Final Report:

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

April 2, 2024

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(RFLP)-typing to refer to genetic variants of Porcine reproductive and respiratory syndrome virus (PRRSV)-2 is both outdated and, more importantly, can lead to misleading or even erroneous conclusions about the relatedness of PRRS viruses.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 farm-level occurrence of 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 142 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 3).
    - These three approaches consistently had the highest reproducibility metrics for the 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 often not genetically distinct from each other
    - Over a 36-month period, the best three approaches produced 115-181 variants in total, but only 27-30 were "common" variants (variants with >50 sequences). 73-84% of sequences belonged to common variants. 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.
- Based on collectively on analysis performed above, we selected the ac.07 approach as the most viable alternative to RFLPs, which is what we utilized for the following aims.

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

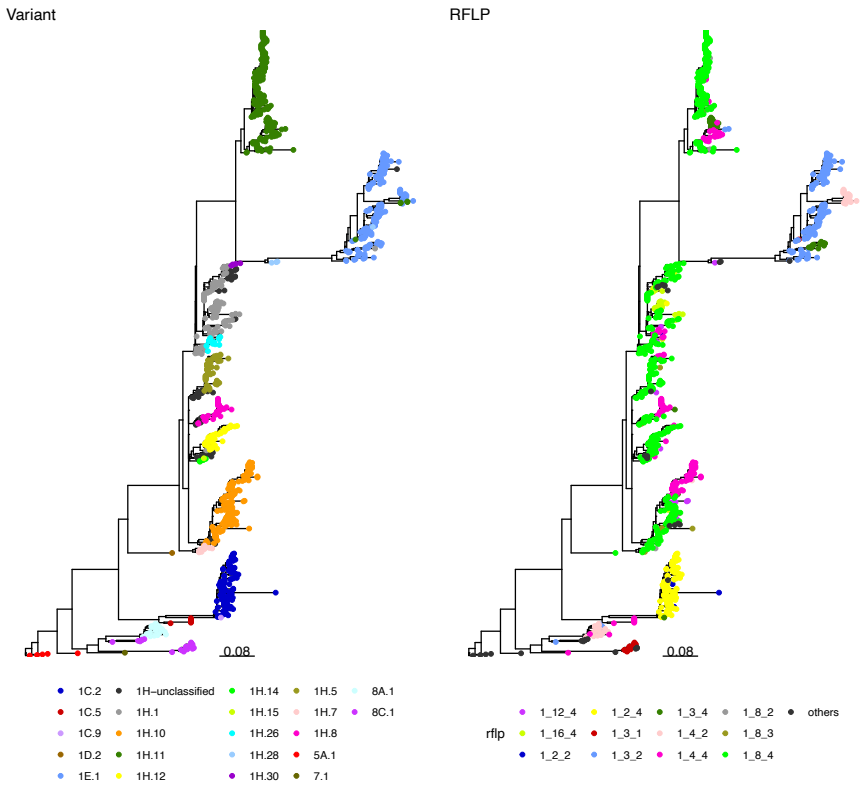
- To assess how the system would function if utilized prospectively, we initiated the classification system in 2018 with 36 months of data (2015-18), then added new data every three months up until September 2024. Each quarter, we classified all new sequences, systematically examined the tree for new variants, and examined existing variants as to whether they had expanded enough to be split. Based on these prospective assessments, we developed criteria of naming new variants that produced robust, reproductible results across time and minimized the need for splitting existing variants. Criteria for naming a new variant are that the group must have ≥5 sequences with robust support of their shared ancestry in the phylogeny (bootstrap value >85 in the tree), and that the genetic distance to the nearest named variant must be >2% (>5% if splitting an existing variant).
- There was a median of 48 active variants/year (26 of which were "common" variants, detected at least 50 times). This compares with 65 active RFLPs per year (25 of which are "common"). There was a median of 19 new variants per year (120 in total), but only ~4 new "common" variants (those that would eventually be detected >50 times). For RFLPs, there was a median of 9 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.
- For the prospective implementation of the classification system (as described in 1b), all codes have been written to automate routine quarterly updates of this system, including identification/naming of any new variants and updating the assignment algorithms to reflect new data.
    - A rudimentary web-tool for sequence assignment was created for the purpose of beta-testing the system with end-users
    - Thus far, data from three different production systems have been examined through the lens of the new classification system. The new system consistently produces a more accurate picture of the relatedness between sequences as compared to RFLPs (for example, see Figure 1).
    - A tentative naming scheme was established that is compatible with the recently L1C sub-groups recently proposed by Yim-im et al. (Yim-im, Anderson et al. 2023). The naming scheme incorporates lineage/sub-lineage, with an integer indicating the variant ID. For example, 1C.1 through 1C.5 are the groups proposed by Yim-im et al. 2023, and we continued the numbering onward from 1C.6 to 1C.9. A similar approach was applied to other sub-lineages (e.g., 1A.1 – 1A.42 includes the large group of viruses that have been dominant since the emergence of the "1-7-4" RFLP type).

Figure 1: *Phylogenetic tree of ORF5 sequences from a single production system in the U.S. Two RFLP types (right: 1-8-4 in green; 1-4-4 in pink) are intermixed in the tree. In contrast, variants (left) provide clear delineation between different clusters of viruses in the tree.*



- 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 (see appendix A for an example) 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.*

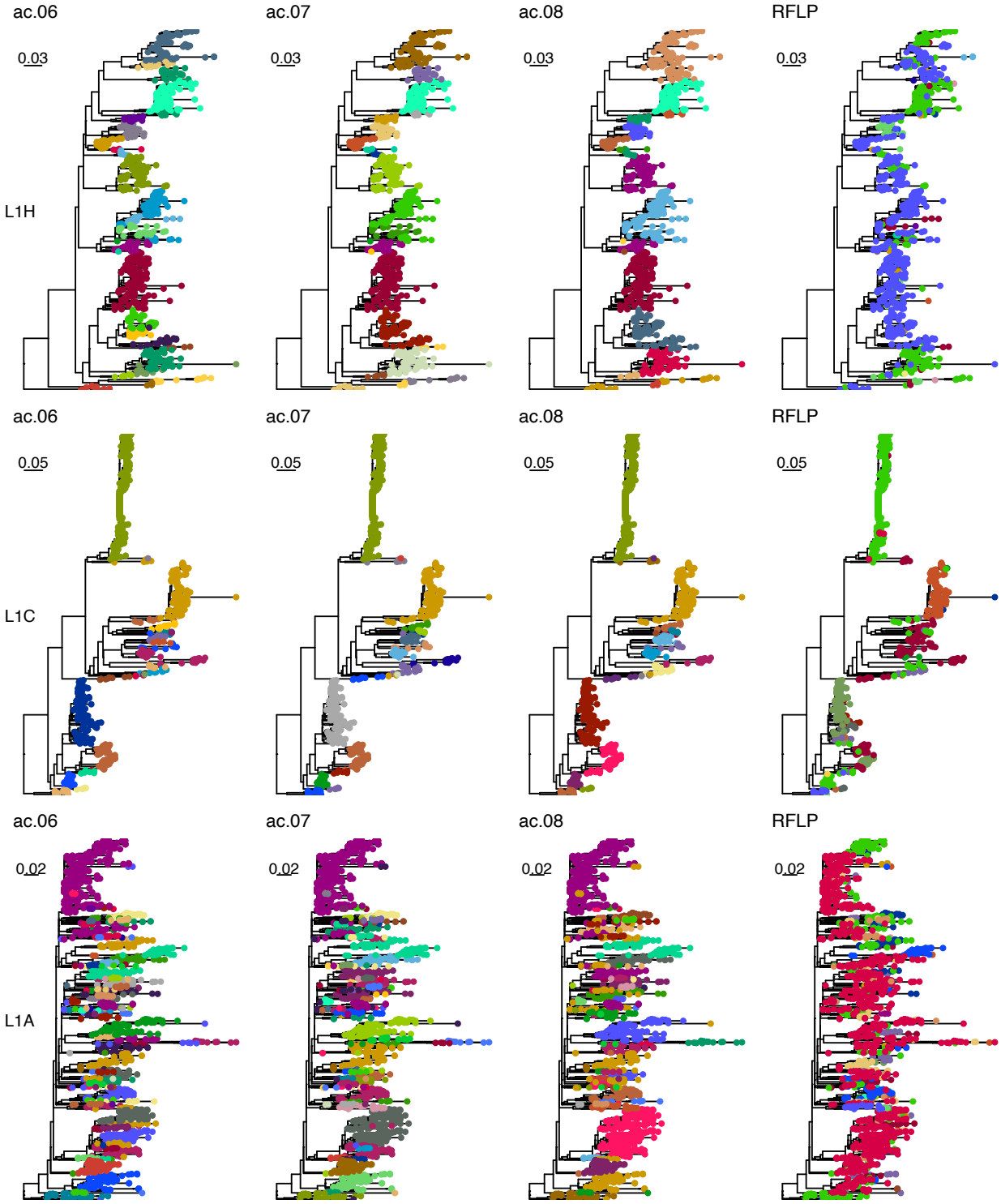
		Lin+RFL	Best alternative methods		
	RFLP	P	ac.06	ac.07	ac.08
Sequences per variant-median (IQR)	6 (1-21)	4 (1-16)	11 (4-25)	11 (4-34)	14 (5-52)
Number variants (over 36 months)	82	142	181	151	115
Number "common" variants (>50 sequences)	16	21	27	29	30
Within-variant genetic distance-mean (IQR, 95th percentile)	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 2. 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.



Detailed Results

Initial characterization of variants

Of the 140 classification approaches initially considered, 31 approaches met the initial criteria of having a median >5 sequences and <15% of sequences belonging to "rare" variants (i.e., fewer than 10 sequences/variant). 30 out of 31 approaches utilized the Average Clade method, wherein variants are defined such that the average genetic distance between sequences belonging to the same cluster must be below a specified threshold (denoted as ac.06, ac.07, etc., with the latter digits representing the genetic distance threshold). These candidates were compared to classifications based on RFLP-typing and Lineage+RFLP, for a total of 33 approaches.

As outlined in detail below, the overall best approaches were selected from amongst these 31 candidates based on the reproducibility of variant classification and the ease of assignment of new sequences. Four approaches, highlighted in green in Figure 3, were selected based on the criteria outlined below. Three of these approaches utilized the same tree-building methods (i.e., strict consensus phylogeny built with de-duplicated sequence data). Summary metrics for these three approaches are shown in Table 1. Phylogenetic trees for the three most common lineages, (L1H, L1C, and L1A) are shown in Figure 2.

Reproducibility of classification amongst different sets of data

We performed several analyses to determine the extent to which the variant classification produced above could be replicated with different sets of data. When variant clustering was performed on duplicate phylogenetic tree from different IQ-tree runs or on a detailed sub-lineage L1C tree with 15 years of data, concordance between the classifications produced for each tree was quantified through the Jaccard index. The Jaccard index ranged between 0.78 and 0.97 for duplicate trees (black in Figure 3a), and between 0.31 and 0.95 for the L1C trees (red in Figure 3a). For the latter, the Jaccard indices improved to 0.83 and 0.95 when considered trees with ≥6% threshold and were notably poor for lower thresholds, indicating a lack of reproducibility when the threshold was set too low.

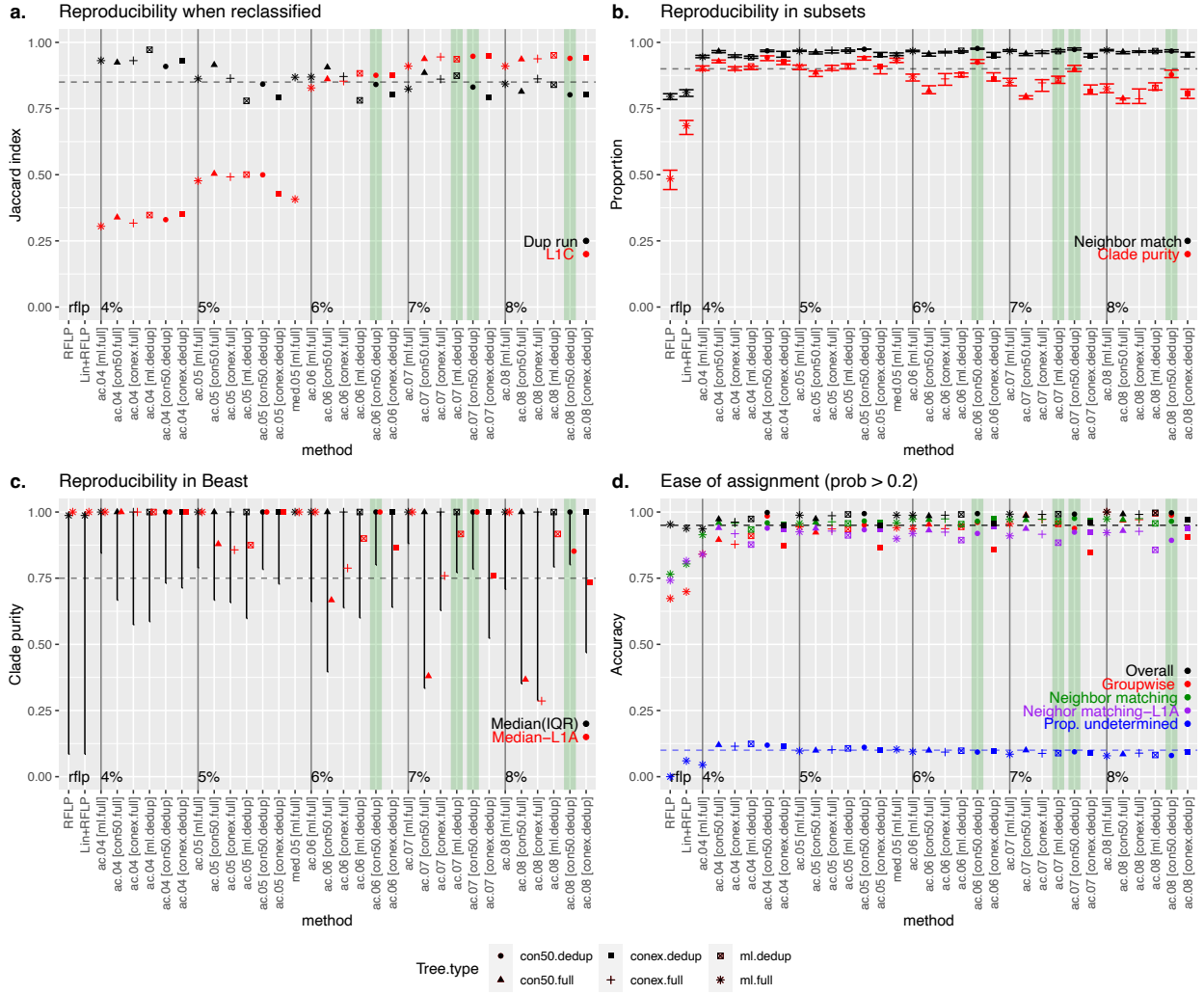
In another reproducibility analysis, trees were constructed with 10 random subsets of the 36-month dataset, and then the variant IDs from the full analysis were annotated onto the trees. We quantified the proportion of sequences whose nearest neighbor had a matching variant ID, as well as clade purity across each of the ten trees. All methods achieved high nearest neighbor matching, with >94% of sequences having a nearest neighbor that had a matching variant ID (black in Figure 3b). Only ~80% of sequences had a matching nearest neighbor when using RFLP or Lineage+RFLP. Median clade purity ranged from 79-94% across approaches, with higher purities of nearly 90% or more achieved by approaches utilizing the strict consensus method on de-duplicated data (red in Figure 3b). Median clade purities were 49% and 69% for RFLP and Lineage+RFLP, respectively.

All approaches were able to capture the *a priori* defined outbreak variant known as L1C-1-4-4, with Jaccard indices of 0.96 – 0.98. RFLP and Lineage+RFLP achieved only a concordance of 0.28 and 0.76, respectively, indicating that these labels did not reliably capture sequences associated with this outbreak.

Although variants were defined on trees built via IQ-tree, Bayesian methods such as BEAST are often considered the most robust approach. Therefore, we assessed whether the variants produced on the IQ-Trees also formed clusters with high purity on Bayesian trees. The median clade purity for variants on time-scaled Bayesian trees was essentially 1.0 in all cases,

but the lower bound of the interquartile range was more variable (black in Figure 3c) and was particularly low for RFLP and Lineage+RFLP.

Figure 3. Comparison of all approaches that produced a median variant size ≥5 sequences/variant. Best performing methods are highlighted in green.



Ease of assignment of new sequences

For each of the 33 classification approach considered in this paper, we trained a random forest algorithm to assign variant IDs to sequences. Model performance was evaluated with an internal test set (most recent 10% of data from the short-term dataset) and an external test set (sequences that were not included in the original analyses). For some sequences, the highest probability ID was quite low, indicating that the model had poor confidence in the assignment. Therefore, we tested two thresholds (prob < 0.2 or prob <0.6) for calling sequences "undetermined" rather than assigning them to an ID. Both uncertainty thresholds resulted in marked improvement in predictive performance. The 0.2 threshold improved the internal test set overall accuracy to 97.4 – 100%, and external test set accuracy to 95.4 – 97.6% (based on nearest neighbor matching). While the 0.6 probability threshold resulted in slightly higher accuracies, it

also resulted in a high percentage (25%) of undetermined sequences in the external test set. With comparable predictive performance, the 0.2 threshold resulted in just 10% of sequences classified as undetermined. Therefore, a probability threshold of <0.2 for calling sequences undetermined was applied to predictions made by the assignment algorithm.

Selection of best classification approaches

For the reproducibility analysis, the best performing approaches were defined as those that achieved a Jaccard index of >0.85 for both re-classification on the duplicate run and on the extended L1C analysis (*criteria 1)*. Nine of 33 approaches met the criteria. For the subset analysis, the best performing approaches were defined as those that achieved both clade purity and nearest neighbor matching of >0.90 (*criteria 2*). 14 approaches met this criteria. For the reproducibility analysis using trees generating by BEAST, which is often considered the gold standard for tree building, the best performing approaches were defined as those in which the lower bound of the interquartile range for clade purity was >0.75 (*criteria 3*). 10 approaches met this criteria. Finally, for ease of classification, we defined the best performing approaches as those with >0.95 overall accuracy in the internal test data as well as >0.90 for both the mean groupwise accuracy in the internal test set and nearest neighbor matching in the external test set (*criteria 4*). 21 approaches met this criteria. Values that missed the threshold by <0.01 were allowed for all criteria. Only one approach (variant.06.ac.dedup.con) met all criteria. An additional three approaches satisfied three of the four criteria, and missed only one criterion by no more than 0.05 (Figure 3). Given that one tree type (strict consensus tree with de-duplicated sequences) accounted for three of four of the best approaches, we proceeded with that tree type for the remaining analyses.

Long-term analysis

The top three variant classification approaches were applied the long-term tree (strict consensus tree with de-duplicated sequences) to evaluate the farm-level occurrence of variants across time (Objective 1a), as well as evaluate number of new variants detected per year, and number of active variants per year (Objective 1b).

Farm-level occurrence of variants

We used to long-term dataset to tabulate the number of farms, production systems, and U.S. states in which each variant was detected (Table 1). These summaries excluded rare variants (<10 sequences), which accounted for 0.8, 1.9, 3.5, 2.4, and 1.8% of sequences, respectively for RFLP, Lineage+RFLP, ac.06, ac.07, and ac.08. Variants (ac.06, ac.07, and ac.08) were found in a median of 8 to 9 farms (max 24), 3 (max 6) production systems, and 2 (max 3) states.

To assess the stability of variant classification during micro-evolution that may occur during virus circulation on farm, 73 farms with at least 4 sequences in a given year were identified from the long-term dataset. From these, 587 sequences were available (4 – 43 sequences per farm, with sequences from some farms spanning multiple years). For each pair of variants that occurred on the same farm, we measured the maximum genetic distance and maximum divergence time between those sequences to help identify situations in which sequences from a single sequence-cluster on a farm were classified as two closely related variants. For ac.06 and ac.07, all variant pairs with maximum genetic distance <0.05 and/or evolution time <2 years were manually inspected on the trees as to whether those sequences

would be more accurately represented by a single variant ID (i.e., they cluster together in the tree). Based on this manual inspection, a merging threshold of <0.02 genetic distance and <3 years divergence time was applied to all pairs of variants from the same farm. Variant pairs from the same farm that met both these conditions were flagged as farm sequence-clusters in which an ID change occurred. An ideal classification system should minimize the occurrence of such ID changes. The percent of farm sequence-clusters with an ID change was 8.7%, 8.7%, and 6.5% for ac.06, ac.07, and ac.08, respectively. In contrast, ~43% of farm sequence-clusters had an RFLP change.

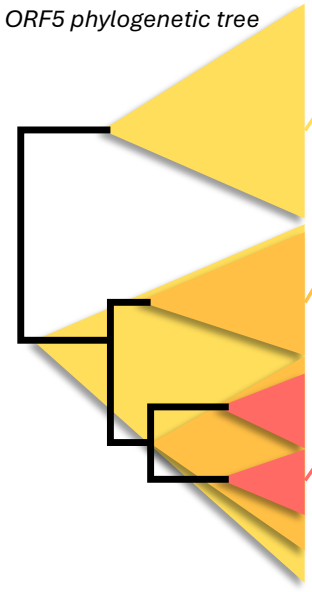
Yim-im, W., T. K. Anderson, I. A. D. Paploski, K. Vanderwaal, P. Gauger, K. Krueger, M. Shi, R. Main, J. Q. Zhang and D. Y. Chao (2023). "Refining PRRSV-2 genetic classification based on global ORF5 sequences and investigation of their geographic distributions and temporal changes." Microbiology Spectrum.

Appendix A: Draft of Fact sheet on PRRSV-2 classification

Draft: 02/20/2024

PRRSV-2 Genetic Classification

- PRRSV-2 classifications are based on genetic relationships in the ORF5 gene.
- **RFLP–typing** is unreliable: unrelated viruses may group together, and closely related ones may be labeled differently.
- **Phylogenetic classification** (lineage/variant) groups viruses into ancestral "families" based on phylogenetic trees.



Lineage

- Within-lineage genetic distance is *typically <11%*.
- Between-lineage genetic distances is *typically >10%*.
- 11 lineages have been described worldwide described up to 2023

Sub-lineage

- Within-sub-lineage genetic distance is *typically <8.5%*.
- Between-sub-lineage genetic distance is *typically >9%*.
- Within the most prevalent lineage in the U.S. (Lineage 1), there are at least 9 sub-lineages (A to J) described up to 2023

Variant

- Lineages and sub-lineages consist of many smaller groups, termed “variants”
- Sequences belonging to the same variant have an *average genetic distance of ~2.5%, but can be as high as ~5%*
- Variants are *typically >3%* different from the closest related variant
- Through time, new variants are defined based on genetic distance from other variants and phylogenetic stability of the grouping on trees**
- The number of active variants at any point of time is dynamic, given their frequent emergence and extinction

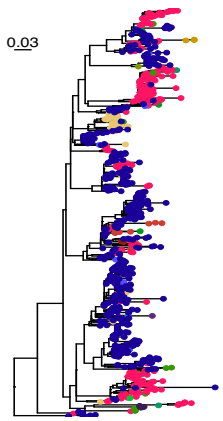
** phylogenetic bootstrap values (>85%), genetic distance from other variants (>2%), and ≥5 sequences.

What *variant* classifications *can* and *can not* tell you…

- ✓ More reliable than RFLPs at determining relatedness, and whether virus A is the same or different than virus B
- ✓ Discriminate between new and previous wild-type viruses in a farm (based on ORF5 gene)
- ✓ More useful for epidemiological investigations, such as determining possible sources of introduction and tracking between-farm spread.

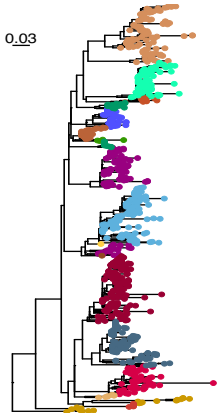
- ✗ No classification system reliably provides information on virulence or clinical picture (apparent virulence likely influenced by co-infections and other external factors)
- ✗ Classifications do not directly translate to immunological cross-protection, although viruses labeled as the same variant are more genetically homologous

Fine-scale genetic variability is better represented by variants than RFLP-typing



Tree colored by RFLP-type

- Viruses grouped by enzyme cut patterns in ORF5
- This sub-lineage 1H tree contains only two major RFLP types, which are inter-mixed on the tree. Thus, these groupings lack significant meaning.



Tree colored by Variant classification

- Viruses grouped by ORF5 sequence similarity (*typically ~2%* genetic distance to each other).
- The same sub-lineage 1H tree has numerous variants that are well-defined in three (limited inter-mixing of colors)

Created by Nakarin Pamornchainavakul and Kimberly VanderWaal, Supported by AASV and USDA