Sponsors

University of Minnesota
College of Veterinary Medicine

College of Agricultural, Food and Environmental Sciences

Extension Service

Swine Center

Editors
W. Christopher Scruton
Stephen Claas

Layout
David Brown

Logo Design
Ruth Cronje, and Jan Swanson;
based on the original design by Dr. Robert Dunlop

Cover Design
Sarah Summerbell

How to interpret and use PRRSV sequence information

Michael P. Murtaugh PhD and Kay S. Faaberg PhD

Department of Veterinary PathoBiology, University of Minnesota

Introduction

PRRS remains a somewhat mysterious disease 12 years after it was first reported as a new syndrome in swine[1]. Prevention and control of PRRS continue to bedevil veterinarians and producers, while new forms of disease and the interaction of the viral agent with other pathogens provide a continuing challenge to accurate disease diagnosis on the farm. A root cause of these difficulties lies in the extensive genetic variation that is characteristic of PRRSV and that is not observed to such a degree in other swine pathogens. Therefore, precise identification of individual PRRSV isolates by determination of the exact nucleotide sequence of the viral genetic material can be a valuable tool for understanding the success or failure of PRRS detection, control, and elimination. Modern technology has made the acquisition of precise genetic information a straightforward process. However, the helpfulness of this information is not always apparent to veterinarians and producers. The objectives of this communication are to provide background knowledge about what PRRSV sequence information is, to explore how it is interpreted and used, and, lastly, to consider limitations and reservations on its use.

What does 'sequence information' mean?

The biological and pathological features and characteristics of PRRSV—i.e., the manner in which it grows, causes disease, elicits an immune response, persists in pigs, and so on—are determined by viral proteins. These proteins are strings of individual building blocks known as amino acids, of which there are about 20 different types. Each type has a distinct chemical structure, so that strings consisting of various types and numbers of amino acids result in proteins with special biological properties and functions. Letters of the alphabet are used to indicate each type of amino acid. A string of letters indicates the exact order of amino acids, the exact type of amino acid at any given position in the string, and the total number of amino acids in the protein. This is the sequence of the protein, and every protein has its own sequence.

The sequence of PRRSV proteins is determined by the genetic material of PRRSV. The genetic material also is a sequence. It is a sequence of four building blocks, called bases, of which there are four types: A, C, G, and T. PRRSV is defined at the genetic level by the specific order of about 15,000 of these four bases. The amino acids of proteins are each encoded by a ministring of three bases. It is just like Morse code: the bases are like dots and dashes which, when put into specific combinations, specify amino acid letters. The consecutive sequence of amino acid letters then spells out the protein word. Because three specific bases determine the identity of one amino acid, a string of 600 bases determines a protein sequence of 200 amino acids. Thus, the type and order of bases in the genetic material determines the sequence of the viral proteins, and the biological and pathological characteristics of PRRSV. "Sequence information" means the sequence of bases in the genetic material, or the corresponding sequence of amino acids that is specified in the resulting protein. **Figure 1** shows the relationship between a nucleotide sequence, amino acid sequence, and PRRSV characteristics.

A nucleotide sequence is organized into discrete packets of information known as genes. A gene contains all the nucleotide sequences for all the amino acids of a protein. PRRSV has 7 or 8 proteins, depending on how they are counted. One of these genes is known as open reading frame 5 (ORF5), and it encodes the information for a protein that resides on the surface of PRRSV and interacts with cells that the virus infects. The sequence of this protein, which is 200 amino acids long and, so, is encoded by 600 bases of nucleotide sequence information, is the portion of the PRRSV genome which is most often used in genetic analysis of PRRSV. Another protein which is sometimes used is called the nucleocapsid and is encoded by ORF 7. It is 123 amino acids long.

The key point about sequence information is that it is a string of bases, or the encoded amino acids, which defines a PRRSV and determines its biological properties.

Variation is present in PRRSV sequences

The sequences, either nucleotide or amino acid, of PRRSV isolated from different farms are rarely identical. Therefore, variation is present in PRRSV sequences. Variation occurs because bases may be changed during virus repli-

Figure 1. Relationships between nucleotide sequence, amino acid sequence, and PRRSV properties

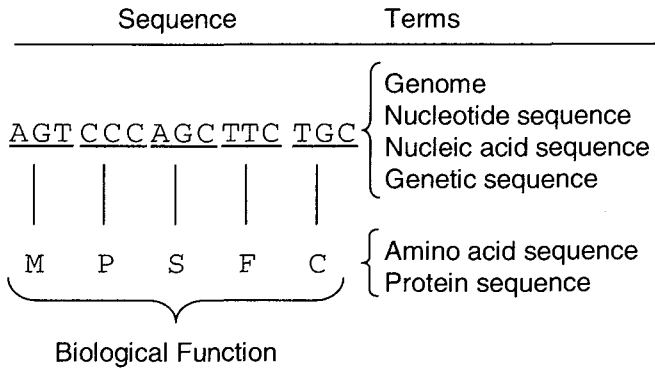
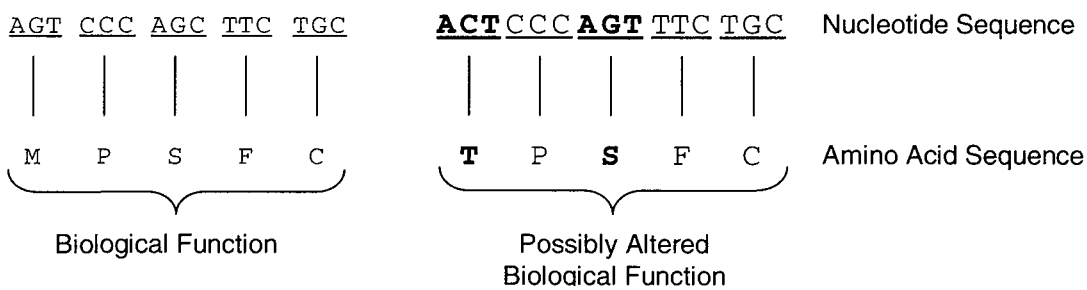


Figure 2. Relationships between nucleotide sequence variation, amino acid sequence variation, and changes in PRRSV properties



The first change in nucleotide sequence, AGT → ACT, resulted in an amino acid change, the second change (AGC → AGT) did not result in an amino acid change. The change may or may not alter the properties of the virus.

cation—a process called mutation—or because viral genetic material can recombine and give rise to new viruses that are a mixture of two preexisting viruses. Changes in nucleotide sequence may give rise to changes in amino acid sequence and result in a virus with different properties. **Figure 2** shows two changes in nucleotide sequence; one results in a change in the amino acid sequence of the encoded protein, and the second is a so-called silent mutation because the protein sequence remains the same. This variation, which occurs as a natural consequence of virus replication, allows us to reconstruct patterns of virus change in the field. More important, it is an engine for biological variation that enables the virus to adapt to changes in its environment.

In either event, the new form is related to the ancestral form from which it was derived and to which it is nearly identical, differing by only one or a few bases or amino acids. Relationships among individual PRRSV isolates can be deduced by comparing sequences to each other. Changes are cumulative; closely related isolates are nearly identical, whereas more distant isolates have the same differences as early isolates, and additional changes which are unique. An example of variation among isolates, showing the change over time from an initial PRRSV sequence, to progressively more distant variants, is shown in **Figure 3**. Isolate 1 changed two times, giving rise to isolate 2 with a single amino acid change (D) and isolate 5, with a single amino acid change (S) at a different position. In the first case, isolate 2 acquired a second mutation, becoming isolate 3, then this isolate acquired several additional changes as indicated in isolate 4. By contrast, isolate 5 gave rise to two independent viruses, isolates 6 and 7, that contained numerous independent mutations. The relationships deduced from the sequences of individual isolates are shown in panel B of Figure 3.

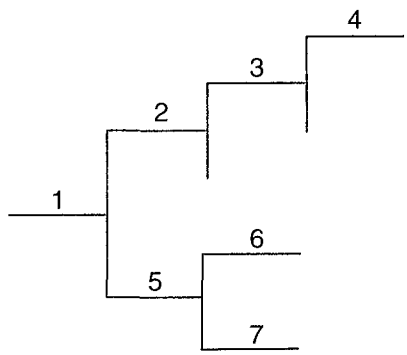
Variation in PRRSV sequences is useful because it permits the reconstruction of patterns of virus change in the field. Thus, it is often possible to determine if virus strains recovered from farms originated from pre-existing isolates or represent new introductions from an outside source. Similarly, analysis of viral sequences may indicate if a strain of virus represents a form of vaccine virus that was reisolated from a vaccinated herd. In theory, changes in the sequence of PRRSV also may lead to biological or pathological changes. However, at present the relationship between nucleotide or amino acid sequence variation and behavior of PRRSV in the field is not known. Therefore, it is not possible now to associate specific genetic characteristics of a PRRSV with specific biological properties or characteristics.

Figure 3. Variation in PRRSV sequences

A.

Isolate	Amino Acid Sequence
1	.....
2	.....D.....
3	.....Q.....D.....
4	.....Q....G..G.....A.....D.....
5	.....S.....G.....
6	.....A...AS.....
7	..GR.....W....VD..S.....F.....NE.....T

B.



(A) Amino acid sequences from 7 individual PRRSV isolates are displayed with conserved, unaltered amino acids shown as dots. Changes are shown as new alphabetic characters. (B) The relationships of the seven isolates as deduced from the pattern of changes observed in panel A.

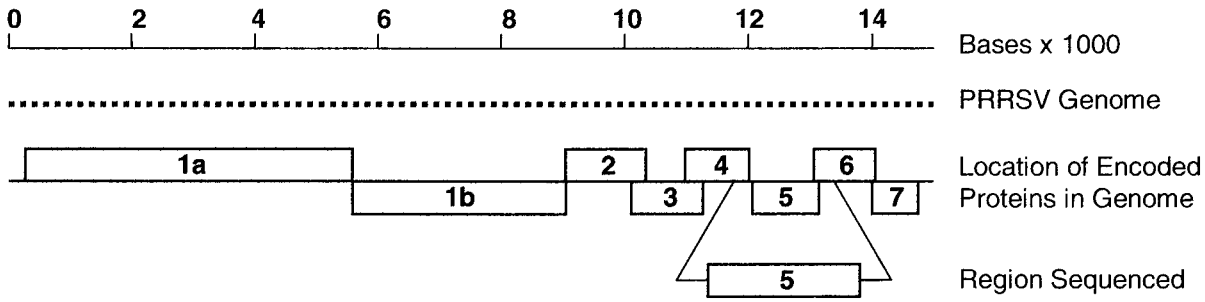
PRRSV diagnostic sequencing

In order to obtain sequence information, the genetic material of PRRSV in samples of blood, fluids, or tissue submitted to a diagnostic laboratory is purified according to standard procedures. The genetic material is subjected to various biochemical procedures and the exact nucleotide sequence is determined on a DNA sequencing machine. Various types of analysis are then performed by computer. Although it is possible to produce an exact sequence of the entire 15,000 bases of the genome, experience has taught that information from a single gene, or roughly 5% of the genome, will provide a suitable genetic characterization of a viral isolate in a timely and cost-effective manner. The most common region used for sequencing is the gene referred to as open reading frame 5 (ORF5) which is 600 bases in length and encodes a 200 amino acid protein knows as the envelope glycoprotein. This protein is of interest because it attaches the virus to macrophages to start a cycle of infection, because it is an important target for neutralizing antibodies, and because it has regions which are known to be highly variable. ORF5 also is the gene used for RFLP typing[2]. RFLP typing was an early method for genetic identification and classification of PRRSV isolates and was particularly useful for distinguishing between RespPRRS vaccine and field isolates. The location of ORF5 in the PRRSV genome is shown in **Figure 4**.

ORF7, which codes for the nucleocapsid protein of PRRSV, may also be used for genetic analysis. It consists of 372 bases and encodes a protein of 123 amino acids. This gene is more highly conserved than is ORF5 and in some cases will suggest that virus isolates are the same, whereas the ORF5 sequence may suggest that the same isolates are not identical.

Once the nucleotide sequence has been obtained it can be analyzed directly or converted into the corresponding amino acid sequence of the encoded protein. Either sequence in and of itself is a mind-numbing string of alphabetic letters with little if any meaning to most veterinarians, producers, and scientists. Useful information is typically extracted by comparison of the test sequence with a database of known vaccine strains, previous farm

Figure 4. Schematic diagram of the PRRSV genome and gene organization



The nucleotide sequence is represented as a dotted line, and the genes, each one of which encodes a viral protein, are shown as numbered boxes. Sequencing is usually performed on the expanded region including ORF5 and portions of ORF4 and ORF6.

isolates, or other relevant sets of sequenced virus isolates. The Minnesota Veterinary Diagnostic Laboratory, for example, maintains a database containing ORF5 sequences from more than 1,000 field and experimental PRRSV samples (as of August 2001) obtained in all years throughout the United States and Canada from 1989 to the present. Other diagnostic laboratories that offer sequencing services have similar databases and provide equivalent comparisons.

At this point in time, the methods for obtaining, analyzing, and reporting sequence information are not standardized among diagnostic laboratories. The exact region of the genome, the number of bases sequenced, the use of nucleotide sequence or amino acid sequence, and computer software programs may vary from one laboratory to another. Although the quality of data is uniformly high, in the present situation it may be difficult to directly compare information obtained from different diagnostic laboratories. Discussions are currently underway to standardize the performance, analysis and reporting of PRRSV sequence in order to maximize its value to the swine industry. Currently, two types of comparative analyses, alignment and dendrogram, are routinely performed.

Interpretation and use of sequence alignments

Sequence alignment consists of ordering a set of sequences and observing the base or amino acid at each position. An example of sequence alignment is shown in **Figure 3A**. Dots are shown in the first position for all of the isolates. This means that they all have the same amino acid. All of the isolates also have a dot at the second position, meaning that they again encode the same amino acid at this position. It can be same or different than the amino acid in the first position. At the third position, isolate 7 has an amino acid different from the consensus amino acid used by the other six isolates. The actual amino acid therefore is shown at this position to indicate that a change has occurred. The use of dots to indicate consensus sequence greatly simplifies the visual identification of changes in sequence, since even among PRRSV isolates which are substantially different, the great majority of sites are conserved.

An alignment report frequently contains additional information, such as the consensus sequence itself. With this information it is possible to reconstruct the exact sequence of each isolate in the alignment. However, this knowledge is of little value in the field since no single amino acid or base is yet known to be responsible for any specific characteristic or trait of a PRRSV. The alignment is also used to provide a percent similarity comparison of two isolates. This is simply the ratio of the number of identical bases or amino acids shared between two isolates divided by the total number of bases or amino acids compared (600 bases, or 200 amino acids, in the case of ORF5).

Experience has shown that isolates that are recovered from pigs tend to have an average of 0.1–0.2% changes in ORF5 as compared to the isolate which was used to infect the pig. This rate stays the same or declines with continued passage of recovered virus through additional pigs. Therefore, isolates that exhibit few differences are interpreted as being closely related, whereas isolates which show substantial differences are interpreted as being independent. Collins (1998) suggested as a general guideline that 10 or fewer differences in nucleotide sequence, out of 960 bases, be interpreted as closely related PRRSV isolates[3]. However, it is not possible to assign strict values of percent similarity that delineate the difference between related and independent isolates.

Sequence alignment is a powerful method for monitoring PRRSV isolates within a herd over time. It can detect and differentiate multiple strains and provide genetic information that can be used to determine the effect of management practices on PRRSV populations. It can assist in determining the source of outbreaks to introduction of

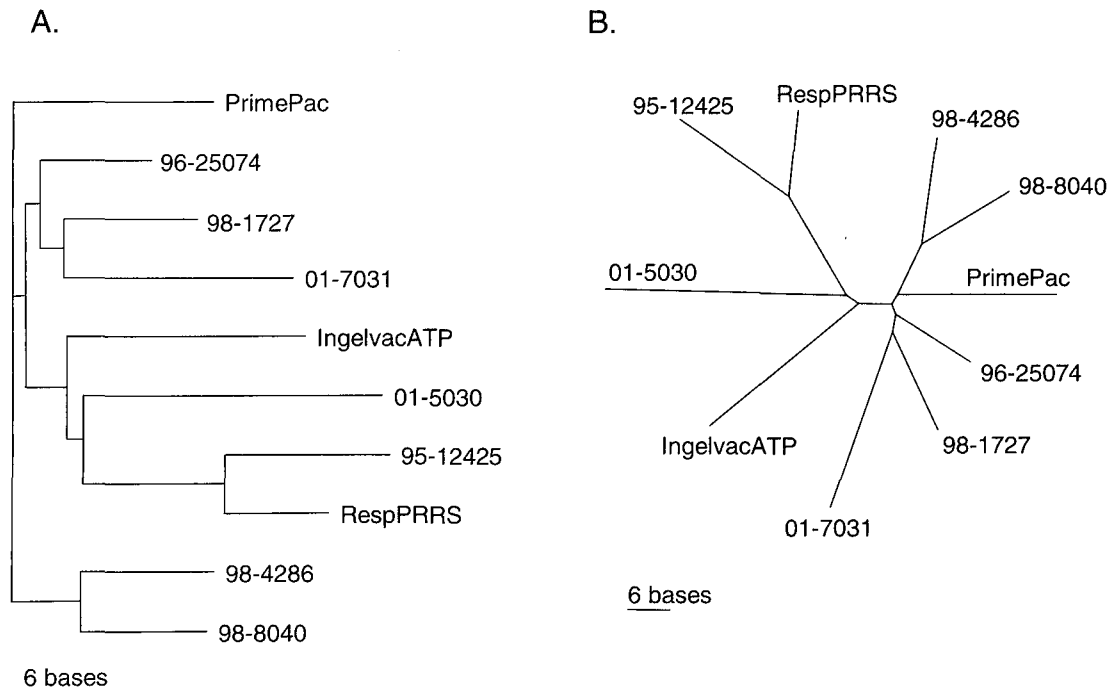
new virus strains, or to reemergence of a preexisting farm strain. The procedure is more sensitive than RFLP typing in detecting changes in PRRSV isolates since it compares hundreds of bases, whereas RFLP typing is based on a total of 18 bases at three locations within the ORF5 nucleotide sequence. Sequence alignment has been used extensively to compare field isolates with vaccine sequences. The interpretation of this analysis should take into account vaccination history of the herd, since vaccine virus, like field isolates, is viremic for a period of weeks after administration and may persist for substantially longer periods. Individual animals may harbor more than one virus strain. Although this appears to be an infrequent occurrence, the likelihood is increased when vaccination is performed in the presence of active disease.

Interpretation and use of dendrograms

Dendrogram analysis is done by aligning sequences of a set of PRRSV samples in an ordered fashion according to the similarity of pairs of sequences. Computer programs compare all possible pairs of sequences, grouping the most similar pair, then comparing this set to all remaining sequences. It continues this process until a tree, the dendrogram, is constructed. These programs assume that all differences among sequences are due to random, independent mutations during the evolution of the virus. Biologists have learned that sequence similarity corresponds to the evolutionary history of an organism. Therefore, dendrograms are commonly thought to reflect ancestral, or phylogenetic, relationships. A dendrogram showing the relationships of the isolates aligned in Figure 3A is given in **Figure 3B**. Isolate 1 is the root, or ancestral, virus from which the remaining viruses emerged over time by a series of sequential changes.

There are several ways to display a dendrogram. Two examples are shown in **Figure 5**. The phylogram portrays the results in the form of a tree turned 90 degrees. More closely related isolates are clustered together on smaller branches of the tree. In the example in **Figure 5A**, the horizontal line distances between isolates is a direct measure of sequence similarity. It is possible that closely related isolates could be less similar than more distantly related isolates. Some computer programs display trees in which all branches have the same length. These trees give the same relationships among a set of isolates, but do not give any information about relative similarity or differences. A second display method, shown in **Figure 5B**, is a radial dendrogram. Comparison of Figures 5A and 5B shows that all of the relationships are identical in the two representations. The chief difference is that no specific ancestral strain is suggested in the radial dendrogram, whereas a phylogram always suggests the presence of a primordial form that gave rise to the present set of samples. In the radial dendrogram, the linear distance

Figure 5. Phylogenetic representations of ORF5 sequences representative of PRRSV variation in the Minnesota Veterinary Diagnostic Laboratory database



(A) Phylogram. (B). Radial dendrogram. The amount of difference between any two samples is proportional to the linear distance (radial dendrogram) or the length of the horizontal connecting lines (phylogram).

between pairs of isolates corresponds to their degree of similarity. Thus, a set of closely related isolates would appear as a cluster of short, barely distinguishable lines at the scale shown.

Dendrogram analysis was originally designed for use in PRRS research to predict how PRRSV isolates were related and to characterize the variation existing in PRRSV around the time of its discovery. The technique is exquisitely sensitive since it can show differences between isolates as small as a single nucleotide. It has been used in the field to determine if the reappearance of PRRSV on a farm is due to the introduction of a new strain or to the reemergence of a strain which existed previously on site. It has also been used to determine if PRRS outbreaks on a farm are due to a single virus clone or to more than one clone. It also has been used to support claims that vaccine has reverted to virulence. As with RFLP typing and alignments, this information can be used to track isolate introduction into a swine system, monitor spread, and, in some cases, to differentiate vaccine and field virus.

Dendrogram analysis is a powerful method for deducing evolutionary relationships when the history of a species is known or when the mechanism of evolution is known to be random mutation. However, nonrandom mutation, such as recombination, may greatly skew phylogenetic software programs and can render dendrograms highly questionable. Recombination is known to occur in PRRSV[4]. Similarly, genetic selection in the form of immunological resistance can induce changes which may not be significant in terms of virus differentiation, but may still show up as different branches on a tree. Because the ORF5 sequences used for PRRSV dendrogram analysis are subject to both recombination and immunological selection, the interpretation of dendrograms should be approached cautiously if the results are inconsistent with other information, particularly sequence alignment. Direct comparison of aligned sequences can reveal the presence of recombination since base or amino acid differences will be highly clustered in specific regions.

Dendrograms cannot be used under any circumstances to assess biological characteristics, like virulence, of PRRSV. They can be used to show the possibility that 2-5-2 field isolates may not be derived from vaccine virus, but such results are not conclusive and may represent examples of recombination.

What PRRSV sequences do not tell

The information contained in a PRRSV genome defines the total biological capabilities of the virus. Genetic variation defines the degree to which expression of specific traits of the virus, such as clinical signs, severity of respiratory or reproductive disease, growth in cell culture, and immunological cross-reactivity, can differ. However, the relationship between genetic sequence and biological traits of PRRSV has not been worked out yet, so it is not possible to predict the real world meaning of differences in nucleotide or amino acid sequence. Therefore, it is inappropriate to use sequence information as an indication of relative isolate virulence, or to make a vaccine selection.

Pigs that harbor PRRSV also vary genetically and biologically. They grow at different rates, have different litter sizes, exhibit different sensitivities to disease, and so on. Some of the variation that is observed in clinical manifestations and disease characteristics within and among herds is due to differences in how individual pigs respond to the virus. The relative contributions of variation among pigs, variation in PRRSV, and the interaction of the virus and the pig, to the total spectrum of PRRS variability is unknown. It is for this reason that speculative predictions about biological or pathological characteristics of a PRRSV isolate based on sequence information are a poor substitute for direct experimental observation and confirmation of relevance in the field.

Sequence alignments and dendrograms are determined by computer programs that use numerous assumptions about data quality and mechanisms by which differences occur. The programs assume that sequence data is robust. Therefore, if the data are not of high quality or contain ambiguities, then the results will be unreliable. One simple test for data quality is to establish that the nucleotide sequence determined by machine is actually translated into a full length ORF5 protein of 200 amino acids. Phylogenetic programs also assume that all differences arise by random mutation. Thus, recombination, in which segments of nucleotide sequence are exchanged or substituted, may result in wholesale changes that drastically alter and misrepresent apparent relationships among isolates. This potential problem can be addressed by sequence alignment analysis if closely related isolates are available for comparison.

Conclusions

The use of sequence information to compare and contrast PRRSV isolates is a valuable tool in the evaluation of disease outbreaks. It can differentiate between viruses resident and newly introduced to a farm, and vaccine and field strains. It is perhaps most effective at the farm or herd level for assessing isolate introduction into a herd and monitoring change over time. Interpretation and use of sequence information depends on the availability of high quality data. Sequence information cannot be used for making inferences about biological properties of PRRSV, and is most useful when alignments and dendrograms are used in combination.

References

1. Keffaber, KK. Reproductive failure of unknown etiology. AASP Newsletter. 1989; 1: 1-10.
2. Wesley, RD, Mengeling, WL, Lager, KM, et al. Differentiation of a porcine reproductive and respiratory syndrome virus vaccine strain from North American field strains by restriction fragment length polymorphism analysis of ORF 5. J. Vet. Diagn. Invest. 1998;10:140-144.
3. Collins, J. Interpreting PRRSV sequencing data. Allen D. Leman Swine Conf. Proceed. 1998; 25: 1-4.
4. Yuan, S, Nelsen, CJ, Murtaugh, MP, et al. Recombination between North American strains of porcine reproductive and respiratory syndrome virus. Virus Research. 1999; 61: 87-98.