RESEARCH REPORT
SWINE HEALTH

Title: Development of next generation sequencing methodology for full genome characterization of porcine reproductive and respiratory syndrome virus (PRRSV) from oral fluids and nasal swabs - **NPB#14-204**

Investigator: Dr. Ben Hause: P.I.

Institution: Kansas State University

Co-Investigators: Dr. Richard Hesse, Dr. Jianfa Bai: co-P.I.

Date Submitted: September 23, 2015

Industry Summary:
Porcine reproductive and respiratory syndrome virus (PRRSV) is one of the most significant swine diseases with near worldwide distribution. In the U.S., open reading frame 5 (ORF5) is commonly sequenced to investigate viral epidemiology. While glycoprotein 5 (GP5) is the major protein on the surface of the virion, the minor glycoproteins GP2a, GP3 and GP4 exhibit similar genetic diversity and are responsible for receptor binding and are important antigens for the immune response. Additionally, the non-structural protein 2 (nsp2) plays key role in viral pathogenicity and deletions in nsp2 have been associated with strains with increased virulence.

The goal of this project was to develop next generation (metagenomic) sequencing methodology to enable full PRRSV genome sequencing directly from clinical samples. The current standard GP5 sequencing for epidemiological investigations takes into account only <5% of the genome and advances in sequencing technology now make it cost-effective to determine a comprehensive picture of PRRSV genetics. Metagenomic sequencing of PRRSV-positive nasal swabs and oral fluids were able to detect PRRSV however read coverage was insufficient to determine complete genomes. In contrast, metagenomic sequencing of PRRSV-positive sera was successful in determining complete PRRSV genomes. Metagenomic sequencing was performed on a collection of 182 PRRSV-positive sera submitted to veterinary diagnostic laboratories. Complete PRRSV genomes were determined from 66 of these samples. Analysis of the viral structural proteins found 4-7 lineages currently circulating in the U.S. This study identified more diversity in the PRRSV structural proteins than previously recognized, possibly due to direct sequencing of clinical samples as opposed to sequencing viruses isolated in cell culture.

The added benefit to metagenomic sequencing of clinical samples is the ability to detect all viruses present in the sample in an unbiased manner. A large number of the serum samples contained porcine parvovirus 2 and porcine parvovirus 3, 4, and 5 were also detected. Importantly, we identified porcine parvovirus 6 (PPV6) for

These research results were submitted in fulfillment of checkoff-funded research projects. This report is published directly as submitted by the project's principal investigator. This report has not been peer-reviewed.

For more information contact:
National Pork Board • PO Box 9114 • Des Moines, IA 50306 USA • 800-456-7675 • Fax: 515-223-2646 • pork.org

the first time in the U.S., in 13.2% of the samples. A novel pestivirus, named atypical porcine pestivirus (APPV), was also identified in 2.1% of the samples. Genetic analysis of APPV found that is highly diverged from livestock pestiviruses bovine viral diarrhea, classical swine fever and border disease virus and is more closely related to a recently partially sequenced bat pestivirus. No information in available on the ability of PPV6 and APPV to cause disease.

The metagenomic sequencing methodology developed in this NPB grant should be of great value to swine producers and veterinarians. The Kansas State Veterinary Diagnostic Laboratory is now offering this test for $300 per sample. Given the ease of sample collection and relative low cost for metagenomic sequencing ($300 currently, prices expected to decrease), metagenomic sequencing will undoubtedly become more commonplace in veterinary diagnostics as producers and veterinarians are routinely paying ~$150 for GP5 sequencing alone. Determination of complete PRRSV genomes will greatly improve our understanding of PRRSV genetics, epidemiology and evolution and will enable more efficacious vaccine development and eradication efforts. The added ability to survey all other viruses present in a sample concurrently significantly improves our diagnostic testing capabilities and enables more comprehensive understanding of disease complexes which will lead to improved control measures.

Keywords: porcine reproductive and respiratory syndrome, sequencing, GP5, metagenomic, evolution

Scientific Abstract:

Porcine reproductive and respiratory syndrome virus causes one of the most significant swine diseases with near worldwide distribution. In the U.S., open reading frame 5 (ORF5) is commonly sequenced to investigate viral epidemiology. While glycoprotein 5 (GP5) is the major protein on the surface of the virion found as a heterodimer with the membrane protein, the minor glycoproteins GP2a, GP3 and GP4 exhibit similar genetic diversity and form a heterocomplex responsible for receptor binding. To increase our understanding of PRRSV diversity and evolution, 66 genome sequences were determined directly from serum samples using viral metagenomic methodology. Phylogenetic analysis identified five, four, seven, seven and six well-supported clades for ORF2a, ORF3, ORF4, ORF5 and ORF6, respectively, which encompassed nearly all strains. Intraclade genetic distance was approximately 0.00-0.12 while interclade distances were 0.10-0.21. Similar genetic diversity was observed for ORF2a, ORF3, ORF4 and ORF5 while ORF6 was more conserved. Topological incongruences were noted in the 3' end of the genome (ORF2a to 3'-terminus) using the genetic analysis recombination detection algorithm with five breakpoints identified with statistical significance ($P<0.05$). Thirteen gene combinations with respect to ORF2a, ORF3, ORF4, ORF5 and ORF6 clade composition were identified. Recombination detection program identified two representatives from these gene combinations as recombinants with high confidence. This study identified more diversity in the PRRSV structural proteins than previously recognized, possibly due to direct sequencing of clinical samples as opposed to selection for growth *in vitro*.

Introduction:

Despite research and intervention for greater than 20 years, porcine reproductive and respiratory syndrome (PRRS) remains one of the most economically significant diseases affecting the U.S. swine industry. Both inactivated and modified live vaccines have long been used in an attempt to control PRRS however often are not effective. The vast amount of genetic variability between PRRS viruses (PRRSV) and the consequent poor match between vaccine strains and field viruses negatively impacts vaccine efficacy. The viral RNA dependent RNA polymerase of PRRSV has a high error rate, leading to rapid viral evolution via accumulation of point mutations. Additionally, multiple PRRSV co-infecting a pig can recombine, leading to exchanges of large sections of the genome. Coupled with either naïve or animals vaccinated with divergent PRRSV that fail to

generate neutralizing antibodies to the challenge virus, these two mechanisms contribute to genetic drift and evolution (Chang et al., 2002; Martin-Valls et al., 2014).

Early after its discovery, the GP5 protein of PRRSV became the focus for PRRS intervention. GP5 is the major membrane glycoprotein of PRRSV. It possesses an epitope that is targeted by virus neutralizing antibodies and consequently is highly variable (Kim et al., 2013). Genetic characterization of PRRSV has almost exclusively focused on GP5, with early restriction fragment length polymorphism (RFLP) methods being replaced with the now conventional GP5 gene sequencing. Despite more than a decade and thousands of GP5 sequences, this information has failed to significantly improve our ability to control PRRSV.

Recent work studying the evolution of PRRSV suggested that immunity to PRRSV is multigenic (Nguyen et al., 2013). Pigs infected with PRRSV elicited antibodies that specifically bound to recombinant polypeptides containing PRRSV ectodomain neutralizing epitopes (a combination of GP5 and the membrane protein M) but their titer did not correlate with neutralizing antibody response and did not neutralize PRRSV infectivity (Li and Murtaugh, 2012). Virus neutralizing antibodies often bind to viral proteins required for cell receptor binding, thus preventing virus from attaching to host cells and initiating infection. The GP5-M heterodimer, which is the dominant peptide on the PRRSV outer membrane, was long thought to be the PRRSV cell receptor binding protein owing to biochemical studies that showed its interactions with porcine sialoadhesin and porcine alveolar macrophages. However recent work suggested a key role for the minor PRRSV glycoproteins GP2a, GP3 and GP4 for receptor binding and consequently genetic variability leading to escape from immunity (Vu et al., 2011). Replacement of the PRRSV minor glycoproteins in a PRRSV infectious clone with those of the related equine arteritis virus (EAV) generated a chimeric virus displaying broad cell tropism characteristic of EAV (Tian et al., 2012). Similarly, a chimeric EAV infectious clone bearing PRRSV N-terminal GP5 and M ectodomains retained its ability to infect EAV permissive cell lines and did not acquire the ability to infect PRRSV permissive cells (Lu et al., 2012). Besides playing key roles in cell binding, neutralizing antibody epitopes have also been identified in the PRRSV minor glycoproteins (Costers et al., 2010).

In addition to the glycoproteins, the non-structural proteins (nsps) of PRRSV play a key role in its pathogenicity. Comparisons of four PRRSv with differing virulence found genetic diversity in both the structural and nsps (Brockmeier et al., 2012). Highly pathogenic PRRSV first identified in Asia in the late 2000's have characteristic large deletions in non-structural protein 2 (nsp2) and have been postulated to be one of the viral genetic determinants contributing to PRRSV pathogenicity (Choi et al., 2014). A targeted deletion of 30 amino acids in nsp2 in a low-virulence PRRSV reverse genetics system however failed to recapitulate the high virulence phenotype, suggesting a multigenic origin (Zhou et al., 2009). Altogether, recent research points towards a critical role of the PRRS minor glycoproteins and nsps for PRRSV pathogenicity and immunity however nearly all the U.S. swine surveillance has been focused on GP5. Clearly more comprehensive genetic surveillance is required for our understanding of how PRRSV genotype correlates to pathogenicity. Cost-effective sequencing on easily obtainable samples will enable this testing.

Objectives:

Objective 1: Develop next generation sequencing (NGS) methodology for PRRS genome sequencing from oral fluids and nasal swabs. Existing NGS methods will be optimized for PRRSV-positive oral fluids and nasal swabs. Methodology will be optimized to maximize sensitivity while maintaining sufficient coverage for complete genome sequencing.

Objective 2: Apply developed NGS methodology on a diverse collection of 100 positive nasal swabs and 100 positive oral fluids. PRRSV-positive oral fluids and nasal swabs collected from a wide geographic area with well characterized clinical histories will be subjected to the NGS protocol to generate full PRRSV genome sequences and identify other viral co-factors present in the samples.

Materials & Methods:

Samples

Metagenomic sequencing was performed on 182 swine serum samples that were real time reverse transcription PCR (RT-PCR) positive for PRRSV. The serum samples were submitted to Kansas State Veterinary Diagnostic Laboratory (KSVLD), Iowa State University Veterinary Diagnostic Laboratory or the South Dakota State University Animal Disease Research and Diagnostic Laboratory for PRRSV RT-PCR. The samples originated from thirteen states (Iowa [n=35], Minnesota [n=39], South Dakota [n=2], Texas [n=9], North Carolina [n=18], Nebraska [n=14], Kansas [n=22], Oklahoma [n=1], Illinois [n=2], Indiana [n=1], Missouri [n=1], Arizona [n=2], and Colorado [n=4]), Mexico (n=4) and unknown (n=28).

Sample Preparation and Sequencing

Serum samples were centrifuged for five minutes (min) at 7500 rpm to remove any large particulates. Samples were prepared for sequencing as previously described (Hause et al., 2015). In brief, one hundred and eighty microliters (µl) of clarified serum was treated with nucleases and incubated at 37°C for 90 min. Nucleic acid was extracted using the Qiagen MinElute Viral Spin Kit (Qiagen) according to manufacturer's directions and eluted in 25 µl of nuclease-free water (Invitrogen). Total RNA was converted to cDNA and amplified using sequence-independent-single-primer amplification. RNA was converted to cDNA using the Superscript III First Strand Synthesis Kit (Invitrogen) and primer consisting of a 20 nt known sequence at the 5' end and a random hexamer at the 3' end. The second strand was made using the Sequenase 2.0 polymerase. Double-stranded cDNA was amplified using a Takara Taq polymerase and a primer consisting of only the 20 nt sequence of the previous primer. Native double-stranded DNA and amplified cDNA was quantified using the Qubit high sensitivity DNA reagent kit (Invitrogen) and diluted to 0.2 nanograms (ng) per µl for library preparation. Libraries were prepared using the Illumina Nextera XT DNA Sample Preparation Kit according to manufacturer's instructions (Illumina). Libraries were sequenced using the Illumina MiSeq and v2 reagents. Paired end reads were demultiplexed and fastq files were created with MiSeq Reporter software (Illumina).

Sequencing Read Assembly and Annotation

Paired end reads for each sample were imported into CLC Genomics Workbench 7.0 software (Qiagen). For each sample, reads were assembled into contigs using the De Novo Assembly function with default parameters. Consensus sequences for assembled contigs were compared against the non-redundant nucleotide database at NCBI using the BLASTn algorithm. The GenBank accession number with the best E-value was used for reference-based assembly using the entire set of paired reads. Consensus sequences were extracted from the reference-based assemblies, open reading frames were determined in CLC Genomics Workbench, and sequences for ORFs 2a to 6 of each genome were extracted, and used in phylogenetic analyses and sequence comparisons.

Phylogenetic Analysis

Nucleotide sequences for ORFs 2 to 6 were aligned in ClustalW under default parameters. Each alignment was then evaluated for the best-fit model of evolution in ModelGenerator v0.85 (Keane et al 2006) against 96 different models of amino acid substitution. The best-fit model was identified using the Bayesian Information Criterion as implemented in ModelGenerator. Phylogenetic trees were reconstructed in MEGA6 under the maximum likelihood approach using the best-fit model of nucleotide substitution. Nodal support was evaluated by 1000 bootstrap pseudoreplicates.

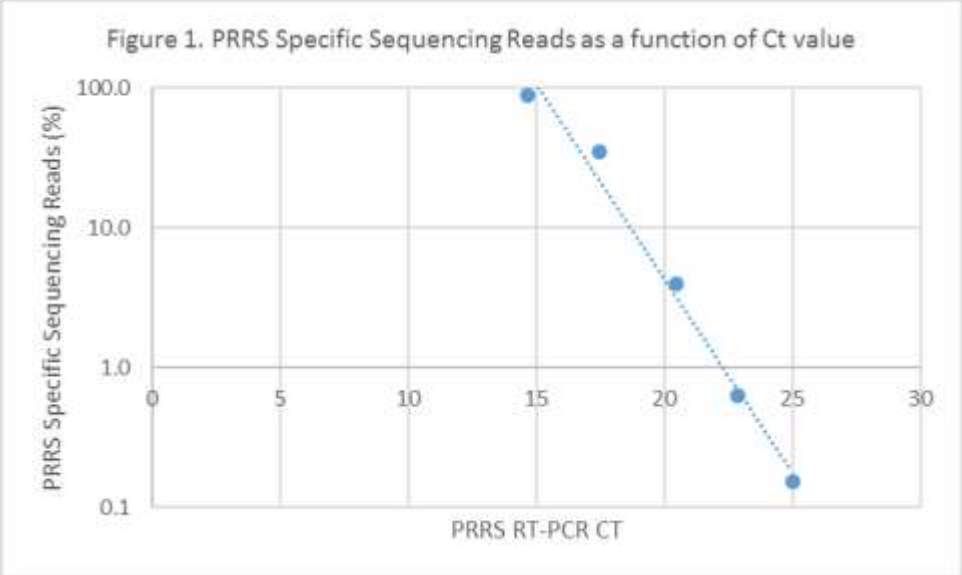
Pairwise Sequence Comparisons

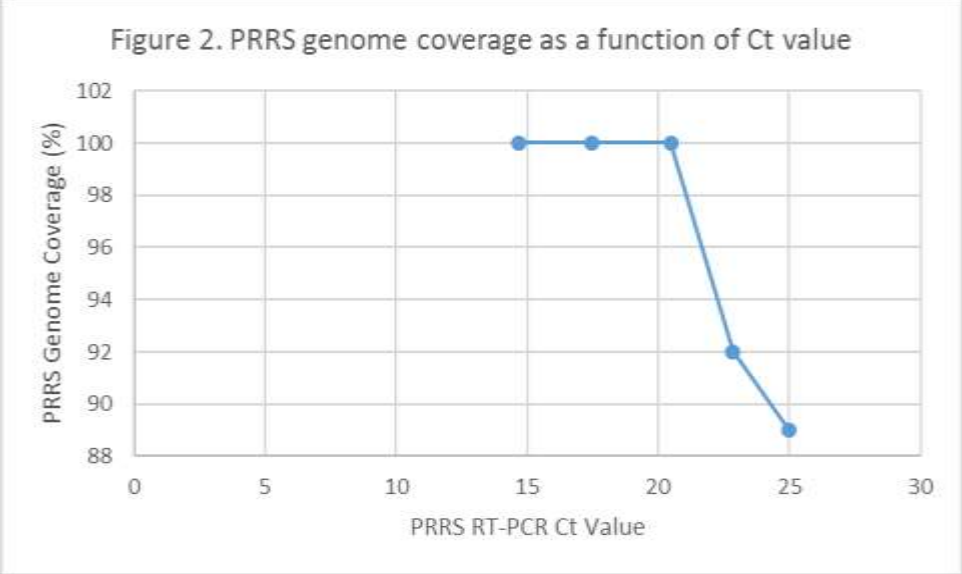
Nucleotide alignments for ORFs 2 to 6 were imported into CLC Genomics Workbench 7.0 (Qiagen) for pairwise comparison of percent identity and number of nucleotide differences. Nucleotide alignments were imported into MEGA6 to calculate the maximum composite likelihood mean within clade and mean between clade distances using the K2 + G model of nucleotide substitution and 100 bootstrap pseudoreplicates.

Results:

Metagenomic sequencing of spiked serum samples

We first tested the NGS methodology on PRRSV-negative porcine serum spiked with a 10-fold dilution series of high-titer, cell culture derived PRRSV. Real time reverse transcription PCR was performed on the samples to determine sample Ct values. NGS libraries were prepared from each sample and sequenced following the developed methodology. As expected, there was a positive correlation with PRRSV specific sequencing reads and viral titer (inversely related to Ct value, Figure 1). Complete or near complete genomes (single contig >14,000 bp covering the coding region of the genome) were determined for samples with a Ct values of 20 or less (Figure 2).





Metagenomic Sequencing of PRRSV-positive oral fluids and nasal swabs

Having satisfactorily developed NGS methodology that is capable of determining full PRRSV genomes, we performed sequencing on 20 nasal swabs with a PRRSV Ct range of 25-35. While PRRSV specific sequencing reads were identified for all samples, maximal genome coverage was 33%. Comparison of Ct values available for all nasal swabs found a minimum Ct value of 25. From Figures 1 and 2, the method's ability to sequence PRRSV decreases dramatically for samples with Ct values greater than 20. We concluded that nasal swabs contain insufficient amounts of PRRSV to determine complete genome sequences. In addition, large numbers of bacterial and non-PRRSV viral sequences were apparent in the samples which detracted from reads directed towards PRRSV. Numerous other viruses were detected in nasal swabs, including bocavirus, hemagglutinating encephalomyelitis virus, parvovirus, porcine cytomegalovirus, sapelovirus, transmissible gastroenteritis virus, picobirnavirus, various small circular DNA viruses, kobuvirus, astrovirus,

enterovirus, rotavirus and adeno-associated virus, suggesting that nasal swabs are suitable specimens for viral surveys using the developed metagenomic sequencing methodology.

Likewise, we performed sequencing on 20 oral fluids with a PRRSV Ct range of 20-29. Results were similar to those observed for nasal swabs. We suspect a combination of low PRRSV titer, viral degradation and competition for sequencing reads by bacteria and other viruses limited our ability to determine PRRSV genome sequences. In addition, large numbers of bacterial and non-PRRSV viral sequences were apparent in the samples which detracted from reads directed towards PRRSV. Numerous other viruses were detected in nasal swabs, including bocavirus, parvovirus, sapovirus, picobirnavirus, various small circular DNA viruses, kobuvirus, astrovirus, enterovirus, rotavirus, adeno-associated virus, porcine adenovirus 5, orthoreovirus, posavirus and porcine parainfluenza virus 1, suggesting that nasal swabs are suitable specimens for viral surveys using metagenomic sequencing.

Given unsatisfactory assay performance on nasal swabs and oral fluids we performed sequencing on PRRSV positive sera, another easily collected diagnostic specimen. Twelve of the 20 samples yielded complete PRRSV genomes. PRRSV Ct values for the samples with full genomes determined ranged from 14-25. For samples where we were unable to determine full PRRSV genomes, sequencing reads were dominated by porcine parvovirus 2 or 6.

As good results were obtained from serum samples, we modified our objectives to focus on sera and proceeded to sequence 182 total PRRSV-positive sera samples. Sixty-six new near full length PRRSV genomes were assembled from viral metagenomic sequencing of 182 serum samples that were PRRSV positive by RT-PCR. Genomic sequences for 16 additional PRRSV strains, including vaccine strains and well known reference strains with full genome sequences representing different lineages of PRRSV based upon previous evolutionary analysis of ORF5 (Shi et al., 2010), were downloaded from GenBank and included in phylogenetic analyses.

Phylogenetic reconstructions based upon the nucleotide (nt) alignments of ORFs 2a – 6 were scrutinized for groups of sequences that were found in well-supported (bootstrap value greater than or equal to 70) relationships across the different ORFs. Clades were defined as the largest group of sequences that were joined at a well-supported node. Trees for each ORF were all found to contain a similar pattern wherein the backbone of the tree, which describes how groups of taxa or strains are related to each other, was not well defined and only had low nodal support. However, several well-supported clades were evident in the phylogenetic reconstructions of each ORF and are described below.

ORF2a

Phylogenetic analysis of ORF2a sequences identified five well-supported clades (Figure 3). All sequences determined directly from clinical samples were included in these five clades with the exception of strain 104194. A number of reference viruses also failed to cluster within the five

clades (SDSU73, CH-1a, NADC-8, PrimePac, FJ-1 and MD-001). Clades ORF2.1 and ORF2.2 contained vaccine strains InglevacATP and InglevacMLV, respectively, and included a number of sequences determined from clinical samples with high similarity to vaccine strains. Clade ORF2.3 included 20 sequences with only approximately 90% identity to previously sequenced PRRSV as determined by BLASTN. Clade ORF2.4 contained reference MN184 reference viruses A, B and C and 12 additional sequences. Twenty-three strains of PRRSV, along with the reference NADC30, encompassed ORF2.5. Mean intraclade distances were less than 0.09. Mean interclade distances were 0.127-0.176 with the exception of 0.062 between ORF2.1 and ORF2.2.

ORF3

Four well-defined clades were identified for ORF3 (Figure 3). Similar to ORF2.1 and ORF2.2, ORF3.1 and ORF3.2 contained strains with high similarity to vaccine strains InglevacATP and InglevacMLV, respectively. Clade ORF3.1 also included reference viruses CH-1a and SDSU73 while clade ORF3.2 contained NADC-8, VR2334 and PRRSV01. Clade ORF3.3 contained 39 strains of PRRSV along with the reference NADC30. Clade ORF3.4 included 16 strains of PRRSV and reference viruses MN184A-C. Due to the large cluster sizes, mean intraclade distances were 0.01-0.122. Interclade distances were 0.095-0.210. As observed for ORF2a, the ORF3 gene for strain 104194 did not cluster with other strains.

ORF4

Phylogenetic analysis identified seven clades for ORF4 (Figure 3). ORF4 intraclade distances (0.01-0.054) were smaller than observed for ORF2a and ORF3 while interclade distances were similar (0.099-0.191). The largest clade was ORF4.3 that included 31 strains and the reference NADC30. Clade ORF4.4 included reference viruses MN184A-C and five strains with 94% identity to MN184. Clades ORF4.5 and ORF4.6 contained 4 and 7 strains, respectively, and were approximately 90-92% identical to MN184. Clade ORF4.7 consisted of eight strains and was most similar to reference virus SDSU73 (90% identity). As seen with ORF2 and ORF3, clades ORF4.1 and ORF4.2 consisted of the same field and vaccine strains. Strain 104194 was the only field strain that did not cluster in any clade.

ORF5

Similar to ORF4, seven clades were identified for ORF5 (Figure 4). Unlike ORF2a, ORF3 and ORF4, a single clade (ORF5.1) contained both vaccine strains InglevacATP and InglevacMLV as well as reference viruses FJ-1, MD-001, PRRSV01, VR2332, NADC-8, PrimePac, SDSU73 and CH-1a. In addition to the reference viruses, the same 10 strains of PRRSV that clustered with the vaccine strains for ORF2a to 4 were included in ORF5.1. Clade ORF5.2 was the largest clade (24 strains) and, similar to ORF5.1, showed a high amount of intraclade diversity. Clade ORF5.3 only contained reference viruses MN184A-C. Clade ORF5.4 consisted of five strains with 91% identity to MN184. Reference virus NADC30 clustered with clade ORF5.5 along with 13 field strains. Clade ORF5.6 was comprised of five strains with 97% identity to a number of viruses isolated in the Midwest in 2007. The eight strains comprising clade ORF5.7 had the highest identity to a 2013 Indiana isolate (96% identity). ORF5 intraclade diversity was 0.001-0.093 and interclade diversity was 0.118-0.182.

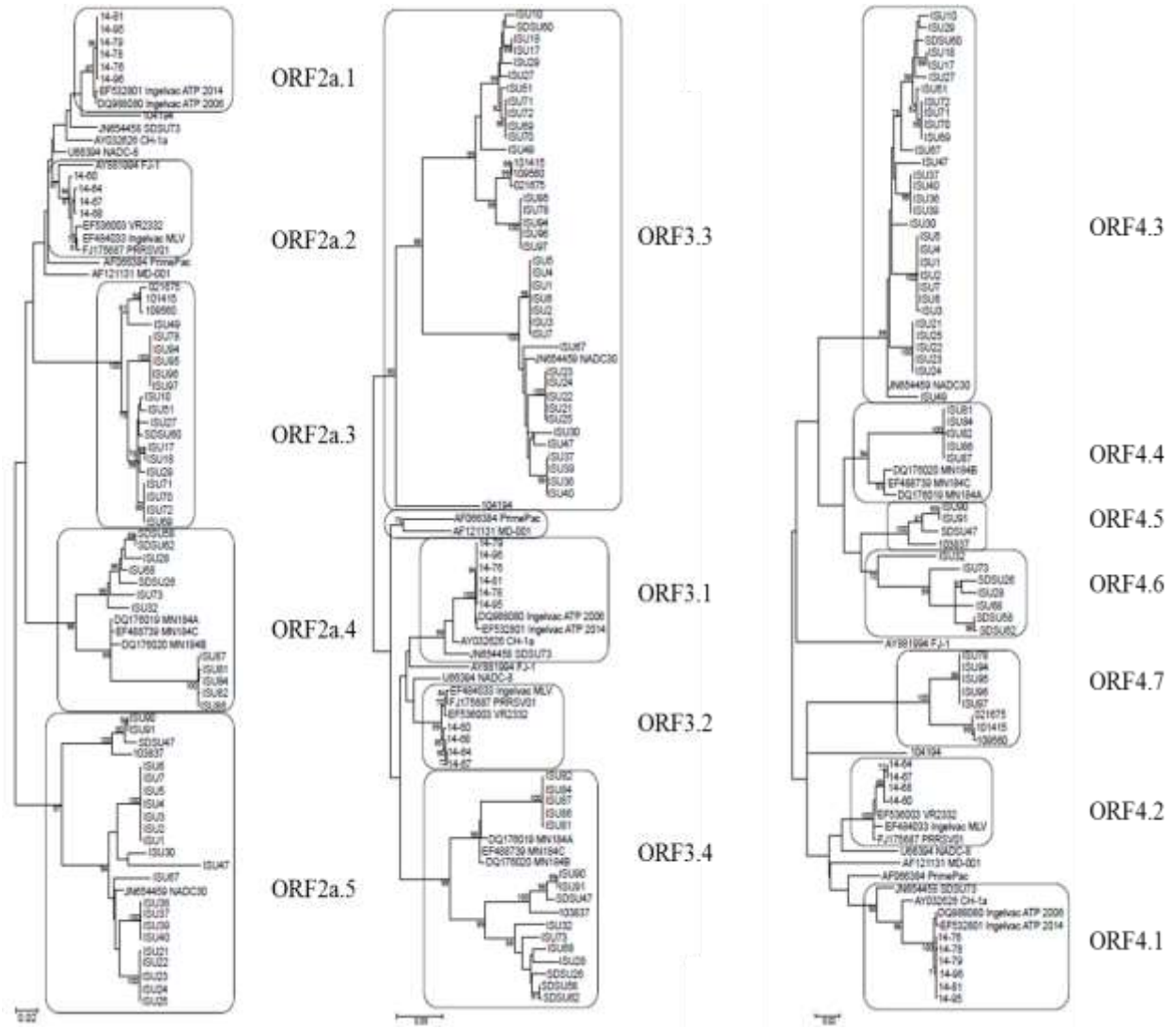


Figure 3. Phylogenetic analysis of ORF2a, ORF3 and ORF4 sequences determined by metagenomic sequencing of PRRSV-positive porcine sera. Maximum likelihood analysis in combination with 1000 bootstrap replicates was used to derive trees based on nucleotide sequences. Clade designations are indicated by boxes and were chosen to maximize strain inclusion based on topological support >70.

ORF6

Phylogenetic analysis of the M gene found six clades (Figure 4). As with the ORF2a.1/ORF3.1/ORF4.1 and ORF2a.2/ORF3.2/ORF4.2 clades, ORF6.1 and ORF6.2 contained InglevacATP and MLV vaccine strains, respectively, along with the same field strains previously identified. These results suggest six of the field viruses (14-95, 14-78, 14-81, 14-76, 14-79 and 14-96) are likely derived from InglevacATP and four field viruses (14-64, 14-67, 14-68 and 14-60) are derived from InglevacMLV. Clade ORF6.3 contained 7 viruses with only 93% identity to previously sequenced PRRSV by BLASTN analysis. Reference viruses MN184A-C clustered in clade ORF6.4 along with 10 field strains. Clade ORF6.5 contained four field strains with only

92% identity to any sequences in GenBank. The final clade, ORF6.6, was the largest with 33 strains and included the reference virus NADC30. NADC30-containing clades for ORF2a, ORF3 and ORF4 similarly contained the largest number of members suggesting that this is the dominant lineage of PRRSV currently circulating in the U.S. Two strains failed to cluster within ORF6.1 through ORF6.6: ISU32 and 104194. The latter strain also failed to cluster within any clades for ORF2a, ORF3 and ORF4. ORF6 intraclade diversity was 0.013-0.056 and interclade diversity was 0.045-0.137 which is less than that observed for the ORF2a through ORF5.

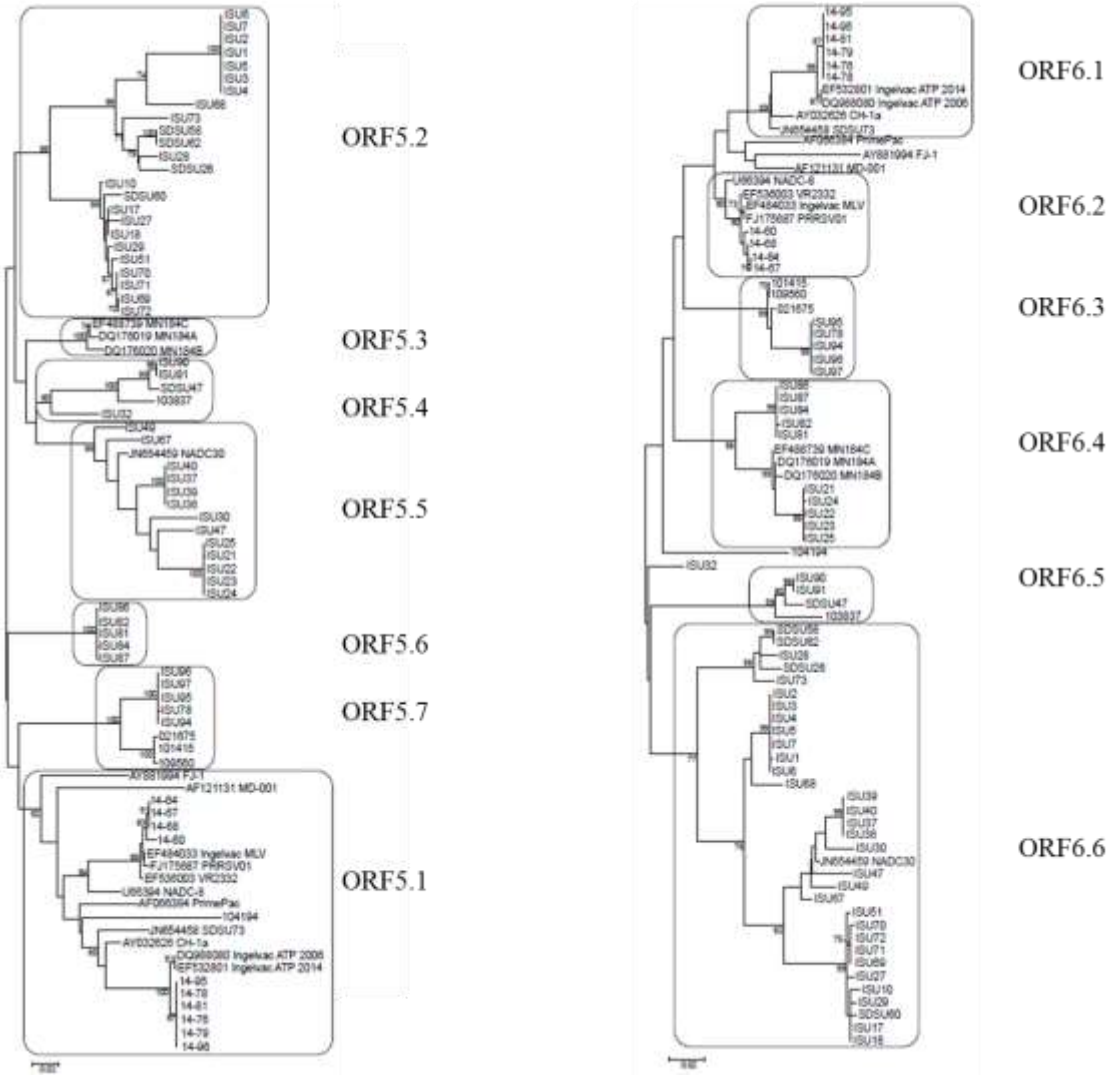


Figure 4. Phylogenetic analysis of ORF5 and ORF6 sequences determined by metagenomic sequencing of PRRSV-positive porcine sera. Maximum likelihood analysis in combination with 1000 bootstrap replicates was used to derive trees based on nucleotide sequences. Clade designations are indicated by boxes and were chosen to maximize strain inclusion based on topological support >70.

Phylogenetic topology

To investigate topological differences between the glycoprotein and matrix phylogenetic trees, individual strain clade combinations were determined. With the exception of strain 104194, genes from clades containing vaccine strains InglevacMLV and IngelvacATP (GP2a.1/GP2a.2, GP3.1/GP3.2, GP4.1/GP4.2, GP5.1 and M.1/M.2) were only found in conjunction with other genes derived from vaccine strains (Table 1, gene combination 1 and 2). Examination of the remaining gene combinations revealed numerous topological incongruences. To determine whether different topologies for the glycoproteins and M gene are due to recombination, representative strains for the identified gene combinations 1-12 were analyzed by the genetic algorithm recombination detection (GARD) software. GARD in conjunction with the Kishino Hasegawa test were used to predict breakpoints and analyze topological incongruences that could be explained by recombination. For this analysis, nt=1 was defined as the A of the start codon for GP2a and the region of the genome analyzed was from GP2a to the 3'-terminus. Five breakpoints with significant topological incongruences were identified (P<0.05). These breakpoints corresponded to the region of GP2a/GP3 overlap, GP5, two in the M gene and N gene, respectively. The strains were next analyzed by the Recombination Detection Program to identify if any of the strains representing gene combinations 1-12 were recombinants. Two strains were identified as recombinants with high confidence (recombination score >0.6). ISU23 (GP2.5, GP3.3, GP4.3, GP5.5, M6.4) was identified as a recombinant between major parent ISU39 (**GP2.5**, **GP3.3**, **GP4.3**, **GP5.5**, M6.6) and minor parent ISU81, (GP2.4, GP3.4, GP4.4, GP5.6, **M6.4**), with breakpoints at bp 2252 within GP5 and bp 3199 within the 3'-UTR (P=2.4x10^{-11}). ISU49 (GP2.3, GP3.3, GP4.3, GP5.5, M6.6) was identified as a recombinant between major parent ISU39 (GP2.5, **GP3.3**, **GP4.3**, **GP5.5**, **M6.6**) and minor parent ISU94 (**GP2.3**, **GP3.3**, GP4.7, GP5.7, M6.3) with breakpoints at bp 26 within GP2a and bp 1102 within GP3 (P=1.1x10^{-20}).

Gene Combination	Representative Strain	Strains (n)	ORF2	ORF3	ORF4	ORF5	ORF6
1	14-79 (Inglevac ATP)	6	2.1	3.1	4.1	5.1	6.1
2	14-64 (Inglevac MLV)	4	2.2	3.2	4.2	5.1	6.2
3	ISU10	11	2.3	3.3	4.3	5.2	6.6
4	ISU49	1	2.3	3.3	4.3	5.5	6.6
5	ISU94	8	2.3	3.3	4.7	5.7	6.3
6	ISU81	5	2.4	3.4	4.4	5.6	6.4
7	ISU32	1	2.4	3.4	4.6	5.4	U
8	ISU28	6	2.4	3.4	4.6	5.2	6.6

9	ISU3	7	2.5	3.3	4.3	5.2	6.6
10	ISU23	5	2.5	3.3	4.3	5.5	6.4
11	ISU39	7	2.5	3.3	4.3	5.5	6.6
12	ISU90	4	2.5	3.4	4.5	5.4	6.5
13	104194	1	U[1]	U	U	5.1	U

Table 1. Glycoprotein and matrix genes clade combinations determined by phylogenetic analysis.

Pairwise Sequence Comparisons

Pairwise nucleotide sequence comparisons of percent identity were made for all ORFs to determine the level of genetic diversity within each ORF (Table 2). In this dataset, ORF3 was the most genetically diverse with 80.4 – 100% identity. ORF5 was the second-most diverse with 82.3-100% identity. ORF2a and ORF4 were only slightly less diverse (83.6-100% identity). ORF 6 was the most conserved ORF, showing 86.7-100% identity.

Error! Not a valid link.

*nucleotides (nt)

Table 2. Sequence characteristics of nucleotide alignments of ORFs 2a - 6 of 81 PRRSV sequences.

Discussion:

Since its emergence in the 1990s, porcine reproductive and respiratory syndrome virus has caused significant economic losses for pig producers worldwide. Early comparative studies of PRRSV structural gene sequences from the American genotype (type 2) identified different levels of diversity within each gene (Meng et al., 1995). More recent analyses have shown that this diversity is the result of mutation by an error-prone RNA polymerase and recombination between co-infecting strains of PRRS (Martin-Valls et al., 2014).

PRRSV epidemiology has focused almost exclusively on using the highly variable nucleotide sequence of ORF5 which encodes the major membrane glycoprotein GP5. A large scale phylogenetic analysis including more than 8,000 ORF5 sequences categorized type 2 PRRSV into nine lineages with less than 10% intraclade and more than 10% interclade genetic diversity (Shi et al., 2010). Although we included many of these reference strains from this study into the current analysis, many of these ORF5 lineage specific references did not cluster into well-

supported lineages for ORF5 or the other structural ORFs. While different analysis methods in the current study may explain part of the inconsistency, it is more likely that the reference strains represent divergent descendants of the ancestral strains that separate into well-supported lineages which are not currently common in the U.S. All of our samples originated in the U.S. in 2014 while those of Shi et al. (2010) included all ORF5 sequences in GenBank until January, 2009. In order to place the new genomes from the current study into the ORF5 lineages, the Shi et al (2010) dataset would have to be reanalyzed with the newly sequenced viral ORF5 sequences. However, by surveying a single gene, recombination cannot be ascertained and the resulting phylogenetic analyses may not accurately represent strain evolution (Rokas et al., 2013). The current work shows within the 66 viral genomes while ORF5 is highly variable (82.26%), ORF3 has greater genetic diversity (80.39%). ORF2a and ORF4 encode proteins that facilitate virus attachment and entry and have very similar diversity values of 83.59% and 83.61%, respectively (Das et al., 2010). This and the wide range of between clade distances for each ORF suggest that during a single year PRRSV strains, even from within a geographic area, are not necessarily closely related.

The metagenomic sequencing methodology developed in this report was successful in determining complete PRRSV genomes directly from serum samples. Given the ease of sample collection and relative low cost for metagenomic sequencing ($300 currently, prices expected to decrease), metagenomic sequencing will undoubtedly become more commonplace in veterinary diagnostics as producers and veterinarians are routinely paying ~$150 for GP5 sequencing alone. Besides delivering a comprehensive picture of PRRSV genetics which can be used to improve eradication and control efforts, metagenomic sequencing can identify all other viruses present in the sample. While all the samples contained PRRSV, a majority of the samples were positive for additional viruses, mainly various species of porcine parvovirus and torque teno suis virus. We were the first to identify porcine parvovirus 6, a new parvovirus described in China in 2014, here. In addition, we identified and characterized a novel, highly divergent porcine pestivirus in this dataset. Further research is necessary to understand what etiological role these viruses play in clinical disease. The ability to both determine complete genome sequencing and profile all viruses in a sample speak to the utility and potential of this methodology to transform our understanding of both PRRSV and virus ecology and diversity.

References

1. Brockmeier SL, Loving CL, Vorwald AC, Kehrli ME, Baker RB, et al. 2012. Genomic sequence and virulence comparison of four type 2 porcine reproductive and respiratory syndrome virus strains. Virus Res 169:212-221.
2. Chang CC, Yoon KJ, Zimmerman JJ, Harmon KM, Dixon PM, et al. 2002. Evolution of porcine reproductive and respiratory syndrome virus during sequential passages in pigs. J Virol 76:4750-4763.
3. Choi HW, Nam E, Lee YJ, Noh YH, Lee SC, et al. 2014. Genomic analysis and pathogenic characteristics of type 2 porcine reproductive and respiratory syndrome virus nsp2 deletion strains isolated in Korea.
4. Costers S, Lefebvre DJ, Van Doorsselaere J, Vanhee M, Delputte PL, Nauwynck HJ. 2010. GP4 of porcine reproductive and respiratory syndrome virus contains a neutralizing epitope that is susceptible to immunoselection in vitro. Arch Virol 155:371-378.
5. Das, P.B., Dinh, P.X., Ansari, I.H., de Lima, M., Osorio, F.A., Pattnaik, A.K., 2010. The minor envelope glycoproteins GP2a and GP4 of porcine reproductive and respiratory syndrome virus interact with the receptor CD163. J Virol 84, 1731-1740.
6. Kim WI, Kim JJ, Cha SH, Wu WH, Cooper V, et al. 2013. Significance of genetic variation of PRRSV ORF5 in virus neutralization and molecular determinants corresponding to cross neutralization among

PRRS viruses. Vet Microbiol 162:10-22.
7. Li J, Murtaugh MP. 2012. Dissociation of porcine reproductive and respiratory syndrome virus neutralization from antibodies specific to major envelope protein surface epitopes. Virology 433:367-376.
8. Lu Z, Zhang J, Huang CM, Go YY, Faaberg KS, et al. 2012. Chimeric viruses containing the N-terminal ectodomains of GP5 and M proteins of porcine reproductive and respiratory syndrome virus do not change the cellular tropism of equine arteritis virus. Virology 432:99-109.
9. Martin-Valls GE, Kvisgaard LK, Tello M, Darwich L, Cortey M, et al. 2014. Analysis of ORF5 and full length genome sequences of porcine reproductive and respiratory syndrome virus isolates of genotype 1 and 2 retrieved worldwide provides evidence that recombination is a common phenomenon and may produce mosaic isolates. J Virol 88:3170-3181.
10. Meng, X.J., Paul, P.S., Halbur, P.G., Morozov, I., 1995. Sequence comparison of open reading frames 2 to 5 of low and high virulence United States isolates of porcine reproductive and respiratory syndrome virus. J. Gen. Virol. 76, 3181-3188.
11. Nguyen VG, Kim HK, Moon HJ, Park SJ, Chung HC, et al. 2013. Evolutionary dynamics of a highly pathogenic type 2 porcine reproductive and respiratory syndrome virus: analyses of envelope protein-coding genes. Transbound Emerg Dis doi:10.1111/tbed.12154
12. Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425, 798-804.
13. Shi, M., Lam, T.T., Hon, C., Murtaugh, M.P., Davies, P.R., Hui, R.K., Li, J., Wong, L.T., Yip, C., Jiang, J., Leung, F.C., 2010. Phylogeny-based evolutionary, demographical and geographical dissection of North American type 2 porcine reproductive and respiratory syndrome viruses. J. Virol. 84, 8700-8711.
14. Tian D, Wei Z, Zevenhoven-Dobbe JC, Liu R, Tong G, et al. 2012. Arterivirus minor envelope proteins are major determinants of viral tropism in cell culture. J Virol 86:3701-3712.
15. Vu HL, Kwon B, Yoon KJ, Laegreid WW, Pattnaik AK, Osorio FA. 2011. Immune evasion of porcine reproductive and respiratory syndrome virus through glycan shielding involves both glycoprotein 5 as well as glycoprotein 3. J Virol 85:5555-5564.
16. Zhou L, Zhang J, Zeng J, Yin S, Li Y, et al. 2009. The 30-amino acid deletion in Nsp2 of highly pathogenic porcine reproductive and respiratory syndrome virus emerging in China is not related to virulence. J Virol 83:5156-5167.