Meulenberg et al., 1993). This highly conserved nucleotide sequence at the end of the leader may be critical for joining the leader to the body of the message (Lai, 1990; Hiscox et al., 1995). However, outside this conserved start codon region, the North American and European PRRSV 5’ NCRs were distinct.

Analysis of ORF1a

The PRRSV 16244B ORF1a encoded 2503 amino acids. This ORF was 107 residues longer than LV and shared only 47% amino acid identity (56–3% nucleotide identity) with LV (Table 1). Sequence alignment of ORF1a polyproteins from 16244B and LV showed amino- and carboxy-terminal conserved domains separated by a central region of lower relatedness. As described by de Vries et al. (1997), six putative NspPs are predicted to result from ORF1a polyprotein processing, Nsp1α, Nsp1β and Nsp2–5. There are major differences in Nsp2 among arteriviruses and between these two PRRSV isolates (Table 2; Fig. 2).

Nsp1α and Nsp1β contained two putative papain-like cysteine proteases (Dougherty & Semler, 1993), with cleavage sites identified at residues Cys166 and Cys494 and putative catalytic residues at His146 and His330. Two papain-like cysteine proteases were also reported in LV (Meulenberg et al., 1993a) and LDV (Godeny et al., 1993), while only one was found in EAV (den Boon et al., 1995). Cleavage at position 166 would separate Nsp1α from Nsp1β, and the predicted cleavage at position 383 (YG/A) would separate Nsp1β from Nsp2 (Fig. 2).

The Nsp2 protein of 16244B was 120 amino acids longer than that of LV. The amino- and carboxy-terminal regions of...
the two proteins, 110 and 145 residues respectively, were 40% identical. Conserved cysteines (residues 437, 493, 524, 529, 1271, 1320 and 1327) were present in ORF1a of the 16244B strain. A proline-rich region of moderate complexity (Wooton & Federhen, 1993), located in the central portion of Nsp2 (amino acids 800–950), exhibited low, non-significant amino acid identity with LV (19%). Two hydrophobic segments were predicted, amino acids 1259–1286 and 1296–1310, and three transmembrane domains were predicted in the carboxy-terminal region (central–internal membrane-face positions 1263–1287, 1274–1310 and 1343–1360).

Like Nsp2 of LV, Nsp2 of 16244B contained a chymotrypsin-like (3C-like) cysteine protease domain (Dougherty & Semler, 1993) between residues 437 and 441. This 3C-like cysteine protease, first identified in EAV (Snijder et al., 1995), is conserved among arteriviruses and might be involved in the cleavage between Nsp2 and Nsp3 (Snijder et al., 1995). The His$_{507}$ residue probably has a catalytic role with the putative cleavage site (G/G) at residue 1363 (Fig. 2).

The Nsp2 protein has been shown to be highly variable among arteriviruses, e.g., LV (861 residues long), LDV (733 residues long) and EAV (401 residues long), with similarity only in the amino- and carboxy-terminal domains (de Vries et al., 1997). The central segments of the protein vary in both length and composition. The net charge of this central segment of Nsp2 from the two PRRSV isolates was −25. In contrast, LDV and EAV have net charges in this region of +5 and +11, respectively.

The Nsp3 protein had a very hydrophobic carboxy terminus and contained a predicted transmembrane domain between residues 1584 and 1729.

The Nsp4 protein contained a chymotrypsin-like (3C-like) serine protease domain (Dougherty & Semler, 1993) between residues 1844 and 1850 (Fig. 2). The putative catalytic residues were His$_{1848}$, Asp$_{1873}$ and Ser$_{1929}$ as described for other arteriviruses (Godeny et al., 1993; Snijder et al., 1996). The predicted 3C-like serine protease cleavage sites were located at residues 1809 (E/G), 2013 (E/G), 2183 (E/G) and 2458 (E/A), which is consistent with the cleavage sites described for LDV, LV and EAV (Godeny et al., 1993; Snijder et al., 1996; Wassenaar et al., 1997).

The Nsp5 protein contained an additional 3C-like serine protease cleavage site at position 2183 (E/G), giving rise to the possibility that Nsp5 is cleaved into two proteins. This site is also present in the other arteriviruses but has not been described previously.

**ORF1a–ORF1b junction region**

Two RNA structures, a heptanucleotide ‘slippery’ sequence and a pseudo-knot, formed a −1 frameshift region similar to that described for LV and other arteriviruses (den Boon et al., 1991; Godeny et al., 1993; Meulenberg et al., 1993a). Overall, PRRSV 16244B showed 98% nucleotide identity with LV in the translational frameshift region (Table 1).

The putative heptanucleotide ‘slippery’ sequence (5’UUUUAAAC 3’) was present in PRRSV 16244B from nucleotides 7689 to 7696, four nucleotides upstream of the UAG stop codon of ORF1a (Fig. 3). The predicted H-type pseudo-knot stem–loop 1, with 13 predicted base-pair bonds, followed the stop codon. The 16244B stem–loop 1 was predicted to be slightly weaker (−82.5 kJ/mol) than the LV stem–loop 1, due to two C→U nucleotide changes at nucleotides 7727 and 7737 (only ten G–C base-pairs). LV has a stronger stem–loop 1 (−91.3 kJ/mol) with 12 predicted base-pair bonds (11 G–C base-pairs). Loop 1 of 16244B was formed by nucleotides 7714–7724 and contributed to the pseudo-knot structure present between nucleotides 7714 and 7759 (Abrahams et al., 1990). Loop 2 (14 nucleotides long) was between positions 7738 and 7751 (one nucleotide shorter than in LV), with a predicted stem–loop 2 of seven base-pair bonds (Fig. 3).

**Analysis of ORF1b**

The PRRSV 16244B ORF1b encoded a polyprotein that was 1463 amino acids in length, which was the same size as LV ORF1b (Table 1). Although the overall amino acid identity between the ORF1b polyproteins of the two PRRSV isolates was 68.3%, with 80% conservative substitutions (64.6% nucleotide identity), the carboxy-terminal 150 residues shared only 42% amino acid identity. The three cleavage sites for the 3C-like protease predicted by van Dinten et al. (1996) in LV and other arteriviruses were also present in 16244B, at residues 646 (E/G), 1087 (E/G) and 1310 (E/G). No putative protease domain was present in ORF1b and, as in EAV (van Dinten et al., 1996), ORF1b is probably cleaved by ORF1a proteases into four proteins, RdRp and putative cleavage products CP2, CP3 and CP4. There are major differences in CP4 among the arteriviruses and between these PRRSV isolates (Table 2, Fig. 2).

RdRp had a core amino acid triplet SDD at residues 511 to 513, as described for LV and other arteriviruses (Meulenberg et al., 1993a; den Boon et al., 1991; Godeny et al., 1993). The four
conserved motifs found in RNA polymerases of positive-stranded RNA viruses (Koonin, 1991) were present in PRRSV 16244B (residues 373–386, 394–411, 443–468 and 506–517). The putative CP2 protein contained the zinc-finger and helicase domains (Fig. 2). The predicted zinc-finger domain was located at residue 653–699. The twelve cysteine and histidine residues of this domain found in other arteriviruses were also conserved in PRRSV 16244B. The predicted RNA helicase domain was located on the carboxy-terminal side of the RdRp, as described for LV and other arteriviruses (Godeny et al., 1993). The six conserved helicase motifs (Gorbalenya et al., 1989) were present in ORF1b of PRRSV 16244B between residues 790 and 1013. The putative nucleoside triphosphate-binding domain was located in the 1(A) motif of the predicted RNA helicase, spanning amino acids 790–807.

The putative CP3 protein of 16244B contained the highly conserved domain of unknown function found in other arteriviruses (Meulenberg et al., 1993; Godeny et al., 1993; den Boon et al., 1991). This protein currently has no known function.

The characteristics of ORFs 2–7 are summarized and compared to both the LV and VR-2332 isolates in Table 3. ORF2 (glycosylated envelope protein GP2) had two predicted glycosylation sites, at positions 178 and 184. ORF3 (envelope protein GP3) had seven predicted glycosylation sites, at positions 29, 42, 50, 131, 152, 160 and 195. The four glycosylation sites predicted in ORF4 (GP4) and ORF5 (glycosylated envelope protein GP5) were identified at positions 37, 84, 120 and 130 (ORF4) and 30, 33, 44 and 51 (ORF5). Also, these two ORFs were separated by 11 nucleotides, while they overlap by one nucleotide in LV (Meulenberg et al., 1993). This peculiarity has also been observed in other North American isolates (Kapur et al., 1996; Meng et al., 1995; Murtaugh et al., 1995). ORF5 also had a predicted signal sequence and three predicted transmembrane domains at residues 62–83, 90–106 and 113–130. ORF6 (membrane protein M) had one predicted glycosylation site and three predicted membrane-spanning regions from the virus isolates belonging to the North American group, which were 95% identical to 16244B (GenBank accession nos AF030306 and AF035409; T. Y. Kuo, L. F. Lee, K. H. Yang, J. T. Chen & S. S. Lai, unpublished results). The putative CP4 protein had a hydrophilic domain between amino acids 1391 and 1441. High variability in hydropathy plots is observed among CP4 proteins of arteriviruses in this region and this protein currently has no known function.

### Analysis of ORF2 to ORF7

ORFs 2–7, which are located at the 3' end (3 kb) of PRRSV 16244B, code for the viral structural proteins (Table 1). The translation products of ORFs 2–6 had characteristics of membrane-associated proteins. They contained amino- and carboxy-terminal hydrophobic regions that might function as a signal peptide sequence and a membrane anchor, respectively, as described for LV (Meulenberg et al., 1993) and other arteriviruses (Godeny et al., 1993; den Boon et al., 1991).

The characteristics of ORFs 2–7 are summarized and compared to both the LV and VR-2332 isolates in Table 3. ORF2 (glycosylated envelope protein GP2) had two predicted glycosylation sites, at positions 178 and 184. ORF3 (envelope protein GP3) had seven predicted glycosylation sites, at positions 29, 42, 50, 131, 152, 160 and 195. The four glycosylation sites predicted in ORF4 (GP4) and ORF5 (glycosylated envelope protein GP5) were identified at positions 37, 84, 120 and 130 (ORF4) and 30, 33, 44 and 51 (ORF5). Also, these two ORFs were separated by 11 nucleotides, while they overlap by one nucleotide in LV (Meulenberg et al., 1993). This peculiarity has also been observed in other North American isolates (Kapur et al., 1996; Meng et al., 1995; Murtaugh et al., 1995). ORF5 also had a predicted signal sequence and three predicted transmembrane domains at residues 62–83, 90–106 and 113–130. ORF6 (membrane protein M) had one predicted glycosylation site and three predicted membrane-spanning regions from the

### Table 3. Comparison of ORFs 2–7 from PRRSV isolates 16244B, VR-2332 and LV

<table>
<thead>
<tr>
<th>ORF</th>
<th>Genome position (nucleotides)</th>
<th>Length (amino acids)</th>
<th>Identity to 16244B (%)</th>
<th>No. of glycosylation sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>12073–12843</td>
<td>256</td>
<td>99</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>12696–13460</td>
<td>254</td>
<td>98</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>13241–13777</td>
<td>266</td>
<td>99</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>13788–14390</td>
<td>200</td>
<td>99</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>14375–14899</td>
<td>124</td>
<td>99</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>14889–15260</td>
<td>123</td>
<td>100</td>
<td>1</td>
</tr>
</tbody>
</table>

Positions of ORFs within the genome of isolate 16244B are shown. Information for the genomes of VR-2332 and LV was taken from Murtaugh et al. (1995) and Meulenberg et al. (1993).
immunological pressure can explain totally the differences observed in Nsp2 from the two PRRSV isolates. Furthermore, based on phylogenetic relationships, we have no evidence for the occurrence of recombination (de Vries et al., 1997) with other arteriviruses (data not shown) and the unique Nsp2 region in all of the arteriviruses matched nothing else in the current sequence databases.

Nsp2, the most variable protein within arteriviruses, is thought to be involved in species-specific functions (de Vries et al., 1997). The low degree of identity and the size differences between Nsp2 of the two PRRSV isolates suggests that variability within this protein may have other than species-specific functions. Whether this involves biological differences observed between virus isolates, such as the degree of pathogenicity in the swine host (Halbur et al., 1995), remains to be determined.

We cannot rule out the possibility that some degree of virus selection took place during the three passages in the MARC-145 cell line. However, other authors have reported nucleotide changes of only 3% and 1% for ORFs 5 and 7 after six passages in MA-104 cells (Suarez et al., 1994). Even a similar degree of change in the Nsp genes, e.g. Nsp2, during three passages would not account for the large differences observed between the North American and European isolates.

Notably, the putative ORF1b protein CP4 shared only 42% amino acid identity with LV, less than 25% identity over the amino-terminal 80 residues with SHFV and no identity at all with either LDV or EAV. The unique region of CP4 matched nothing else in the current genomic databases. The lack of CP4 similarity between PRRSV and LDV was surprising, since LDV is considered to be a closely related arterivirus (Meulenberg et al., 1993a). Interestingly, CP4 was very different in the two PRRSV strains, while the two LDV strains (P and C) had a high level of identity in CP4 (89%; Palmer et al., 1995), which could be explained by their being two prototypic PRRSV genomes.

Together, these genomic differences confirm that North American and European PRRSV isolates represent distinct virus genomic groups. Further analysis of PRRSV NspS (ORF1a and ORF1b) may prove useful for grouping PRRSV isolates.

R. Allende and T. L. Lewis contributed equally to this work. We thank Judith A. Galeota for excellent technical assistance, Claudio Afonso, John Neilan, Edan Tulman and Laszlo Zsak for helpful comments and research assistance. Luis Rodriguez for manuscript review comments and Phil Green, Brent Ewing and David Gordon at the University of Washington for sharing their programs (Phred, Phrap and Consed, available at http://www.genome.washington.edu). R.A. was supported by project 96-35204 (NRICGP-USDA) and A.A. by project no. 1715 (National Pork Producers Council).

References

Genomic differences between North American PRRSV isolate 16244B and European isolate LV

Here we have analysed the entire genome of a pathogenic North American PRRSV isolate, 16244B, and compared it with the European isolate LV. The North American and European PRRSV strains had the same genome organization but major differences were found in the 5’ and 3’ NCRs and in the non-structural coding region of ORF1a and ORF1b (Nsp2 and CP4), confirming the existence of distinct North American and European PRRSV genomic groups.

Surprisingly, the Nsp2 protein shared only 32% amino acid identity with LV and was 120 residues longer. In contrast, Nsp2 of LDV strains P and C were more than 80% identical (Palmer et al., 1995). Neither error-prone replication nor

outer–inner membrane faces at positions 18–34, 41–60 and 70–89 (Jones et al., 1994). A similar pattern is present in other North American isolates (Kapur et al., 1996; Murtough et al., 1995) as well as in LV and other arteriviruses (Meulenberg et al., 1993a; den Boon et al., 1991; Godeny et al., 1993). ORF7 (nucleocapsid protein N) also had only one predicted glyco- sylation site. The amino-terminal half of the polypeptide was composed of 25% basic amino acids (arginine, histidine and lysine), which is also the case for ORF7 of LV (Murtough et al., 1995; Meulenberg et al., 1993a; den Boon et al., 1991; Godeny et al., 1990). A comparison of ORF5 and ORF7 from 16 European and three North American PRRSV isolates suggested distinctive European and North American genotypes (Suarez et al., 1996). A similar interpretation was also made by Le Gall et al. (1998) upon comparison of ORF7 from 15 European and 13 North American PRRSV isolates.

3’ NCR and poly(A) tail

The 3’ NCR of PRRSV 16244B was 150 nucleotides in length, 36 nucleotides longer than that of LV (Meulenberg et al., 1993a), and shared 76% nucleotide identity (Table 1). There was a 26 nucleotide insertion and a one nucleotide deletion when compared with LV. Also, 16244B used a UGA stop codon nine nucleotides 5’ of the LV stop codon (Fig. 1b). This same insertion, deletion and different stop codon position have also been found in the North American PRRSV isolate VR-2332 and ten other European PRRSV isolates from the Netherlands (Conzelmann et al., 1993), Spain (Plana Duran et al., 1997) and UK (Drew et al., 1997).

The poly(A) tail, varying in length from 14 to 23 nucleotides (mean length 17 nucleotides) among the clones analysed, was similar to LV (Meulenberg et al., 1993a). The eight nucleotides immediately upstream of the poly(A) tail (Fig. 1b) were identical to LV. There is a high degree of conservation in these nucleotides adjacent to the poly(A) tail among arteriviruses, which suggests that this motif may be important for either RdRp recognition or replication initiation or both (Godeny et al., 1993).

Genomic differences between North American PRRSV isolate 16244B and European isolate LV

Here we have analysed the entire genome of a pathogenic North American PRRSV isolate, 16244B, and compared it with the European isolate LV. The North American and European PRRSV strains had the same genome organization but major differences were found in the 5’ and 3’ NCRs and in the non-structural coding region of ORF1a and ORF1b (Nsp2 and CP4), confirming the existence of distinct North American and European PRRSV genomic groups.

Surprisingly, the Nsp2 protein shared only 32% amino acid identity with LV and was 120 residues longer. In contrast, Nsp2 of LDV strains P and C were more than 80% identical (Palmer et al., 1995). Neither error-prone replication nor


Received 1 July 1998; Accepted 22 October 1998