Progress Report:

Fine-scale classification of PRRSV-2: Moving past RFLPs to improve sequence interpretation for disease control and management

October 25, 2023

Primary investigator

Kimberly VanderWaal
1365 Gortner Avenue, St. Paul, MN 55421
kvw@umn.edu

Investigative core team

Cesar Corzo	University of Minnesota
Albert Rovira	University of Minnesota
Igor Paploski	University of Minnesota
Mariana Kikuti	University of Minnesota
Derald Holtkamp	Iowa State University
Daniel Linhares	Iowa State University
Giovani Trevisan	Iowa State University
Michael Zeller	Iowa State University
Jianqiang Zhang	Iowa State University
Tavis Anderson	NADC, USDA Agricultural Research Center

PROBLEM STATEMENT

It is increasingly apparent that using restriction fragment length polymorphisms (RFLP)-typing to refer to genetic variants of Porcine reproductive and respiratory syndrome virus (PRRSV)-2 is both outdated and, more importantly, can lead to misleading or even erroneous conclusions about the relatedness of PRRS viruses. The shortcomings of RFLPs have long been recognized and a recent AASV-administered survey found that 88% of surveyed swine practitioners are in favor of moving away from RFLP-typing, but only if there is viable alternative. As yet, no alternatives have been pursued. Lineages and sub-lineages provide more biologically meaningful classification for PRRS viruses, but do not have the level of granularity often required for on-farm management and outbreak investigations of PRRSV – which are major reasons for sequencing conducted by swine veterinarians. With recent advances in computational power and the creation of national-scale sequence databases (such as the Morrison Swine Health Monitoring Project [MSHMP] and the Swine Disease Reporting System [SDRS]), we are now in a position to address long-recognized issues with RFLP-typing and find better solutions.

The purpose of this research project is to evaluate the feasibility of implementing alternative nomenclature systems for fine-scale sub-typing of PRRSV, one that is expandable to new genetic diversity that emerges as consequence of virus evolution.

OBJECTIVES

1) Evaluate and compare alternative systems for classifying and naming PRRSV-2 variants
   a) Refine variant definition based on farm-level patterns of occurrence
   b) Assess adaptability of classification system to accommodate expanding genetic diversity at national scales

2) Develop procedures for prospective implementation and expansion that would meet the needs of diagnostic labs and practitioners. Any newly developed system would aim to be scalable and reproducible (i.e., powered by tools that can easily accessed/ implemented by individuals, VDLs, MSHMP, or SDRS, providing the same results everywhere).

MATERIALS AND METHODS

Data Source and phylogenetic reconstruction

Sequence data were obtained from the Morrison Swine Health Monitoring Project (MSHMP), which is a voluntary initiative operated by University of Minnesota that monitors PRRS occurrence in farms belonging to 37 production systems, accounting for >50% of the U.S. sow population. Participating production systems also share PRRSV ORF5 sequences that are generated as part of routine monitoring and outbreak investigations in breeding, gilt developing units, growing and finishing herds.

Sequences were divided into short- and long-term datasets. The short-term dataset, which included three years of sequence data (6749 sequences from Jul. 1, 2018 – Jun. 30 – 2021), was utilized for comparing different classification methods in classifying PRRSV genetic variants that concurrently co-circulate within U.S. swine populations. The long-term dataset, which included ~11 years of sequence data (28,965 sequences from Jan. 1, 2010 – Sep. 30, 2021) was used to evaluate the frequency of emergence of new PRRSV variants. Sequences were aligned and IQ-Tree2 was used to build phylogenies based on the maximum-likelihood, strict consensus, and extended consensus methods. Phylogenies were either constructed with the full or de-duplicated set of sequences.

Variant classification

Several tree-based clustering approaches were applied to the phylogenies using the *TreeCluster* package available in Python; clusters of genetically related sequences identified in the trees were referred to as "variants." Multiple relatedness thresholds (2 – 8%) were compared for each clustering method. In total, 142 approaches were compared: 23 *TreeCluster* methods applied to each of three tree types (maximum-likelihood, strict consensus, and extended consensus) built on two datasets (full and de-duplicated), plus RFLP and Lineage+RFLP.

SIGNIFICANT RESULTS

1) Evaluate and compare alternative systems for classifying and naming PRRSV-2 variants

- We rigorously compared 140 approaches that utilized different approaches to cluster ORF5 sequences into genetic "variants" based on their relatedness. Of these, only 31 approaches produced variants with a median of ≥5 sequences/variant.
- Selection of best approaches: We further identified ***three approaches that produced highly reproducible results***, both when classifying sequences across different subsets of data and for assigning new sequences to a variant ID (Table 1, Figure 1).
    - These three approaches consistently had the highest reproducibility metrics for six metrics assessed in various analyses.
    - For example, when variant IDs were annotated onto trees built with 10% subsets of data, the mean clade purity (proportion of sequences in a phylogenetic clade that belong to the same ID) was 88-93% for the top three approaches, whereas clade purity for RFLP and Lin+RFLP was 49 and 69%, respectively.
    - All three approaches captured viruses associated with the so-called L1C-1-4-4

variant, with >96% concordance. Use of RFLP and Lin+RFLPs to label this outbreak variant only achieved a 28% and 76% concordance, respectively.

- Genetic characterization of top three approaches (Table 1):
    - Mean within-variant genetic distance was 2.1 – 2.5%
    - Median genetic divergence between closely related variants was 2.5-2.7%. This compares to 0.5% for RFLP, showing that RFLP-types are not genetically distinct from each other
    - Over a 36-month period, these three approaches produced 115-181 variants in total, but only 27-30 were "common" variants (variants with >50 sequences). For RFLP and Lin+RFLP, respectively, there were 82 and 142 IDs in total, but only 16 and 21 "common" IDs.

a) Refine variant definition based on farm-level patterns of occurrence

- To assess the stability of variant classification during micro-evolution that may occur while a virus circulates on a farm, 73 farms with at least 4 sequences in a given year were identified from an 11-year dataset available through MSHMP. From these, 587 sequences were available (4 – 43 sequences per farm).
- An ideal classification system should minimize the occurrence of ID changes within sequence-clusters (identified from phylogenetic trees) that are clearly associated with circulation of a single virus on a farm.
- The percent of farm sequence-clusters with an ID change was 6.5 - 8.7% for the best three approaches. In contrast, ~43% of farm sequence-clusters had an RFLP change.

b) Assess adaptability of classification system to accommodate expanding genetic diversity at national scales

- As a first step, we evaluated the number of new variants per year across 11 years of data to better understand the scalability and routine updating that will be required for a classification system to accommodate expanding genetic diversity.
- For the top three approaches, there was a median of 19-37 new variants per year, but only 3 to 5 new "common" variants (those that would eventually be detected >50 times). For RFLPs, there was a median of 24 new IDs and 0 new "common" IDs per year. The low number of new RFLPs demonstrates that this classification is not scaling well to newly emerging PRRSV diversity.

2) Develop procedures for prospective implementation and expansion that would meet the needs of diagnostic labs and practitioners.

- A key feature of any new classification system is the ability to assign variant IDs to new sequences as they are generated by diagnostic labs.
- Therefore, we trained a machine learning algorithm that can take a sequence and assign it to the appropriate variant ID.
    - The trained algorithm achieves >96% accuracy when assigning sequences that are entirely external to the original dataset (i.e., sequences present in the UMN VDL dataset, but not in the MSHMP dataset used to create the variant classifications).
    - 8-10% of these external sequences could not be assigned reliably to a variant ID, likely because those variants were not present in the MSHMP dataset. This could be improved by using a more representative national dataset, such as SDRS, that

would yield a more complete view of PRRSV diversity in the U.S.

- Next steps include discussions of these results with the PRRSV nomenclature working group, the AASV PRRS Committee, major diagnostic laboratories, and practitioners. If a version of this new nomenclature is adopted, then we will work with USDA NADC to build an html-based platform for prospective implementation. We will also develop educational materials and engage in outreach activities to help stakeholders understand and utilize a new system.

Table 1. *Summary metrics for the best performing approaches for variant classification. Key differences between the performance of different approaches are shown in red and green.*

		Lin+RFLP	Best alternative methods		
	RFLP		ac.06	ac.07	ac.08
Sequences per variant-median (IQR)	6 (1-21)	4 (1-16)	11 (4-25)	11 (4-34)	14 (5-52)
Number variants (over 36 months)	82	142	181	151	115
Number "common" variants (>50 sequences)	16	21	27	29	30
Within-variant genetic distance-mean (IQR, 95th percentile)	4.3% (0.9-7.1, 9.9%)	2.5% (0.8-3.8, 6.6%)	2.1% (1.2-2.6, 4.3%)	2.3% (1.2-3.0, 4.4%)	2.5% (1.3-3.3, 5.3%)
Genetic divergence from closest-related variant-median (IQR)	0.5% (0.2-1.2%)	0.7% (0.2-1.9%)	2.5% (2.5-4.5%)	2.5% (1.6-5.0%)	2.7% (1.7-5.1%)
Assignment accuracy-internal	95.3%	93.8%	99.4%	99.2%	99.7%
Assignment accuracy-external	76.5%	80.4%	96.5%	97.6%	96.5%
% farm sequence-clusters with ID change	43.30%	NA	8.70%	8.70%	6.50%

Discussion of how results can be applied by practitioners

While phylogenetic analysis is still the gold standard for interpretation of sequence data, practitioners and field epidemiologists often find it timelier and more convenient to have a label in which they can refer to a given genetic variant as part of everyday communication and outbreak investigations. Currently, the naming method used by the industry to discriminate between sequences is RFLP-typing, sometimes in combination with an additional label corresponding to phylogenetic lineage. However, only 12 lineages and sub-lineages have been described and these are too coarse for on-farm decision-making, and using RFLP-types to refer to PRRSV-2 viruses is both outdated and often leads to misleading or even erroneous conclusions (e.g., viruses assigned to the same RFLP-type often are not closely related, and vice versa). Over 50% of survey respondents indicated that we should find an alternative to RFLPs, and an additional 38% are in favor of moving away from RFLPs, but only if there is a viable replacement that is easy to implement at the lab- and slat-level.

Our intent is not to replace lineages, as we do believe that this larger classification is useful for explorations of phenotype as well as tracking the macro-evolutionary dynamics of PRRSV. Thus, we propose to incorporate lineage into the labels utilized in the new fine-scale naming system, which will be developed with inputs from stakeholders. While having a better classification system will not solve PRRS, one is clearly needed and has been requested by practitioners for many years. A better classification system will facilitate communication about outbreaks, tracking of emerging and endemic variants across time and space, and provide the basis to group viruses into "strains" to which we can begin to measure phenotypic variation.

Figure 1. Phylogenetic trees for L1H (top row), L1C (middle row), and L1A (bottom row),which were the most common lineages during the study period. Colors in the first, second, third, and fourth columns represent classifications with the ac.06, ac.07, ac.08, and RFLP methods. Colors denoting RFLP-type are carried over across all three lineages for RFLP-types, but colors do not carry over for the other methods shown.