The PRRSV-2 Saga: Evolutionary and Epidemiological Dynamics of Porcine Reproductive and Respiratory Syndrome Virus 2 in the United States

A dissertation submitted to the College of Veterinary Medicine of

the University of Minnesota

by

Nakarin Pamornchainavakul

In partial fulfillment of the requirements for the degree of Doctor of Philosophy

Advisors: Kimberly VanderWaal

January 2024

copyright Nakarin Pamornchainavakul, 2024

Acknowledgements

I would like to express my heartfelt thanks to everyone who supported me throughout my PhD journey: my advisor, thesis committee, colleagues, friends, lecturers, and administrative staffs at the University of Minnesota; collaborators at the University of Edinburgh; and my lovely family at home and Thailand.

Advisor: Kimberly VanderWaal

Thesis Committee: Montserrat Torremorell (Chair), Samantha Lycett, Cesar Corzo, and Declan Schroeder

VanderWaal's current and former laboratory members: Igor Paploski, Dennis Makau, Julia Baker, Kaushi Kanankege, Anna Munsey, Umanga Gunasekera, J. Trevor Vannatta, and Rahul Bhojwani

Collaborators: Albert Rovira, Maxim Cheeran, Mariana Kikuti, Jing Isaac Huang, Andrea Doeschl-Wilson, Rowland Kao, and Daniel Balaz

Family: Kanokpan and Rosalind Tsriwong, Pigpen, and Lilly Pepper

Funding:

1. The joint US-UK NIFA-NSF-NIH-BBSRC Ecology and Evolution of Infectious Disease awards [EEID project #2019-67015-29918 and BB/T004401/1]
2. The University of Minnesota College of Veterinary Medicine Signature Programs [grant #MIN-62-133]
3. The Critical Agricultural Research and Extension Program [grant #2018-68008-27890]
4. The intramural research program of the U.S. Department of Agriculture, National Institute of Food and Agriculture, Data Science for Food and Agricultural Systems Program [grant #2023-67021-40018]
5. The University of Minnesota Swine Disease Eradication Center (SDEC) and the Swine Health Information Center (SHIC) as the funding agency for MSHMP
6. The Royal Thai Government Scholarship

Dedication

I dedicate this dissertation to my Family, Democracy, and Star Wars.

Abstract

Porcine reproductive and respiratory syndrome (PRRS) has inflicted substantial economic losses on the US swine industry over the past three decades, driven by the main etiological agent, PRRSV-2, which continuously evolves and spreads despite control efforts. Enhancing disease control measures necessitates an understanding of evolutionary dynamics of PRRSV. By leveraging virus genetic data and bioinformatics tools, this dissertation aims to unravel how PRRSV-2 has adapted, persisted, and disseminated within the U.S. Chapter 1 provides a background of the disease, the virus itself, and the existing knowledge gaps. Chapter 2 employs nationwide PRRSV-2 genetic and geographic data to uncover the patterns of disease spread and the dynamics of the virus population within the U.S. In Chapter 3, we conduct an in-depth investigation into between-farm transmission of an emerging PRRSV-2 sub-lineage within a specific, swine-dense region, using genetic and animal movement data. Chapter 4 utilizes data from the largest active PRRS monitoring program in the U.S. to forecast the potential emerging variants. Finally, in Chapter 5, we pinpoint the origin of a novel PRRSV-2 variant through an advanced analysis of whole-genome sequences.

Chapter 2 revealed a cyclical pattern of sub-lineages contributing to the overall PRRSV-2 population and a shift across time in major hotspots for inter-regional spread. In Chapter 3, we narrow our focus to intra-regional spread by applying molecular epidemiological tools to construct farm-to-farm transmission networks for an emerging PRRSV-2 sub-lineage. These networks allowed us to examine factors contributing to between-farm spread and highlighted the significance of live animal movement, while recognizing that most transmission events remained unexplained. Both Chapters 2 and 3 characterize the periodic emergence of novel genetic variants of PRRSV-2, and anticipating such emergence events could aid in more strategic disease control. Chapter 4 demonstrated the utility of phylogenetic branching patterns and putative antigenic

differences as early indicators of variant emergence. Finally, in Chapter 5, we expand the discussion of variant emergence from the ORF5 gene to the whole genome perspective. Analysis of whole-genome sequences unveiled a recombinant ancestor for an emerging variant of concern and emphasized the role of genomic recombination in PRRSV-2 evolution. Ultimately, our findings address novel insights into PRRSV-2 evolution and epidemiology at various geographic scales, providing beneficial guidance for targeted and early-response PRRS mitigation strategies in the U.S.

Table of contents

List of tables

List of figures

“Prelude”

Chapter 1: Introduction

Porcine reproductive and respiratory syndrome (PRRS) stands as a significant threat to the swine industry, both in the United States and on a global scale (Lunney et al., 2010; VanderWaal & Deen, 2018). This formidable challenge is attributed to the porcine reproductive and respiratory syndrome virus (PRRSV), the causative pathogen, which interrupts the innate immune system (Calzada-Nova et al., 2011), weakens adaptive immunity (Rahe & Murtaugh, 2017), induces cell death (Costers et al., 2008; Malgarin et al., 2021), and triggers inflammatory responses (Alex Pasternak et al., 2020; J. Li et al., 2017; Thanawongnuwech et al., 2004) in infected pigs. These multifaceted effects lead to a range of clinical manifestations, including late term abortion and pre-mature farrowing in sows, high mortality in piglets, and failure to thrive in growing pigs (The OIE AD HOC group on porcine reproductive respiratory syndrome, 2008). These clinical impacts during PRRS outbreaks have a direct and adverse effect on pig production and income, impacting not only individual farm-level economics (Valdes-Donoso et al., 2018) but also national-scale production, with losses exceeding US $600 million per year in the U.S. alone (Holtkamp et al., 2013; Neumann et al., 2005). Although many control efforts such as herd immunization, pig flow management, and farm biosecurity have been attempted to eliminate the disease over the past three decades (Arruda et al., 2016; Corzo et al., 2010), PRRS remains the most problematic infectious disease in the North American swine industry. PRRS can be found endemically or epidemically in almost any part of the world where pigs are raised (*Porcine Reproductive and Respiratory Syndrome: OIE - World Organisation for Animal Health*, n.d.), which makes it amongst the most prevalent swine diseases in the modern era (Lunney et al., 2010). Here, we provide a brief overview of PRRSV covering its fundamental genomic structure, evolutionary mechanisms, diversity, and potential immune escape strategies. Finally, this chapter concludes by highlighting various

unresolved areas of knowledge and sets the stage for subsequent chapters in this dissertation, which seek to address and fill these gaps in understanding.

PRRSV genome and functions

Undoubtedly, one of key success factors of PRRS persistence in swine populations is the nature of the virus itself. This enveloped virus, classified within the *Arteriviridae* family, harbors a single-stranded positive-sense RNA genome. Its structural and functional organization closely resembles that of other viruses belonging to the *Nidovirales* order, a group which encompasses well-known pathogens, including Coronaviruses (Adams et al., 2016; Saberi et al., 2018). Three-quarters of the 15 kb PRRSV genome is comprised of open reading frames (ORFs) 1a and 1b genes, which code for non-structural proteins (NSPs), while the remaining portion of the genome is the nested set of sub-genomic mRNAs consisting of ORFs 2 – 7 genes, which synthesize virion structural proteins (Meulenberg et al., 1993; Pasternak et al., 2006). NSPs are necessary for viral replication since they form a replication and transcription complex that modulates both genomic and sub-genomic replicative cycles (Snijder et al., 2013). Notably, studies suggest some NSPs are relevant to viral pathogenicity. For example, discontinuous deletions of 30 amino acids in NSP2 is a unique marker for the identification of highly pathogenic PRRSV (HP-PRRSV, lineage 8) (Guo et al., 2018; Tian et al., 2007), NSPs 3 – 8 possibly contain major virulence factors (Y. Fang & Snijder, 2010; Kwon et al., 2008), and changes in NSPs 9 and 10 are associated with an increase in fatality in piglets infected by HP-PRRSV (Y. Li et al., 2014). Structural proteins consist of the nucleocapsid (N) protein that encloses the virus genetic material and the envelope transmembrane proteins. The major components of PRRSV envelope are the glycoprotein 5 (GP5) and the non-glycosylated M protein encoded by ORFs 5 and 6 genes, respectively, that form a disulfide-bonded heterodimer, whereas GP2, GP3, and GP4 from ORFs 2 – 4 genes form the minor glycoprotein complex (Wissink et al., 2005). These virion surface proteins directly interact with host receptors on macrophages and

PRRSV's target cells, and thus are involved in priming an infection and virus neutralization by the host (Popescu et al., 2017). Thus, genetic variation across the genome can affect virulence and anamnestic immune responses.

PRRSV evolutionary mechanisms

Several evolutionary mechanisms drive PRRSV genetic diversity. Like other RNA viruses, its RNA polymerase lacks proofreading ability. This creates mutated offspring that are 1 – 2 nucleotides different from their parent in every generation (Duffy, 2018; Schneider & Roossinck, 2001; Vignuzzi & Andino, 2012). PRRSV is estimated to have amongst the highest nucleotide substitution rates ($6.6 \times 10^{-3} – 1.3 \times 10^{-2}$ substitutions/site/year) compared with other RNA viruses ($10^{-3} – 10^{-5}$ /site/year) (Hanada et al., 2005; Paploski et al., 2021). At a larger scale, PRRSV genetic diversity is also created through recombination, a mechanism by which portion of the genome is exchanged between viruses. RNA viruses' sub-genomic negative-sense RNA synthesis normally adopts copy choice recombination, which is guided by similar nucleotide sites between RNA templates (Simon-Loriere & Holmes, 2011; Snijder et al., 2013). Those template switching sites likely facilitate genomic recombination within a viral population but less frequently produces non-homologous recombinants unless two distinct viral variants infect the same cell (Murtaugh et al., 2010). Non-homologous PRRSV recombination does seem to contribute to the genetic diversification as recombination events have been repeatedly detected between distinct wild-type or even vaccine strains (Anping Wang, Qi Chen & Darin Madson, Karen Harmon, Phillip Gauger, Jianqiang Zhang, 2019; Eclercy et al., 2019; Murtaugh et al., 2010; Shi, Holmes, et al., 2013; H. Zhao et al., 2017). Viral genetic diversity is also determined by the balance between viral fitness and selection (Lauring & Andino, 2010), which is complicated for PRRSV given that fitness depends on how quickly the virus population responds to selection pressures such as host immunity.

PRRSV genetic diversity and classifications

PRRSV is classified into two species based on their origin and nucleotide identity: *Betarterivirus suid 1* (type 1) and *Betarterivirus suid 2* (type 2) from European and North American origins, respectively (Walker et al., 2021). In the U.S., the type 2 virus, also known as PRRSV-2, is the predominant viral species responsible for most PRRS outbreaks (Paploski et al., 2019; Shi, Lam, Hon, Hui, et al., 2010). PRRSV-2 diversity has been conventionally quantified through variation in ORF5 gene, a 603-nucleotide gene encodes the viral GP5 (Paploski et al., 2019; Shi, Lam, Hon, Hui, et al., 2010; Wesley et al., 1998). GP5 is one of the major envelope proteins (Gorbalenya et al., 2006) involved in *in-vivo* neutralization (Wissink et al., 2003), infectious virion assembly (Wissink et al., 2005), and host cell entry (Delputte & Nauwynck, 2004). Many studies have found that ORF5 gene evolves under diversifying selection (Chen et al., 2016; Costers et al., 2010; Hanada et al., 2005; Paploski et al., 2019; Storgaard et al., 1999), leading to high genetic variation. Such immunogenic importance coupled with its high genetic variability between viruses (Kim et al., 2013; Paploski et al., 2019; Shi, Lam, Hon, Murtaugh, et al., 2010) make ORF5 gene a good marker for PRRSV-2 genetic diversity. Therefore, ORF5 is the most common gene that is sequenced and deposited to either public or private nucleotide sequence databases.

Based on ORF5 gene, genetic variation of PRRSV-2 was originally characterized by restriction fragment length polymorphisms (RFLP). Briefly, RFLP typing provides a three-digit code according to the cutting pattern of three restriction enzymes applied to ORF5 gene PCR products (Cha et al., 2004). While RFLP patterns have been conventionally used for reporting PRRSV epidemic strains due to its rapidity and inexpensiveness, grouping viruses by RFLP type may contradict the actual genetic distance between viruses and obscure ancestral relationships amongst variants (Larochelle et al., 2003). In 2010, PRRSV-2 was systematically classified into nine lineages (lineages 1 to 9)

based on phylogenetic analysis of ORF5 gene sequences in publicly available databases (Shi, Lam, Hon, Murtaugh, et al., 2010). Pairwise genetic distances are typically <11% for sequences within the same lineage, with typically >10% divergence between lineages (Paploski et al., 2021; Yim-im et al., 2023). This laid the foundation for robust classification and opened the door to various integrative methods centered on virus evolution and so-called phylodynamic analyses. Major branches of phylodynamics include phylogeography to reconstruct patterns of geographical spread of the pathogen, coalescent theory to describe changes in ancestral population sizes through time, and selective pressure analysis to identify amino acid sites that mutate under selection (Lemey et al., 2009; Volz et al., 2013; Z. Yang et al., 2000). Bayesian phylogeographic analysis, for example, revealed an introduction of Canadian PRRSV-2 into the Midwest USA in the late 1990s, which caused a major genetic shift in the dominant circulating strains in the U.S. (Shi, Lemey, et al., 2013). Incorporation of spatial information into phylodynamic models demonstrated differences in population growth, evolutionary, and geographic dispersal rates between endemic and emerging strains in the U.S. Such data critically can inform disease intervention using risk-based approaches (Alkhamis et al., 2017). Recently, diversification of PRRSV-2 lineages was thoroughly explored, and the most prevalent lineage, lineage 1 (L1), was further divided into sub-lineages (L1A to L1H). Multiple L1 sub-lineages have been shown to co-circulate in the U.S., following a general pattern of sequential turnover of the dominant sub-lineage through time (Paploski et al., 2019, 2021).

Although most studies of PRRSV-2 diversity have focused on ORF5 gene given its significance and availability, genetic distances solely using this gene do not always consistently translate to cross-protection or neutralization phenotypes (J. Li & Murtaugh, 2012; Martínez-Lobo et al., 2011). Furthermore, NSP2 has the highest genetic diversity and contains insertions-deletions that may be markers for highly pathogenic strains (Yoshii et al., 2008; F. Yu et al., 2020). ORF7 gene

encodes the most immunodominant antigen, the N protein, which is a suitable candidate for serological tests (Hao et al., 2011). Thus, as with any organism, a whole genome perspective is needed to fully understand the evolutionary history of PRRSV-2. For many prokaryotes and viruses, phylogenetic incongruence is often observed between their genomic and sub-genomic trees mostly caused by horizontal gene transfer (HGT) (Chan et al., 2013; Q. Zhang et al., 2017), mediated in this case by recombination. There is disagreement of strain grouping based on ORF5 and ORF7 gene tree topologies, and mosaic genome structure (Martin-Valls et al., 2014) indicates that frequent recombination in PRRSV through HGT can confound phylogenetic interpretation. Unfortunately, the number of WGS submitted to public databases is very limited in comparison with ORF5 gene. For these reasons, it is challenging to apply phylodynamic analysis to PRRSV WGS data.

PRRSV evolution and immune evasion

Genetic variation reveals not only evolutionary dynamics of PRRSV but also viral adaptation to host immune response. Without immunization, an infected pig may suffer from respiratory disease complex because of immunosuppression since the viral infection occurring in macrophage impairs both innate and adaptive immune induction (Basta et al., 2000; Meier et al., 2003; Thanawongnuwech et al., 2000). Even with immunization, a pig may not be cross-protected against heterologous strains (Montaner-Tarbes et al., 2019; Murtaugh & Genzow, 2011). Thus, disease control efforts using herd immunization can be hindered by diversification of PRRSV, which continually creates novel strains. One way PRRSV can escape antibody neutralization is by modifying its envelope proteins by glycosylating an asparagine (N) amino acid, which alters protein conformation (Ansari et al., 2006; Marshall, 1974). When this phenomenon occurs in glycoprotein-rich regions adjacent to neutralizing epitopes, *N*-linked glycosylation sites may mask an epitope and alter antibody accessibility (Ye et al., 2000). Other viruses such as human and simian

immunodeficiency viruses, influenza virus, and hepatitis C virus also possess this glycan shielding property in their heavily glycosylated surface proteins (M. Zhang et al., 2004). For PRRSV-2, reintroduction of *N*-linked glycosylation sites into either GP3 or GP5 of a strain that was naturally susceptible to neutralization protected the virus from antibody neutralization (Vu et al., 2011). From an epidemiological perspective, a new set of *N*-linked glycosylation sites on GP5 was also found to coincide with the emergence of novel sub-lineages or epidemic variants in the U.S., emphasizing the role of this mechanism in PRRSV-2 macro-evolutionary dynamics (Paploski et al., 2022).

As of now, conferring a neutralizing antibody against PRRSV remains the ultimate goal in vaccination due to the limited understanding of cell-mediated immunity (Loving et al., 2015). Amongst the various PRRSV antibody epitopes, the GP5 epitope B has been widely suggested as a target for broad neutralization (Pirzadeh & Dea, 1997; Popescu et al., 2017; L. Yang et al., 2000). Apart from that, an alternate signal peptide cleavage site in some PRRSV GP5 creates the non-neutralizing or decoy epitope, epitope A, that attracts most of the antibodies interacting with GP5 in the early stage, thereby weakening and delaying the actual virus neutralization (L. Fang et al., 2006; Ostrowski et al., 2002). Hypervariable region 2 (HVR-2) of GP5 has been proposed to be a third epitope, epitope C, in which antibody binds and probably blocks epitope B accessibility. Unlike epitope B which is thought to be broadly neutralizing, epitope C is a target for homologous neutralization, and mutations in this epitope may help the virus evade neutralization by homologous antibodies (Popescu et al., 2017). These mechanisms are presumably associated with the variable efficacy of vaccines against PRRS.

Knowledge gaps

In summary, PRRSV genetic data, encompassing whole genomes or specific genes like ORF5 gene, provide a valuable resource for generating

insights that can improve PRRS mitigation strategies. Past research has demonstrated the utility of such genetic information in various ways, such as quantifying virus variation (Paploski et al., 2019, 2021; Shi, Lam, Hon, Murtaugh, et al., 2010), investigating the geographical spread and associated risk factors (Alkhamis et al., 2017; Makau, Alkhamis, et al., 2021; Shi, Lemey, et al., 2013), and estimating the dynamics of positively selected sites within the virus genome, which may evolve under immune pressure (Costers et al., 2010; Delisle et al., 2012; Do et al., 2016; Hanada et al., 2005; Hao et al., 2011; Hu et al., 2009; Paploski et al., 2019, 2021). Nonetheless, fundamental questions remain unanswered. For example, we have yet to explore and update the historical and current patterns of contemporary PRRSV spread at a national level. In detail, what are the major factors driving the spread, especially between farms? Can we predict the emergence of PRRSV? Furthermore, what benefits can we derive from implementing PRRSV whole genome sequencing in outbreak investigations? These unaddressed queries offer an opportunity to leverage existing PRRSV genetic data to enhance our understanding of this critical issue.

Here, this dissertation endeavors to bridge such knowledge gaps within the field by employing a comprehensive analysis of PRRSV genetic data and associated metadata. Our overarching aim is to illuminate the patterns of PRRSV adaptation, persistence, and spread within the U.S. swine industry, spanning from a national to local scale. We performed a diverse array of approaches tailored to the distinct data characteristics and the specific research objectives of each chapter (Figure 1.1). Throughout the following chapters, we concentrate primarily on PRRSV-2 lineage 1 (L1), as it has been the predominant circulating lineage within the U.S. during the study period (Trevisan et al., 2022; Paploski et al., 2019). Chapter 2 applies phylodynamic analyses based on the most extensive ORF5 gene sequencing data within the U.S. to date to illustrate a nationwide geographical spread and demographic dynamics of PRRSV-2 L1. Chapter 3 delves into potential drivers of the spread documented in Chapter 2 by

unraveling the intricate web of novel PRRSV-2 L1 variant transmission between farms situated in a swine-dense region and the key determinants associated with these connections, utilizing an integrative approach that combines ORF5 gene data with animal movement data. In Chapter 4, we showcase our ability to predict potential emerging PRRSV-2 L1 variants, similar to those emergence events described in Chapter 2 and 3, by harnessing massive ORF5 gene data sourced from an active national PRRS monitoring database. Finally, Chapter 5 underscores the utility of whole genome sequencing as a critical piece of evidence, shedding light on the evolutionary mechanisms behind the emergence of PRRSV-2 L1 variant in the Midwestern U.S. The methodologies and findings presented in these studies are poised to serve as alternative tools and offer invaluable insights, ultimately contributing to the advancement of PRRS prevention and control strategies at various levels within the United States.

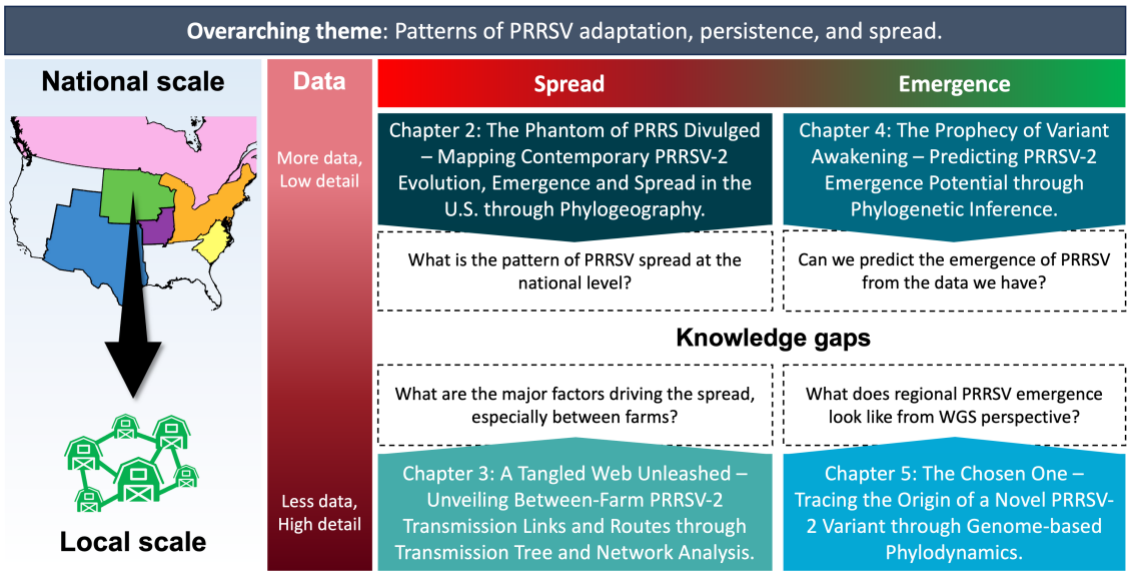


Figure 1.1: Theoretical framework of the dissertation showing how the chapters address each knowledge gap at different geographical scales and data availability under the overarching theme.

“The Phantom of PRRS Divulged”

Chapter 2: Mapping Contemporary PRRSV-2 Evolution, Emergence and Spread in the United States through Phylogeography.

Material adapted from a published article in Pathogen 12(5) (2023), doi: 10.3390/pathogens12050740

Mapping the Dynamics of Contemporary PRRSV-2 Evolution and Its Emergence and Spreading Hotspots in the U.S. Using Phylogeography.

Nakarin Pamornchainavakul, Igor A. D. Paploski, Dennis N. Makau, Mariana Kikuti, Albert Rovira, Samantha Lycett, Cesar A. Corzo, and Kimberly VanderWaal

2.1: Introduction

PRRS first appeared in the United States in the states of North Carolina, Minnesota, and Iowa in 1987 – 1988 (Keffaber, 1989). After more than three decades, 20 – 30% of sow farms throughout the U.S. pig industry still experience PRRSV-2 outbreaks each year (The Morrison Swine Health Monitoring Project, 2022). The persistence of PRRSV-2 is characterized by the cyclical emergence of new genetic variants of the virus (Paploski et al., 2021), whose spread is then facilitated by continuous movement of pigs between herds as part of multi-site, vertically integrating production systems. New variants typically emerge in a particular geographic area and then disseminate widely in the industry through routine animal movements (Lee et al., 2017; Makau et al., 2022; Pileri & Mateu, 2016; VanderWaal et al., 2020). The location of emergence and subsequent patterns of spread are crucial information guiding where prevention measures should be strengthened to limit pathogen dispersal. Phylogeography, which utilizes evolutionary relationships amongst viral genetic sequences to reconstruct ancestral locations and migrations, can be a useful tool to address these questions.

Since 2010 when PRRSV-2 lineages were firstly defined (Shi, Lam, Hon, Murtaugh, et al., 2010), lineage 1 (L1) appearance has increased from <40% to >60% of sequenced viruses, while other lineages have been decreasing over time (Paploski et al., 2019). It is worth noting that the majority of non-L1 viruses detected through sequencing in the past decade (20 – 30% of sequenced viruses) are L5 and L8 (Trevisan et al., 2022), which have been widely used as commercial live-attenuated vaccines, whereas wild-type non-lineage 1 sequences account for <1% of sequences. In addition, lineage 1 has diverged into several subpopulations, i.e., sub-lineages L1A to L1H (typical pairwise genetic distances <8.5% for sequences belonging to the same sub-lineage, and >9% divergence between sub-lineages (Paploski et al., 2021). The prevalence of different sub-lineages varies geographically and temporally, with cyclic population expansions and contractions of particular groups of viruses observed continuously (Paploski et al., 2021).

Although patterns of L1 sub-lineage emergence and turnover are well described (Paploski et al., 2019, 2021), the geographic origin and spreading hotspots of contemporary L1 viruses have not been determined. Such information is crucial for PRRS prevention, management, and containment. The most recent analysis of PRRSV-2 phylogeography in the U.S. used sequences up to 2011 (Shi, Lemey, et al., 2013), prior to when L1 became the dominant lineage in the U.S. Thus, this earlier work might not be a representative of the current epidemiologic situation. To fill these gaps, our objective is to provide an updated analysis of the spatio-temporal dynamics of PRRSV-2 L1 spread and population growth over three decades in the U.S., with particular emphasis on understanding the spatiotemporal dynamics underpinning the continued diversification of L1 into numerous sub-lineages. We inferred phylodynamics of each L1 sub-lineage and identified potential selection pressures associated with phylogenetic divergence that may help explain the overall L1 phylogeography. Given the large size of the aggregated U.S. PRRSV ORF5 gene sequence

dataset, we also evaluated the robustness of our results across various sub-sampling strategies that aimed to generate spatial-temporal representative subsets for analysis. This study advances our understanding on large-scale geographic expansions and evolutionary dynamics of the virus which may help answer why PRRSV-2 persistently circulates within the U.S

2.2: Materials and Methods

Data sources

Three data sources, the National Center for Biotechnology Information GenBank (NCBI, 1991 – 2021), the University of Minnesota Veterinary Diagnostic Laboratory (UMN VDL, 2004 – 2021) and the Morrison Swine Health Monitoring Project (MSHMP, 2010 – 2021), were accessed in October 2021 to gather PRRSV-2 ORF5 gene sequences collected in the U.S. Briefly, the UMN VDL has generated virus sequences from throughout the U.S. as part of services rendered to primarily US-based industry clients. MSHMP is a voluntary program established in 2011 that archives a variety of swine disease data from over 35 participating production systems, capturing data from more than 50% of the U.S. breeding population (*MSHMP History | College of Veterinary Medicine - University of Minnesota*, n.d.). Outbreaks identified and reported to MSHMP may have an accompanying ORF5 gene sequence. Sequences reported to MSHMP are typically produced by University of Minnesota, Iowa State University, South Dakota State University, or Kansas State University VDLs. All available sequences from Canada (only found in the NCBI GenBank) were also included given that routine transport of hogs between the U.S. and Canada (Economic Research Service & USDA, 2022) might lead to transboundary PRRSV transmission.

Inclusion criteria for sequences were completeness of sequence (> 580 nucleotides) and availability of sample collection date and location (i.e., state in the U.S. where a sample was collected or, in the case of the UMN VDL, state

where the client submitting the sample was located). Primary deduplication by sequence identification number was done between the MSHMP and the UMN VDL datasets. Subsequently, the aggregated dataset of 40,365 ORF5 gene sequences were aligned using pairwise local alignment applied in MAFFT v.7.310 (Katoh, 2002), and the alignment was used to build an approximately-maximum-likelihood tree using FastTree v.2.1.10 (Price et al., 2010). PRRSV-1 sequences detected in the tree were excluded before realigning only PRRSV-2 ORF5 gene sequences onto a PRRSV-2 reference (GenBank accession no. NC_038291.1) using pairwise codon alignment via VIRULIGN (Libin et al., 2019). Repeated sequences from localized outbreaks caused by highly related viruses (defined by same exact collection date, collected within the same state, and 100% nucleotide similarity) were deduplicated. After filtering out sequences containing ambiguous nucleotides (n = 4,644), gaps (n = 530), or with signals of potential recombination (n = 4) consistently detected by all seven methods implemented in in RDP5 (D. P. Martin et al., 2021), wherein a fully exploratory (all sequences compared to all others) recombination analysis was performed using the methods RDP (D. Martin & Rybicki, 2000), GENECONV (Padidam et al., 1999), MaxChi (J. M. Smith, 1992), BootScan (D. P. Martin et al., 2005), SiScan (Gibbs et al., 2000), Chimaera (Posada & Crandall, 2001), and 3Seq (Lam et al., 2018), the curated 29,554 PRRSV-2 ORF5 gene sequences were classified into (sub-)lineage by measuring pairwise distance between a sequence to each (sub-)lineage's anchors, as described elsewhere (Paploski et al., 2019, 2021). Non-L1 sequences were excluded from further analysis. Ultimately, only the alignment of 19,395 PRRSV-2 L1 ORF5 gene sequences was used for further analyses (Figure 2.1).

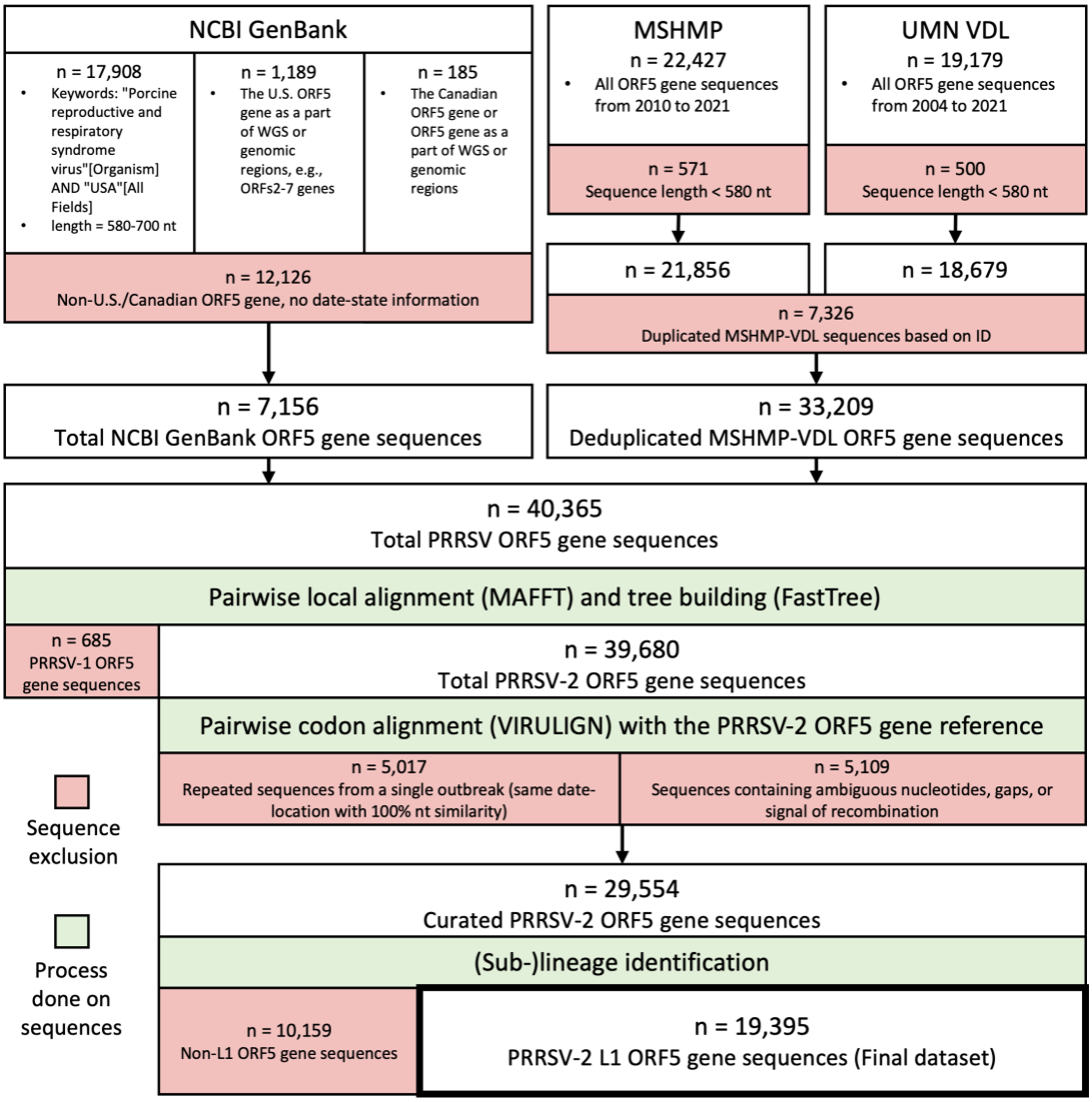


Figure 2.1: PRRSV ORF5 gene sequences gathering and filtering process.

Subsampling

Geographic regions utilized for discrete-space phylogeography were adapted from the Swine Health Information Center's (SHIC) regions, with boundaries creating divisions between major pork producing areas in the U.S. Due to a high number of available sequences, SHIC's region 3 (Midwest) was

subdivided into two regions according to the spatial distribution of the pig population (National Agricultural Statistics Service, 2019; The Swine Health Information Center, 2022). Regions included the Southwest (SW, n = 1,057 sequences), Upper Midwest (UMW, n = 10,384 sequences), Central Midwest (CMW, n = 827 sequences), Northeast (NE, n = 473 sequences), East (E, n = 6,554 sequences), and Canada (CAD, n = 80 sequences). The Western (W, n = 12 sequences) and Southeastern (SE, n = 8 sequences) U.S. were not included in the analysis since less than 20 sequences were available from those regions, likely reflecting low pig populations (National Agricultural Statistics Service, 2019). Given that one data source is a Minnesota-based diagnostic lab, data availability was highly biased towards the Upper Midwest. Imbalances in the number of sequences per region can create analytical biases that influence ancestral state reconstructions produced by phylogeographic models, wherein the model is falsely confident that the overrepresented region is the ancestral region. To diminish this bias and for computational feasibility, the L1 alignment (n = 19,395) was sampled 5 times (500 sequences per subset) using a spatio-temporal stratified uniform sampling method with year and region strata. This approach strives to equalize the number of samples included from each region/year, hence an equal number of sequences were drawn from the available sequences per region for each calendar year ranging between 1991 to 2021, except for some years that a few (<5) sequences or no sequence was available from particular regions (Figure 2.2). Through this approach, the median sequences/region/year was three (IQR = 0 – 4) and the total unique sequences utilized across the five random data sets was 1,765, or 9.1% of all L1 sequences.

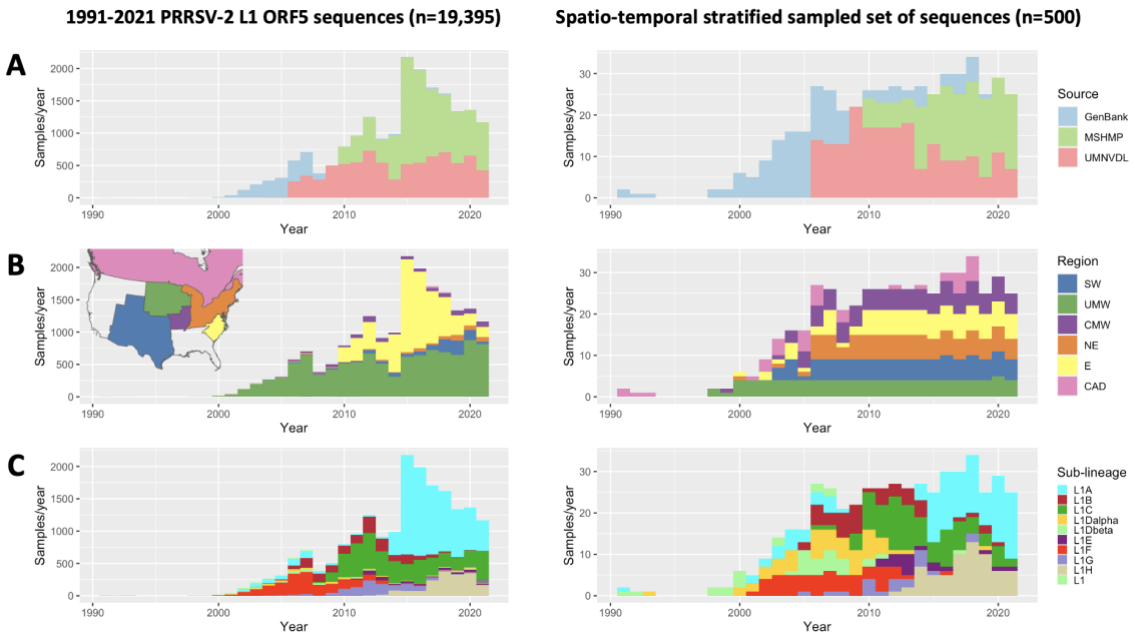


Figure 2.2: Temporal distribution of the full PRRSV-2 L1 dataset (Left) and an example of spatio-temporal stratified subsampled dataset (Right) colored by (A) source, (B) sampling location (region), and (C) pre-determined sub-lineage.

As there is some uncertainty about the most appropriate sub-sampling strategy for addressing sampling bias (De Maio et al., 2015; Kalkauskas et al., 2021), two other subsampling approaches were also conducted five times each. For uniform spatial stratified sampling, sub-sampling was performed as described above but ignoring year. For proportionate stratified sampling, the number of sequences included per region was set to be proportionate to the relative pig population in each region and state, as reported in the 2017 agricultural census (National Agricultural Statistics Service, 2019). A comparison of the spatio-temporal representation of all three subsampling approaches is shown in supplementary materials. (Supplementary Figures S2.1, 2.2)

Phylogeographic analyses

Origin and frequency of inter-regional spread of PRRSV-2 Lineage 1 and its sub-lineages were estimated from the subsampled sequences via Bayesian

discrete phylogeography, also called discrete trait analysis (DTA). This analysis treats the sampling location as a discrete trait, then estimates ancestral locations and geographic migrations of the virus from the trait transitions across the viral evolutionary tree through continuous-time Markov chain (CTMC) modeling (Lemey et al., 2009). To do that, time-scaled phylogenetic trees with phylogeographic inference were reconstructed from each subsampled set using BEAST v.1.10.4 (Suchard et al., 2018). The models used for the analysis were the general time reversible with gamma plus invariant site heterogeneity (GTR + I + G) as the nucleotide substitution model, the uncorrelated relaxed clock (Drummond et al., 2006) with log-normal distribution as the molecular clock model, and the non-reversible CTMC for the asymmetric discrete trait substitution model (Lemey et al., 2009). We additionally inferred viral population dynamics by specifying the Bayesian Skygrid's Gaussian Markov random field (GMRF) model (Gill et al., 2013) as a coalescent prior. Prior to Bayesian analysis, the temporal signals of each subsampled set were checked with TempEst v.1.5.3 (Rambaut et al., 2016) by analyzing the root-to-tip distances in maximum likelihood trees built by IQ-TREE (Trifinopoulos et al., 2016). Bayesian analyses were run with 300 million Markov chain Monte Carlo (MCMC)'s chain length. The first 10% of samples from the MCMC chain were discarded as burn-in, and the remaining trees were summarized as a maximum clade credibility (MCC) tree via TreeAnnotator v.1.10.4 (Drummond & Rambaut, 2007) and visualized in the Nextstrain (Hadfield et al., 2018) platform. Phylogeographic and population dynamics outputs were visualized using the ggplot2 (Ginestet, 2011) package in R (R Core Team, 2019). All these approaches were repeated for each set of sequences belonging to each of seven sub-lineages identified on the MCC trees, including L1A, L1A(2) (secondary re-emergence of L1A viruses (Paploski et al., 2021)), L1BG (monophyletic clade comprising L1B and L1G), L1C, L1E, L1F, and L1H. In total, 50 DTA runs (15 runs from three subsampling strategies performed on L1 overall, plus 35 runs from the sub-lineage analyses for which

we only used spatio-temporal subsampling) were performed to infer spatial-temporal dynamics of the virus. The key phylogeographic results were combined from five runs, e.g., range of median times to the most recent common ancestor (tMRCA), range of 95% highest posterior densities (HPDs), and range of probabilities in the ancestral region.

Maximum likelihood discrete-space phylogeography (Sagulenko et al., 2018) and Bayesian structured coalescent approximation (Müller et al., 2017) were explored as alternative analytical approaches and to evaluate the sensitivity of our results to the modeling platform utilized. The first method, maximum-likelihood phylogeography, basically infers ancestral traits across an evolutionary tree's internal nodes by treating migration between discrete locations in the same way as genetic mutation given a time-reversible model (Sagulenko et al., 2018), and is a non-Bayesian version of the DTA described above. Bayesian structured coalescent approximation extends the Bayesian coalescent model to allow migration between subpopulations within a structured population and approximates the internal nodes' trait probability (Müller et al., 2017, 2018). The same five spatio-temporal stratified sampled sets used for DTA were reanalyzed by these methods via TreeTime v.0.8.5 (Sagulenko et al., 2018) in the Nextstrain's augur v.10.1.1 pipeline (Hadfield et al., 2018) (maximum-likelihood approach), and the marginal approximation of the structured coalescent (MASCOT) v.2.1.2 (Müller et al., 2018) package in BEAST v.2.5.1 (Bouckaert et al., 2019), respectively. Since TreeTime does not require massive computational power to run a large dataset, the full set of PRRSV-2 L1 ORF5 gene sequences (n = 19,395) was also analyzed by TreeTime. Detailed parameter settings are displayed in supplementary data (Supplementary Table S2.1). Ancestral locations and their transitions through time from all phylogeographic analysis were summarized by the Babel v.0.4.0 package in BEAST v.2.5.1 (Bouckaert et al., 2019).

Selective pressure analysis

Possible selective pressures along PRRSV-2 L1 evolutionary history were identified using the branch-site test of positive selection (Z. Yang & Nielsent, 2002). We used a branch-site test because branches can be tied to an inferred geographic region, and we wanted to test the hypothesis that selection pressures vary between different regions. The analysis was performed on each of five spatio-temporal subsampled MCC trees from DTA and their original sequence alignment. Implemented in aBSREL (adaptive branch-site random effects likelihood) v.2.3 software, all tree branches were tested through the exploratory analysis in which optimal ω (non-synonymous to synonymous substitutions ratio or dN/dS) of each branch was inferred by the small-sample Akaike Information Criterion (AIC_c), then compared to the null model ($\omega \leq 1$) by the likelihood ratio test (LRT) (M. D. Smith et al., 2015). If a branch has the inferred $\omega > 1$ and LRT p-value < 0.05, the virus is estimated to evolve under episodic positive (diversifying) selection at that branch.

2.3: Results

Origin of PRRSV-2 L1 and its sub-lineages

Results from five DTAs on different spatio-temporal stratified subsampled sets (n = 500 each) suggest that PRRSV-2 L1 in the U.S. originated in Canada (100% posterior probability) during the late 1980s (median time to the most recent common ancestor (tMRCA) = 1986 ;[range of 95% highest posterior densities (HPDs) = 1981 – 1989]). Although sub-lineages L1A and L1F concurrently diverged from the primitive L1 in the late 1990s (range of median L1A tMRCA = 1995 – 2000 ;[range of 95% HPDs = 1992 – 2002] and range of median L1F tMRCA = 1995 – 1998 ;[range of 95% HPDs = 1993 – 2000]), L1F was likely imported from Canada (93 – 100% probability), while L1A emerged within the Upper Midwestern U.S. (97 – 100% probability). Within a few years, two additional sub-lineages, L1C and L1BG, diverged (range of median L1C

tMRCA = 1998 – 1999 ;[range of 95% HPDs = 1995 – 2002] and range of median L1BG tMRCA = 2000 – 2002 ;[range of 95% HPDs = 1997 – 2003]). L1C was a sister clade of L1F, and L1BG was a sister of L1A on every MCC tree. Like their sisters, the origins of L1C and L1BG were potentially in Canada (97 – 100% probability) and the Upper Midwest (93 – 100% probability), respectively. L1 further diverged into another two sub-lineages, L1E and L1H, in the 2000s. L1E was a relatively small clade directly rising from the basal L1 around mid- to late 2000s (range of median L1E tMRCA = 2003 – 2007 ;[range of 95% HPDs = 2000 – 2009]) and was estimated to have emerged in the Upper Midwest (80 – 98% probability). On the contrary, L1H was a recent sister of L1C and L1F, and its approximated time and place of origin were late 2000s (range of median L1H tMRCA = 2008 ;[range of 95% HPDs = 2006 – 2010]) in Canada (100% probability). The second and larger wave of L1A (L1A(2)) branched out from the original L1A in early 2010s (range medians L1A(2) tMRCA = 2009 – 2012 ;[range of 95% HPDs = 2007 – 2013]). Although the most likely origin of this emergence appears to be the Upper Midwest (52 – 99% probability across analyses based on five subsets of data), the probability that the source was the Eastern U.S. (4 – 47% probability) could not be ruled out (Figure 2.3).

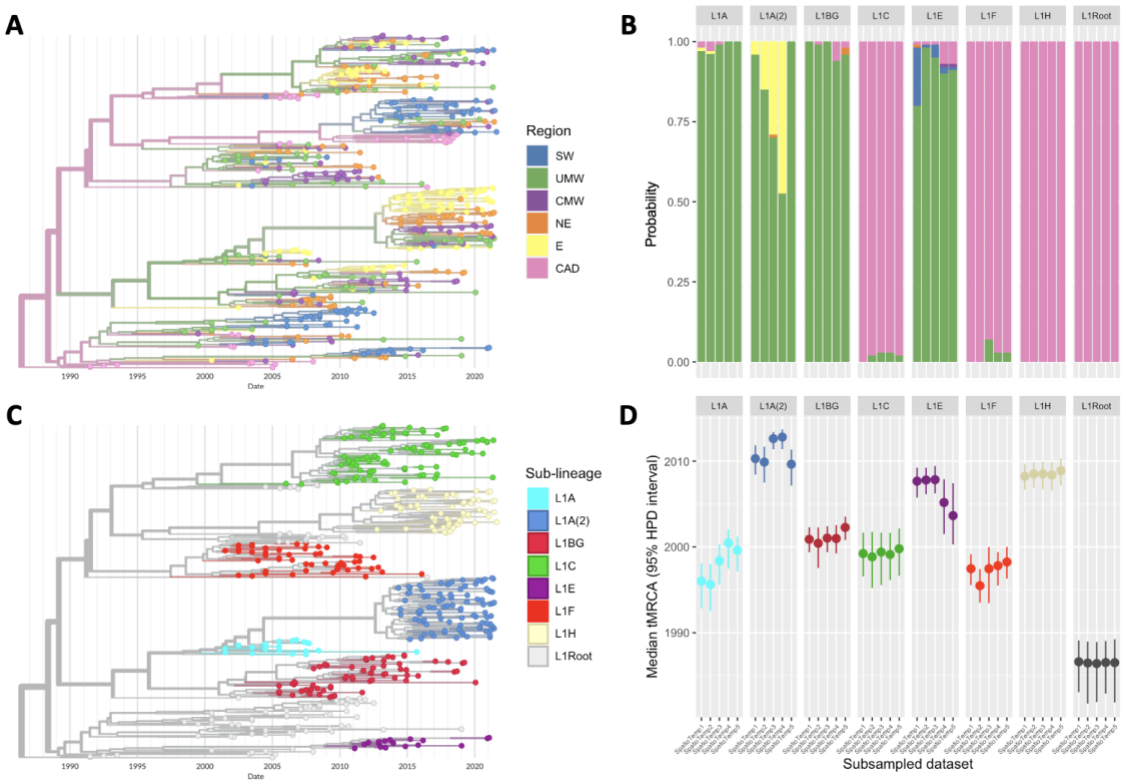


Figure 2.3: Key results from the discrete trait analysis (DTA) on spatio-temporal stratified sampled sets. (A) The time-scaled phylogenetic tree of one subsampled set with tip colored by sampling region and internal branch colored by inferred ancestral region. (B) Probability (0 – 1) on region of origin for each L1 sub-lineage and overall L1 from all runs. (C) The same timed-scaled tree with tip colored by classified sub-lineage. (D) Median tMRCA with 95% HPD interval of L1 and its sub-lineages from the same runs.

When comparing the results generated by different subsampling techniques, estimations of both time and location of origin from the spatial stratified subsampled sets were markedly similar to that from the spatio-temporal stratified subsampling. Unlike the first two techniques, estimations based on the proportionate stratified samples indicate that the L1 virus and all sub-lineages were likely originated in the Upper Midwestern U.S., the region with the proportionately largest pig population and thus the most sequences included in the analysis (48% of sequences were from the Upper Midwest). Inferred tMRCAs

from population-based subsampling were less consistent compared to the spatial-based subsamples (Supplementary Figure S2.3).

The virus phylogeography inferred by DTA differed when using other modeling approaches despite the same inputs (i.e., all models utilized the same subsets produced from spatio-temporal stratified sub-sampling). For instance, the origin of L1BG, L1E, and L1A(2) was more uncertain in maximum-likelihood-based TreeTime analysis than DTA; in some cases, three regions were considered almost equally likely as the putative origin (sub-lineages L1A(2) and L1BG), or different runs produced different results (sub-lineage L1E). With that being said, estimations from TreeTime were far more similar to the DTA than the results from MASCOT. In the latter, there were high levels of uncertainty on the region of origin for most sub-lineages, and the Eastern U.S. was frequently inferred to be a potential origin of L1 overall and most sub-lineages. DTA, TreeTime, and MASCOT, however, similarly estimated that the origin of L1H was in Canada (100% probability). Focusing on the time of emergence, estimates of tMRCAs and inferred nucleotide substitution rates were similar across all methods (Supplementary Figure S2.4).

Both geographic region and time of origin estimated from the non-subsampled set (n = 19,395) of the L1 virus using TreeTime corresponded well to the inference from the spatio-temporal stratified samples by DTA. The only disagreement was the origin of the sub-lineage L1F, where the ancestral location of the full dataset was inferred to be the Upper Midwest instead of Canada, possibly as a result of overrepresentation of the Upper Midwest in this dataset (Figure 2.4 and Supplementary Figure S2.5).

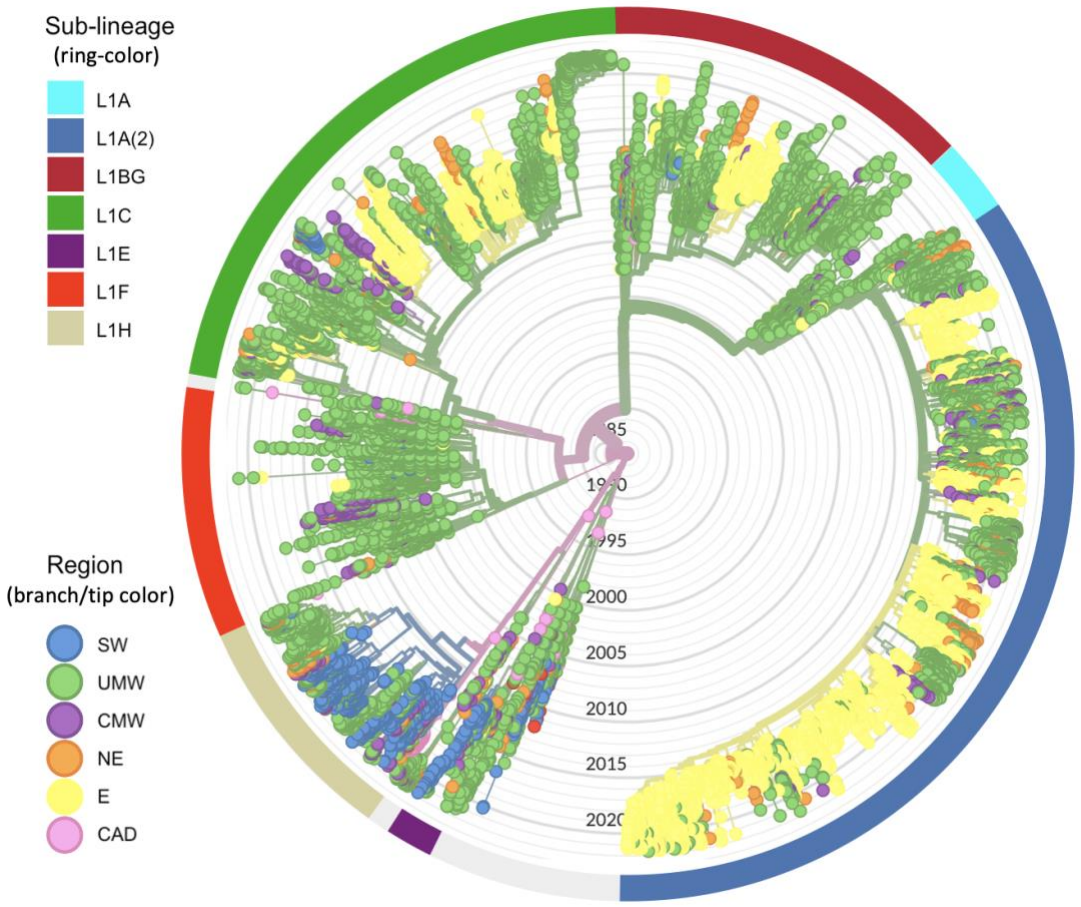


Figure 2.4: Maximum likelihood time-scaled phylogenetic tree of the full L1 dataset (n = 19,395) estimated by TreeTime. Tips and branches are colored by sampling region and inferred ancestral region, respectively. Exterior ring is colored according to L1 sub-lineages based on phylogenetic grouping.

Inter-regional spread and spreading hotspots

The source regions contributing to frequent inter-regional spread (based on the median number of transitions between regions across five runs) inferred by the phylogeographic analysis were considered hotspots for inter-regional spread. DTA analysis on spatio-temporal stratified samples suggests that the Upper Midwestern U.S. was the main hotspot for inter-regional spread and was the origin of frequent dissemination events to every region except Canada. The

common destinations of such events were the Central Midwest (~126 transitions since L1 emergence until 2021) and the Northeast (~111 transitions). Canada and the Eastern U.S. can also be considered as hotspots though they mainly spread the virus to only a single adjacent region such as the Upper Midwest (~75 transitions) and the Northeast (~71 transitions), respectively (Figure 2.5A). Summarizing the proportion of branches inferred to exist in each region across time, L1 was primarily circulating in Canada in the 1990s and was likely introduced to the U.S. through the Upper Midwest during the early 2000s. After around 2005, Canada faded out as a major player in U.S. inter-regional spread (though we have relatively few Canadian sequences in later years to fully understand its role in later periods). These phylogeographic patterns were relatively stable across different phylogeographic approaches and subsampling techniques, albeit with some variation. For instance, MASCOT approximated that the virus spread from the East to both the Northeast and the Upper Midwest, and that there was no apparent flow from Canada to the U.S., which was inconsistent with the patterns inferred by the other approaches (Supplementary Figure S2.6).

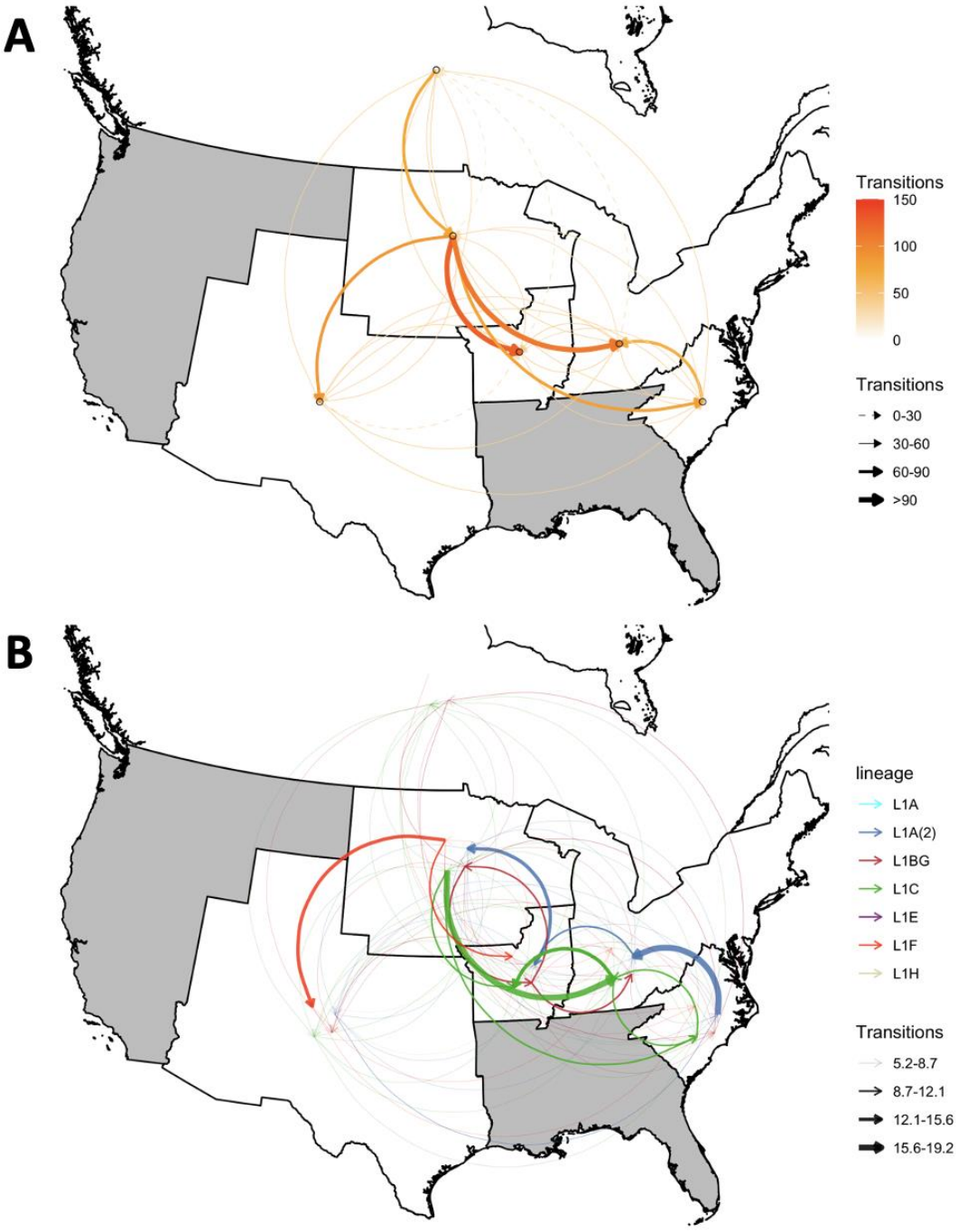


Figure 2.5: Inter-regional spread of PRRSV-2 L1 in the U.S. estimated by DTA on spatio-temporal stratified sampled sets. (A) Median between-region transitions of the L1 lineage overall. Color shade and thickness of arrow represent estimated number of transitions. (B) Median between-region transitions of each L1 sub-lineage. Color of arrow represents sub-lineage, whereas the thickness represents the number of transitions.

At the sub-lineage level, the direction and extent of inter-regional spread were unique for each viral subpopulation. L1C and L1A(2) were the predominant sub-lineages with the highest number of samples across all subsampled datasets (Figure 2.2 – 2.4). Notwithstanding, inferred patterns of inter-regional spread were in opposite directions. The DTA phylogeography estimated that the Upper Midwest was a spreading hotspot for L1C, which disseminated the virus mainly to the Northeast, Central Midwest and East. L1C spread was also found in the opposite direction from the East to Northeast to Central Midwest. In contrast, L1A(2) displayed a more unidirectional northwesterly flow from the East to Northeast to Central Midwest to Upper Midwest. With a smaller number of inferred transitions events due to its smaller population size, L1BG mostly circulated back and forth between two Midwestern regions with some spillover to the Northeast, whereas L1F largely spread from the Upper Midwest to the Southwest, with some introductions to the Central Midwest (Figure 2.5B). Even though the extent of L1H spread was not comparable to other sub-lineages' spread, all phylogenetic trees show that L1H viruses circulated primarily within the Southwestern U.S., which was also its spreading hotspot according to the DTA inference. The smallest sub-lineages were L1A and L1E, for which no distinct pattern of inter-regional spread was discerned.

Population dynamics, mutation, and selection pressure

PRRSV-2 L1 population dynamics in the U.S. were summarized from median effective population sizes inferred by the Bayesian Skygrid analysis on different spatio-temporal stratified sets. The virus population rapidly grew within the first decade (1990 – 1998), after which population dynamics were characterized by a wave-like pattern wherein the effective population size slightly decreased and then increased to a new peak approximately every ~6 years until the present (Figure 2.6). Over 30 years, there were four peaks in the effective

population size of L1, each of which was driven by different sub-lineages. The first L1 peak (1998) occurred prior to the peaks of any of the analyzed sub-lineages, and may have resulted from older sub-lineages, such as L1D, that were classified as primitive L1 in this study (Paploski et al., 2021). Subsequent peaks coincided with the epidemic-like wave of L1F which appears to have pushed the overall L1 population to the second peak in 2005. The third wave peaking in 2012 involved L1F, L1BG and L1C, but increases associated with this peak appear to be most driven by a rapid increase in L1C. The most recent peak (2017) was mainly driven by sub-lineages L1A(2) and L1H that emerged in the late 2000s but did not experience rapid growth until the mid 2010s (Figure 2.6). The Skygrid analyses of inferred population sizes were remarkedly consistent across different subsampling techniques, nearly completely overlaying on top of each other except during the early time period (<2000), where the size of the first peak in population size varied by subsampling strategy (Supplementary Figure S2.7).

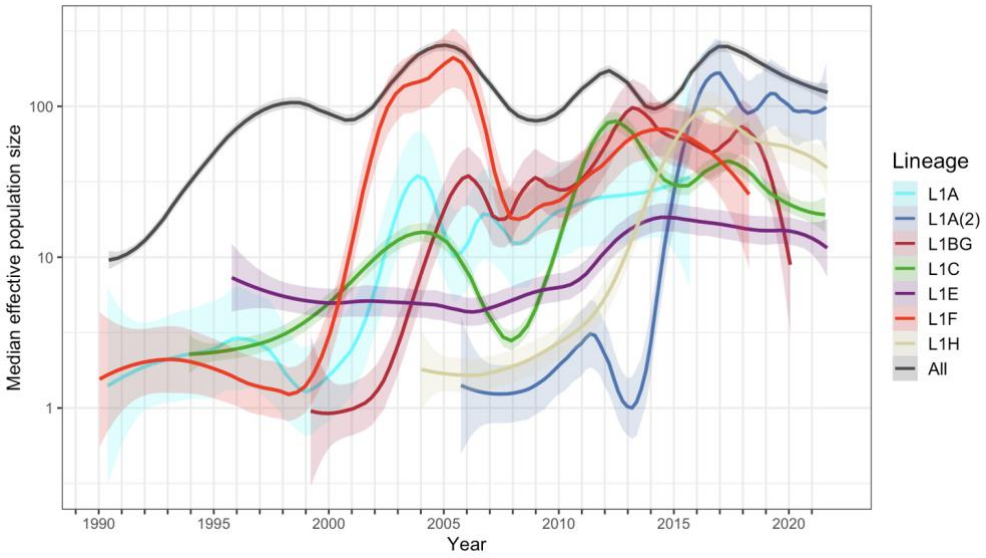


Figure 2.6: PRRSV-2 L1 population dynamics in the U.S. estimated by Bayesian Skygrid analyses on spatiotemporal stratified sampled sets. Lines with shaded bands are LOESS smoothing curve with 95% confidence interval of the median log effective population sizes from five runs of each sub-lineage and overall L1. Lines are colored according to L1 sub-lineage.

The median nucleotide substitution rate for L1 overall, which were computed from the spatio-temporal MCC trees, was 8.6×10^{-3} (IQR = 7.7×10^{-3} – 1.0×10^{-2}) substitutions per nucleotide site per year (s/n/y). There was no significant difference between sub-lineages or geographic regions, nor was there a clear temporal pattern in branch-specific substitution rates.

The selective pressure analysis on the spatio-temporal stratified sequences and trees estimated that ORF5 gene evolved mostly under near-neutrality regardless of lineage or region (median and mode $\omega = 1$, IQR = 0.38 – 2.11). Although the adaptive branch-site model's ω showed that the ancestral viruses at several trees' branches also evolved under both extremely high purifying and positive selection, only three branches from three different subsampled trees can be considered as positively selected (p-value < 0.05) according to the likelihood ratio test (LRT). These rare episodic positive selection events displayed no clear spatial or temporal pattern. One was identified at an internal branch of primitive L1 in 1992 prior to the virus being introduced from Canada, another was a terminal branch of primitive L1 in 2000 in the East that originated from Canada, and third was a terminal branch of sub-lineage L1C in 2014 in the East without any associated inter-regional transition. The positive selected branches did not have higher evolutionary rates compared to others.

2.4: Discussion

In this work, we aggregated several large-scale PRRSV-2 L1 ORF5 gene sequence datasets from the U.S. and Canada to reconstruct historical patterns of lineage emergence, inter-regional spread and population dynamics using phylogeography. Such a large dataset allowed us to perform different subsampling techniques and phylogeographic approaches, and test the sensitivity of our inferences to the sub-sampling technique and modeling approach utilized (De Maio et al., 2015; Liu Pengyu AND Song, 2022). We found evidence that L1 viruses overall, as well as sub-lineages L1C, L1F, and L1H,

potentially originated in Canada, while sub-lineages L1A, L1A(2), L1BG, and L1E emerged within the Upper Midwestern U.S.

However, the initial region of origin might not always be the hotspot most responsible for frequent inter-regional spread. Moreover, the hotspot and spreading pattern of each viral subpopulation varied. For example, the opposing directionality of dissemination of two current predominant sub-lineages, L1C and L1A(2), demonstrated the varied phylogeographic history of different subpopulations. This complexity was also found in the temporal demographic inference in which the contribution of each co-existing sub-lineage to the overall L1 population dynamics has fluctuated through time.

The majority of our inferences as well as previous phylogeographic analysis published in 2013 (Shi, Lemey, et al., 2013) indicate that L1 viruses in the U.S. originated in Canada. This does not seem to be due to lack of sequence data in the U.S. before 1998 (the earliest detected L1 sequence in the U.S.) (Han et al., 2006), given that there is abundant sequence data available (>230 sequences, all non-L1) from throughout the U.S. since 1989 (Andreyev et al., 1997; Kapur et al., 1996; Meng et al., 1995; Meulenberg et al., 1993; Wesley et al., 1998). The detection of L1 sequences in the 1990s in Canada as early as 1991, but not in the U.S., (Gagnon & Dea, 1998; Mardassi et al., 1995; Pirzadeh et al., 1998; Rodriguez et al., 1997) further supports the conclusions from our the phylogeographic analysis. Many of the early Canadian sequences were not included in the analysis where sub-sampling was proportional to pig population size; the preponderance of Upper Midwestern sequences and exclusion of early Canadian sequences in those sub-samples resulted in the Upper Midwestern U.S. being the inferred origin of L1 and all its sub-lineages, though we believe this to be an analytical artefact of the overrepresentation of this region. Thus, we believe that spatial-temporal stratified sub-sampling was the best approach to overcome the sampling biases in our dataset.

Performing the analysis with the spatial-temporal sub-sampled sequences in MASCOT, which uses a forwards/backwards algorithm (Pearl, 1982) with a structured coalescent model (Müller et al., 2017), resulted in drastically different results in which all L1 viruses except L1H were inferred to originate in the East. Such a pattern would be possible only if L1 virus did truly exist in the East undetected during or before the 1990s but had never been sampled or sequenced. In fact, according to our data, at least 40 samples from the Eastern region were sequenced between 1992 – 1999, but none of them was classified as L1. This further supports our conclusion that the inference from spatio-temporal stratified samples by DTA was the most reliable and robust approach for our data.

While the Canadian origin of lineage L1 is well-supported, the potential role of Canada in subsequent inter-regional spread was inconclusive given the available dataset. Introductions of the primitive L1, L1F (circa 1998 – 2001), L1C (circa 2000 – 2005), and L1H (circa 2009 – 2011) from Canada to the U.S. were highlighted, while reverse dissemination from the U.S. back to Canada was rare. This unidirectional pattern is not surprising given the U.S. has imported millions of feeder and finishing pigs per year from Canada, but exported only a few thousand pigs back to Canada annually (Economic Research Service & USDA, 2022). Changes in the U.S. and Canadian pig industry structure that led to this phenomenon have been described elsewhere (Brisson, 2014; Haley, 2004), but it is worth noting that 90% of imported pigs are moved into the Upper Midwest (Whiting, 2008). L1H, the most recent sub-lineage originating from Canada, appears to have been introduced to the U.S. in the late 2000s when the Canada-to-U.S. pig imports were at peak (8 to 10 million heads/year) (Economic Research Service & USDA, 2022). Once introduced to the U.S., the Upper Midwest appears to be the epicenter of inter-regional spread, likely because it is a hub for interstate pig shipments given the number of harvest plants and proximity to corn and soybean crops used in feed (Cabezas et al., 2021; Shields

& Mathews, 2003). Notably, this general pattern of virus movement from Canada to the Upper Midwest, following by inter-regional spread within the U.S. mirrors what has been observed for swine influenza (Nelson et al., 2017; Scotch & Mei, 2013).

The question of whether Canada continues to disseminate viruses into the U.S. cannot be fully answered due a reduced number of Canadian L1 ORF5 gene sequences available in the NCBI GenBank database in recent decades. Reductions in the number of PRRSV-2 sequences available in GenBank also occurred in the U.S. After 2010, when sequencing technology became more accessible and affordable, more ORF5 gene sequencing was likely requested by field veterinarians for diagnostic purposes, which could explain the reduced number of publicly available sequences. Thus, data sharing via the GenBank, especially with sensitive metadata we needed, i.e., sampling date and location (state), has been decreasing due to confidentiality concerns.

Phylogeographic and phylodynamic analyses on each particular L1 sub-lineages not only revealed varying patterns of inter-regional spread and population dynamics, but also connected the dots between the virus's evolutionary history, historical PRRS outbreaks, and potential factors facilitating disease spread. Apart from the first emergence caused by the primitive L1 viruses (L1D), L1F was the earliest well-described sub-lineage and contributed to the second wave of the effective population size of L1, which peaked in the mid-2000s. A virus belonging to this sub-lineage was first isolated from a sample collected in 2001 from a severe PRRS case in southern Minnesota, and this isolate is widely referred to as MN184 (Johnson et al., 2004) according to the 1-8-4 ORF5 gene restriction fragment length polymorphism (RFLP) pattern (Kapur et al., 1996). L1F viruses were mainly detected in the Midwestern regions with multiple introductions to the Southwest (Han et al., 2006). Interstate swine movement records in 2001 further supported this, showing that the Southwest was the second-most frequent destination for pig movements from the Upper

Midwest (Shields & Mathews, 2003). Currently, L1F appears almost extinct in the U.S., with less than two sequences detected each year since 2019 (Figure 2.4).

Between 2007 and 2008, two virulent PRRSV-2 strains responsible for regional outbreaks in Iowa and Minnesota were isolated. They were initially designated as 1-18-2 (P. Yeske & Murtaugh, 2008) and NADC-30 (Brockmeier et al., 2012) strains, which were eventually classified as part of sub-lineages L1BG and L1C, respectively. These coincided with the third wave of L1 population expansion in the early 2010s that was influenced by the rise of those sub-lineages along with the sub-lineage L1F. Unlike the previous wave, the estimated inter-regional spread of L1C and L1BG more heavily involved the eastern part of the country in addition to the Upper Midwest. Patterns of spread completely shifted during the most recent wave in the late 2010s, in which L1A(2) and L1H were the primary contributors. L1A(2) was the infamous and virulent 1-7-4 strain first detected during 2014 – 2015 (Alkhamis et al., 2016; van Geelen et al., 2018). Our analysis confirmed unpublished reports that its spreading hotspot was the Eastern U.S., where the virus spread widely before spilling to the Midwest (Morrison, 2015).

In contrast, L1H (80% of the clade members were RFLP 1-8-4 (Paploski et al., 2021)) appears to be an endemic virus mostly confined in the Southwest; we found little published epidemiological and virological characterizations of this particular group of viruses. The impact of recent and widespread outbreaks in the Upper Midwest in 2020 – 2021, caused by the novel L1C-1-4-4 (Kikuti, Paploski, et al., 2021) variant, was not captured by our analysis despite a number of available ORF5 gene sequences. This may be because the novel L1C variant, as of 2021, was not yet widely spread at national-scales to the point that significantly contributed to the overall population and spatio-temporal dynamics.

Inter-regional, long-distance spread of PRRSV-2 has been an unavoidable consequence of vertical integration of U.S. swine production systems, and L1

viruses have evolved in parallel with an expansion of U.S. swine production. After entering the country in perhaps the late 1990s, circulation of L1 viruses would be almost impossible to contain to the Midwestern U.S. since hog and pig inventories were rising strikingly in new areas of the country, including the Eastern state of North Carolina and several Southwestern states (Oklahoma, Colorado, Texas, and Utah) during the same period (Key & McBride, 2011). In 2013, Shi, et al. estimated that L1 viruses largely circulated only between Canada and the Midwestern regions (Lake States, Corn Belt, and Northern Plains), meanwhile the East (Appalachia) was one of the spreading hotspots for non-L1 lineages (Lineages 5 – 9) (Shi, Lemey, et al., 2013). Over the course of time with increasing L1 prevalence, we captured the geographical shift in the L1 spread. The Eastern region became another spreading hotspot for particular sub-lineages (L1C and L1A(2)) during the third and the fourth waves of L1, likely because it is now the second largest swine producer after the Midwest (National Agricultural Statistics Service, 2019) and the number of outgoing animal shipments are comparable or sometimes higher than the Midwest (Sellman et al., 2022). These spatial dynamics of spread may not be exclusive to PRRSV-2 but may be applicable for other swine diseases. Swine influenza, for example, appears to have been introduced from Canada into the Midwest, and new spreading hotspots in the East or other regions are also apparent in phylogeographic analyses of swine influenza in the U.S. (Nelson et al., 2017; Scotch & Mei, 2013). Such a correspondence indicates that dissemination patterns of multiple pathogens in the U.S. swine industry have common determinants, likely related to structural and demographic organization of the industry.

The U.S. swine industry has been characterized by multi-site pig production for several decades (Harris, 1992; Tokach et al., 2016) and animal movement is a key component that keeps the production flow uninterrupted (Ramirez et al., 2011). Long-distance transport of live animals, cull hogs (Blair &

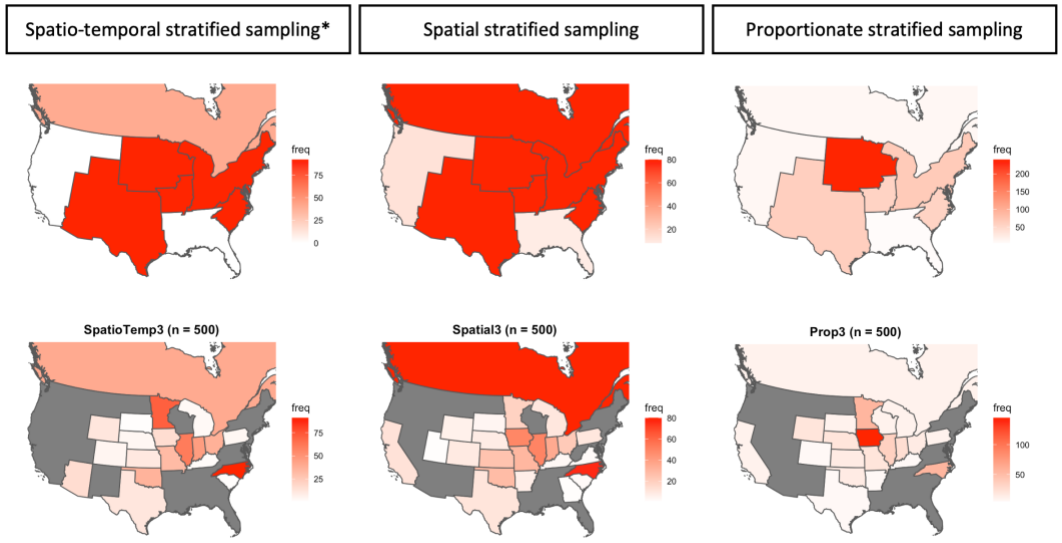
Lowe, 2019), feed, personnel, and equipment are driven by uneven distribution of feed resources, production phases, and slaughtering facilities amongst geographic regions, and/or by contractual relationships between sites (Ramirez et al., 2011). All of these logistics are potential risk factors that help transmit PRRSV-2 via either direct or indirect contacts (Pileri & Mateu, 2016). In addition, variation in protective measures such as biosecurity practices and vaccination may unpredictably change the patterns of viral spread, though we were not able to assess these here. Further information surrounding field samples, such as production type of an infected farm, immunization status, and farm-related transportations before an outbreak, could be useful factors to include in phylogeographic regressions that can be employed within the DTA framework (Faria et al., 2013; Lemey et al., 2014) to estimate factors associated with inter-regional transmission. The fraction of sequenced samples relative to the number of actual cases could improve sub-sampling techniques and avoid biases related to unequal sampling/sequencing efforts between regions (P. Liu et al., 2022).

We found no support for the hypothesis that selection pressures varied spatially. Episodic diversifying selection along the L1 virus phylogenetic tree was rarely detected. The few branches with evidence of episodic selection were neither related to the (re-)emergence of L1 sub-lineages nor to the expansion of any virus variant. This is not surprising, as such analyses are typically used to assess virus adaptation to a different host species (Benfield et al., 2021; Berry et al., 2019; Caraballo et al., 2021; Cariou et al., 2022; MacLean et al., 2021; Spielman et al., 2019). We, accordingly, concluded that there is little evidence for episodic selection in PRRSV-2 L1 evolutionary population dynamics. The estimation of episodic selection takes into account changes in overall genetic changes within ORF5 gene across each branch of the tree. Since many parts of ORF5 gene are strongly conserved, using this branch-site model potentially obscures positive selection occurring at specific amino acid residues, e.g.,

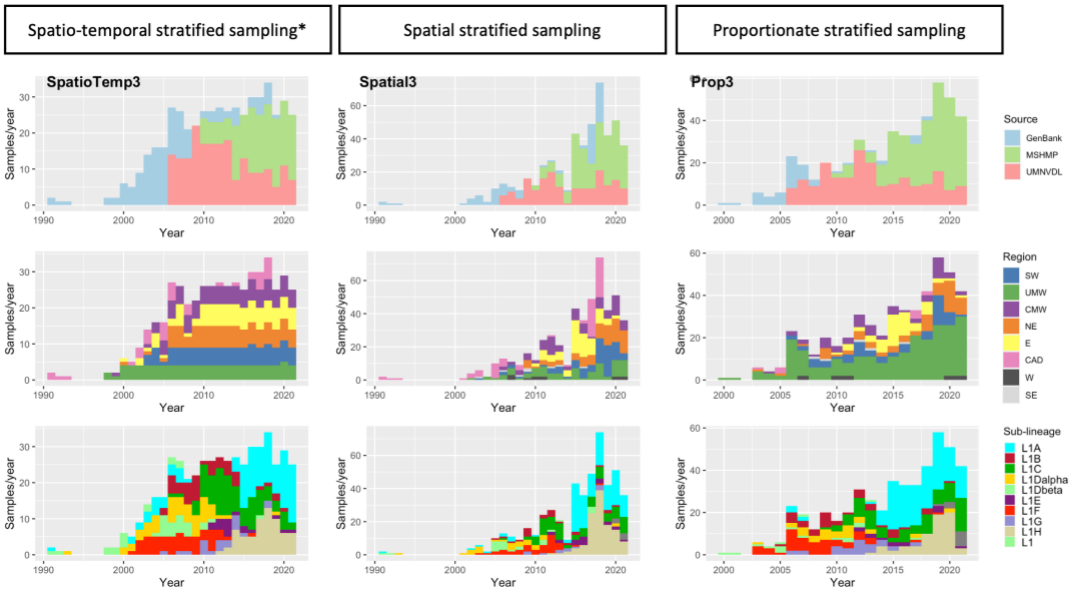
antigenically important amino acid sites which have been detected using site-specific models (Delisle et al., 2012; Paploski et al., 2019).

Ultimately, this chapter overviews PRRSV-2 L1 dynamics at the national level, particularly focusing on changes in inter-regional spread patterns over three decades. Understanding these dynamics is essential for the implementation of large-scale PRRS control strategies. However, due to constraints in data availability, the potential drivers of this spread were not fully ascertained in this analysis. The subsequent chapter delves deeper into the intra-regional dynamics of PRRSV-2 L1, enabling the estimation of spread pathways down to the farm level. This higher resolution data provides an opportunity to uncover transmission determinants within the region.

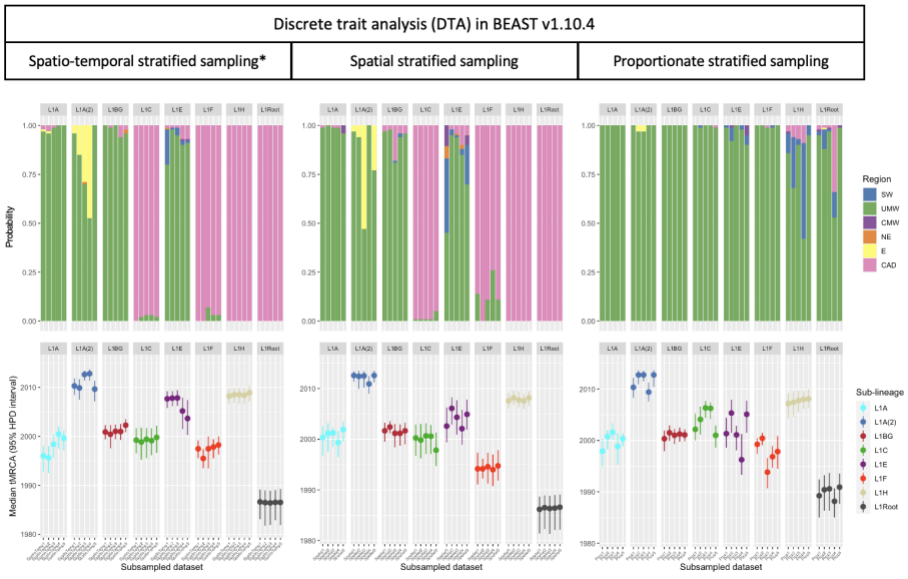
2.5: Supplementary Materials



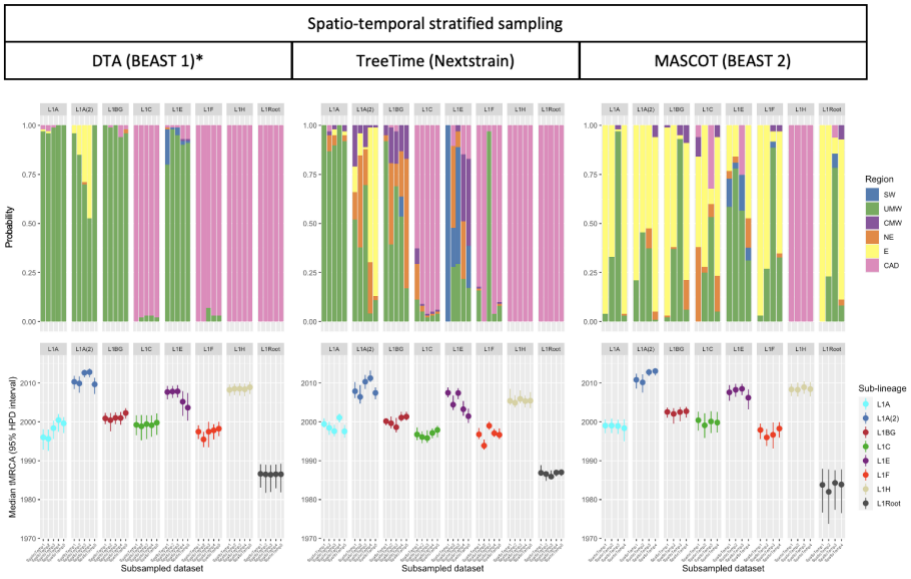
Supplementary Figure S2.1: Geographical distribution of ORF5 gene sequences from different subsampling approaches.



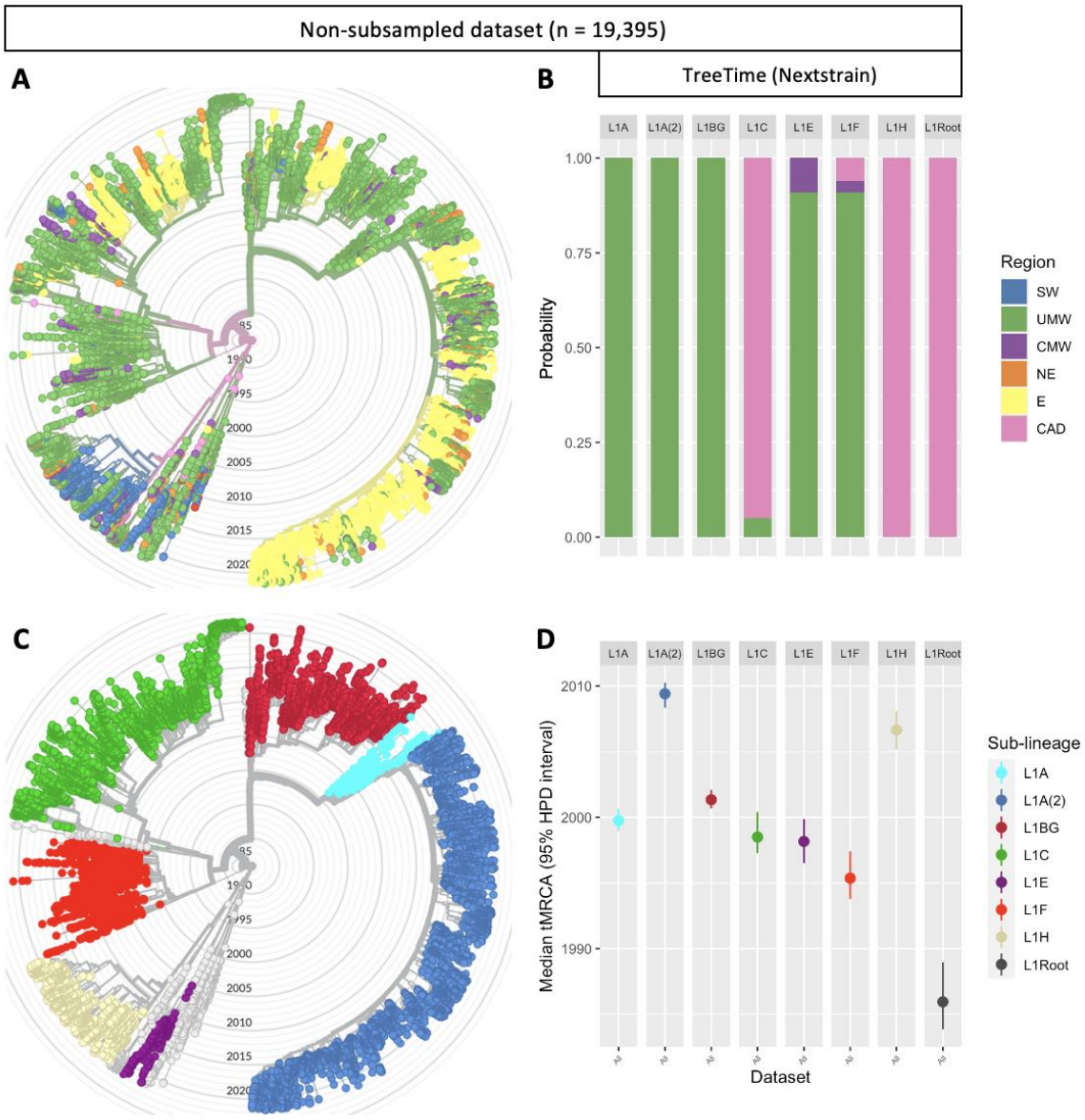
Supplementary Figure S2.2: Temporal distribution of ORF5 gene sequences from different subsampling approaches colored by source, sampling region, and pre-identified sub-lineage.



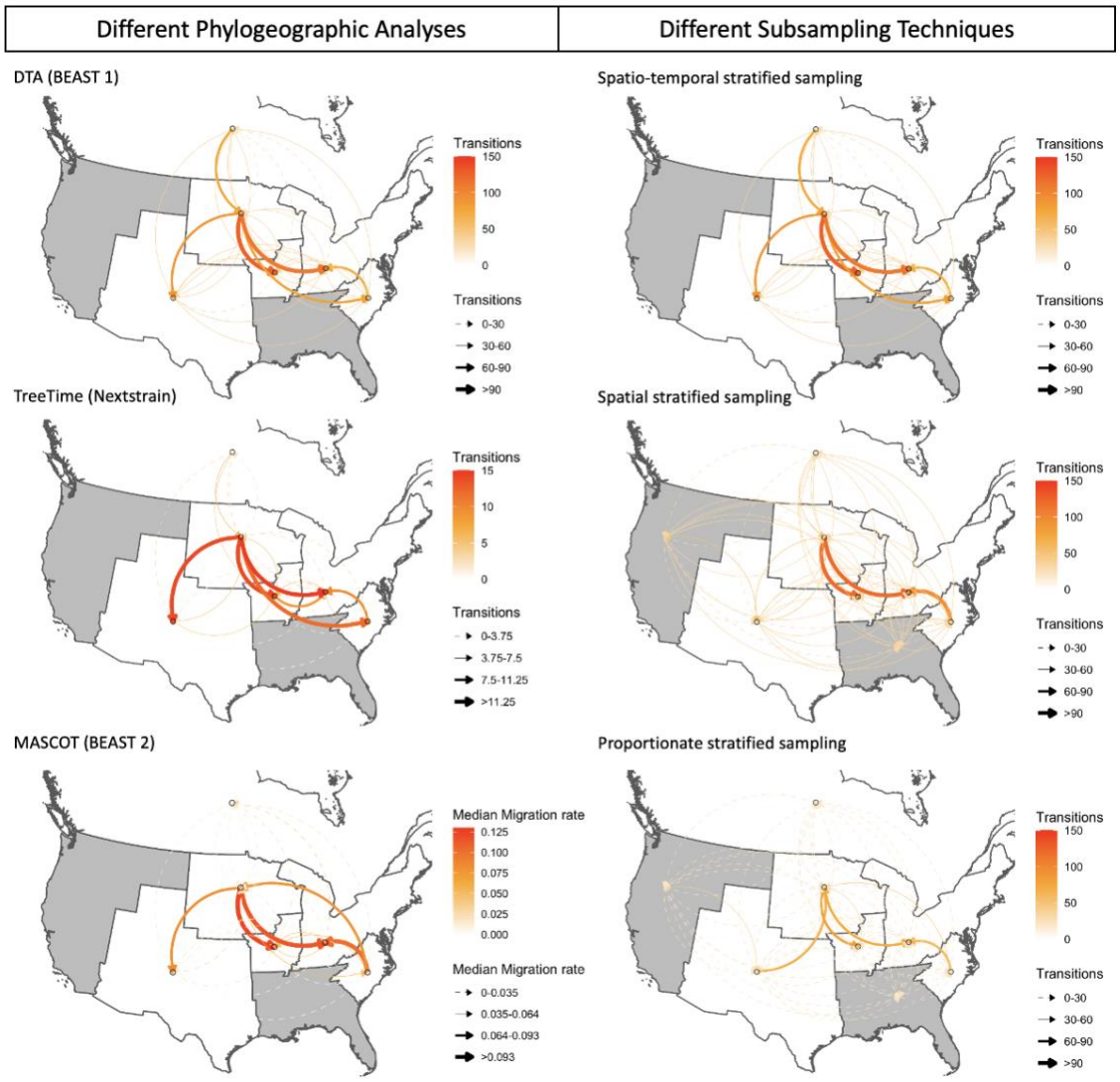
Supplementary Figure S2.3: Probability of region of origin (top) and Median tMRCA with 95% HPD interval (bottom) of each L1 sub-lineage and overall L1 from all runs compared between different subsampling techniques.



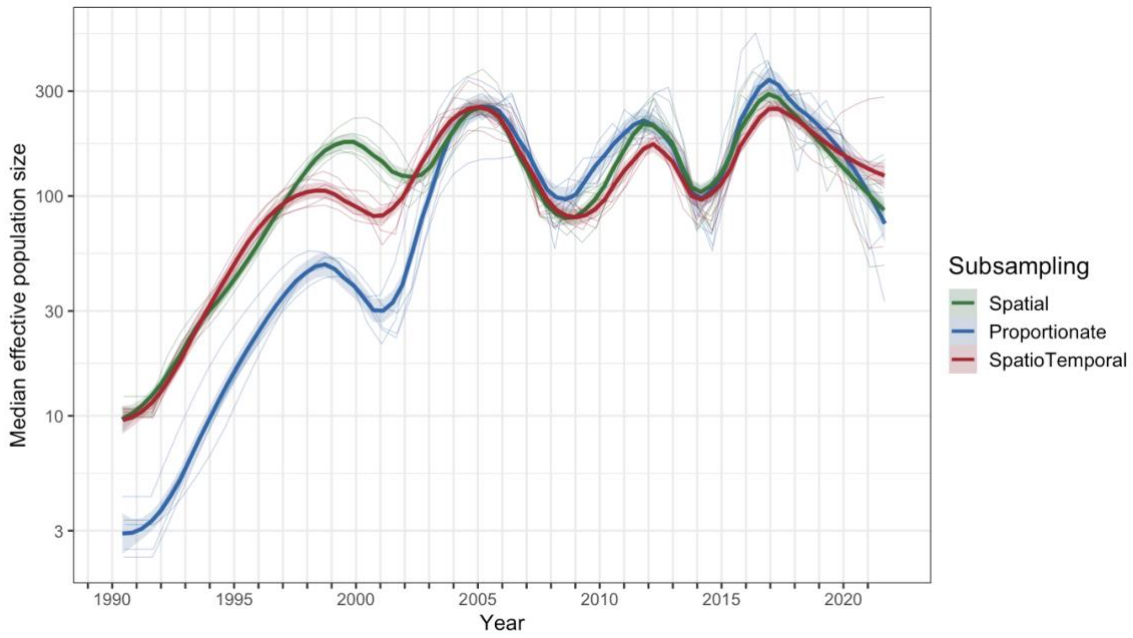
Supplementary Figure S2.4: Probability of region of origin (top) and Median tMRCA with 95% HPD interval (bottom) of each L1 sub-lineage and overall L1 from all runs compared between different phylogeographic approaches.



Supplementary Figure S2.5: Key results from TreeTime analysis on the full L1 dataset. (A) The time-scaled phylogenetic tree with tip colored by sampling region and internal branch colored by inferred ancestral region. (B) Probability (of region of origin of each L1 sub-lineage and overall L1. (C) The similar timed tree with tip colored by classified sub-lineage. (D) Median tMRCA with 95% HPD interval of L1 and its sub-lineages.



Supplementary Figure S2.6: Comparison of inter-regional spread of PRRSV-2 L1 in the U.S. between different phylogeographic analyses and subsampling techniques in map and arrows format.



Supplementary Figure S2.7: Comparison of L1 population dynamics estimated by Bayesian Skygrid analysis on datasets from different subsampling techniques. Thin lines in the background are median effective population size of each run. Thick lines with bands are LOESS smoothing curve with 95% probability interval of the median population sizes from five runs of each subsampling technique. Color of line and band on the plot represents subsampling technique.

Supplementary Table S2.1: Parameter settings in different phylogeographic approaches.

Software	**BEAST v1.10.4**	**TreeTime v0.8.5**	**BEAST v2.5.1**
Datasets used	All subsampled datasets (n = 15, 500 seqs each)	Spatio-temporal subsampled datasets (n = 5, 500 seqs each) and non-subsampled dataset (19,395 seqs)	Spatio-temporal subsampled datasets (n = 5*, 500 seqs each)
Nucleotide substitution model	GTR + I + G	GTR	GTR + G
Molecular clock model	Uncorrelated relaxed clock, Log-normal distribution	Strict clock	Uncorrelated relaxed clock, Log-normal distribution
Model for phylogeography	Non-reversible CTMC (DTA)	GTR (mugration model)	Marginal approximation of the structured coalescent (MASCOT)
Coalescent model (population dynamics)	Bayesian GMRF Skygrid	Constant coalescent rate	Structured coalescent with constant effective population size and migration rate
MCMC chain length	300 million	No MCMC running	300 million

*Only 4 out of 5 datasets could be accomplished in BEAST 2 without error. (Potential errors: https://taming-the-beast.org/tutorials/Mascot-Tutorial/)

"A Tangled Web Unleashed"

Chapter 3: Unveiling Between-Farm PRRSV-2 Transmission Links and Routes through Transmission Tree and Network Analysis.

Material adapted from a published article in Evolutionary Applications (2023), doi: 10.1111/eva.13596

Unveiling invisible farm-to-farm PRRSV-2 transmission links and routes through transmission tree and network analysis.

Nakarin Pamornchainavakul, Dennis N. Makau, Igor A. D. Paploski, Cesar A. Corzo, and Kimberly VanderWaal

3.1: Introduction

In Chapters 2, we captured long-term evolutionary dynamics and inter-regional spread of contemporary PRRSV in the U.S. that are highly informative for PRRS prevention and control on a national level. Nevertheless, with limited metadata of such country-wide collected genetic sequences, characteristic of disease transmission as a major force behind the virus dynamics was not fully investigated. PRRSV is not only transmitted by direct contact between pigs but also indirectly through contact with contaminated fomites, iatrogenic farm practices, or aerosols (Pileri & Mateu, 2016), all of which may contribute to the spread of the virus between farms. Transport of PRRSV-positive semen or animals are also mechanisms for introducing the virus to other herds over long distances (C. Nathues et al., 2016; Thakur et al., 2015). In addition, contaminated trucks, equipment, or personnel sharing can transmit the disease via indirect contact (S. Dee et al., 2002; S. A. Dee et al., 2004). At short distances, PRRSV is possibly transmitted through aerosols based on experimental/semi-experimental studies and air sampling near infected farms, though field evidence remains unclear (Arruda et al., 2019). Attempts to estimate the relative contributions of these routes to the overall PRRSV-2 transmission

have been made using a simulation model that utilized disease incidence and between-farm contacts as an input (Galvis et al., 2021). However, exploitation of the genetic relatedness amongst viruses found on different farms may provide a clearer, empirical-based picture of disease transmission (Firestone et al., 2019).

Based on experimental studies (Charpin et al., 2012; Rose et al., 2015) and mathematical modeling (Nodelijk et al., 2000), the European PRRSV genotype (PRRSV-1) has an estimated R_0 of 2-5, meaning that an average infected pig transmits the virus to 2-5 other pigs (assuming an immunologically naïve population). In contrast to R_0, the effective reproduction number (R_e, R_t or R) relaxes the assumption of the population being fully susceptible, and is defined as the average number of secondary cases that are infected by a single infectious individual regardless of immune status of the population (Nishiura & Chowell, 2009). R is often used to measure disease transmissibility for endemic diseases or unfolding epidemics, and importantly, can be used to help quantify the impact of control measures. However, neither R_0 nor R measured at animal-level can explain between-farm transmissibility, which drives PRRSV persistence at a regional scale and is the level at which control measures are implemented.

Using regional-scale PRRSV-2 genetic data coupled with available information related to farm characteristics and contact between farms, our objective was to infer farm-to-farm transmission links, estimate farm-level transmissibility as defined by reproduction numbers (R), and identify associated risk factors for transmission. We analyzed a set of PRRSV-2 ORF5 gene sequences collected from swine farms along with animal movement data in a swine-dense region of the United States to fill knowledge gaps on the between-farm transmissibility and pathways of spread. Farm-level R and potential pathways were estimated by integrating transmission tree inference of PRRSV-2 sequences with network-based statistical inference. Our results not only illuminate a clearer picture of PRRSV-2 dynamics in a major swine producing region of the U.S., but also demonstrate a novel approach to quantify the

between-farm transmissibility of PRRSV-2 that can be expanded to evaluate the effectiveness of control measures across space and time.

3.2: Materials and Methods

Data selection

Most data used in this study were obtained from the Morrison Swine Health Monitoring Project (MSHMP) database, which was established to track progress on PRRS control in the U.S. and aimed to be the national hub for voluntary data sharing between swine veterinarians from different production systems (*MSHMP History | College of Veterinary Medicine - University of Minnesota*, n.d.). The project has archived farm-level data such as farm location, herd size, disease incidence, and pathogen genetic sequences from more than half of the US breeding population (Paploski et al., 2019). For this particular study, we focused on a major swine-dense farming region (confidential data) in the US in which 70% of farms (n = 2,724) belong to two multi-site swine production systems that participate in MSHMP. A swine production system is a commercial entity consisting of multiple swine production sites connected by either ownership, management, or contractual agreements (Kinsley et al., 2019; Makau, Paploski, et al., 2021). Production systems are very insular, with nearly 100% of animal movements occurring between farms within the same system (Kinsley et al., 2019).

Phylogenetic analysis of PRRSV-2 ORF5 gene sequences has been used for virus lineage and sub-lineage classification (Paploski et al., 2019; Shi, Lam, Hon, Murtaugh, et al., 2010). Based on the frequency of samples submitted to MSHMP, lineage 1A (L1A) was found to be a predominant PRRSV-2 sub-lineage in the US since 2014 (Paploski et al., 2019). Moreover, a previous study also suggested that the L1A virus was either introduced to or emerged in our study area in early 2013 and started expanding within the region in 2014 (Makau, Alkhamis, et al., 2021). Focusing our analysis on the emergence and spread of

L1A viruses from 2014-2017, we queried 1,515 voluntarily submitted ORF5 gene sequences from the study area that were identified as L1A in the MSHMP database. Field veterinarians typically requested ORF5 gene sequencing for routine diagnosis during a PRRS outbreak (particularly at the beginning) after case confirmation by RT-PCR. The sequences were aligned using MAFFT (Katoh, 2002) and screened for potential recombination using RDP4 (D. P. Martin et al., 2015). Subsequently, a maximum likelihood phylogeny was reconstructed from the alignment using RAxML (Stamatakis, 2014) with the GTRCAT nucleotide substitution model and transfer bootstrap clade support computation (Lemoine et al., 2018) from 1,000 bootstrapped trees. To prevent potential bias caused from multiple sequences available from the same farm, sequences from a single farm that formed monophyletic clades on the tree were subsampled, retaining the median dated sequence. The filtered dataset contained 943 sequences derived from 651 farms between 2014-2017.

In order to focus the analysis on groups of sequences that were more likely to be epidemiologically linked, we identified the largest three clusters of closely related sequences from the phylogenetic tree, then conducted further analyses for each cluster separately (Figure 3.1A). Groups of monophyletic sequences were systematically defined as a cluster with Cluster Picker (Ragonnet-Cronin et al., 2013) when their bootstrapped clade support was >70% and the maximum genetic distances within-group was <4.5%.

The animal movement data used in this study was directly obtained from the two participating production systems, which electronically recorded all farm-to-farm movements of pigs, totaling 283,959 movement events amongst 2,724 farms during 2014-2017. Premise ID, geographic coordinates, farm production types of the origin and destination farms, and shipment date of each event were prepared for the network analysis. Production types included sow farm (14.7%), nursery (17.5%), finisher (67.1%), and boar stud (0.7%). A sow farm is a premise that comprises at least breeding and farrowing sows or gilts population. A

nursery farm is a premise that only raises pigs from newly weaned to grower stage. A finisher farm is a premise that feeds grower pigs until they reach the market weight. A boar stud is a premise where boars are raised for semen collection.

Transmission network inference

Phylogenetic temporal signals of the three clusters were checked from maximum likelihood sub-trees using TempEst (Rambaut et al., 2016); R^2 and correlation coefficient above 0.5 and 0.4 respectively were considered to be evidence of sufficient temporal signal in the data for construction of time-scaled phylogenies. We constructed time-scaled phylogenies for the clusters in BEAST 1.10.4 (Drummond & Rambaut, 2007) using a Monte-Carlo Markov Chain (MCMC) length of 10 million. The cluster's ORF5 gene alignment and sampling date were inputted and run with the GTR+I+G substitution model selected based on Bayesian information criterion computed by ModelFinder (Kalyaanamoorthy et al., 2017), the uncorrelated relaxed clock model (Drummond et al., 2006), and the time-aware Bayesian skyride model (Minin et al., 2008). Time-scaled phylogenetic trees were constructed by the maximum clade credibility (MCC) method, excluding the first million burn-in MCMC states using TreeAnnotator 1.10.4 (Drummond & Rambaut, 2007). To infer transmission networks from the MCC trees, we conducted transmission tree analysis in *TransPhylo* 1.4.5 in R (R Core Team, 2019). Amongst several available packages for transmission tree inference, TransPhylo had the most appropriate assumptions and options that best aligned with our data, notably in that it allows for incomplete sampling of cases and ongoing (endemic) outbreak scenarios (Didelot et al., 2017; Duault et al., 2022). Given a time-scaled phylogeny, molecular clock, and assuming a stochastic branching epidemiological process, TransPhylo uses a Bayesian approach to create a network indicating who infected whom and inferring the number of unsampled individuals in the transmission chain connecting a pair of sampled cases (Didelot et al., 2017). To parameterize the model priors, we

assumed that the observed transmission processes were part of an ongoing outbreak (i.e., the outbreak was not resolved by the last sampling point), and that the PRRS mean generation time was 14.5 days (Charpin et al., 2012). The analyses were executed with 50,000 MCMC iterations (Figure 3.1B).

The outputs of the transmission tree analysis include inferred pig-to-pig transmission chains, with information on the origin, recipient, and infection window coinciding with the sampling date on the first and second sequence. These outputs can be conceptualized as a dynamic directed network. Since the direction of transmission at the animal-level may or may not reflect directionality at the farm-level, we transformed the trees into undirected networks, then used the igraph package (Csardi & Nepusz, 2006) to compute the shortest pathlength (SPL) between all the sampled pairs, disregarding directionality, to capture the most feasible pig-to-pig infection chain between each pair of samples. Although TransPhylo can take into account within-host evolutionary dynamics, this was not implemented in our model given that the scale of variation occurring at the between-farm level likely is far greater than variation arising from within-host evolution. Also, our infection chains (at the pig level) were not sufficiently well sampled to be able to discern within-host evolutionary processes.

Data integration

In order to translate the transmission network from the animal-level to the farm-level, we used the farm ID associated with each sequence to create a farm-level transmission network, assuming that the virus must have moved between farms somewhere along the pig-to-pig infection chain inferred in the transmission network analysis. However, direct farm-to-farm transmission cannot be assumed between farms connected in the transmission network, as it is possible that the infection chain has passed through an unsampled intermediate farm. Therefore, we used animal movement data to identify “known” direct transmission events, wherein direct farm-to-farm transmission can be reasonably assumed if an

animal movement occurred between the farms during the infection window. While there are multiple modes of between-farm contact that could lead to transmission between farms, pairs of farms connected via animal movement are a form of contact for which we have quantifiable data on direct contact between farms. We used the infection chain lengths of inferred transmission events that were concurrent with documented between-farm animal movement to define a maximum threshold in the length of pig-to-pig infection chains, below which direct transmission between farms (regardless of the presence of a movement) is feasible. To do this, we created a time-stamped dynamic movement network and extracted the shortest path between the samples existing during the infection window (sample dates of the earlier and later sequences collected from two different farms) of every sampled pair. Between each pair of samples, the infection chain length (ICL) was designated as the inferred number of unsampled pigs in the shortest path connecting each pair in the transmission network, and the movement pathlength (MPL) was designated as the number of steps (i.e., farms) that must be passed through to connect two farms in the movement network (Figure 3.1C). Two farms were considered “unreachable” if no path existed that connected the two farms during the infection window. Viable paths in the movement network must follow the directionality and the temporal sequence in which movements occurred. The Pearson’s correlation coefficient between ICL and MPL of each cluster was calculated. Since the farm-level R can be calculated only from direct farm-to-farm transmission, we assumed that direct farm-to-farm transmission had occurred in any case where the infection chain was shorter than the median of ICL at MPL of 1 (regardless of the presence of documented animal movement) and used the resulting farm-to-farm transmission events to create a farm-level transmission network.

Considering uncertainties in the PRRSV generation time used to infer animal-level transmission and the length of time prior to sample collection that PRRSV-2 could be circulating in a farm, we performed two sensitivity analyses.

First, we re-specified the generation time to 11.6 and 17.4 days (plus and minus 20% from the originally estimated 14.5 days). Second, we extended the infection window start-date to 3 and 6 months before the original window. This allows the farm to be infected earlier than the sample was collected. We then repeated all procedures for each setting.

Farm-level R estimation

All the farm-representative sample pairs that had infection chain lengths less than the threshold (median of ICL at MPL of 1) and sampling date interval less than 1 year were considered candidate transmission pairs for R estimation. In some cases, one recipient farm may have multiple potential sources of transmission present in the candidate list. Thus, for each recipient, we selected the single source with the shortest ICL as the most probable. Farm-level R values for different phylogenetic clusters and sensitivity analytic settings were then computed by counting the number of recipients per each source and summarized into descriptive statistics (Figure 3.1D). Details and code for our approaches are available at https://github.com/author's identifier/FarmR.

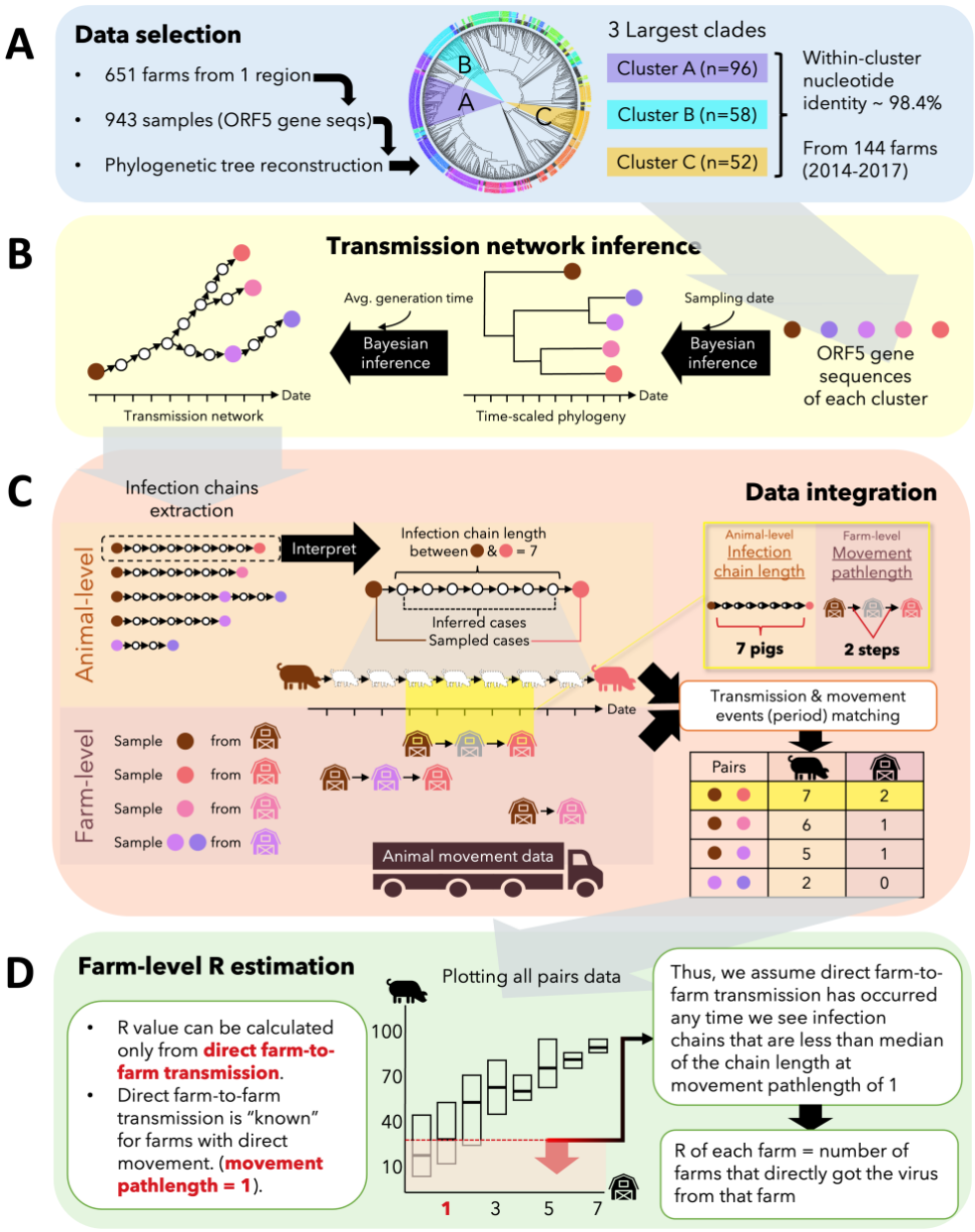


Figure 3.1: Workflow of farm-to-farm network reconstruction and R estimation. The three biggest phylogenetic clusters of the lineage 1A PRRSV-2 were selected for the analysis (A). Each cluster's ORF5 gene sequences were used to reconstruct the time resolved phylogenetic tree and transmission tree (B). Pig-to-pig infection chains were extracted from the transmission tree then matched with animal movement data (C). Infection chain length was used to estimate direct farm-to-farm transmission links that were combined into a farm-level transmission network. Farm-level effective reproduction number (R) was calculated from the network (D).

Analysis of potential factors associated with farm-to-farm PRRSV-2 transmission.

Farm-level metadata, including location and production type, were utilized to investigate potential modes of transmission between farms beyond animal movement. The distance between a pair of farms was calculated from their geographical coordinates. The approximate longest distance that viable PRRSV-2 can be found in aerosols, 10 km (Arruda et al., 2019; Otake et al., 2010), was set as a threshold for classifying farms in close enough proximity for potential local-area spread. PRRSV-2 transmission through contaminated semen was presumed when a transmission pair contained boar-sow farms as the source-recipient. Altogether, the mode of transmission for pairs that were not connected via animal movement was designated into the following non-mutually exclusive categories: farm proximity related factor, transmission by contaminated semen, and undetermined. Undetermined may include a wide variety of transmission modes for which we do not have data, such as movement of equipment, feed, and personnel.

Farm-to-farm transmission pairs used for the R value calculation were transformed into directed transmission networks for each cluster. We applied multivariable exponential random graph models (ERGMs) to the networks to identify factors significantly associated with the occurrence of an inferred transmission link using the ergm 4.3.2 package in R (Hunter et al., 2008). ERGMs are a type of statistical regression that treat network topology as a response and edge/node attributes as predictors. The output of ERGMs includes the odds that a particular attribute influences the network structure. In our case, the analysis indicates which factor is significantly associated with the existence of a transmission link between two farms. The model was initially constructed with the best fit structural covariate (in-stars; frequency of star-like network structures in which several nodes connect to the same central node without connection with each other) that accounted for the underlying architecture of the directed network. Put simply, this was a baseline model that captures basic

characteristics of network structure but did not account for how node- or dyad-level attributes influenced which nodes were connected.

Additional node- or edge-level covariates were then added to the baseline model. Node-level covariates included a farm's production type, herd size, season in which the sequence was sampled, farm density, and in- and out-degree in the movement network. Sampling time was classified into season by month, i.e., winter (December to February), spring (March to May), summer (June to August), and fall (September to November). Farm density was summarized for a 10 km radius around each farm, and was computed from regional pig farms' coordinates using the PointdensityP package (Evangelista & Beskow, 2019). In-degree and out-degree (number of farms that the focal farm received or sent animals to in a six month period) were computed using the igraph package (Csardi & Nepusz, 2006) from a 6-month period that was temporally matched with the sample. Six months was selected based on previous work demonstrating that degree metrics calculated in swine movement networks reach stability when six months of data are aggregated (Makau, Paploski, et al., 2021). These node/edge attributes were incorporated into the model using several different ERGM terms. Categorical node-level covariates, such as production type and sampling season were incorporated with the terms *nodefactor* (e.g., some production types are generally more likely to form transmission links), *nodematch* (e.g., transmission links are more likely to be found between nodes with the same production type), and *nodemix* (e.g., accounting for differential frequencies with which transmission links form between farms of different production types). The *absdiff* (e.g., two farms with similar herd size are more likely to have transmission link) and *nodecov* (e.g., farms in higher density areas are more likely to form transmission links) terms were used for continuous node-level covariates.

Three edge-level covariates were also included in the model: geodesic distance (km), MPL (steps in the movement network) between the sampled pair,

and sampling date interval (days). We calculated geodesic distance (km) using the geosphere package in R (Hijmans et al., 2019), then dichotomized using 10 km as a cut-off value (0 = more than 10 km, 1= less than 10 km). Based on the distribution of MPL, we assigned MPLs of <=3 steps as possible movement connections (1), while >3 as not (0). Sampling date interval was included to control for temporality in the model (i.e., samples collected close in time were more likely to have short infection chain lengths and be linked in the transmission networks). The *edgecov* term (e.g., transmission links are more likely found between farms that have animal movement connections) was used for both categorical or continuous edge-level covariates.

We performed an AIC-based stepwise approach to build multivariable ERGM models from all predictors. Coefficients (log-odds), probabilities (inverse logit), and p-values of predictors were reported from the most parsimonious model that was <2 delta-AIC from the model with the lowest AIC.

3.3: Results

Descriptive analysis and phylodynamics

The three largest phylogenetic clusters of PRRSV-2 lineage 1A, denoted as cluster A (n = 96 sequences), B (n = 58 sequences), and C (n = 52 sequences) were included in the analyses (Table 3.1 and Supplementary Figure S3.1). All three clusters contained sequences identified from farms that all belonged to a single production system. Pairwise genetic distance within each cluster was 1.6-1.8%, and the distance between clusters was >2.6%. An average of ~1.4 (SD: 0.85) ORF5 gene sequences were available from each farm that were part of clusters A-C. More than half of the samples were collected from sow herds, followed by nursery, finisher, and boar farms (Table 3.1). In three cases, serially collected samples from three farms were classified into different clusters. Maximum likelihood trees constructed for each cluster exhibited a strong temporal signal indicated by high correlation coefficients between root-to-tip

divergence and tip date, ranging from 0.73 to 0.79 with acceptable R-squared values (0.53-0.63). Based on inference from Bayesian time-scaled phylogenies, mean viral evolutionary rates were relatively consistent across three clusters (7.2-9.8 × 10^{-3} substitutions/site/year) and the time to the most recent common ancestors (tMRCA) for each cluster were in early 2014 (Table 3.1). The Bayesian Skyride analysis suggests that the effective viral population of clusters A and C sharply increased from mid-2014 until early 2015, then cluster C's population decreased after summer 2015, whereas cluster A plateaued and decreased in late 2016. The effective population of Cluster B originally the smallest, but gradually rose starting early 2016 and was comparable to cluster A in late 2017 (Figure 3.2B).

Table 3.1: Data structure, genetic relationship, temporal signal of the selected clusters' ORF5 gene sequence samples, and key statistics from their time resolved phylogenetic trees.

		Cluster A	Cluster B	Cluster C
Data description	Number of samples	96	58	52
	Average pairwise identity (%)	98.2	98.4	98.5
	Bootstrap support at ancestral node (%)	86.3	93.9	71.1
	Sampling date range (dd/MM/yyyy)	18/6/2014 - 28/11/2017	3/12/2014 - 29/12/2017	18/6/2014 - 27/9/2017
	Number of farms	72	42	36
	Number of sow farms (%)	41 (56.9%) *	22 (52.4%) *	18 (50%)
	Number of nursery farms (%)	16 (22.2%) **	15 (35.7%) **	9 (25%)
	Number of finishing farms (%)	13 (18.1%)	5 (11.9%)	7 (19.4%)
	Number of boar studs (%)	2 (2.8%)	0 (0%)	0 (0%)
	Number of unidentified farms (%)	0 (0%)	0 (0%)	2 (5.6%)
Temporal signal (Root-to-tip divergence ~ time)	Correlation coefficient	0.78	0.73	0.79
	R^2	0.61	0.53	0.63
Bayesian timed phylogeny estimation	Mean Rate (substitutions/site/year)	7.20×10^{-3}	7.25×10^{-3}	9.82×10^{-3}
	Mean tMRCA	2014.3	2014.5	2014.2
	95% HPD interval	[2014.0, 2014.5]	[2014.1, 2014.8]	[2013.9, 2014.5]

*Two sow farms submitted samples belonging to different clusters (A in 2015-2016 and B in 2017).

**A nursery farm submitted samples belonging to different clusters (A in 2016-2017 and B in 2017).

Inferred infection chains and animal movements

Transmission tree analysis estimates the number of unsampled cases between a pair of samples, referred to as the infection chain length (ICL). Movement pathlength (MPL) is the shortest number of animal movement steps that corresponds to the sampled farms in the infection chain during the infection window (dates of sample A and B). Linear regression shows that ICL and MPL, combined across all clusters, were significantly correlated (R^2=0.29, Pearson's r = 0.53, $p < 0.001$), indicating that pairs of farms that were more distant to each

other in the movement network also had longer inferred infection chains. According to the sensitivity analysis that varied the assumed mean generation time and infection window length, this correlation was robust to uncertainties in the generation time and to extending the infection window used for calculating movement pathlengths (Supplementary Figure S3.2). Given that movement pathlengths of 1 represents “known” direct contact between farms, we assumed that direct farm-to-farm transmission was reasonably likely for any pair of farms where the number of pigs in an infection chain was less than median of ICL at MPL of 1. This threshold was applied to identify candidate transmission pairs across all pairs (regardless of presence of movement). Accordingly, the threshold ICL for farm-to-farm transmission for clusters A, B, and C were 35, 28, and 46 pigs, respectively.

Farm-level R and transmission events

We were able to infer 80, 45, and 49 farm-to-farm transmission events in cluster A, B, and C, respectively. Farm-level effective reproduction numbers (R) were computed for each source farm that appeared in the list of candidate transmission pairs (events). Across all clusters, R had a median of 1, with interquartile ranges (IQR) of 1-2 for cluster A and B, and 1-2.5 for cluster C. This indicates that an infected farm typically infects 1 to 2 additional farms. The number of farms having R > 1 in cluster A, B, and C were 20, 10, and 12, respectively. The highest observed Rs for cluster A and C (R = 5) were all observed between February to March 2015, whereas the cluster B’s highest farm-level R (R = 4) was observed in March 2017 (Figure 3.2C). The median R did not change when the generation time and infection window for capturing time-matched movements were varied, with the exception of one scenario for cluster C (generation time of 17.4 days and 3-6 months relaxed timeframes) for which the median R was 2 (Supplementary Figure S3.4).

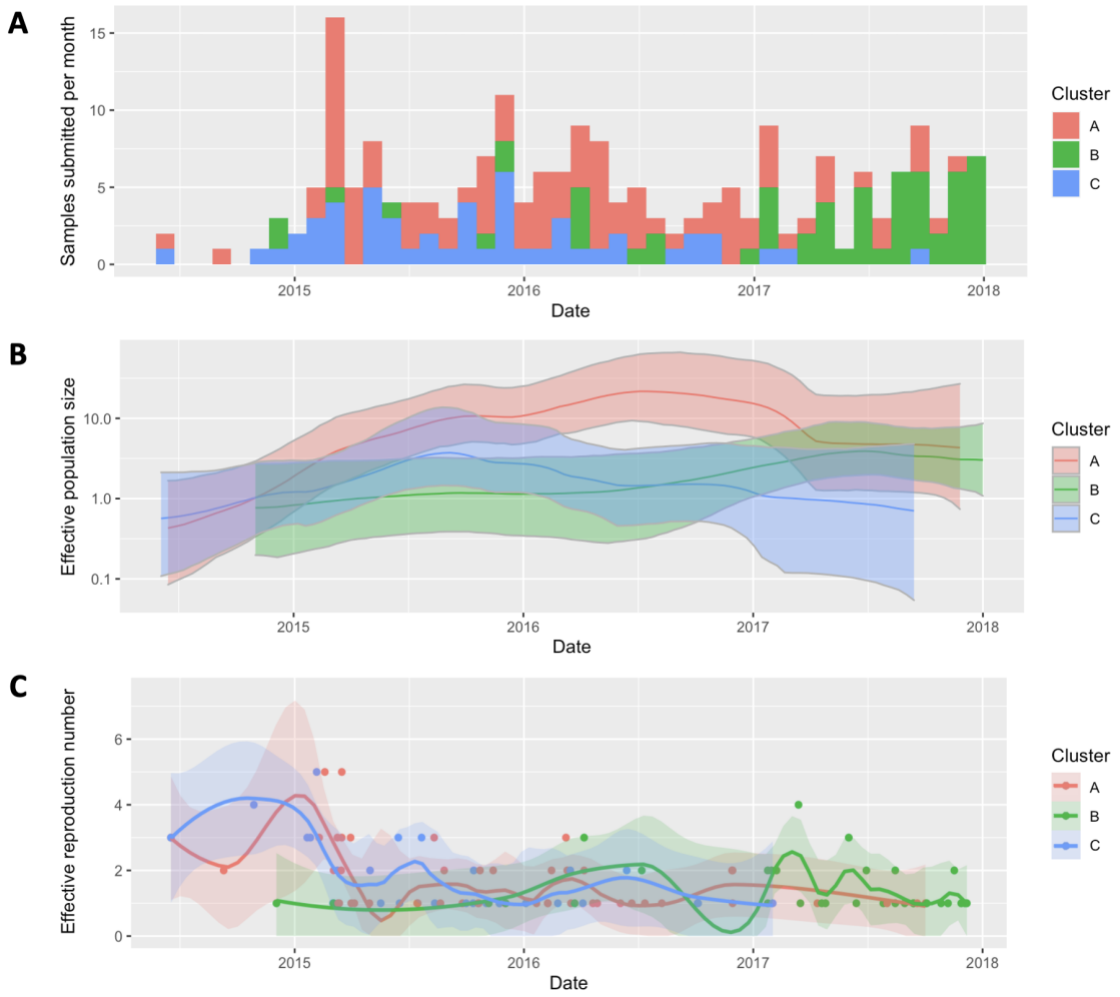


Figure 3.2: PRRSV-2 dynamics of the three genetic clusters during 2014-2017. Number of ORF5 gene sequences submitted per month (A). Median effective viral population size with the 95% highest posterior density (HPD) estimated by Bayesian skyride analysis (B). Scatterplot of the effective reproductive number of individual farms, dated according to source farm's sampling date, with LOESS curves overlayed to visualize temporal trends (C).

Overall, over 80% of the farm-to-farm transmission events had no corresponding animal movement linking the two farms. For the events that were not linked by animal movements, 8.3 to 23.7% of the transmission events across clusters involved farms located less than 10 km apart, whereas the longest inferred transmission range was over 100 km for every cluster (Figure 3.3). Most

transmission events occurred between sow herds (32.7%) followed by sow-to-nursery (13.8%) and nursery-to-sow (12.6%). Relative to the directionality of pig production flows (pigs move from sow farms to nurseries to finishing farms), the direction of transmission could be downstream (71.3%) or upstream (28.7%). Interestingly, two boar studs in cluster A had relatively high R (2 and 3), and the recipients were four different sow farms. The source of infection to these boar studs appeared to be sow and finishing farms.

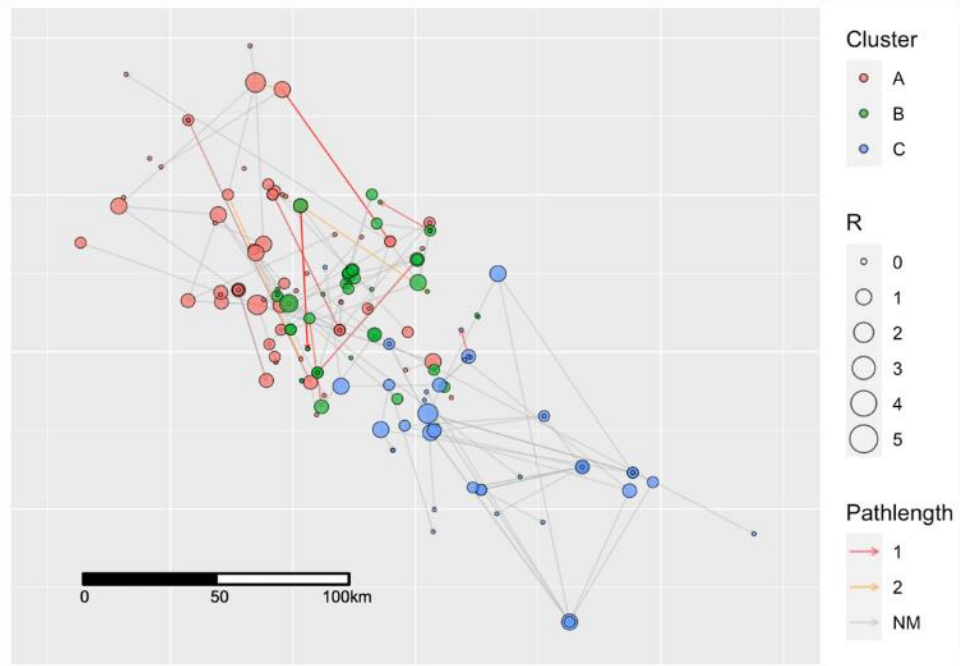


Figure 3.3: Spatial representation of the estimated transmission network of farm-to-farm PRRSV-2 transmission from 2014-2017. Node color represents designated phylogenetic clusters (A to C). Node diameter size corresponds to an individual farm's effective reproduction number (R). Edge color represents movement pathlengths (NM; No movement). Samples with unknown farm location (n = 6) were dropped from the network.

Factors associated with transmission.

We created directed networks from inferred farm-to-farm transmission events, then fit exponential random graph models (ERGMs) (Hunter et al., 2008) to identify and measure factors associated with transmission links. Clusters A and B's networks were explained by the same best-fit model, and Cluster C's model differed by a single variable (Table 3.2). Our best-fit models suggest that farms within a 10 km radius of one another were >10 times less likely to have a transmission link (odds = 0.03-0.1) than farms located more than 10 km apart. The odds of having a transmission link increased by 22 (cluster A) to 37 times (cluster B) if there were animal movements (MPL 1 to 3 steps) between two farms. The log-odds of a transmission link decreased by 0 to 0.01 for each additional day of sampling date interval, meaning that the odds of two farms forming a transmission link was halved after ~70 days. Pairs of farms in areas with similar farm density had an increased log-odds of 0.02 for cluster A and 0.04 for cluster B. Particularly for cluster A, transmission links were more likely (odds = 1.9) between farms whose samples were collected in different seasons (Table 3.2). Multicollinearity among the predictors was not detected in any model according to the ERGMs' variance inflation factor.

Table 3.2: ERGMs' predictors of each cluster's network with coefficients reported on the log-odds (odds) scale.

	Cluster A			Cluster B			Cluster C		
	Coefficient	SD	P-value	Coefficient	SD	P-value	Coefficient	SD	P-value
Structural term: In-stars	-2.59 (0.08)	0.29	<0.001***	-2.10 (0.12)	0.45	<0.001***	-2.12 (0.12)	0.40	<0.001***
Animal movement (y/n)	3.10 (22.20)	0.41	<0.001***	3.62 (37.34)	0.71	<0.001***	1.42 (4.14)	0.79	0.074
Farm proximity < 10 km (y/n)	-2.42 (0.09)	0.34	<0.001***	-2.27 (0.10)	0.45	<0.001***	-3.41 (0.03)	0.56	<0.001***
Difference in farm densities surrounding the source / recipient farms (continuous)	-0.02 (0.98)	0.01	0.029*	-0.04 (0.96)	0.01	0.004**	-0.02 (0.98)	0.01	0.055
Sampling date interval (days)	-0.01 (0.99)	0.00	<0.001***	-0.01 (0.99)	0.00	<0.001***	0.00 (1.00)	0.00	<0.001***
Same sampling season (y/n)	-0.66 (0.52)	0.33	0.049*	0.60 (1.82)	0.40	0.131			
Same Farm type (y/n)							0.55 (1.73)	0.34	0.104

3.4: Discussion

In this study, we implemented novel integrative approaches that utilized routinely collected PRRSV-2 genetic sequences, animal movement records, and farm metadata to strengthen our ability to trace the spread of PRRSV-2 between farms. We used this approach to infer farm-to-farm transmission links, quantify the farm-to-farm transmissibility of the virus by estimating farm-level effective reproduction numbers, and identify factors associated with disease transmission. In summary, the analysis suggests that most infected farms transmitted the virus to one additional farm, though several potential super-spreader events (R substantially above 1) were observed. Our R value estimation, however, likely underestimated the true spread due to potential shortcomings associated with sequencing data generation. Although most transmission events could not be attributed directly to animal movement, movement was a crucial risk factor associated with between-farm transmission links. In addition, the odds of an inferred transmission link between two farms reduced by 50% by ~70 days after the sampling date (likely when the outbreak was recognized) at the source farm.

This suggests that most onward transmission from farms occurs within the first two months or so.

Historically, restriction fragment length polymorphism (RFLP) typing, percent genetic distance comparisons, and phylogeographic reconstruction have been used to assess or track PRRSV spread in different situations (Alkhamis et al., 2017; J. Liu et al., 2021; Rosendal et al., 2014). The core principle of these analyses is an association between shared genotype or phylogenetic relatedness with other attributes of the farms that could help disentangle the likely route of transmission, such as measurable host contacts, spatial adjacencies, temporal continuity. Here, we further expand the concept by converting time-scaled phylogenies into high resolution transmission networks that infer who infected whom and how many animals were potentially involved in the infection chain between a pair of samples (based on the generation time and molecular clock of the virus). The estimated evolutionary rates and trends in viral population growth of the observed clusters are consistent with analysis of PRRSV-2 sub-lineage L1A drawn from nationwide ORF5 gene databases (7.62-7.72 × 10^{-3} substitutions/site/year with the peak of population size in 2016) (Paploski et al., 2021). In addition, the detection of multiple clusters through time on individual farms emphasizes the possible role of re-infection by closely related viral variants, which could potentially contribute to disease persistence at the local or regional scales.

Pig-to-pig transmission networks, however, cannot be interpreted in the same manner as person-to-person transmission for a human disease (Hatherell et al., 2016) since pig populations are highly discretized into relatively homogenous sub-populations (i.e., farms). Indeed, between-farm rather than between-animal transmission is more important for understanding regional spread and pathogen persistence. To make sense of that, the animal-level network was transformed into the farm-level network, and time-matched animal movement data was used to inform the maximum length of animal-level infection

chains that would be consistent with direct farm-to-farm transmission (as opposed to longer infection chains that may be more likely to have passed through an unsampled intermediate farm). Many of the inferred transmission pairs had no recorded connections via animal movement, highlighting the fact that other transmission modes play a role in between-farm spread. We were limited by data availability on other transmission routes, and therefore we assumed that the infection chain lengths for other modes of transmission would be similar to that of animal movement mediated transmission, which is a limitation to our approach.

Even though live animal movement within the study production systems was well-documented, most between-farm transmission events could not be explained by movement. This phenomenon possibly emerges for three reasons. First, live animal movement prior to an outbreak (particularly movements from a PRRS-positive farm) may be viewed as a “smoking gun” or a primary suspect for the route of disease introduction. In such cases, field veterinarians may not submit a sample for sequencing given that the source of outbreak appears obvious. Second, other undocumented transmission routes may be playing a substantial role in between-farm transmission. For example, while boar studs are seldom infected with PRRS, sequences associated with two outbreaks at boar studs had higher than average farm-level Rs. Sow farms were the recipients in all cases, likely suggesting transmission via contaminated semen. Other kinds of transport may also disseminate the virus, such as fomites transported by contaminated vehicles or equipment, between farms can contribute to long-distance transmission (S. Dee et al., 2002; S. A. Dee et al., 2004). For example, our analysis suggests that the virus was transmitted from a positive nursery farm to a finishing farm (103 km apart) and then the recipient transmitted the virus back to the origin within a few months, but the matched movement event was only detected in the first event (nursery to finisher). The latter event is one of several inferred transmission links where the directionality is opposite to the

unidirectional flow of animals through farms typical of a vertically integrated pig production system (Lee et al., 2017; Passafaro et al., 2020) (Supplementary Figure 3.5). That being said, we cannot precisely conclude the contribution of each mode of transmission without a complete sequence database for all farms and concrete data associated with other transmission modes, such as semen samples and delivery history, all between-farm traffic records, or contemporary environmental samples.

Our results indicate that PRRS outbreaks have a farm-level R of ~1. This finding is consistent with previous estimations of R using PRRS incidence data throughout the U.S. from 2009 to 2016 (Arruda, Alkhamis, et al., 2017), though a seasonal pattern of super-spreader events (R > 1) was not clearly detected in our network. The timing of super-spreader events, however, was concurrent with the increases in the effective population size shown by the phylodynamic analysis and in disease incidence depicted by frequency of sample submission (Figure 3.2A). Taken together, this suggest that expansions in regional transmission were at least partly coincident with super-spreader events as opposed to multiple one-to-one transmission events. More generally, applying our integrative approach can enhance disease monitoring efforts by providing fine-scale epidemiologic assessment such as estimating patterns of spread of PRRSV-2 variants in a particular region or production system, or comparing between-farm transmissibility before and after control interventions.

The sensitivity analysis showed that altering either the mean generation time and the timeframe for identifying time-matched movements only slightly affected the infection chain length used as a threshold for direct transmission (Supplementary Figure S3.3) and estimated farm-level R values (Supplementary Figure S3.4). This means, first, biological variation in animal-level factors that may influence generation time, such as host contact rates and host-pathogen interactions (Q. H. Liu et al., 2018), may not substantially contribute to variation in farm-level transmissibility. Second, while there was a concern that animal

movements that occurred prior to sample collection on the farm may allow for transmission beyond the infection window, our results from sensitivity analysis were not substantially altered by extending this timeframe to include movements that had occurred three to six months earlier than the original infection window. This is potentially because animal movements between specific pairs of farms often recur at regular intervals through time (Makau, Paploski, et al., 2021), such that few new and unique connections were identified by extending the infection window.

Possible between-farm PRRSV transmission routes are well documented (Arruda et al., 2019; S. Dee et al., 2002; S. A. Dee et al., 2004; C. Nathues et al., 2016; Thakur et al., 2015), but their contribution to the regional disease endemicity is somewhat vague. Our results further point towards animal movement as an important but not sole mode of transmission. The ERGM analysis highlights the role of animal movement as a primary risk factor for transmission (although the coefficient for animal movement in Cluster C was only trending towards significance). Once animals were shipped from an infected farm, the shipping destination's risk of becoming infected is many folds higher than farms that do not receive animals from an infected source. We also hypothesized that local area spread of PRRSV-2 (less than 10 km) might explain some transmission events which animal movement cannot. Surprisingly, the network analysis revealed that a short distance between the pair of farms was a protective factor for the occurrence of a transmission event between farms. Apart from mechanical transmission routes such as infectious fomites or personnel sharing amongst neighbors, this result suggests that between farm PRRSV-2 transmission rarely occurs via the airborne route, which agrees with the conclusions of the previous reviews and phylogeographic analysis (Arruda et al., 2019; Makau, Alkhamis, et al., 2021).

Albeit other predictors in the best fit ERGMs had statistically significant effects on the transmission networks, interpreting those in terms of mode of

transmission is challenging. High pig farm density has been underlined as a risk factor for PRRSV spread and persistence in several studies (Arruda, Vilalta, et al., 2017; Jara et al., 2021; Makau, Alkhamis, et al., 2021). Our model suggests that a pair of farms located in areas with similar density are more likely to transmit the virus to one another. Though it is unclear, one plausible explanation is if farm density represents an aspect of biosecurity investment. Locations in low density areas are often chosen for farms where biosecurity may be of particular importance, such as nucleus, multipliers, breeding herds, and boar studs, whereas there is often less investment in biosecurity at farms in higher density areas. Exclusively for cluster A, farm pairs sampling the virus in different seasons were significantly associated with the transmission link. This could be perhaps a function of the time of year (fall to winter) when disease incidence was primarily increasing (Trevisan et al., 2020). That being said, we do not have a good explanation for all the associations documented by our model, and it perhaps may be an artifact of an undocumented confounding factor.

We also found that transmission links were more likely if the time interval between samples was shorter. We included this factor to account for the temporality in epidemiological process. However, the results also provide insights into the farm-level generation time (i.e., the lapse of time between the primary and a secondary farm involved in a transmission pair), assuming that sample collection dates are at least somewhat aligned to disease detection dates. The results of the ERGM suggest that the likelihood of such onward transmission is drastically reduced after approximately two months, even though the average sow farm takes 6-10 months to bring an outbreak under control (Linhares et al., 2014; Sanhueza et al., 2019). This would make intuitive sense if between-farm transmission is correlated with prevalence or shedding on a farm, both of which would likely decrease after the initial acute phase of an outbreak. Such information may be useful in managing and mitigating the risk posed to other farms by farms experiencing PRRSV-2 outbreaks.

While these findings provide a unique window into the dynamics of between-farm PRRSV-2 transmission, one limitation of the analysis is that it focused on the voluntarily submitted ORF5 gene of a specific sub-lineage within a single, albeit large, swine producing region in the U.S., and we do not know the extent to which results can be generalized to other circulating PRRSV-2 variants or to other regions or countries that may have distinct farming practices. In addition, although we cannot fully elucidate PRRSV-2 evolutionary dynamics without full length genomes, genetic variation and phylogenies estimated from the ORF5 gene can yield comparable results when compared to genomic-level analysis (Frias-De-Diego et al., 2021) and constraining the analysis to highly related genetic variants (>98% pairwise identity) decreases the likelihood that evolutionary analyses would be confounded by recombination.

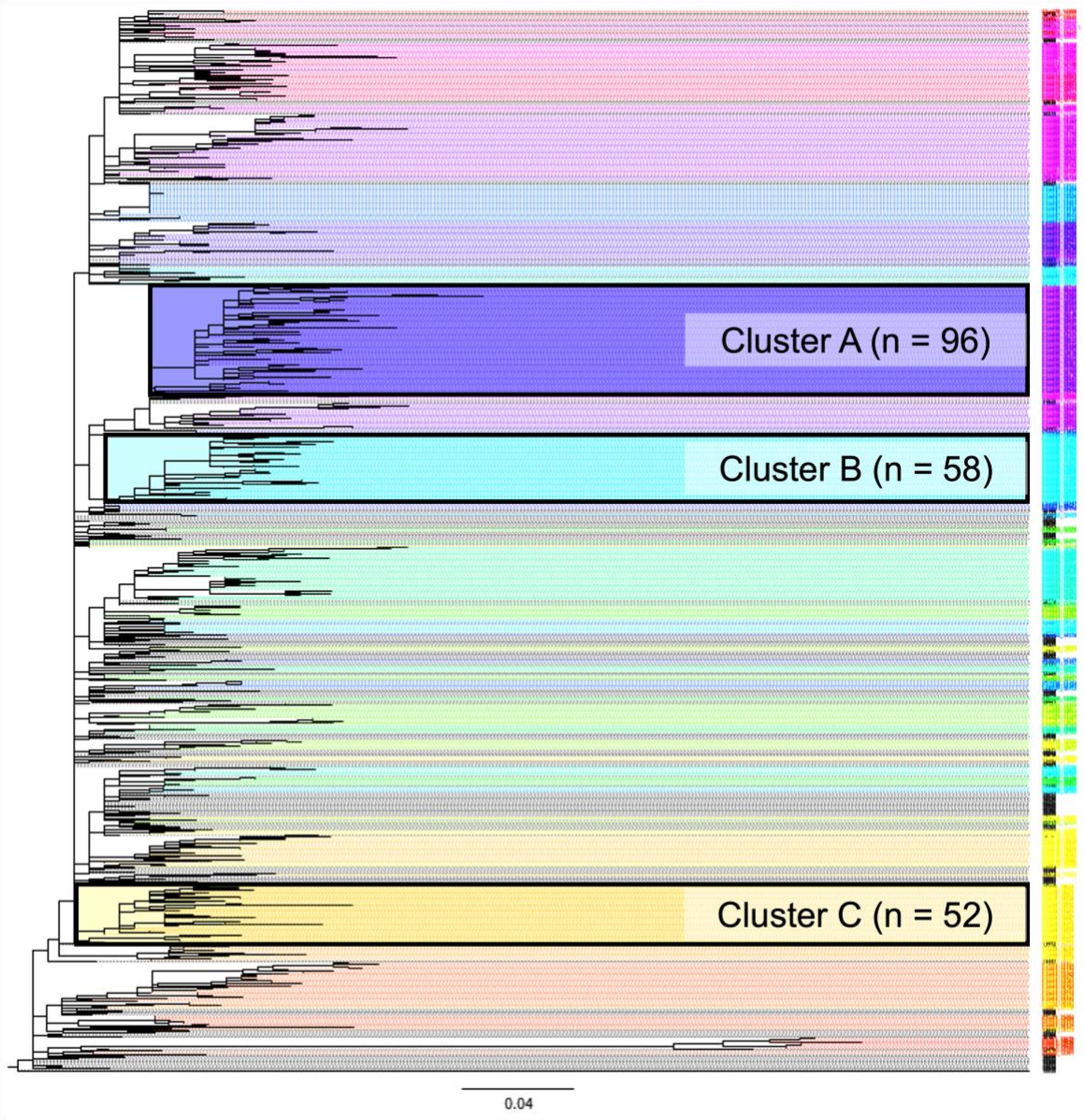
Although this study fully focused on PRRS epidemiology, the approach we designed can be applied to assess other infectious disease transmissions comparable to our context. Swine pathogens with genetic marker that represents their evolution, such as HA and NA genes of swine influenza virus, spike (S) gene of porcine epidemic diarrhea virus, or whole genome sequence of African swine fever virus, likely fit with our analysis since they are or will be circulating in the same farming system as PRRSV-2. Our principle that estimates the transmission between units (farms) comprising a group individuals (animals) can also be extended to analyze the potential transmission risks of other livestock or even human diseases if the unit, between-unit connection (animal movement in our case), and transmission routes are clearly defined.

Although the primary focus of this study was on PRRS epidemiology, the approach we devised has broader applicability and can be effectively employed to assess comparable infectious disease transmissions in similar contexts. Swine pathogens harboring genetic markers indicative of their evolution, such as the HA and NA genes of the swine influenza virus, the spike (S) gene of the porcine epidemic diarrhea virus, or the whole genome of the African swine fever virus,

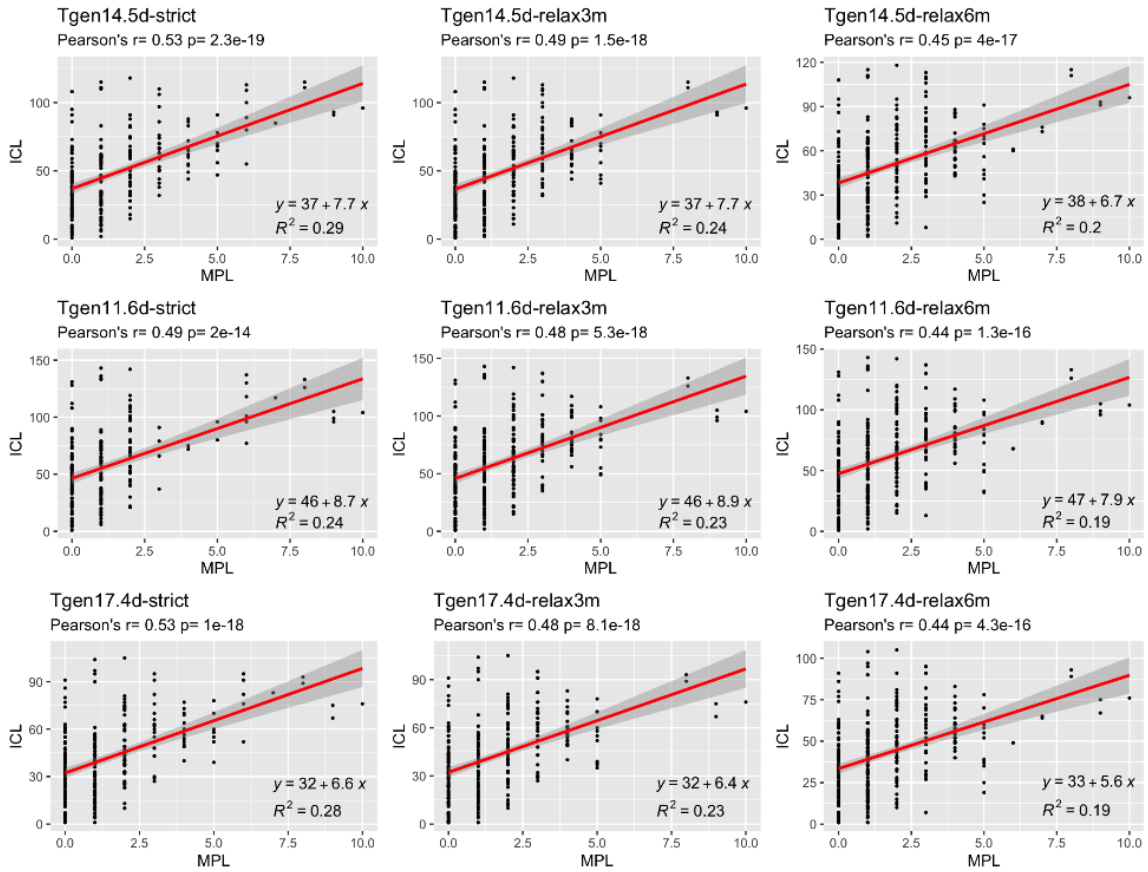
are likely compatible with our analytical framework, as they either circulate or are anticipated to emerge within the same farming system as PRRSV-2. Moreover, our fundamental principle, which estimates transmission between units (farms) comprising groups of individuals (animals), can be readily extended to evaluate the potential transmission risks of other livestock diseases or even human diseases, provided that the units, between-unit connections (as exemplified by animal movement in our case), and transmission routes are clearly defined.

The phylogenetic clusters outlined in this chapter exemplify a PRRSV-2 variant that successfully emerged and spread within a specific region over a defined period. Moving forward to the next chapter, the study identifies hundreds of similar emerging variants across the U.S. by systematically analyzing a decade's worth of PRRS monitoring data. Moreover, the study suggests that the phylogenetic attributes of PRRSV-2 ORF5 gene could serve as an early indicator for the potential emergence of new variants through predictive modeling.

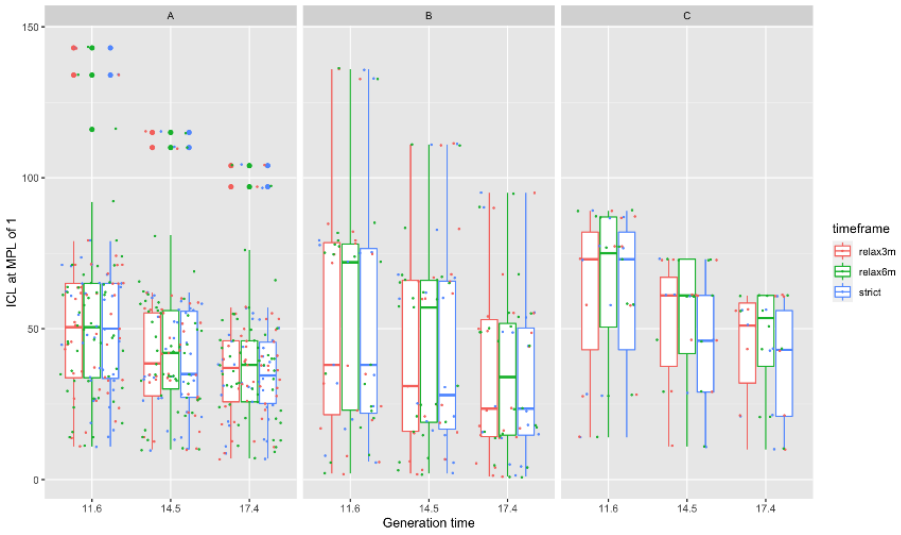
3.5: Supplementary Materials



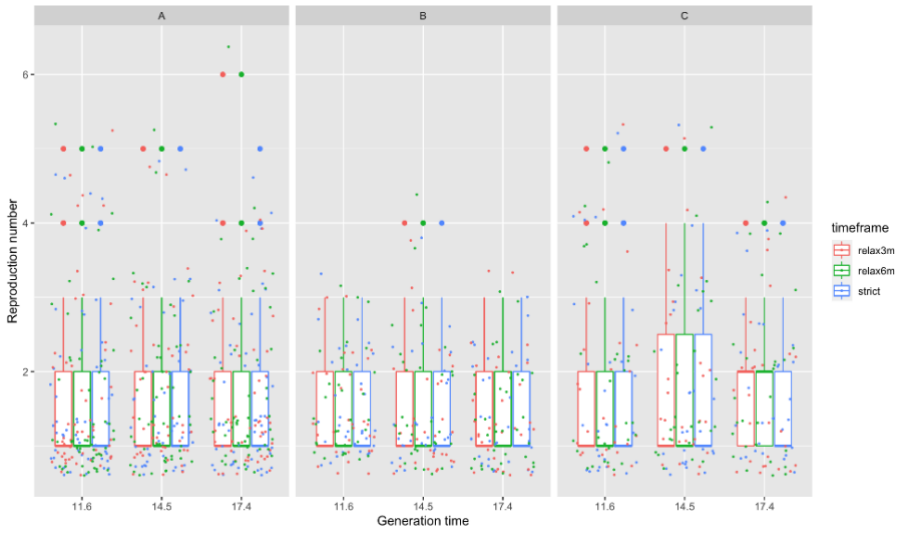
Supplementary Figure S3.1: Maximum likelihood tree of 943 ORF5 gene sequences collected from 651 farms between 2014 – 2017. Tips were colored based on phylogenetic clusters defined by clade support > 70% and maximum within group genetic distance < 4.5%. The largest three clusters (A to C) used in this study were framed in bold black rectangle.



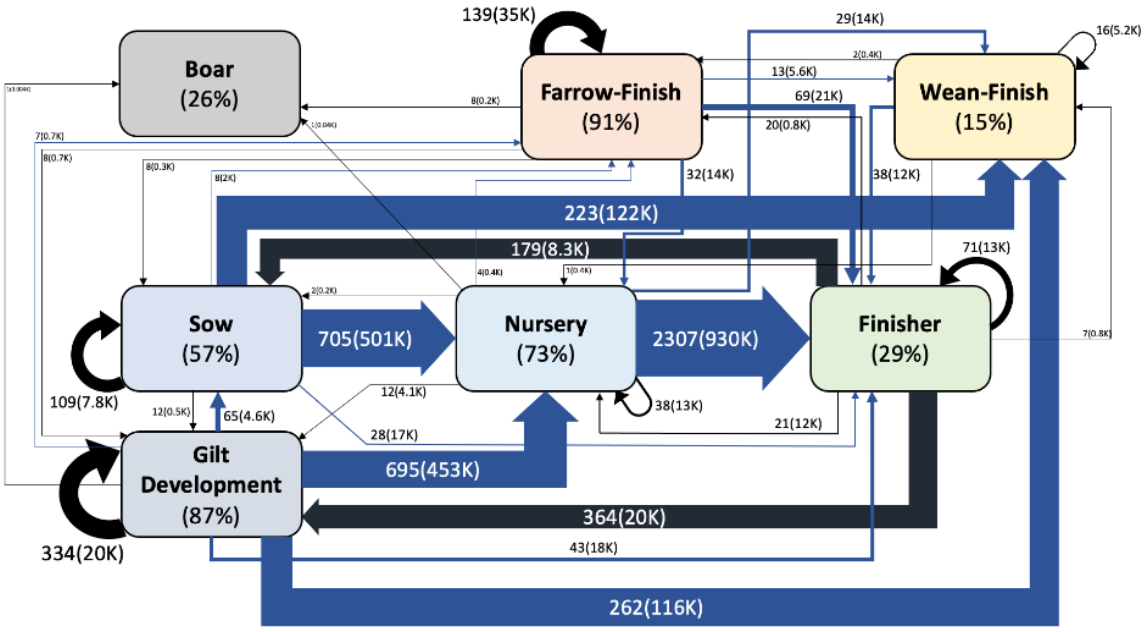
Supplementary Figure S3.2: Sensitivity analysis of the correlation between infection chain length (ICL) and movement pathlength (MPL) of all clusters varied by PRRS mean generation time (T_{gen}); 11.6, 14.5, and 17.4 days, and the time frame for capturing matched animal movement events; strict (onset to terminus of an inferred infection chain), relax3m (3 months prior to the onset to terminus), relax6m (6 months prior to the onset to terminus).



Supplementary Figure S3.3: Infection chain length (ICL) at movement pathlength (MPL) of 1 of each cluster (A to C) varied by PRRS mean generation time (T_{gen}); 11.6, 14.5, and 17.4 days, and the time frame for capturing matched animal movement events; strict (onset to terminus of an inferred infection chain), relax3m (3 months prior to the onset to terminus), relax6m (6 months prior to the onset to terminus).



Supplementary Figure S3.4: Farm-level effective reproduction number (R) of each cluster (A to C) varied by PRRS mean generation time (T_{gen}); 11.6, 14.5, and 17.4 days, and the time frame for capturing matched animal movement events; strict (onset to terminus of an inferred infection chain), relax3m (3 months prior to the onset to terminus), relax6m (6 months prior to the onset to terminus).



Supplement Figure S3.5: Monthly animal movement among pig sites in the study production system from January 2014 to December 2017. (Box) The number in parentheses beneath production type represents the mean proportion of active sites. (Arrow) The arrow thickness and the number within the arrow demonstrate the mean number of monthly shipments with the mean number of pigs in parentheses. The directions of movement in comparison to the production flow are represented by arrow's color (Blue; toward the flow, Black; against the flow or movement to the same production type)

“The Prophecy of Variant Awakening”

Chapter 4: Predicting PRRSV-2 Emergence Potential through Phylogenetic Inference.

Material adapted from a manuscript prepared for Transboundary and Emerging Diseases

Predicting Potential PRRSV-2 Variant Emergence Through Phylogenetic Inference.

Nakarin Pamornchainavakul, Mariana Kikuti, Igor A. D. Paploski, Cesar A. Corzo, and Kimberly VanderWaal

4.1: Introduction

Infectious disease emergence and re-emergence has posed significant challenges to human well-being over the centuries. Despite advancements in technology, the ability to preemptively prepare for such unexpected events remains limited unless there is a high degree of emergence predictability. These threats extend beyond human and zoonotic diseases that directly impact human health; they also include livestock diseases that undermine food security. Today, PRRSV is endemic in swine herds worldwide (Valdes-Donoso & Jarvis, 2022) and remains a significant concern due to its enormous economic consequences (Holtkamp et al., 2013; Lunney et al., 2010; H. Nathues et al., 2017; Neumann et al., 2005). PRRSV genetic variants involved in contemporary outbreaks are distinct from the early virus (Paploski et al., 2021), reflecting the rapid mutation rate of the virus (Hanada et al., 2005). Some of these variants have been associated with distinct virulence or epidemic characteristics, presenting atypical clinical manifestations (Ruedas-Torres et al., 2021) or increased disease spreadability (Kikuti, Paploski, et al., 2021).

As shown in Chapters 2 and 3, the PRRSV-2 viral population is characterized by co-circulation and turnover of distinct genetic variants, and the

routine emergence and epidemic-like spread of novel variants across space and time (Pamornchainavakul, Makau, et al., 2023; Pamornchainavakul, Paploski, et al., 2023; Paploski et al., 2021). The genetic relationship between viral clades, as demonstrated through ORF5 gene phylogenies, generally reflects overall genome-level relationships, particularly in cases where recombination is absent within a given set of samples (Frias-De-Diego et al., 2021; Pamornchainavakul et al., 2022). Over time, the operational taxonomic unit (OTU) of ORF5 gene has undergone revisions in its classification methodology, transitioning from restriction fragment length polymorphism (RFLP) patterns (Wesley et al., 1998) to lineage and sub-lineage classification based on phylogenetic analysis (Paploski et al., 2019, 2021; Shi, Lam, Hon, Murtaugh, et al., 2010). Sub-lineages constitute the smallest phylogeny-based OTU, with typically less than 8.5% nucleotide dissimilarity within the group (Paploski et al., 2019, 2021), and are made up of finer-scale genetic clades, here referred to as “variants,” that potentially exhibit heterogeneous virulence or epidemiological impacts (Kikuti, Paploski, et al., 2021; Pamornchainavakul et al., 2022).

In addition to implementing biosecurity measures, vaccination plays a crucial role in mitigating clinical PRRS outbreaks on farms that have tested positive for the virus (Holtkamp et al., 2011; The OIE AD HOC group on porcine reproductive respiratory syndrome, 2008). While the precise mechanisms of immunity against PRRSV, particularly regarding neutralizing antibodies, have yet to be fully understood (Murtaugh & Genzow, 2011; Nan et al., 2017), the genetic diversity within epitopes found on GP5 is recognized as a factor influencing immunological cross-protection. Consequently, in the past decade, there has been a growing trend for selecting virus strains for immunization that are either homologous or genetically more similar to field strains. Examples include the use of field viruses as autogenous inoculums to homogenize immunity within a herd (so-called live virus inoculation), the development of new commercial vaccines that are based on the currently most prevalent phylogenetic lineages (Chamba et

al., 2019; P. E. Yeske et al., 2021), and the use of killed vaccines matched to the amino acid sequence of particular epitopes (DeBuse, 2010). However, the effectiveness of using a “homologous” vaccine to confer optimal protection remains a topic of debate (Charerntantanakul, 2012; X. Li et al., 2014; Murtaugh & Genzow, 2011; Nan et al., 2017; Proctor et al., 2022; Roca et al., 2012), but one key challenge for such approaches is the continual emergence of new genetic variants (Paules et al., 2019; Wei et al., 2020).

Efforts to predict viral strain emergence have been successfully developed for certain human contagious diseases, with the aim of minimizing future outbreaks through informed vaccine strain selection. One pioneering example is the prediction of human seasonal influenza, where the fitness of different genetic variants is inferred from the branching patterns of each node on a phylogenetic tree—a metric known as the local branching index (LBI) (Neher et al., 2014a). Subsequent advancements have enhanced short-term prediction accuracy by incorporating LBI with tree shape and epitope features (Hayati et al., 2020). More recently, it has become possible to predict emergence of SARS-CoV-2 lineages by evaluating key amino acid substitutions and spatio-temporal prevalence data from millions of genomes, without the need for complete phylogenetic tree reconstruction (Obermeyer et al., 2022). Surprisingly, such informative prediction techniques have not been explored for PRRSV-2, despite the continuous generation of large amounts of genetic data and corresponding metadata through ongoing monitoring and surveillance. In this study, we leveraged a decade's worth of PRRSV-2 ORF5 gene sequences from one of the largest swine disease monitoring databases in the United States. Our objective was to systematically classify PRRSV-2 variants, assess their epidemiologic success over time in respect to population growth, geographic expansion, and genetic diversification, and develop predictive models that can be used to estimate a variant’s future emergence potential. By identifying variants of interest at a particular point of

time, our proposed model offers a proactive approach and provides an additional tool for achieving more precise PRRS control.

4.2: Materials and Methods

Data

PRRSV-2 ORF5 gene sequences collected from January 1, 2010, to June 30, 2021, were obtained from the Morrison Swine Health Monitoring Project (MSHMP), which is an ongoing monitoring program that archives, analyzes, and reports data related to major swine diseases. MSHMP monitors over 50% of the U.S sow population, and curates all PRRSV ORF5 gene sequences generated by MSHMP participants. Sequences are obtained directly from participants or from the main veterinary diagnostic laboratories where participants typically submit their diagnostic samples (University of Minnesota, Iowa State University, South Dakota State University, and Kansas State University). In U.S. swine production systems participating in MSHMP, ORF5 gene sequence data is typically generated when field veterinarians request ORF5 gene sequencing after confirmation of a PRRS outbreak by RT-PCR; often, sequences are generated for nearly every outbreak occurring on breeding farms within a production system. The phylogenetic lineage or sub-lineage of each sequence was subsequently determined based on its pairwise nucleotide distance to the reference sequences for each lineage (Paploski et al., 2019, 2021). Lineage 1 (L1) has been a predominant group of PRRSV-2 circulating in the U.S. during the recent decade (Pamornchainavakul, Paploski, et al., 2023; Paploski et al., 2019, 2021). The second and the third most common groups, namely L5 and L8, were largely associated with commonly used live attenuated vaccine strains (Cheng et al., 2022; Trevisan et al., 2022; Kikuti, Sanhueza, et al., 2021; Paploski et al., 2019). Hence, we used only L1 ORF5 gene sequences for the analysis of PRRSV-2 genetic variants. For this study, 20,700 complete length (603 nucleotides) L1 ORF5 gene sequences were compiled and then aligned using

the local alignment method in MAFFT v.7.310 (Katoh, 2002). All sequences had sampling date information and most sequences had corresponding spatial metadata including the US state (74.8% of all sequences) and county (66.2% of all sequences) (Supplementary Figure S4.1).

Phylogenetic reconstruction and variant assignment

Our goal was to identify early phylogenetic indicators that were predictive of a genetic variant's future epidemiological success (see below for metrics of success). Therefore, we approached this analysis by defining a series of sliding windows (Figure 4.1A) over which to quantify early indicators, and then correlate these indicators to the variant's future success in a follow-up period of time. The ORF5 gene alignment was used to reconstruct retrospective phylogenies across different windows of time. For each observation time (*t*), set as every six months starting from 1st January 2011 to 1st July 2020, we built two sets of "pre-trees" (time-scaled phylogenetic trees of sequences collected within the previous 12 or 24 months before time *t*) and four sets of "post-trees" (time-scaled trees created from the same sets of sequences in the pre-trees plus sequences collected within the following 12 or 24 months after time *t*) (Figure 4.1A – B). Some trees at the beginning and the end of the study period could not be built due to truncation of sequences 24 months before or after time *t*. From resampled alignments generated by PHYLIP's Seqboot v.3.69 (Felsenstein, 2009), each tree was initially built by FastTree v.2.1.10 (Price et al., 2010) using the maximum likelihood (ML) method, the GTR + CAT substitution model (generalized time-reversible with each site's rate approximation), and 100 bootstrap replicates. The ML tree bootstrap clade supports were then converted into the transfer bootstrap expectation (TBE) using BOOSTER v.0.1.1 (Lemoine et al., 2018), as transfer bootstraps typically yield better results for phylogenetic analyses with large datasets and rapidly evolving viruses (Lemoine et al., 2018). We defined PRRSV-2 variants based on the patristic distance, i.e., the sum of the shortest branch length connecting two taxa on the tree. The "Avg Clade" method of

TreeCluster v.1.0.3 (Balaban et al., 2019) was applied to each ML tree, which classified sequences into variants where a variant was defined as a monophyletic clade with an average pairwise patristic distance of <2% regardless of the clade support (Figure 4.1C). Using TreeTime v.0.9.2 (Sagulenko et al., 2018), branch lengths in each ML tree were re-estimated to generate two time-scaled phylogenetic trees, one tree using the default strict molecular clock model with highly diverging tips pruned and the other tree using the uncorrelated relaxed clock model without tree pruning. The highly diverging tips are tips for which residuals exceed four interquartile distances of the residual distribution in the least-square root-to-tip distance versus sampling date regression (Sagulenko et al., 2018).

Early indicators

The early indicators, which were considered as potential parameters in the predictive model, were either retrieved or calculated from the set of pre-trees (Figure 4.1C). First, we located the most recent common ancestor (MRCA) of each variant on the tree (i.e., variant's ancestral node and branch) using Biopython v.1.81's Bio.Phylo toolkit (Talevich et al., 2012) in Python (Van Rossum et al., 2009). Thereafter, we calculated four key categories of parameters related to the variant's ancestor, including ancestral branch length, local branching index (LBI), nucleotide substitution rate, and putative antigenic distinctiveness from contemporary most-prevalent variants. The ancestral branch length is a length of branch from the ancestral node to the closest deeper node in the original ML tree, and thus provides a metric of genetic divergence from other sequences in the tree. LBI is the sum of the tree length in each node's neighborhood, exponentially weighted by distance from the focal node (Neher et al., 2014b). Using Nextstrain's "augur lbi" command (Hadfield et al., 2018; Neher et al., 2014b), the LBI of each variants' ancestral node (ancestral LBI) was computed from the strict clock time-scaled tree with the tau (τ) parameter, which controls the size of neighborhood measured in units of the average pairwise

distance in the samples (Neher et al., 2014b), equal to 0.0625 times the average pairwise patristic distance of each particular tree, as recommended by Neher et al., 2014 (Neher et al., 2014b). Average pairwise patristic distance was calculated by “cophenetic.phylo” function in R’s ape v.5.6.2 (Paradis & Schliep, 2019; R Core Team, 2019). Nucleotide substitution rates for each variants’ ancestral branch (ancestral rate) were extracted from the relaxed clock time-scaled trees. We also averaged the substitution rates across all branches within a variant’s clade (average clade rate). Lastly, putative antigenic distinctiveness of each variant was measured in two ways based on the variant’s ancestral GP5 sequence (translated ORF5 amino acid sequence), with the hypothesis that variant’s whose sequences differ in antigenically relevant ways from the most prevalent variants at the time may be better able to escape population immunity present against those more prevalent variants. A variant’s ancestral sequence was inferred as part of the relaxed clock time-scaled tree building using the “ancestral” function in TreeTime v.0.9.2 (Sagulenko et al., 2018). Putative antigenic distinctiveness was measured as (1) ancestral amino acid distance—a pairwise amino acid distance (“dist.aa” function in R’s ape v.5.6.2) (Paradis & Schliep, 2019; R Core Team, 2019) between the ancestral GP5 to the consensus GP5 from all samples collected in the same calendar year, and (2) ancestral N-glycosylation pattern similarity—Jaccard similarity between potential N-glycosylation sites (positions having N-X-S/T sequons) (Gavel & Heijne, 1990) on the ancestral GP5 and the most frequent N-glycosylation pattern found in all samples of the recent calendar year. The Jaccard index ranges between 0 and 1, with lower values indicating fewer N-glycosylation sites in common and putatively greater antigenic dissimilarity. In total, these six parameters were considered as candidate early indicators.

Measures of success

For any given timepoint, the pre- and post-trees constituted separate phylogenetic reconstructions, so the first step of measuring success in the post-

tree was to identify the clade that corresponded to variants identified on the pre-tree. Because phylogenetic construction is an imperfect best-estimate of true underlying evolutionary relationships, topological differences between the pre- and post-tree meant that not all variant's present in the pre-tree were readily identifiable as monophyletic clades in the post-tree. We considered a pre- and post-tree variant to be matched if their members (i.e., the sequences present in both the pre- and post-tree analyses) were highly overlapping (>75% Jaccard similarity, indicating that 75% of sequence pairs belonged to the same variant in both the pre- and post-tree analyses).

Success of a variant was estimated from the new descendants of a variant in the post-tree and was characterized across three aspects—population expansion, spatial distribution, and genetic diversity (Figure 4.1D). We quantified population expansion of each variant by computing the absolute and relative increases in number of taxa from the pre- to post-tree. Spatial distribution of the variant was also estimated as the absolute and relative increases in number of states, and number of counties, in which the variant was detected. Additionally, pairwise geographical distance between county centroids were calculated between sequences belonging to the same variant. The maximum pairwise distance (as well as the 95th percentile to mitigate the effect of outliers) was extracted to approximate the geographic range of a variant in each pre- and post-tree. To measure changes in geographic extent, the absolute and relative increases of pairwise county distance (based on either the maximum or 95th percentile) was calculated for each post-tree variant compared to its geographic extent based on the original members from the pre-tree. Genetic diversity was measured as pairwise nucleotide distance ("dist.dna" function with K80 evolutionary model in R's ape v.5.6.2) (Paradis & Schliep, 2019; R Core Team, 2019) amongst all members of a variant, and the 95th percentile was used as the representative nucleotide distance of the variant (e.g., 95% of sequences belonging to a variant have a nucleotide distance of less that *x* distance). Then,

the absolute and relative increases in nucleotide distance were calculated between the pre- and post-tree. In total, 12 features were considered as potential measures of variant success.

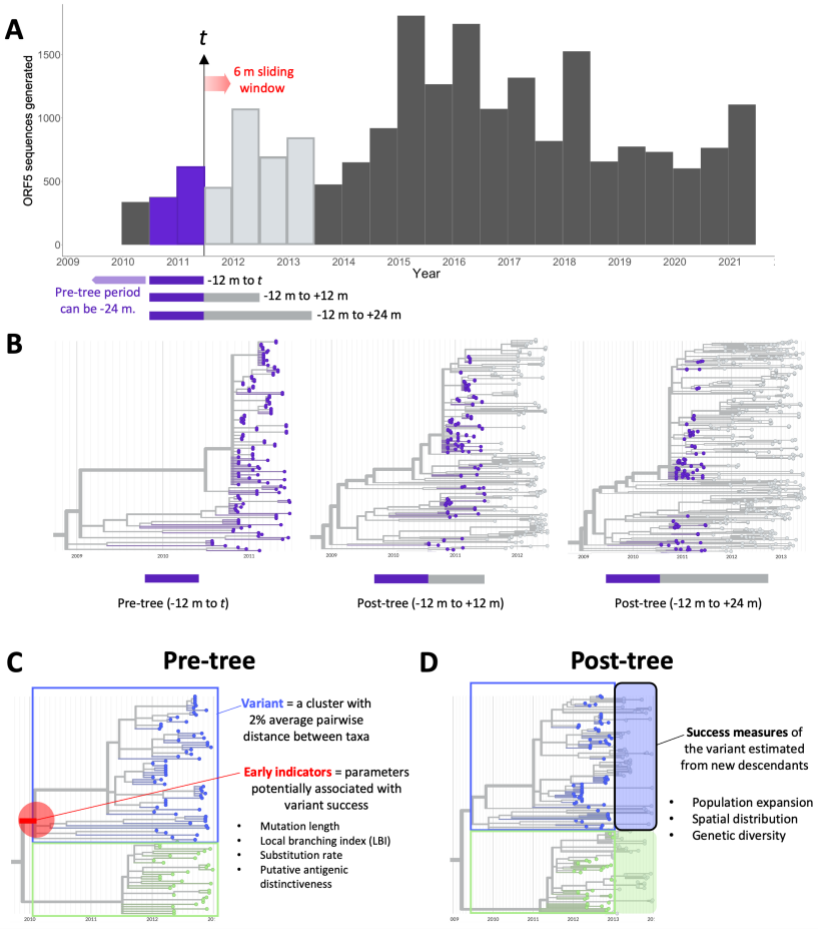


Figure 4.1: Conceptual framework of data generation for systematic predictive modeling. (A) Temporal distribution of PRRSV-2 L1 ORF5 gene sequences. As an example, observation time (*t*) is shown in July 2011 (vertical arrow) with its corresponding pre-tree (purple bars) and post-tree (purple and grey bars) periods. (B) Example pre- and post- timed phylogenetic trees inferred from sequencing data presented in plot A. Tips in purple show sequences from the pre-tree that are present in both post-trees. (C) Information computed in an example pre-tree, including designated variants (colored rectangle frames) and early indicators (red circle shows the ancestral node of the blue variant). (D) Success measures (colored oblong shape) calculated from variants' new descendants in the post-tree.

Predictive modeling

For each of four scenarios (12 or 24 months before and after *t*), a matrix of Spearman's correlation coefficients (ρ) was computed and visualized between all six early indicator candidates, using "ggpairs" function from the R's GGally v.2.2.0 and ggplot2 v.3.4.3 packages (Ginestet, 2011; Schloerke et al., 2022), to assess collinearity given that multivariable models can be severely impacted at collinearity of $|\rho| > 0.7$ (Dormann et al., 2013). All possible sets of non-collinear candidates were used as predictor variables. Given that successful variants appeared to be rare (early data exploration showed that the distribution of success metrics was highly right skewed), and because the numerical range of success metrics likely varied depending on the size of phylogenetic tree at different periods of time, a matched case-control study was applied using the observation time (*t*) of each scenario as a matched set (stratum). For each of 12 measures of success, variants whose success measure fell in the top 95th percentile were classified as a successful variants or "cases", whereas variants in the lower 75th percentile were classified as non-successful variants or "controls". Three controls were randomly selected from the same pre-tree for every case. Using "clogit" function in the survival package v.3.5.0 in R (Gail et al., 1981; R Core Team, 2019; Therneau, 2021), we fitted conditional logistic regression models on the training dataset using the first eight years (2011 – 2018) or approximately 80% of the data. Cases and controls selected from the last two years (2019 – 2020) of the data were used as a test set to validate the predictive model performance. To perform prediction on the test set, as described elsewhere, we first derived the average threshold value for each predictor that minimized misclassification rate in the training dataset, and these values were used to generate predictions for the test dataset (Reid & Tibshirani, 2014).

Amongst choices of models that differ by set of predictors (early indicators), response variables (measure of success), and scenario (length of time periods considered), only the models from the training set that had p-values

<0.05 for likelihood ratio tests were kept for further assessment, indicating that these models performed significantly better than a null model. For each aspect (population growth, geographic extent, and genetic diversification) and each follow-up period (short- versus long-term success [12 vs 24 months]), we selected the measure of success for each aspect that has the highest performance, as measured by mean concordance, (i.e., analogous to area under the ROC curve (AUC) for binary responses(Carrington et al., 2020)), based on the model fitted on the training set and the highest mean balanced accuracy based on the prediction on the test set. Then, we selected the n-month pre-tree model that maximized concordance and balanced accuracy for each n-month post-tree and selected measure of success. Coefficients, odds ratio, and p-values of each predictor in the final models, and model performance including sensitivity (SE), specificity (SP), positive predictive value (PPV), negative predictive value (NPV), F1-score and balanced accuracy (BA) on the test sets were reported. Furthermore, we applied these final models to predict the success of all observed variants (the full dataset) at each time point, aiming to evaluate the predictive performance on the data beyond the scope of the matched case-control design.

4.3: Results

A total of 74 unique sets of time-scaled phylogenetic trees with a median size of 4,247.5 (IQR = 2,688 – 6,712.75) taxa were reconstructed to obtain the pre-trees and the post-trees at each 6-month window observation time (*t*) throughout 2011 – 2020 for all four scenarios (12 or 24 months before and after *t*). Classified by 2% average pairwise patristic distance, the median number of variants per tree was 151 (IQR = 96 – 204), the median size of a variant was 12 (IQR = 5 – 30) taxa per variant, and the median bootstrap clade support of variant in all trees was 84% (IQR = 63 – 97). Only 58% of all the pre-tree variants could be matched (>75% Jaccard similarity between variants' members) with a post-tree variant. The number of matched variants varied from 53% to 63% of the

total pre-tree variants for each scenario. According to Welch two sample t-test, the bootstrap clade support of pre-tree variants with post-tree matches [mean = 85 (IQR = 80 – 100) %] was significantly higher ($p < 0.001$) than that of the unmatched variants [mean = 68 (IQR = 51 – 89)%]. An average of 11.7, 24.8, 5.3, and 13.9% of the total post-tree variants were new variants (no tips derived from the pre-tree period) for the 12-*t*-12, 12-*t*-24, 24-*t*-12, and 24-*t*-24 scenarios, respectively.

The candidate early indicators of variant success were obtained from the pre-tree variants' ancestral nodes (branch length, LBI, substitution rate, amino acid distance, and N-glycosylation similarity) and the whole variant clade (average clade rate). LBI was the only parameter that could not be computed for all ancestral nodes due to excessive branch length (higher than four interquartile distances from the clock model regression) in several time-scaled trees. We thus computed LBI only from the time-scaled trees for which the problematic branches were pruned. This resulted in a small proportion of variants (1.1% of all 4,323 matched variants) sharing the same ancestral LBI because their ancestral nodes were collapsed together during the tree pruning step. Amongst all six candidate indicators, severe collinearity ($|\rho| > 0.7$) was only detected between ancestral rate and average clade rate in the overall data and in every scenario (Figure 4.2). Therefore, two models were separately fitted for each measure of success (response) from the remaining four candidate predictors plus either ancestral rate or average clade rate.

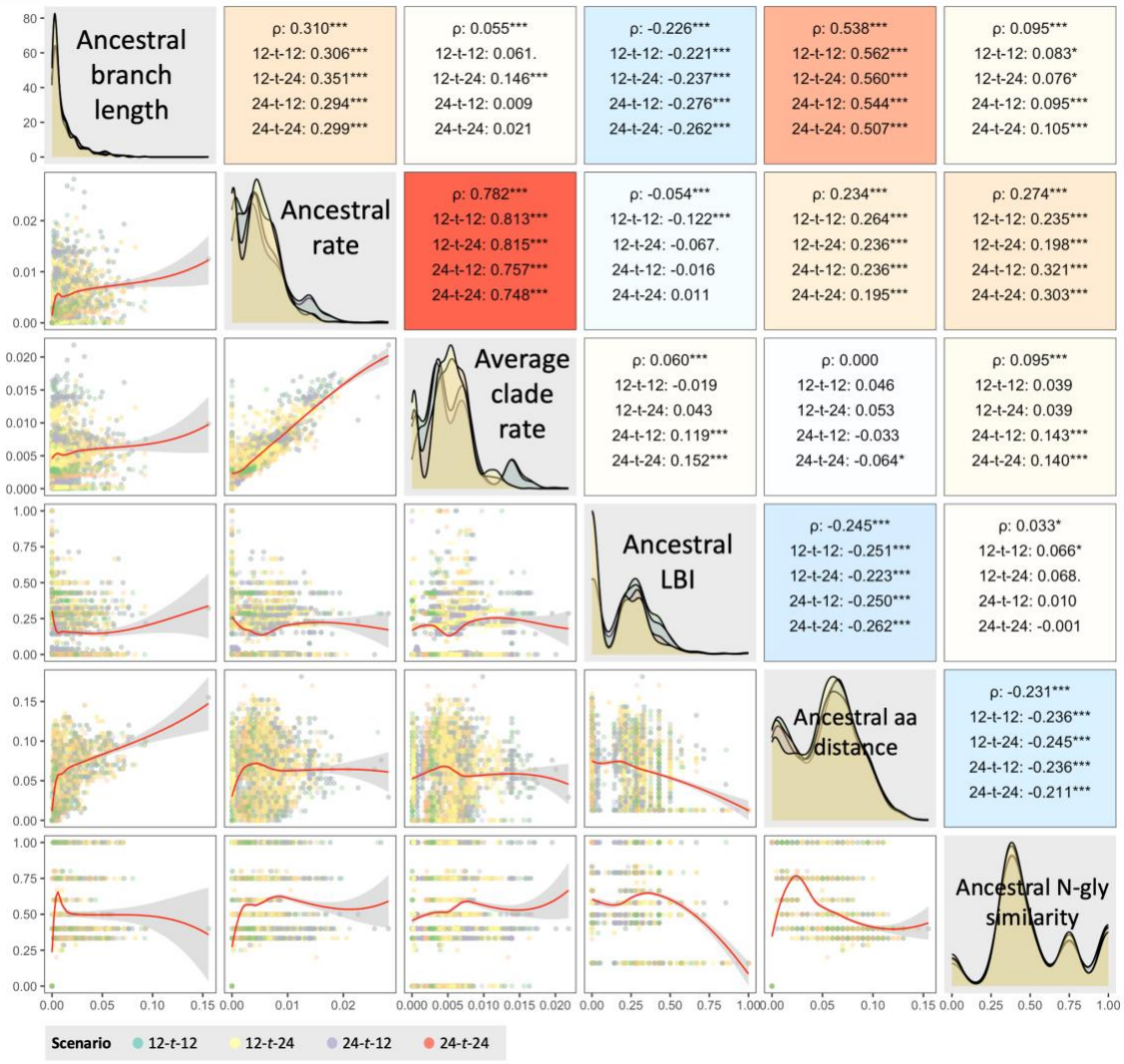


Figure 4.2: Matrix of Spearman's correlation coefficients (ρ) between all candidate early indicators for the overall data and each prediction scenario data with background color corresponding to the strength of correlation from 1 (red) to -1 (blue) (upper panel), their data density plots (diagonal), and bivariate scatterplots colored by the scenario with LOESS curves fitted (red line) and associated 95% confidence intervals (grey polygon) (lower panel).

Three aspects of variant success, comprising population expansion, spatial distribution, and genetic diversification, were calculated as absolute or relative increases of the measures when comparing between the matched post-

and pre-tree variants. Although we did control for impact of tree size or time on measures of success (e.g., time periods with greater sequencing effort could influence how many additional taxa a variant could increase by in the post-tree) by using a matched case-control design combined with a conditional logistic regression, the numerical distribution of success metrics were relatively consistent through time and even across different scenarios for most measures of success. For population expansion, the successful post-tree variants were typically at least twice the size of the original pre-tree variant [median relative increase in number of taxa = 400% (IQR = 283.3 – 804.7)] or had at least 20 more taxa than the pre-tree variant [median absolute increase in number of taxa = 70 (IQR = 44.5 – 112.5)] (Figure 4.3A). For genetic diversification, successful variants increased in their genetic diversity from the pre- to -post tree by the median distance of 0.01 (IQR = 0.008 – 0.018) or one site per 100 nucleotides, while the diversity of non-successful variants decreased by the median of 0.004 (IQR = 0.002 – 0.009) or 0.4 site per 100 nucleotides (Figure 4.3A). Geographic expansion metrics (except for number of states) were also well stratified between successful and non-successful variants, particularly when considering measures based on estimated geographical distance; successful variants often doubled their geographic extent, with distances increasing by up to 1000 km or more, whereas non-successful variants frequently displayed no increase whatsoever (Figure 4.3B). Such a high increase in geographic extent likely implies that successful variants are ones that have jumped between major swine producing regions (Pamornchainavakul, Paploski, et al., 2023).

A Venn diagram visualized by the R's Vennerable v.3.0 package (Swinton, 2023) was used to tabulate the number of variants that achieved success in one or more of the population, geographic, or genetic diversification aspects (Figure 4.3C). Interestingly, across scenarios, more than half (53 – 67%) of the successful variants in the population aspect (based on any of the population measures) were also successful based on their geographic dispersion. In

contrast, 48 – 60% of the successful variants based on genetic diversification did not successfully expand geographically or population-wise. Only 3 – 5% of successful variants achieved “success” across all three aspects. With that being said, most variants were not successful in any aspects (68 – 74%) or in only one aspect (17 – 23%) (Figure 4.3C).

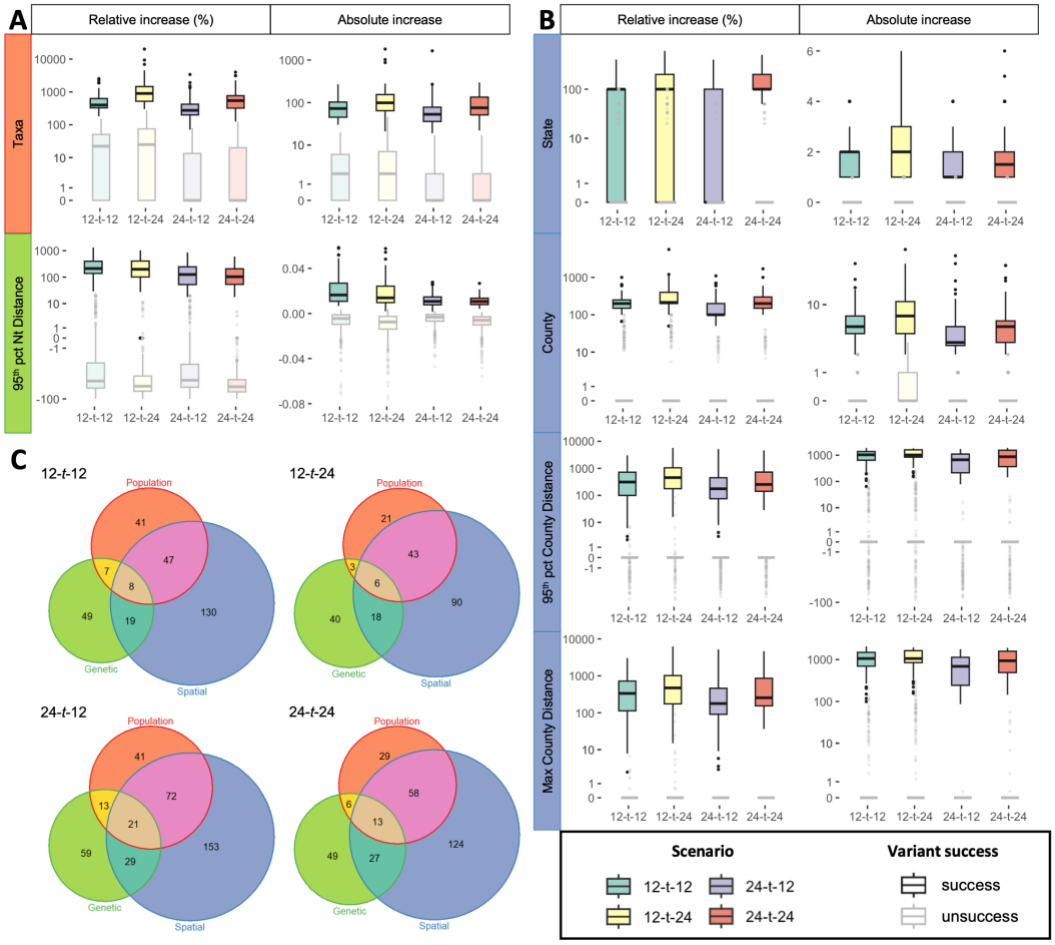


Figure 4.3: Aspects of success, success measures, and distribution of values for success vs. unsuccess of each measure. (A) Distribution of success metrics for population expansion (orange) and genetic diversity (green). (B) Distribution of success metrics for spatial distribution (blue). (C) Venn diagrams tabulating the number of variants that achieved success in one or more of the population, geographic, or genetic diversification aspects (not including success in relative increase in number of states).

Across the different combinations of candidate early indicators, measures of success, and temporal scenarios, 96 conditional logistic regression models were fitted on the training datasets. Only 41 models were considered significantly better than a null model ($p < 0.05$) based on the likelihood ratio test. From these models, we chose to perform subsequent predictive modeling on the success measures of that yielded the highest mean concordance (model fitting on the training set) and balanced accuracy (prediction on the test set, i.e., 2019 – 2020 data) for population expansion, spatial distribution, and genetic diversification. For population expansion, spatial distribution, and genetic diversification, the selected success measures were absolute increase in number of taxa (AbIn.Taxa), relative increase in maximum between-county geodesic distance (ReIn.MaxCounty.Dist), and absolute increase in 95th percentile pairwise nucleotide distance (AbIn.95Nt.Dist), respectively. According to the concordance and balanced accuracy metrics, utilizing data from the previous 12 months provided the overall most accurate predictions for all the selected successes in the subsequent 12 months. Similarly, using the data from the previous 24 months generally yielded the highest predictive performance for most successes in the following 24 months, except for the model predicting the ReIn.MaxCounty.Dist which had better performance when trained and tested on the previous 12 months of data to predict outcomes in the subsequent 24 months (Table 4.1). Thus, of all model's fitted as part of this analysis, six models are presented in Table 4.1, focusing on the three selected success metrics and primarily the 12-*t*-12 short-term and 24-*t*-24 long-term scenarios. Across these models, concordance for the training data was > 0.7 in all cases and balanced accuracy for predictions on new data ranged from 0.58 to 0.79. Predictive performance of other models was not as high.

The multivariate conditional logistic regression analysis showed that only one or two predictors out of the five examined were significantly associated ($p < 0.05$) with success in each model. Raw odds ratios are based on a 1 unit change

in the predictor variable, whereas the entire range of many of our variables was far less than one. To make odds ratios more interpretable based on the observed range of each predictor, we calculated an adjusted odds ratio based on upper and lower quartiles of each predictor in the training data, while keeping other predictors constant. This can be interpreted as how many more times a variant is to be successful when moving from the first to the third quartile values of an early indicator variable. For population expansion, variants in the upper quartile for ancestral LBI had approximately 12 – 13 times higher odds of being successful (AbIn.Taxa > 20 or having at least 20 more taxa in the follow-up period) in the next 12 or 24 months compared to variants in the lower quartile.

Regarding spatial distribution, variants with slower ancestral substitution rate (first quartile: 1×10^{-3} substitutions/nucleotide site/year (s/n/y)) had far greater chance of successfully extending their geographic distribution (ReIn.MaxCounty.Dist) in the next 12 months compared to variants with higher estimated substitution rates (third quartile: 7×10^{-3} s/n/y). When focusing on a 24 month rather than 12-month follow-up period, there was a significant association a variant's branch length and amino acid distance and the odds of increasing its geographic extent. Moving from the lower quartile (ancestral branch length of 0.003 – or ~1.8 nucleotides diverged from its next phylogenetic common ancestor) to the upper quartile (0.017 – or ~10.3 nucleotides diverged) quartiles of observed ancestral branch length, the odds of being the successful variant decreased by a factor of 10. Conversely, transitioning from the lower (0.025 – or ~5 amino acids different from the most prevalent PRRSV-2 GP5 sequence) to the upper (0.075 – or ~15 amino acids different) quartiles of observed ancestral amino acid distance resulted in 14.5 times greater odds of being a successful variant.

For genetic diversification, absolute increase in 95th percentile pairwise nucleotide distance (AbIn.95Nt.Dist) showed a significant association with branch length in the short-term period, while being associated with amino acid distance

in the long-term period. Specifically, variants with an ancestral branch length of 0.018 (upper quartile – or ~10.9 nucleotides diverged from its inferred phylogenetic ancestor) had 4.9 times higher odds of experiencing high levels of genetic diversification in the next 12 months compared to variants with a length of 0.003 (lower quartile – or ~1.8 nucleotides diverged). Moreover, variants with an ancestral amino acid distance of 0.075 (upper quartile – or ~15 amino acids different from the most prevalent PRRSV-2 GP5 sequences) were 6.8 times more likely to undergo diversify in the following 24 months than variants with a distance of 0.03 (lower quartile – or ~6 amino acids different). (Table 4.1)

The best fit models displayed fair to good predictive performance on the test set, depending on the success aspect and scenario. Notably, the prediction of genetic diversity variant success in the next 12 months achieved the highest rankings in both balanced accuracy and F1 score (BA = 0.79, F1 = 0.62). Following closely was the prediction of success in population expansion over the next 24 months (BA = 0.75, F1 = 0.57). Moreover, the prediction of variant success in population expansion within the next 12 months, as well as success in spatial dispersion and genetic diversity within the next 24 months, exhibited similar performance (BA = 0.67, F1 = 0.5). However, the prediction of spatial success over a 12-month timeframe was notably poorer in comparison (BA = 0.58, F1 = 0.43) (Table 4.1). When comparing model performance on predictions made on all the matched variants observed throughout the study period to model performance on the test set, it was found that the balanced accuracy of all the models was slightly lower (BA = 0.56 – 0.74). However, the F1 score showed a significant decrease of more than half (F1 = 0.15 – 0.27) due to a high proportion of false positives (Supplementary Figure S4.2 and Supplementary Table S4.1). For unmatched variants for which the future success could not be measured, the models predicted that 21% to 28% of these would be successful in terms of genetic diversity, while 59% to 68% of the variants were predicted to be successful in population or geographic expansion (Supplementary Figure S4.3).

Table 4.1: The best fit model for each success aspect and predicted period

Success aspect (Measure of success)	Scenario	Model p-value (LRT)	Predictor	P-value	Adjusted OR	Predictor's lower – upper quartiles	Training set	Test set					
							Concordance	SE	SP	PPV	NPV	F1	BA
Population expansion (Absolute increase in number of taxa: Abln.Taxa)	12-*t*-12	0.002	Branch length	0.098	0.12	0.00 – 0.02	0.875	0.75	0.58	0.38	0.88	0.500	0.667
			Average clade rate	0.808	0.4	0.00 – 0.01							
			LBI	0.015*	12.12	0.03 – 0.333							
			Amino acid distance	0.159	3.15	0.020 – 0.075							
			N-gly Similarity	0.410	1.79	0.333 – 0.750							
	24-*t*-24	0.001	Branch length	0.120	0.08	0.002 – 0.013	0.810	1.00	0.50	0.40	1.00	0.571	0.750
			Ancestral rate	0.446	3.65	0.002 – 0.007							
			LBI	0.030*	12.67	0.000 – 0.286							
			Amino acid distance	0.652	1.42	0.030 – 0.075							
			N-gly Similarity	0.939	1.09	0.333 – 0.750							
Spatial distribution (Relative increase in maximum between-county geodetic distance: Reln.MaxCounty.Dist)	12-*t*-12	0.006	Branch length	0.587	1.34	0.003 – 0.018	0.729	0.75	0.42	0.30	0.83	0.429	0.583
			Ancestral rate	0.035*	0	0.001 – 0.007							
			LBI	0.116	6.04	0.026 – 0.333							
			Amino acid distance	0.725	1.38	0.020 – 0.075							
			N-gly Similarity	0.081	3.24	0.333 – 0.750							
	12-*t*-24	0.029	Branch length	0.044*	0.1	0.003 – 0.017	0.792	0.50	0.83	0.50	0.83	0.500	0.667
			Ancestral rate	0.136	0.17	0.002 – 0.008							
			LBI	0.293	5.13	0.026 – 0.322							
			Amino acid distance	0.022*	14.47	0.025 – 0.075							
			N-gly Similarity	0.620	1.39	0.333 – 0.750							
Genetic diversity (Absolute increase in 95th percentile pairwise nucleotide distance: Abln.95Nt.Dist)	12-*t*-12	< 0.001	Branch length	0.004*	4.92	0.003 – 0.018	0.917	1.00	0.58	0.44	1.00	0.615	0.792
			Ancestral rate	0.603	0.42	0.001 – 0.007							
			LBI	0.758	0.69	0.026 – 0.333							
			Amino acid distance	0.909	1.12	0.020 – 0.075							
			N-gly Similarity	0.779	1.27	0.333 – 0.750							
	24-*t*-24	0.006	Branch length	0.119	2.07	0.002 – 0.013	0.857	1.00	0.33	0.33	1.00	0.500	0.667
			Ancestral rate	0.880	0.8	0.002 – 0.007							
			LBI	0.958	0.95	0.000 – 0.286							
			Amino acid distance	0.045*	6.78	0.030 – 0.075							
			N-gly Similarity	0.535	2.09	0.333 – 0.750							

4.4: Discussion

In this study, we utilized over 10 years (2010 – 2021) of PRRSV-2 ORF5 gene sequencing data representing PRRS circulation across the U.S. to retrospectively evaluate the predictability of potential emerging variants across time. Each phylogenetic-based viral variant was systematically traced through time over both short-term (12 months) and long-term (24 months) periods in order to first calculate putative early indicators from retrospective phylogenetic inference and then quantify various aspects of epidemiologic success during the follow-up period. Primarily, we found that variants that were classified as successful through population growth likely were also classified as successful through geographic expansion, but typically did not show notable genetic diversification. Across all models presented in Table 4.1, the early indicators which were significantly associated with variant success at least once included local branching index (LBI), branch length, mutation rate of the ancestral branch, and amino acid distance from the most prevalent contemporary GP5 sequences. The best predictive performance was achieved in the models that predicted long-term population growth using local branching index (LBI) and short-term genetic diversification using ancestral branch length. When applied to new data, these models successfully captured a significant number of successful variants with good sensitivity though relatively low specificity. The positive predictive value of the predictions was poor, as many predicted successful variants turned out to be false positives.

More generally, virus emergences can be assessed using diverse criteria, including abundance, adaptability, host range, and diversity (Wasik & Turner, 2013). For example, reproductive success, gauged by an effective reproduction number (R_e) above 1, indicates a virus's ability to emerge and spread in a population (DeFilippis & Villarreal, 2000; Geoghegan et al., 2016). Thus,

measures of emergence success commonly employ prevalence and growth rate for predictions (Hayati et al., 2020; Neher et al., 2014c; Obermeyer et al., 2022). For PRRSV-2, measuring the impact of the emergence of novel variants goes beyond case numbers, and should encompass spatial spread and variant genetic diversity, which play a crucial role in determining disease control efficacy. Our study showed emerging variants, as indicated by higher detection rates (population increases), also displayed extensive geographic expansion but lacked substantial genetic diversification within the given timeframe. These findings align with the idea that highly abundant variants have a greater likelihood of geographic dissemination compared to others, especially through routes such as infected animal transport (C. Nathues et al., 2016; Thakur et al., 2015). Additionally, these results suggest the presence of genetic bottlenecks and founder effects (H. Li & Roossinck, 2004; McCrone & Lauring, 2018) in PRRSV-2, wherein the widespread transmission of genetically similar viruses likely stems from the initial emergence of a highly fit variant or a founding population. Therefore, when defining the concept of emergence, it is essential not to restrict it solely to genetic diversity, particularly when using our criteria of diversity over the following 12 or 24 months.

Various parameters such as LBI (Neher et al., 2014a), epitope features (Hayati et al., 2020), phenotypic data (Huddleston et al., 2020), and consensus sequence of the current viral population (Barrat-Charlaix et al., 2021) have been shown to have implications in forecasting seasonal influenza A virus (IAV). However, it remains unclear whether these early indicators are useful when applied to PRRSV-2 data, as these two viral species differ significantly in their evolutionary dynamics. For instance, IAV branching patterns, which affect LBI values, are shaped by short-lived viral variants that quickly go extinct, and selective sweeps caused by frequent antigenic drift in IAV, resulting in a comb-like or ladder-like genealogy tree (Grenfell et al., 2004; Poon et al., 2013). In contrast, the persistent co-circulation and sequential dominance of various

PRRSV-2 sub-populations (Pamornchainavakul, Paploski, et al., 2023; Paploski et al., 2021) give rise to phylogenetic clades with a bush- or star-like structure (short internal branches and long external branches). Nevertheless, our analysis demonstrates that LBI was the only indicator that showed significant predictive power in forecasting both short-term and long-term population expansion of PRRSV-2 in the best fit models, and LBI was significant in 83.3% of all the models with significant predictor(s). This suggests that the underlying assumption of LBI, which is that rapid branching patterns is associated with high fitness of the inferred ancestor (Neher et al., 2014c), may also apply to PRRSV-2 phylodynamics.

Understanding of immune epitopes, antigenic properties, and genotype-phenotype correlation of PRRSV-2 remains incomplete and cannot be directly inferred from sequencing data (J. Li & Murtaugh, 2012; Loving et al., 2015; Martínez-Lobo et al., 2011). This lack of knowledge makes it challenging to incorporate such features into prediction models. To address this issue, we developed two indicators to capture the putative distinctiveness of a variant compared to the current most prevalent GP5 protein at a given point in time: GP5 amino acid distance and N-glycosylation pattern similarity. The GP5 amino acid distance parameter is relatively similar to the best predictor used to forecast IAV reported by Barrat-Charlaix et al. (Barrat-Charlaix et al., 2021). Variants with more divergent GP5 proteins were >6 times more likely to become more genetically diverse, and >14 times more likely to undergo spatial expansion than less divergent variants. We hypothesize that this metric captures, in part, the extent to which epitopes found on GP5 may differ from those recognized by the prevailing immunity in the population, hence more divergent variants may be able to better evade pre-existing immunity at the population level. That being said, the nature of our data did not allow us to know whether the emerging variants infected the same animals (or even farms) that were previously exposed to the prevailing GP5.

N-glycosylation pattern similarity parameter focuses on specific amino acid sites involved in potential glycan shielding, which is one immune evasion mechanism utilized by PRRSV-2 (Ansari et al., 2006; Faaberg et al., 2006). These sites have evolved under positive selection pressure (Delisle et al., 2012; Do et al., 2016; Hu et al., 2009; Paploski et al., 2019) and are believed to be associated with emergence at the sub-lineage level (Paploski et al., 2021, 2022). However, N-glycosylation pattern similarity was not significantly associated with a variant's success in any of our predictive models. Although previous work suggested that N-glycosylation pattern changes sometimes coincided with PRRS epidemic events, the patterns were not stable within a sub-lineage (only 40 – 60% of sequences in sub-lineage shared a N-glycolsyltion pattern) (Paploski et al., 2022). Thus, N-glycosylation patterns may change too frequently to attribute a single pattern to a particular variant, as we did here.

Branch length of the ancestral node is a more fundamental component of phylogenetic trees compared to other parameters. It also was the sole significant predictor in the best-fit model predicting success in genetic diversity in the short term. Specifically, variants whose inferred ancestors had undergone greater evolutionary changes (longer branch lengths) were more likely to genetically diversify shortly after. We hypothesize that these rapidly evolving variants had not yet reached a state of fitness stability and hence continued diversifying during the early stage of emergence. Ultimately, the disparity in significant predictors between the short-term and long-term models, along with the notably superior performance of the short-term model, led us to conclude that branch length stands out as the most robust predictor for success in genetic diversity. Branch length was also a significant predictor alongside GP5 amino acid distance in the model predicting spatial success in long-term. The association of branch length to the spatial success (OR < 1) was opposite to its association to the genetic success (OR > 1) which was consistent with the observation that these two success measures are indicators of different aspects of the underlying

epidemiological dynamics (i.e., variants that expand geographically did not tend to undergo substantial diversification, Figure 4.3).

Ancestral and average substitution rates of variants were included as candidate early indicators since substitution rate is a metric estimating how fast the virus has evolved and was previously observed to be dependent on population size dynamics in particular conditions (Goldstein, 2013). However, ancestral substitution rate was a significant predictor only in the short-term model for spatial success, which performed poorly and yielded poor predictions. Thus, the substitution rate extracted from the time-scaled trees proved to be a relatively irrelevant indicator for all definitions of variant success.

An essential task for this research was defining a PRRSV-2 variant. Given the large dataset and the need to build numerous trees with different subsets of data, the primary phylogenetic tree building method we utilized (FastTree ML) was the most plausible approach, offering an adequately reliable overview of the genetic relationships among the viruses, albeit not the most accurate tree-building approach (Price et al., 2010; Zhou et al., 2018). Accordingly, variant classifications based on the tree's patristic distance (TreeCluster's Avg Clade (Balaban et al., 2019)) may differ if an alternative tree building method had been used. In our analysis, we used an average patristic distance cut-off of 2% to define variants, which proved suitable because the variant size and clade support remained consistent across various trees and scenarios and is in-line with thresholds conventionally used to define PRRSV sequences as homologous or heterologous (Murtaugh, 2012). However, this approach has limitations as it can lead to abrupt appearance of new variants in the follow-up period that appear to be >2% from any of the original sequences. New variants accounting for approximately 5 to 25% of total variants per observation time based on our results and have potentially significant implications. These occurrences could potentially signify the introduction of exotic variants, the re-emergence of under-the-radar variants absent from the current sequencing data, or the cyclic

emergence of new sub-lineages that occurs every 1 to 4 years (Paploski et al., 2021).

Another limitation of this research relates to the interpretation of a variant's successful geographic expansion. Numerous external factors (i.e., animal movement) contribute to the spatial dissemination of a virus that are not measurable from viral phylogenies nor related to a virus's phenotype. In addition, sampling locations were not known for all sequences, and not all pig producing regions of the U.S. were equally well represented in the dataset (Supplementary Figure S4.1). More comprehensive spatial data could improve model predictions, as could incorporating regional variability in the prevailing GP5 sequences and emergence success.

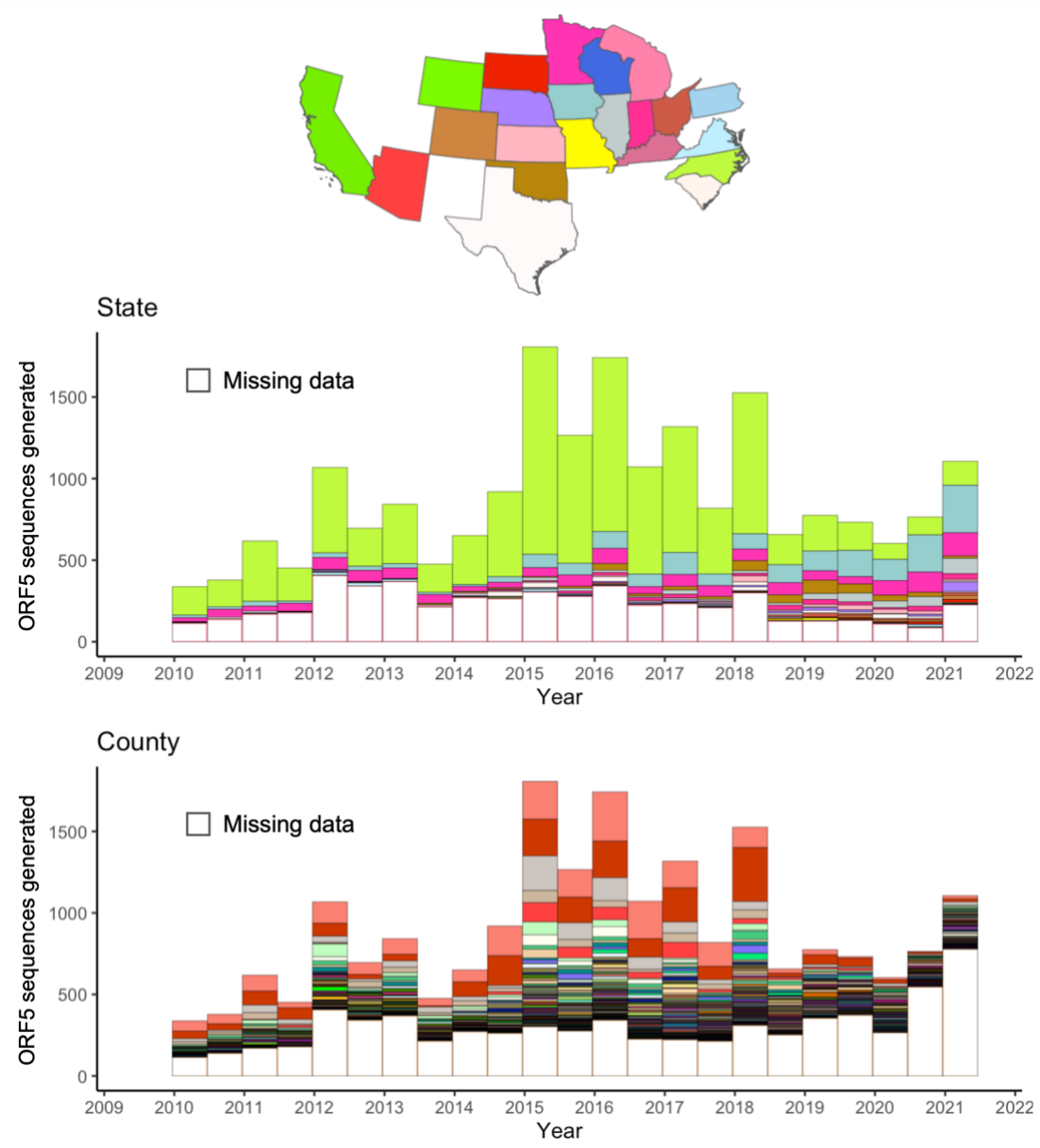
We identified certain early indicators that are associated with predicting various aspects of success. However, when implementing these models to all matched variants beyond the selected variants in the case-control design, one key issue undermining model performance was the low positive predictive values (PPV). This was unsurprising, given that the training case-control dataset had a significantly higher proportion of successes than the overall data. Because of the low PPV, predictions of variant success could be better interpreted as identifying those variants with high emergence potential (with not all variants realizing their potential) as opposed to a projection of what will happen with certainty. In addition, population and spatial emergence success was commonly predicted for mismatched variants, whose actual success could not be measured. If predictions are generated prospectively, we suggested filtering out such variants by removing variants with low clade support (< 75%) before making predictions.

Despite the low positive predictive value, our models successfully identified a variant with nine taxa in the initial tree (based on 12 months of data from 2019 – 2020), which ultimately led to the impactful emergence of the novel L1C-1-4-4 outbreak in the Midwestern U.S. This prediction was made as early as

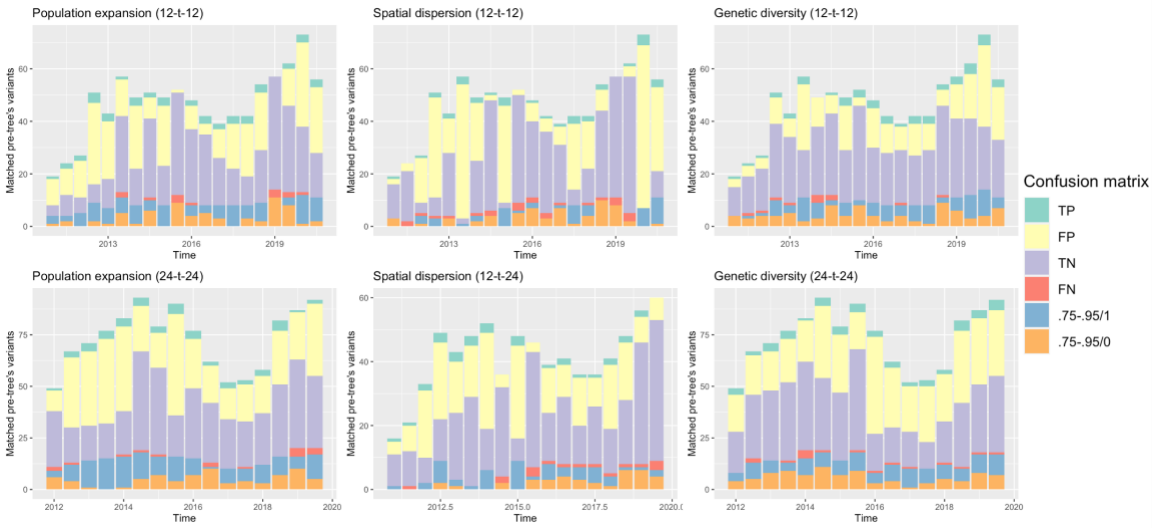
January 2020, more than six months prior to the first official notice of the outbreak in fall 2020 (Kikuti, Paploski, et al., 2021). The models correctly anticipated that this variant would exhibit both population growth and geographic expansion but would not undergo significant genetic diversification, aligning with our hypothesis of a high-fitness variant. Nevertheless, we acknowledge that our work represents just the initial stage of developing methods for prediction of the emergence of PRRSV-2 variants, highlighting informative phylogenetic-based early indicators for a variant’s emergence. It is crucial to continue improving our approach by incorporating better spatial-related metadata, expanding the training and the test sets with more data in the future, and exploring additional potential predictors.

This chapter demonstrates an example of utilizing ORF5 gene phylogenetic inference, representing PRRSV-2 genetic diversity, for predicting emerging variants. Nevertheless, the genetic characteristics contributing to the success of a particular variant are not solely confined to the ORF5 gene. In the following chapter, a collection of whole genome sequences from a newly emerged PRRSV-2 variant is employed to comprehensively investigate the potential evolutionary mechanisms responsible for the variant's success.

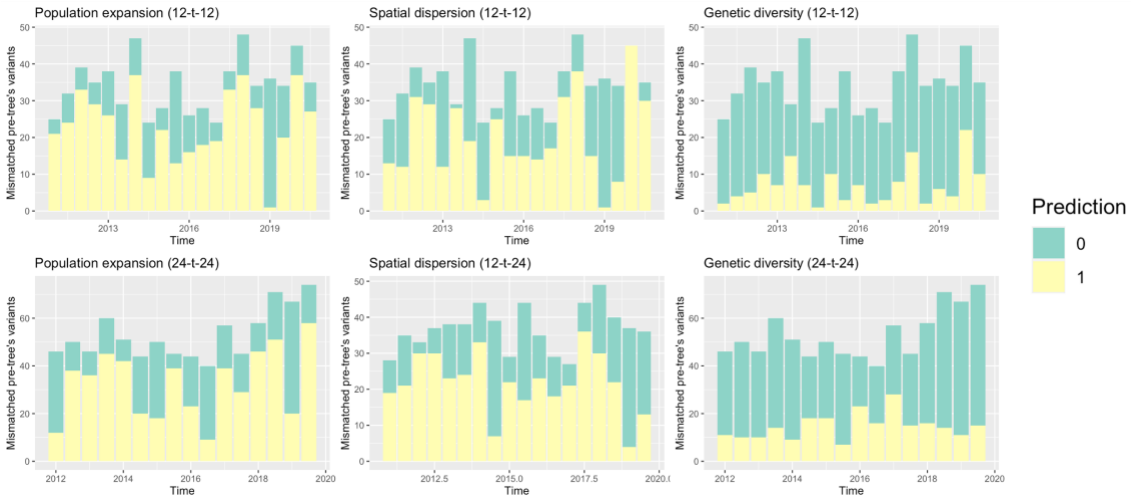
4.5: Supplementary Materials



Supplementary Figure S4.1: Temporal distribution of PRRSV-2 L1 ORF5 gene sequencing data with spatial data (state and county) availability. Color of states in the top map correspond to colors in the distribution plot at state level (middle).



Supplementary Figure S4.2: Predictive performance of the best fit model of each success aspect and predicted period on the full dataset (2011 – 2020) demonstrated by confusion matrix components (TP: True positive, FP: False positive, TN: True negative, FN: False negative, .75 –.95/1: Intermediate between success and unsuccess that was predicted to be positive, .75 –.95/0: Intermediate between success and unsuccess that was predicted to be negative).



Supplementary Figure S4.3: Number of positive (1: yellow) and negative (0: turquoise) predicted mismatched pre-tree's variants that the true success could not be measured.

Supplementary Table S4.1: Predictive performance of the best fit model of each success aspect and predicted period on the full dataset (2011 – 2020).

Success aspect	Scenario	PPV	NPV	SE	F1	SP	BA
Population expansion	12-*t*-12	0.13	0.97	0.78	0.22	0.56	0.67
	24-*t*-24	0.1	0.97	0.76	0.18	0.52	0.64
Spatial distribution	12-*t*-12	0.09	0.95	0.58	0.15	0.54	0.56
	12-*t*-24	0.1	0.96	0.65	0.18	0.58	0.62
Genetic Diversity	12-*t*-12	0.16	0.98	0.81	0.27	0.66	0.74
	24-*t*-24	0.11	0.97	0.77	0.19	0.52	0.65

“The Chosen One”

Chapter 5: Tracing the Origin of a Novel PRRSV-2 Variant through Genome-based Phylodynamics

Material adapted from a published article in Frontiers in Veterinary Science 9 (2022), doi: 10.3389/fvets.2022.846904

Measuring How Recombination Re-shapes the Evolutionary History of PRRSV-2: A Genome-Based Phylodynamic Analysis of the Emergence of a Novel PRRSV-2 Variant.

Nakarin Pamornchainavakul, Mariana Kikuti, Igor A. D. Paploski, Dennis N. Makau, Albert Rovira, Cesar A. Corzo, and Kimberly VanderWaal

5.1: Introduction

In the beginning of 2020s, pig producers in the Midwestern U.S. experienced atypical production losses caused by a fast-spreading variant of PRRSV-2 (Kikuti, Paploski, et al., 2021). From early 2020 to September 2021, 355 genetically similar viruses were detected (i.e., > 98% nucleotide identity based on ORF5 gene). Based on data from the Morrison Swine Health Monitoring Project (MSHMP), which tracks the infection status of ~50% of the U.S. breeding herd, 294 pig sites belonging to 15 different production systems, and ~12% of breeding farms in the region had been impacted (Kikuti, Geary, et al., 2021). The virus involved in this outbreak is referred to as a novel L1C-1-4-4 variant, as it falls within the L1C sub-lineage based on phylogenetic relatedness (Paploski et al., 2019) and mostly possesses a 1-4-4 cut-pattern based on conventional restricted fragment length polymorphism (RFLP)-based classification. Both classifications are based on ORF5 gene sequences (Wesley et al., 1998). While the exact case definition used is based on >98% nucleotide identity on ORF5 gene (Kikuti, Paploski, et al., 2021), here we refer to this variant simply as L1C-1-4-4. Although ORF5 gene has been widely used for virus

classification or epidemiological assessment since it is highly variable and immunologically relevant (Shi, Lam, Hon, Murtaugh, et al., 2010), it only represents ~4% of the genome and may not represent the full evolutionary history of the virus, particularly when recombination is involved. For example, when phylogenetic trees are constructed using whole genome sequences, the L1C-1-4-4 variant nests within a clade of viruses that are classified as sub-lineage L1A based on the ORF5 gene (Kikuti, Paploski, et al., 2021; Schroeder et al., 2021), suggesting that recombination may be confounding the virus' genealogical tree topology. Because there is no WGS-based nomenclature for classifying PRRSV-2 genetic sub-types, here we refer to WGSs according to their ORF5 gene lineage classification and recognizing the limitation that genetic relatedness might not hold true when looking at other regions of the genome.

The rapid spread and high production impact of the newly emerged L1C-1-4-4 variant has drawn concern from the industry as similar events occurred in the past and the contributing factors to emergence of these strains are poorly understood. Given that PRRSV-2 in the U.S. is characterized by the cyclic emergence of new strains, and turnover in the dominant sub-lineage every few years (Paploski et al., 2021), this emerging variant may continue increasing in prevalence, bringing further issues to the U.S. swine industry. However, different aspects of PRRSV evolutionary history documented in previous chapters were estimated based on genetic diversity of ORF5 gene sequences generated from past outbreaks, which limits our ability to fully discern evolutionary processes associated with strain emergence. For example, PRRSV-2 virulence and antigenic determinants are multigenic, meaning that clinical presentation characteristics are influenced by a variety of genes throughout the viral genome (Ruedas-Torres et al., 2021). Thus, understanding the origin of this variant from a whole genome perspective is a crucial step in response to this outbreak that had been occurred during the study period (2021 – 2022), and may help elucidate evolutionary processes associated with strain emergence more broadly and lead

to potential preventive interventions at the farm level. Here, we estimate divergence times, mutation rates, and parental strains of the novel L1C-1-4-4 variant using genomics-based approaches. More generally, to better understand the role of recombination in shaping PRRSV-2 phylogenies, we also quantify the frequency of potential inter- and intra-lineage recombination events.

5.2: Materials and Methods

Data

A convenience sample of PRRSV-2 L1C-1-4-4 variant whole genome sequences (WGS) were obtained from the Veterinary Diagnostic Laboratory at the University of Minnesota (see (Kikuti, Paploski, et al., 2021) for details). These samples were from multiple production systems in the Midwestern U.S. participating in the Morrison Swine Health Monitoring Project. Sample selection criteria for WGS included having ORF5 gene sequences within the emerging variant's phylogenetic clade (<2% genetic distance to at least one other sequence classified as L1C-1-4-4) and cycle threshold (Ct) value ≤25 for reverse transcription polymerase chain reaction (RT-PCR) using VetMAX™ NA and EU PRRSV Reagents (Thermo Fisher Scientific, MA, USA). Oral fluids and processing fluids samples were excluded due to the low success rate for whole genome sequencing (Gagnon et al., 2021; J. Zhang et al., 2017). At least one ORF5 gene sequence from each participating system was whole genome sequenced. For systems that had two or more ORF5 gene L1C-1-4-4 sequences identified during this period, the earliest and the most recent samples were selected for WGS sequencing. As described in Kikuti *et al.* (Kikuti, Paploski, et al., 2021), the selected samples were sequenced using Clontech SMARTer RNA Pico v2 kit on illumina MiSeq v3 (Illumina, CA, USA). Of the total 19 WGSs (GenBank accession numbers OL963961 – OL963979), one isolate classified as the novel L1C-1-4-4 variant based on the above criteria was from an outbreak limited to a single production system in 2018, while the others were collected

from multiple systems during the current epidemic (i.e., 2020 – 2021). We aligned the WGSs with all available PRRSV-2 WGSs from the U.S. that were publicly available and included date meta-data from NCBI GenBank (n = 232), ranging between 1995 and 2018 using MAFFT (Katoh, 2002) and manual curation. Genetic distances were calculated between all sequences using seqcombo (G. Yu, 2021).

Recombination detection

As a first step for screening sequences for recombination, the alignment was imported to RDP5 (D. P. Martin et al., 2021) for recombination detection. A putative recombination event was flagged when it was detected by at least 4 of 7 methods: RDP (D. Martin & Rybicki, 2000), GENECONV (Padidam et al., 1999), MaxChi (J. M. Smith, 1992), BootScan (D. P. Martin et al., 2005), SiScan (Gibbs et al., 2000), Chimaera (Posada & Crandall, 2001), and 3Seq (Lam et al., 2018). We performed the analysis as a two-pronged approach. First, we specifically explored recombination in the novel L1C-1-4-4 variant group, which was set as a query against all GenBank WGS references. Second, the alignment was fully scanned (with no reference and query groups defined) to estimate the location of recombination hotspots within the genome. A recombination hotspot is defined as a genomic position in which the frequency of putative recombination exceeds neutral expectations (>99% confidence interval of the local density plot created by a permutation test) (D. P. Martin et al., 2015); genomic regions between hotspots are inferred to have low rates of recombination. Thus, the locations of hotspots can be used to subdivide the genome into fragments, where each fragment is relatively free of frequent within-fragment recombination and thus can be used for further phylogenetic analysis (Aiewsakun et al., 2020; Brito et al., 2018).

Maximum likelihood phylogenies were built from each WGS fragment using W-IQ-TREE (Trifinopoulos et al., 2016) with automated substitution model

selection and 1,000 bootstraps. The consensus trees were assessed to: (1) check the temporal signal under a molecular clock assumption using TempEst (Rambaut et al., 2016), and any fragment whose phylogenetic reconstruction did not show a sufficient temporal signal was excluded from further time-scaled analyses. (2) Down-sample the dataset based on pairwise distances from the novel L1C-1-4-4 variant using *ape* version 5.5 (Paradis & Schliep, 2019) applied in R (R Core Team, 2019). Only the 50 most closely related sequences to each distinct fragment of the novel L1C-1-4-4 variant were retained, yielding a total of 142 sequences for further analysis.

Time-scaled phylogenetic reconstruction

The time to the most recent common ancestor (tMRCA) and substitution rate of each fragment were estimated by Bayesian inference with Markov chain Monte Carlo (MCMC) applied in BEAST v.1.10.4 (Suchard et al., 2018). According to IQ-TREE's substitution model test, we chose the general time reversible (GTR) with empirical base frequencies and gamma plus invariant site (G + I) heterogeneity model for all fragments. An uncorrelated relaxed clock (Drummond et al., 2006) with log-normal distribution and the Gaussian Markov random field (GMRF) skyride (Minin et al., 2008) were specified as molecular clock model and coalescent prior, respectively. The alignments with these model settings were run with 500 million generations of MCMC. Maximum clade credibility (MCC) trees of each fragment were built using TreeAnnotator v.1.10.4 (Drummond & Rambaut, 2007).

Discrete trait analysis

The frequency of inter- and intra-lineage recombination between ORF5 gene and other fragments was approximated through WGS-fragment phylogenies using discrete trait analysis in BEAST. For each WGS fragment, the ORF5-based lineage (Shi, Lam, Hon, Murtaugh, et al., 2010) or sub-lineage (Paploski et al., 2019) of each sample was assigned as a discrete trait, and the

ancestral trait of each internal node was inferred. Ancestral transitions between traits (i.e., the label the sequence received based on its ORF5 gene lineage) in the WGS-fragment phylogenies can be interpreted as putative recombination between the ORF5 gene and other WGS regions (i.e., instances where sequences are no longer clustered with other sequences that share the same ORF5-lineage label). Potential recombination in the WGS-fragment phylogenies were estimated from the number of trait (lineage) transitions with Bayes factors (BF) support obtained from an asymmetric substitution model with Bayesian stochastic search variable selection (BSSVS). Other parameters were set as the software default. The analyses were run with MCMC length of 500 million each. Ancestral states annotated on MCC trees were visualized using FigTree v.1.4.4 (Rambaut, 2018). Lineage and sub-lineage transitions were reported with BF computed by SpreaD3 (Bielejec et al., 2016). High numbers of transitions between inferred ORF5-lineage in the phylogenies of other WGS-fragments would provide support that recombination is more common, and that shared phylogenetic ancestry based on ORF5 gene lineage identity is scrambled on the whole genome due to putative recombination. Low transitions suggest that recombination events that leave descendants detected by surveillance activities are relatively rare, and that shared ancestry based on ORF5 gene lineages are relatively stable across the genome.

5.3: Results

The 18 novel L1C-1-4-4 WGSs associated with the 2020 – 2021 outbreak displayed a 98.2 to 99.9% nucleotide identity. The 2018 virus, which was included for whole genome sequencing based on its high similarity on ORF5 gene, showed a 96.5 to 97.4 % pairwise similarity to the 2020 – 2021 L1C-1-4-4 whole genomes. The greatest difference (<90 % similarity) between the 2020 – 2021 group and the 2018 virus was in nsp9 to nsp10 in the ORF1b gene (Supplementary Figure S5.1).

Recombination profile

All 19 WGSs had a relatively similar recombination profile, with at least 6 putative recombinant regions in common across the viral genome. The minor parents (i.e., the parent contributing the shorter part of the overall sequence) of several of these recombinant regions were viruses that clustered with other viruses classified as sub-lineage L1C in their ORF5 gene. For all 19 WGSs, a large recombinant region was identified in nsp2. The recombination detection algorithms implemented using RDP5 were not able to identify a feasible minor parent in the alignment of the 232 GenBank sequences for the nsp2 recombinant region and for a short recombinant region in ORF2 gene. The detectable minor parents of other events were identified as viruses belonging to the L1H (in nsp1) and L1C (in nsps2 – 9, and ORFs5 – 6 genes) sub-lineages based on their ORF5 gene variation (Figure 5.1B). Major differences in the recombination pattern of the 2020 – 2021 (n = 18) and the 2018 (n = 1) samples were found in the nsp9 to nsp12 of ORF1b gene, where the parents of the 2020 – 2021 sequences were other L1C, while the 2018's parents were mostly unknown (Figure 5.1B). In agreement with the estimated location of recombinant genomic regions, recombination breakpoints of this variant are located in the following genomic regions: nsp1 flanking regions, insertion and deletion (indel) sites of nsp2 (F. Yu et al., 2020), inside nsp9, ORF1ab-ORF2 junction, and ORF2 and ORF5 genes flanking regions (Figure 5.1A, C).

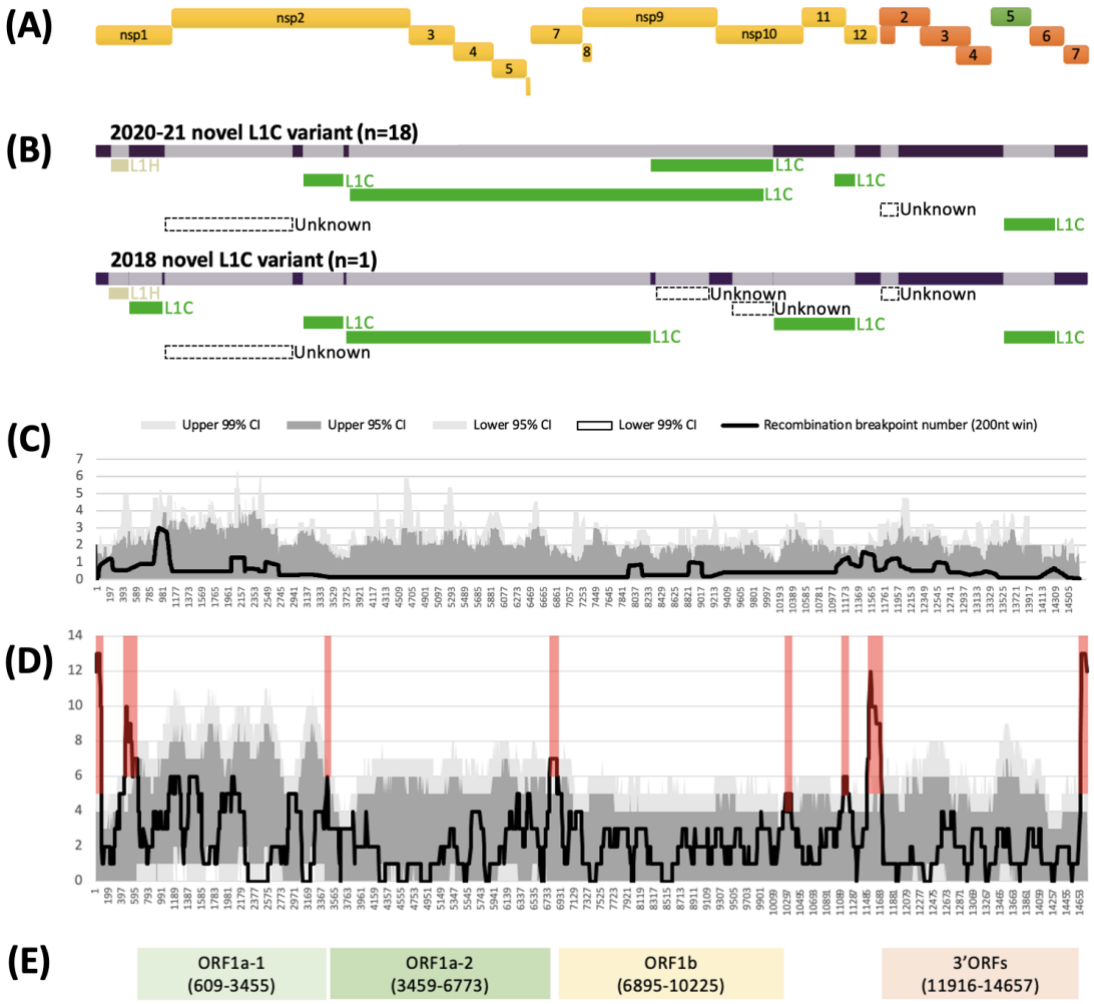


Figure 5.1: Recombination profile of the novel L1C-1-4-4 viruses in relation to PRRSV-2 genomic organization. (A) PRRSV-2 genomic organization. (B) Putative recombinant regions and minor parents of the 2020 – 2021 (n = 18) and the 2018 (n = 1) L1C-1-4-4 variants. The long bar across the top represents the viral genomic backbone. The smaller bars below represent putative minor parents labelled according to the ORF5-based sub-lineages. (C) Recombination breakpoint distribution of the novel L1C-1-4-4 WGSs as queries against other PRRSV-2 WGSs. (D) Overall recombination breakpoint distribution of the 251 PRRSV-2 WGSs. Recombination hotspots defined by the local density plot are highlighted in red. (E) Genomic fragments with low within-fragment recombination rates used for phylogenetic analyses. Nucleotide positions in the alignment are shown in the parenthesis.

Several recombination hotspots were detected when employing an all-to-all approach for identifying recombination events (Figure 5.1D). We used these hotspots to extract four fragments within which there was a low frequency of recombination, namely ORF1a-1 (nucleotide position in the alignment (nt) 609 to 3,455), ORF1a-2 (nt 3,459 to 6,773), ORF1b (nt 6,895 to 10,225), and 3' ORFs (nt 11,916 to 14,657) after their genomic annotation (Figure 5.1E). A significant temporal signal was found in the best-fitting rooted maximum likelihood trees for all fragments, though unlike the other fragments, the temporal signal of ORF1a-2 was only significant when using the correlation and R-squared-based rooting methods (Supplementary Table S5.1).

Table 5.1: Ancestral date and evolutionary rate estimates of the novel L1C-1-4-4 variants and other PRRSV-2

WGS fragment	**Overall (n = 161)**		**2020-2021 novel L1C-1-4-4 (n = 18)**		**2018-2021 L1C-1-4-4 (n = 19)**	
	tMRCA* (95% HPD)	**mean rate** (95% HPD)**	**tMRCA* (95% HPD)**	**ancestral branch rate** (95% HPD)**	**tMRCA* (95% HPD)**	**ancestral branch rate** (95% HPD)**
ORF1a-1	Oct 1988 (Mar 1983, April 1992)	3.81×10^{-3} (2.69×10^{-3}, 4.98×10^{-3})	Nov 2018 (Feb 2018, Sep 2019)	2.15×10^{-2} (2.00×10^{-3}, 6.79×10^{-2})	Nov 2017 (Jan 2016, Jun 2018)	1.22×10^{-2} (7.00×10^{-4}, 4.60×10^{-2})
ORF1a-2#	Aug 1546 # (Jan 1194, Jan 1782)	4.07×10^{-4} (3.29×10^{-4}, 4.86×10^{-4})	May 2003 # (Jun 1993, Mar 2014)	8.96×10^{-4} (1.22×10^{-4}, 2.39×10^{-3})	Jun 1997 # (May 1987, Jan 2009)	2.92×10^{-3} (4.61×10^{-4}, 6.78×10^{-3})
ORF1b	Oct 1985 (Feb 1979, Jun 1991)	2.40×10^{-3} (1.67×10^{-3}, 3.07×10^{-3})	Jan 2019 (Apr 2018, Nov 2019)	8.82×10^{-3} (3.05×10^{-3}, 1.52×10^{-2})	NA (the 2018 taxon does not group with others)	NA (the 2018 taxon does not group with others)
3'ORFs	Jul 1987 (Apr 1981, May 1992)	2.55×10^{-3} (1.91×10^{-3}, 3.23×10^{-3})	Dec 2018 (Feb 2018, Sep 2019)	5.56×10^{-3} (6.64×10^{-4}, 1.17×10^{-2})	May 2017 (July 2014, May 2018)	1.60×10^{-3} (7.94×10^{-4}, 2.51×10^{-3})
ORF5	Nov 1989 (Oct 1984, May 1994)	3.20×10^{-3} (2.34×10^{-3}, 4.09×10^{-3})	Dec 2018 (Mar 2018, Nov 2019)	5.15×10^{-3} (2.65×10^{-4}, 1.27×10^{-2})	Sep 2017 (Aug 2015, Jun 2018)	2.04×10^{-3} (3.42×10^{-4}, 3.99×10^{-3})

*Time to the most recent common ancestor

**Evolutionary rate (substitutions/nucleotide site/year)

#Estimates may be anomalous due to relatively poor temporal signal in this fragment

Evolutionary rate and ancestral date

Amongst time-scaled phylogenies of PRRSV-2 genomic fragments, mean evolutionary rates ranged from 2.40 – 3.81 x 10^{-3} substitutions/site/year, with the exception of ORF1a-2, which had the mean evolutionary rate 10 times lower than that of the rest of the genome (Table 5.1). ORF1a-2's temporal signal, as estimated by Tempest, was more uncertain which may be caused by the low evolutionary rate, ultimately resulting in an eccentric ancestral date estimation that may not be reliable. We thus excluded this fragment from the interpretation of time-scaled trees. The 161 viruses included in this analysis (L1C-1-4-4 variant and the most closely related GenBank sequences across each fragment) had median tMRCAs ranging from 1985 to 1989. The 2020 – 2021 novel L1C-1-4-4 samples (n = 18) form a monophyletic clade sharing a common ancestor in all WGS fragments' trees. Their tMRCA was dated from late 2018 to early 2019. If the 2018 sequence whose ORF5 gene had high nucleotide identity to the 2020 – 2021 L1C-1-4-4 samples was included, tMRCA of the complete set of novel L1C-1-4-4 variants (n = 19) was estimated to be in 2017. The 2018 sequence was a basal taxon to L1C-1-4-4 clade in most fragments except the ORF1b tree, where the 2018 taxon was separated and embedded in a clade consisting of viruses labeled as the L1A sub-lineage, suggesting that the 2018 virus experienced a separate recombination event that did not occur in the 2020 – 2021 sequences. The nucleotide substitution rate at the ancestral branch of the novel L1C (inclusive of the 2018 sequence) was lower in some fragments than the overall mean rate, whereas the ancestral branch of the more recent 2020 – 2021 epidemic samples had a higher rate than the mean (Table 5.1).

Inter- and intra-lineage recombination

Phylogenetic clustering of samples on each WGS-fragment tree, labeled according to ORF5 gene lineages, are visually well-aligned with ORF5-based lineage classification. However, there were instances where clustering of

sequences by ORF5 gene lineage did not translate perfectly to other WGS-fragments, which suggested the possibility of genomic recombination outside ORF5 gene. In the ORF5 gene tree, there was no significant mixing between lineages/sub-lineages except between sub-lineage L1G ancestors and L1B descendants; sub-lineage L1G is thought to have descended from L1B (Paploski et al., 2021), so L1B-L1G mixing in the tree might be due to some misclassification of closely related sequences. This pattern was also apparent in the 3' ORFs fragment (ORF5 gene is embed in this larger fragment), though the clade containing the novel L1C-1-4-4 group became the closest sister to a clade containing the majority of L1A in the 3' ORFs tree. Intermixture of lineage groupings was more apparent in the three ORF1 fragments, suggesting some level of recombination between these genomic regions and ORF5 gene. Although most taxa remained grouped by their ORF5 gene classification, numerous ancestral recombination were observed between lineage 1 sub-lineages. This observation was supported by a high number of transitions between traits (i.e., ORF5 gene lineage label), Bayes factors (Figure 5.2A), and ORF1ab tree topology (Figure 5.2B). The novel L1C-1-4-4 variant's evolutionary history was part of that phenomenon since it was a descendant of the major L1A clade in ORF1a-1 tree. An L1A virus collected in early 2018 (MN073102) was its closest related taxon in all ORF1 trees regardless of whether the L1C-1-4-4 clade was embedded in a larger L1C or L1A clade (Figure 5.2B), suggesting that this virus had a similar evolutionary and recombination history throughout this genomic region.

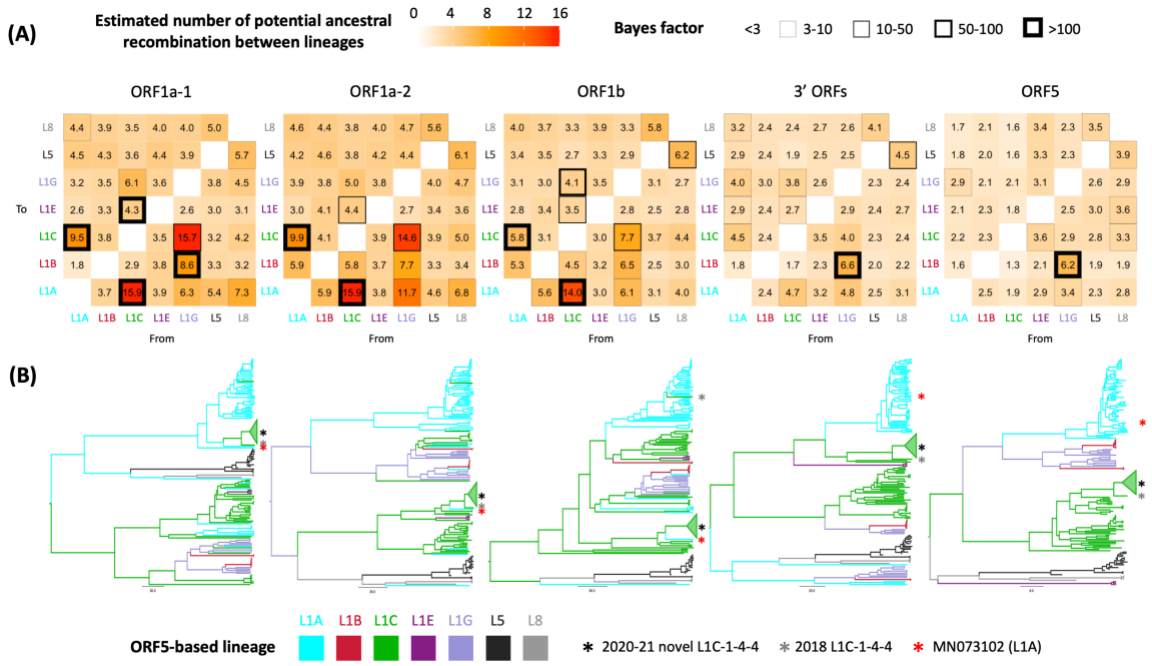


Figure 5.2: Discrete trait analysis of PRRSV-2 lineage/sub-lineage recombination. (A) Heat map showing number of potential ancestral recombination between lineages/sub-lineages of each genomic fragment estimated from the trait transitions. Cell border thickness represents Bayes factor (BF) support for each recombination. (B) Bayesian MCC trees colored by ancestral ORF5-based lineage or sub-lineage. Asterisks locate the phylogenetic position of taxa of interest.

5.4: Discussion

Exploratory analysis of the genome and evolutionary history of viruses causing atypical outbreaks is a key step to understanding their origin. Here, we analyze a set of whole genome sequences (WGSs) from an emerging PRRSV-2 variant and contextualize its evolution using publicly available WGSs from the U.S. swine industry. Our results suggest that the 2020 – 2021 epidemic associated with the novel L1C-1-4-4 viruses arose from a recombinant ancestor of which most genomic parts derived from viruses whose ORF5 genes were classified as sub-lineage L1C. An ancestor of those viruses was estimated to have emerged around late 2018 to early 2019 with a slightly higher mutation rate

than the average rate. Two samples from 2018, classified as the L1C-1-4-4 variant and L1A (MN073102) based on ORF5 gene, are the closest relatives of the 2020 – 2021 epidemic variants, with phylogenetic placement varied according to which genomic was examined. The observed shift in phylogenetic clustering of the L1C-1-4-4 variant from L1C in the ORF5-based tree to L1A in ORF1A-based tree, combined with the inferred frequency of recombination estimated from the discrete trait analysis, highlights the role of recombination within L1 sub-lineages in shaping PRRSV-2 genetic diversity.

Interpretations from our analysis should consider some key limitations. First, samples from the NCBI database may not represent the diversity of the U.S. PRRSV-2 population since whole genome sequencing (WGS) is not a routine practice for disease surveillance because of its cost and availability. In addition, viruses with atypical clinical presentations in the field are more likely to undergo WGS. Thus, our recombination analysis only suggests the most likely parents or close relatives of the novel L1C-1-4-4 from amongst published sequences, which itself may be biased. In fact, the recombination detection was affected by data availability, as evidenced by several unknown parents of the novel L1C recombinant. Second, a fully recombination-free fragment, which is an ideal input for phylogenetic analysis, does not exist in the alignment because breakpoints are distributed across the genome. We alternatively used WGS-fragments with low frequencies of recombination to avoid recombination that may confound the genealogical tree. Genomic positions of such fragments nicely fit with three main protein coding regions of PRRSV-2 and other nidoviruses: ORF1a, ORF1b, and the nested set of multiple ORFs at the 3’-terminal (3’ ORFs) (Saberi et al., 2018). Last, the novel L1C-1-4-4 variant is defined by ORF5 genetic relatedness rather than clinical manifestation, and comparable data quantifying clinical aspects of disease were not available across data sources. Thus, an association between L1C-1-4-4’s virulence and its evolution/recombination cannot be concluded from our study.

The inferred number of putative recombination events (trait transitions) from the discrete trait analysis reflect inter- and intra-lineage recombination between ORF5 gene and other genomic regions (i.e., ORF5 gene lineage was used as the discrete trait). From this analysis, we observe that recombination between lineages was rare, though this may be an artefact of the fact that the majority of included sequences belonged to a single lineage. However, recombination between sub-lineages within lineage 1 are more frequent, though still relatively uncommon. This corresponds to the mechanism of RNA recombination whereby the RNA polymerase is prone to switch from one RNA template to another that has a similar nucleotide sequence (Simon-Loriere & Holmes, 2011).

Additionally, recombination requires co-infection of the same cell, and viral prevalence will influence the likelihood that an animal is co-infected with two distinct viruses simultaneously. The prevalence of sub-lineages is temporally variable (Paploski et al., 2019), which likely shapes opportunities for co-infection. Sub-lineages L1A, 1C, and 1H had the highest effective viral population sizes at the approximate tMRCA of the novel variant (Paploski et al., 2019). Thus, the ancestor of the novel L1C-1-4-4 variant appears to have acquired each genomic portion from divergent viral subpopulations that were prevalent at the time. Recombination scanning along with phylogenetic tree analysis suggests that the majority of the 2020 – 2021 novel L1C-1-4-4 genomic fragments still derived from L1C viruses, while the proximal part of ORF1a gene, mostly nsp2, is genetically closer to viruses whose ORF5 gene is classified as L1A rather than L1C. This evidence coupled with the fact that nsp2 is the most variable gene in PRRSV-2 genome (Yoshii et al., 2008) explains why the novel L1C viruses clustered with viruses classified as L1A at the ORF5 gene level in the WGS tree in previous studies (Kikuti, Paploski, et al., 2021; Schroeder et al., 2021).

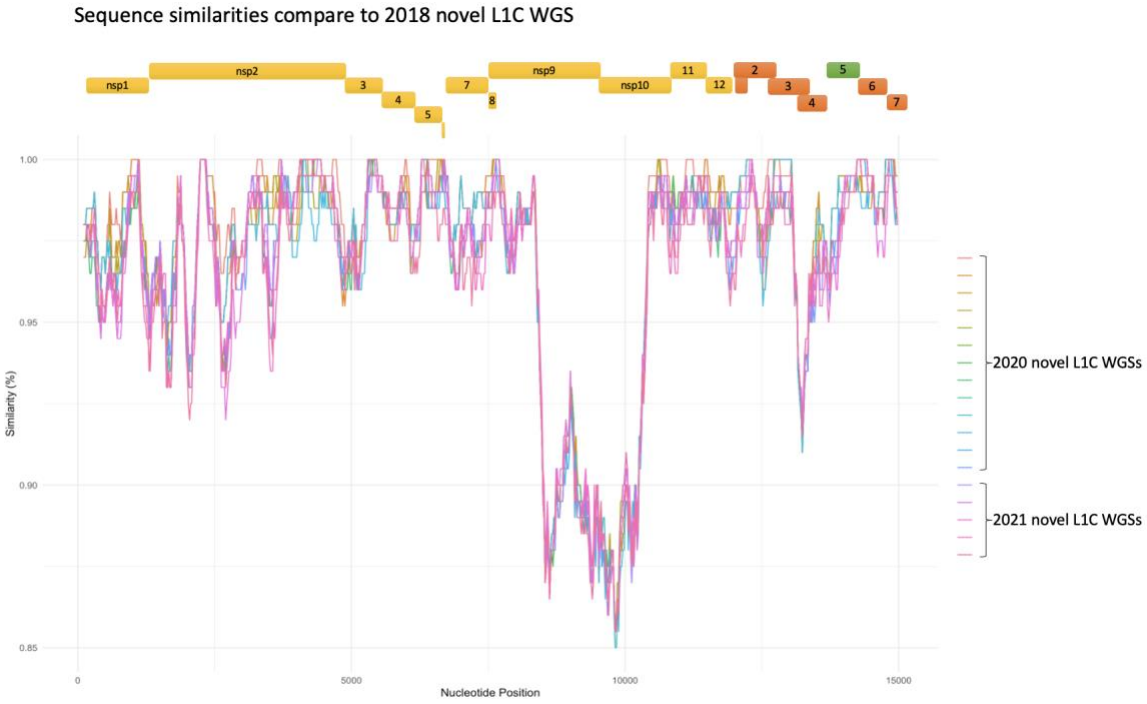
The 2018 L1C-1-4-4 sample was included in this study because it carries an ORF5 gene closely related to the sequences associated with the 2020 – 2021

epidemic and was recovered from the same geographic area. However, some genomic parts as well as real world outbreak circumstances differ from the 2020 – 2021 epidemic. To our knowledge, there was no widespread PRRS outbreaks or heightened concern across the industry in connection with the 2018 virus, though anecdotally, field veterinarians noted that this particular virus transmitted readily between farms belonging to the same company and was challenging to control. The differential recombination profiles between the 2018 and 2020 – 2021 L1C-1-4-4 viruses suggested by the RDP5 analysis were consistent with more robust phylogenetic analyses, which indicate that recombinant parents and phylogenetic position of the 2018 virus's ORF1b are different from the 2020 – 2021 sequences. Altogether, we hypothesize that both diverged in 2017 from the same recombinant ancestor that had a L1A-like ORF1a-1 fragment. The 2018 virus appears to be a result of an additional recombination event that appeared to leave very few progenies in our dataset. Other descendants kept evolving with or without recombination until they reached optimal fitness or a tipping point for exponential growth and became the 2020 – 2021 variant that is associated with the current outbreak.

An assessment of whether the acquisition of different WGS-fragments through recombination had a viral fitness benefit that allowed this variant to spread widely is beyond our limited understanding of the genetic determinants of pathogenicity and antigenicity. Therefore, we do not know the extent to which recombination contributed to the emergence or atypical clinical presentation of this virus. A study on SARS-CoV-2, a distant relative to PRRSV-2 in the same Nidovirales order, suggests the possibility that multi-strain recombination strengthens virulence (Haddad et al., 2021). For PRRSV-2, all four genomic fragments we analyzed harbor at least one virulence-related gene. Mutations in nsp2, a part of both ORF1a-1 and ORF1a-2 fragments, are associated with target cell tropism of PRRSV-2 (Song et al., 2019) and high fever in the host (Du et al., 2021). RNA-dependent RNA polymerase (RdRp), a crucial component

determining virus replication efficiency and pathogenicity (K. Zhao et al., 2018), is encoded by nsp9 in ORF1b gene. Most of the 3'-terminal ORFs are transcribed and translated into the virus structural glycoprotein that directly interacts with either the target cell or host immune response (Das et al., 2011; Van Breedam et al., 2010). Hypothetically, being able to rapidly shift antigenic phenotype through recombination may potentially confer a fitness advantage if it allows the virus to better evade population immunity. Genetic change in one of these genomic parts might be a key success of the novel L1C-1-4-4 variant but would need to be investigated by experimental studies such as targeted mutagenesis. However, our analysis better quantifies the contribution of recombination to PRRSV-2 genetic diversity and evolution, and points to the role of co-circulation of multiple variants within the same farm that may create conditions for recombination and selection for traits beneficial to the virus.

5.5: Supplementary Materials



Supplementary Figure S5.1: Similarity plot between the 2018 (reference) and the 2020-21 (queries) L1C 1-4-4 WGSs

Supplementary Table S5.1: Temporal signal of PRRSV-2 WGS fragment trees

WGS fragment	Best-fitting root method							
	Heuristic residual mean squared		Residual mean squared		Correlation		R-squared	
	r*	R^2	r	R^2	r	R^2	r	R^2
ORF1a-1	0.3817	0.1457	0.3817	0.1457	0.3875	0.1502	0.3875	0.1502
ORF1a-2	-0.0256	6.56×10^{-4}	-0.0256	6.56×10^{-4}	0.2677	7.17×10^{-2}	0.2677	7.17×10^{-2}
ORF1b	0.3438	0.1182	0.3722	0.1385	0.4447	0.1978	0.4447	0.1978
3'ORFs	0.06546	4.29×10^{-3}	0.06546	4.29×10^{-3}	0.3411	0.1164	-0.3905	0.1525

*r = Pearson's correlation coefficient between root-to-tip divergence and time

“Epilogue”

Chapter 6: Conclusion

In order to elucidate the patterns of PRRSV adaptation, persistence, and spread within the United States, we utilized a diverse range of methods coupled with extensive existing data, offering novel and informative insights throughout this dissertation. Key findings of all studies are summarized in Figure 6.1. In Chapter 2, our phylogeographic analysis of continent-wide ORF5 gene sequences revealed that PRRSV-2 L1 likely originated from Canada in the late 1990s, subsequently diverging into multiple sub-lineages. Over nearly three decades, different sets of these sub-lineages made sequential contributions to the overall PRRSV-2 population, with periodic peaks approximately every six years. Notably, we observed a gradual shift in the hotspot for inter-regional spread, transitioning from the Upper Midwest to the Eastern United States. This change in spread patterns within each sub-lineage was likely influenced by the expansion of hog inventories in the U.S. In Chapter 3, we integrated transmission tree inference using regional ORF5 gene sequences with animal movement data and network analysis, to unravel farm-to-farm transmission pathways of PRRSV-2. The results indicated that most infected farms spread the virus to an average of one other farm per year ($R = 1$), with sporadic occurrences of super-spreader events ($R > 1$) within endemic areas. Regarding the network analysis, live animal movement was highly associated with transmission link between farms, while farm proximity related factors, including airborne spread, did not appear to play a major role in shaping transmission networks. However, a significant proportion (over 80%) of transmission events remained unexplained, attributable to the scarcity of available data concerning alternative routes. In Chapter 4, we utilized ORF5 gene data from an active PRRS monitoring database to systematically assess various aspects of PRRSV-2 variant success, serving as a proxy for their emergence, while identifying informative early indicators of such success. Our predictive modeling unveiled that a swift phylogenetic branching pattern,

quantified by the local branching index (LBI), and putative antigenic difference to the current virus population, were significant indicators of future emergence. Additionally, our study demonstrated that variants displaying robust population growth also tended to undergo geographic expansions, often without significant genetic diversification. In Chapter 5, the availability of current PRRSV-2 whole genome sequences allowed us to explore the complete origin of the novel epidemic variant, L1C-1-4-4. Recombination detection analysis coupled with phylodynamic inference consistently suggested that this novel variant was a descendant of a recombinant ancestor characterized by recombination at the ORF1a gene between two segments that would have otherwise been classified independently under the ORF5 gene as L1C and L1A, two of the recent predominant sub-lineages in the U.S. Interestingly, we found that heterologous recombination events which leave detectable numbers of descendants are not common, underscoring the role of genomic recombination as one of the key mechanisms driving the emergence of high fitness variants.

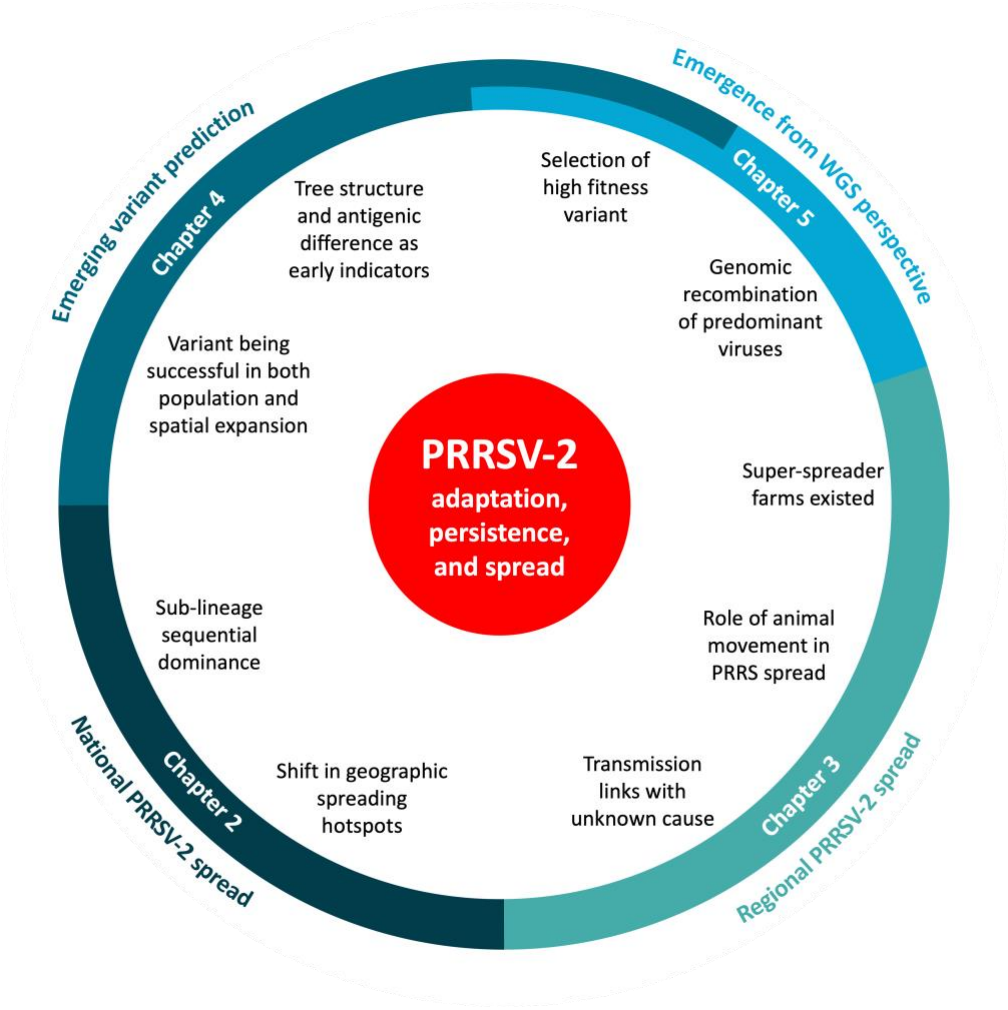


Figure 6.1: Key findings of each chapter in relation to the overarching theme of the dissertation (Center Venn diagram)

The findings gleaned from this dissertation expand our understanding of PRRSV-2, with a particular focus on its evolutionary and epidemiological dynamics. This knowledge forms a foundational base necessary for shaping future PRRS prevention and control strategies in the U.S. On a national scale, the cyclical population dynamics of PRRSV-2, as highlighted in Chapter 2, can guide the timing and focus of prevention and control efforts, while recognizing the shift in inter-regional spread patterns raises situational awareness of the national

disease dynamics. On a regional level, the integrated approaches for tracking farm-to-farm transmission links proposed in Chapter 3 can serve as an alternative tool in PRRS surveillance and outbreak investigation. Additionally, the findings in Chapter 3 augment our understanding of between-farm spread risk factors, which are highly informative for future disease mitigation plans. More applicable, using phylogenetic-based early indicators discerned in Chapter 4 can aid in narrowing down the upcoming variants of concern and better inform targeted intervention approaches. Finally, the potential role of genomic recombination in the emergence of the concerning variant, exclusively detailed in Chapter 5, emphasizes the necessity of continuous whole-genome sequencing for future surveillance. This approach is vital to gain a comprehensive understanding of PRRSV-2 evolution and to effectively monitor potential emerging variants.

All studies within this dissertation explicitly shared similar major limitations, determined by data availability that was contingent on the specific scope of each study. From Chapter 2 to Chapter 4, our analytical findings were primarily based on the genetic relationships of PRRSV-2 ORF5 gene sequences, which are commonly sequenced for routine PRRS diagnosis. While we were able to achieve a decent representation of the virus population both nationwide (Chapters 2 and 4) and regionally (Chapter 3) over an extended time frame, it is important to acknowledge that focusing solely on the ORF5 gene, which constitutes only ~ 4% of the genome, though highly significant, may potentially overlook other crucial factors associated with other genes across the entire genome, as elaborated in Chapter 1. Conversely, while whole genome sequences of the novel variant responsible for the recent outbreak in Chapter 5 could be sourced from multiple swine production systems, the availability of historical PRRSV-2 whole genome sequences in the U.S. in existing databases was constrained. This limitation implies that the existing whole genome

sequences may not provide as extensive a representation of the virus population as the ORF5 gene.

The availability of metadata accompanying PRRSV-2 genetic data significantly shaped our approach and outcomes in each study. In instances where the sample size, data sources, study area, or timeframe expanded, the completeness of available metadata for advanced analysis diminished. To illustrate, in Chapter 2, the precision of geographic location and sampling time data extracted from three distinct databases varied considerably, reflecting differences in data collection, processing, and storage methodologies according to the specific goals of each source. This variability limited our ability to conduct advanced phylodynamic analyses, which typically rely on additional sample attributes to estimate potential risk factors associated with disease spread. Such an issue similarly influenced the interpretation of the results in Chapter 3, where most inferred transmission events remained unexplained due to the limitations of available metadata. This spotlights the critical need to broaden data collection efforts to include details of other farm activities that may feasibly contribute to disease transmission in addition to animal movement.

Beyond the data aspect, access to additional pertinent knowledge and technologies can enhance our research endeavors. In Chapter 4, for instance, the ability to predict the 3D structure (Jumper et al., 2021) or identify immune cell epitopes (Gutiérrez et al., 2015) of PRRSV-2 GP5 could substantiate the significance of putative antigenic differences based on GP5 amino acid sequences in the success of emerging variants. Profiling viral quasispecies – i.e., a group of genetically closely related mutants within a single cell or sample (Domingo et al., 2012; Gregori et al., 2016) – from whole genome sequence samples of an outbreak or an experiment, can enhance our comprehension of the intricate mechanisms behind high fitness virus selection and recombination, as proposed in Chapters 4 and 5. This becomes particularly valuable under

varying circumstances, especially in the presence of uneven immune pressure, a factor we consider as a significant driving force in the evolution of PRRSV-2.

Given the current limiting factors, there exist numerous opportunities for future studies to enhance our foundational work through the utilization of superior data, analytic tools, or approaches. On a national scale, integrating historical records of interstate animal movement and other transportation aspects within the swine industry into phylogeographic analysis using a generalized linear model (Streicker et al., 2010) may assist in uncovering the risk factors steering the inter-regional spread dynamics of PRRSV-2. Refinement of the local disease transmission network could be achieved by providing a more comprehensive representation of PRRSV-2's evolutionary history through whole-genome sequences. Exploring alternative machine learning techniques could also improve the accuracy of predictive modeling for emerging PRRSV-2 variants, given that the characteristics and quantity of novel early indicators align with the requirements of such techniques (Sarker, 2021).

The recurring emergence of new variants and the insufficient comprehension of immunity against PRRSV-2 (Loving et al., 2015; Lunney et al., 2016; Murtaugh & Genzow, 2011) have stood as the two foremost challenges in the quest to eradicate this devastating disease from the U.S. The perspectives on PRRSV-2 evolution and epidemiology presented in this dissertation apparently bridge substantial knowledge gaps associated with the first challenge, potentially laying the groundwork for a better understanding of the second. However, the implementation of targeted PRRS prevention and control strategies, informed by most of our findings, may remain unattainable until effective pig immunizations against PRRSV-2 are developed.

Bibliography

Adams, M. J., Lefkowitz, E. J., King, A. M. Q., Harrach, B., Harrison, R. L., Knowles, N. J., Kropinski, A. M., Krupovic, M., Kuhn, J. H., Mushegian, A. R., Nibert, M., Sabanadzovic, S., Sanfaçon, H., Siddell, S. G., Simmonds, P., Varsani, A., Zerbini, F. M., Gorbalenya, A. E., & Davison, A. J. (2016). Ratification vote on taxonomic proposals to the International Committee on Taxonomy of Viruses (2016). *Archives of Virology*. https://doi.org/10.1007/s00705-016-2977-6

Aiewsakun, P., Pamornchainavakul, N., & Inchaisri, C. (2020). Early origin and global colonisation of foot-and-mouth disease virus. *Scientific Reports*, *10, 15268*. https://doi.org/10.1038/s41598-020-72246-6

Alex Pasternak, J., MacPhee, D. J., & Harding, J. C. S. (2020). Fetal cytokine response to porcine reproductive and respiratory syndrome virus-2 infection. *Cytokine*, *126*. https://doi.org/10.1016/j.cyto.2019.154883

Alkhamis, M. A., Arruda, A. G., Morrison, R. B., & Perez, A. M. (2017). Novel approaches for Spatial and Molecular Surveillance of Porcine Reproductive and Respiratory Syndrome Virus (PRRSv) in the United States. *Scientific Reports*. https://doi.org/10.1038/s41598-017-04628-2

Alkhamis, M. A., Perez, A. M., Murtaugh, M. P., Wang, X., & Morrison, R. B. (2016). Applications of Bayesian phylodynamic methods in a recent U.S. porcine reproductive and respiratory syndrome virus outbreak. *Frontiers in Microbiology*, *7*(FEB). https://doi.org/10.3389/fmicb.2016.00067

Andreyev, V. G., Wesley, R. D., Mengeling, W. L., Vorwald, A. C., & Lager, K. M. (1997). Genetic variation and phylogenetic relationships of 22 porcine reproductive and respiratory syndrome virus (PRRSV) field strains based on sequence analysis of open reading frame 5. *Archives of Virology*, *142*(5). https://doi.org/10.1007/s007050050134

Anping Wang, Qi Chen, L. W., & Darin Madson, Karen Harmon, Phillip Gauger, Jianqiang Zhang, G. L. (2019). Recombination between Vaccine and Field Strains of Porcine Reproductive and Respiratory Syndrome Virus. *Emerging Infectious Diseases*, *25*(12), 2335–2337.

Ansari, I. H., Kwon, B., Osorio, F. A., & Pattnaik, A. K. (2006). Influence of N-Linked Glycosylation of Porcine Reproductive and Respiratory Syndrome Virus GP5 on Virus Infectivity, Antigenicity, and Ability To Induce Neutralizing Antibodies. *Journal of Virology*. https://doi.org/10.1128/jvi.80.8.3994-4004.2006

Arruda, A. G., Alkhamis, M. A., VanderWaal, K., Morrison, R. B., & Perez, A. M. (2017). Estimation of time-dependent reproduction numbers for porcine reproductive and respiratory syndrome across different regions and production systems of the US. *Frontiers in Veterinary Science*. https://doi.org/10.3389/fvets.2017.00046

Arruda, A. G., Friendship, R., Carpenter, J., Greer, A., & Poljak, Z. (2016). Evaluation of control strategies for porcine reproductive and respiratory syndrome (PRRS) in swine breeding herds using a discrete event agent-based model. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0166596

Arruda, A. G., Tousignant, S., Sanhueza, J., Vilalta, C., Poljak, Z., Torremorell, M., Alonso, C., & Corzo, C. A. (2019). Aerosol detection and transmission of porcine reproductive and respiratory syndrome virus (Prrsv): What is the evidence, and what are the knowledge gaps? *Viruses*. https://doi.org/10.3390/v11080712

Arruda, A. G., Vilalta, C., Perez, A., & Morrison, R. (2017). Land altitude, slope, and coverage as risk factors for Porcine Reproductive and Respiratory Syndrome (PRRS) outbreaks in the United States. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0172638

Balaban, M., Moshiri, N., Mai, U., Jia, X., & Mirarab, S. (2019). TreeCluster: Clustering biological sequences using phylogenetic trees. *PLoS ONE*, *14*(8). https://doi.org/10.1371/journal.pone.0221068

Barrat-Charlaix, P., Huddleston, J., Bedford, T., & Neher, R. A. (2021). Limited Predictability of Amino Acid Substitutions in Seasonal Influenza Viruses. *Molecular Biology and Evolution*, *38*(7). https://doi.org/10.1093/molbev/msab065

Basta, S., Carrasco, C. P., Knoetig, S. M., Rigden, R. C., Gerber, H., Summerfield, A., & McCullough, K. C. (2000). Porcine alveolar macrophages: Poor accessory or effective suppressor cells for T-lymphocytes. *Veterinary Immunology and Immunopathology*. https://doi.org/10.1016/S0165-2427(00)00237-3

Benfield, C. T. O., Hill, S., Shatar, M., Shiilegdamba, E., Damdinjav, B., Fine, A., Willett, B., Kock, R., & Bataille, A. (2021). Molecular epidemiology of peste des petits

ruminants virus emergence in critically endangered Mongolian saiga antelope and other wild ungulates. *Virus Evolution*, *7*(2). https://doi.org/10.1093/ve/veab062

Berry, I. M., Eyase, F., Pollett, S., Konongoi, S. L., Joyce, M. G., Figueroa, K., Ofula, V., Koka, H., Koskei, E., Nyunja, A., Mancuso, J. D., Jarman, R. G., & Sang, R. (2019). Global Outbreaks and Origins of a Chikungunya Virus Variant Carrying Mutations Which May Increase Fitness for Aedes aegypti: Revelations from the 2016 Mandera, Kenya outbreak. *American Journal of Tropical Medicine and Hygiene*, *100*(5). https://doi.org/10.4269/ajtmh.18-0980

Bielejec, F., Baele, G., Vrancken, B., Suchard, M. A., Rambaut, A., & Lemey, P. (2016). SpreaD3: Interactive Visualization of Spatiotemporal History and Trait Evolutionary Processes. *Molecular Biology and Evolution*, *33*(8), 2167–2169. https://doi.org/10.1093/molbev/msw082

Blair, B., & Lowe, J. (2019). Describing the cull sow market network in the US: A pilot project. *Preventive Veterinary Medicine*, *162*. https://doi.org/10.1016/j.prevetmed.2018.11.005

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., Matschiner, M., Mendes, F. K., Müller, N. F., Ogilvie, H. A., Du Plessis, L., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., … Drummond, A. J. (2019). BEAST 2.5: An

advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, *15*(4). https://doi.org/10.1371/journal.pcbi.1006650

Brisson, Y. (2014). The changing face of the Canada hog industry. *Statistics Canada*, *96-325–X*(005).

Brito, B., Pauszek, S. J., Hartwig, E. J., Smoliga, G. R., Vu, L. T., Dong, P. V., Stenfeldt, C., Rodriguez, L. L., King, D. P., Knowles, N. J., Bachanek-Bankowska, K., Long, N. T., Dung, D. H., & Arzt, J. (2018). A traditional evolutionary history of foot-and-mouth disease viruses in Southeast Asia challenged by analyses of non-structural protein coding sequences. *Scientific Reports*, *8, 6472*. https://doi.org/10.1038/s41598-018-24870-6

Brockmeier, S. L., Loving, C. L., Vorwald, A. C., Kehrli, M. E., Baker, R. B., Nicholson, T. L., Lager, K. M., Miller, L. C., & Faaberg, K. S. (2012). Genomic sequence and virulence comparison of four Type 2 porcine reproductive and respiratory syndrome virus strains. *Virus Research*, *169*(1). https://doi.org/10.1016/j.virusres.2012.07.030

Cabezas, A. H., Sanderson, M. W., Lockhart, C. Y., Riley, K. A., & Hanthorn, C. J. (2021). Spatial and network analysis of U.S. livestock movements based on Interstate Certificates of Veterinary Inspection. *Preventive Veterinary Medicine*, *193*. https://doi.org/10.1016/j.prevetmed.2021.105391

Calzada-Nova, G., Schnitzlein, W. M., Husmann, R. J., & Zuckermann, F. A. (2011). North American Porcine Reproductive and Respiratory Syndrome Viruses Inhibit Type I

Interferon Production by Plasmacytoid Dendritic Cells. *Journal of Virology*, *85*(6). https://doi.org/10.1128/jvi.01616-10

Caraballo, D. A., Lema, C., Novaro, L., Gury-Dohmen, F., Russo, S., Beltrán, F. J., Palacios, G., & Cisterna, D. M. (2021). A novel terrestrial rabies virus lineage occurring in south america: Origin, diversification, and evidence of contact between wild and domestic cycles. *Viruses*, *13*(12). https://doi.org/10.3390/v13122484

Cariou, M., Picard, L., Guéguen, L., Jacquet, S., Cimarelli, A., Fregoso, O. I., Molaro, A., Navratil, V., & Etienne, L. (2022). Distinct evolutionary trajectories of SARS-CoV-2-interacting proteins in bats and primates identify important host determinants of COVID-19. *Proceedings of the National Academy of Sciences*, *119*(35), e2206610119. https://doi.org/10.1073/pnas.2206610119

Carrington, A. M., Fieguth, P. W., Qazi, H., Holzinger, A., Chen, H. H., Mayr, F., & Manuel, D. G. (2020). A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. *BMC Medical Informatics and Decision Making*, *20*(1). https://doi.org/10.1186/s12911-019-1014-6

Cha, S. H., Chang, C. C., & Yoon, K. J. (2004). Instability of the restriction fragment length polymorphism pattern of open reading frame 5 of porcine reproductive and respiratory syndrome virus during sequential pig-to-pig passages. *Journal of Clinical Microbiology*. https://doi.org/10.1128/JCM.42.10.4462-4467.2004

Chamba, F., Sui, J., Conarchy, B., Zhang, X., Kesl, L., Ma, S., Ruth, D., & Venegoni, A. (2019). Experimental safety and efficacy of a unique MLV PRRSV vaccine:

PRRSGard®. *50th Annual Meeting of the American Association of Swine Veterinarians 2019*, 170–174.

Chan, J. M., Carlsson, G., & Rabadan, R. (2013). Topology of viral evolution. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1313480110

Charerntantanakul, W. (2012). Porcine reproductive and respiratory syndrome virus vaccines: Immunogenicity, efficacy and safety aspects. *World Journal of Virology*, *1*(1). https://doi.org/10.5501/wjv.v1.i1.23

Charpin, C., Mahé, S., Keranflech, A., Belloc, C., Cariolet, R., Le Potier, M. F., & Rose, N. (2012). Infectiousness of pigs infected by the Porcine Reproductive and Respiratory Syndrome virus (PRRSV) is time-dependent. *Veterinary Research*. https://doi.org/10.1186/1297-9716-43-69

Chen, N., Trible, B. R., Kerrigan, M. A., Tian, K., & Rowland, R. R. R. (2016). ORF5 of porcine reproductive and respiratory syndrome virus (PRRSV) is a target of diversifying selection as infection progresses from acute infection to virus rebound. *Infection, Genetics and Evolution*. https://doi.org/10.1016/j.meegid.2016.03.002

Cheng, T.-Y., Campler, M. R., Schroeder, D. C., Yang, M., Mor, S. K., Ferreira, J. B., & Arruda, A. G. (2022). Detection of Multiple Lineages of PRRSV in Breeding and Growing Swine Farms. *Frontiers in Veterinary Science*, *9*. https://doi.org/10.3389/fvets.2022.884733

Corzo, C. A., Mondaca, E., Wayne, S., Torremorell, M., Dee, S., Davies, P., & Morrison, R. B. (2010). Control and elimination of porcine reproductive and respiratory syndrome virus. *Virus Research*. https://doi.org/10.1016/j.virusres.2010.08.016

Costers, S., Lefebvre, D. J., Delputte, P. L., & Nauwynck, H. J. (2008). Porcine reproductive and respiratory syndrome virus modulates apoptosis during replication in alveolar macrophages. *Archives of Virology*, *153*(8). https://doi.org/10.1007/s00705-008-0135-5

Costers, S., Vanhee, M., Van Breedam, W., Van Doorsselaere, J., Geldhof, M., & Nauwynck, H. J. (2010). GP4-specific neutralizing antibodies might be a driving force in PRRSV evolution. *Virus Research*. https://doi.org/10.1016/j.virusres.2010.08.026

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems*.

Das, P. B., Vu, H. L. X., Dinh, P. X., Cooney, J. L., Kwon, B., Osorio, F. A., & Pattnaik, A. K. (2011). Glycosylation of minor envelope glycoproteins of porcine reproductive and respiratory syndrome virus in infectious virus recovery, receptor interaction, and immune response. *Virology*, *410*(2), 385–394. https://doi.org/10.1016/j.virol.2010.12.002

De Maio, N., Wu, C. H., O'Reilly, K. M., & Wilson, D. (2015). New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. *PLoS Genetics*, *11*(8). https://doi.org/10.1371/journal.pgen.1005421

DeBuse, N. (2010). Application of MJ PRRS (TM) vaccine for PRRS control and elimination. *American Association of Swine Veterinarians Annual Meeting 2010*, 247–250. https://www.researchgate.net/publication/281218963_Application_of_MJ_PRRS_TM_vaccine_for_PRRS_control_and_elimination#fullTextFileContent

Dee, S. A., Deen, J., Otake, S., & Pijoan, C. (2004). An experimental model to evaluate the role of transport vehicles as a source of transmission of porcine reproductive and respiratory syndrome virus to susceptible pigs. *Canadian Journal of Veterinary Research = Revue Canadienne de Recherche Vétérinaire*.

Dee, S., Deen, J., Rossow, K., Wiese, C., Otake, S., Joo, H. S., & Pijoan, C. (2002). Mechanical transmission of porcine reproductive and respiratory syndrome virus throughout a coordinated sequence of events during cold weather. *Canadian Journal of Veterinary Research*.

DeFILIPPIS, V. R., & VILLARREAL, L. P. (2000). An Introduction to the Evolutionary Ecology of Viruses. In *Viral Ecology*. https://doi.org/10.1016/b978-012362675-2/50005-7

Delisle, B., Gagnon, C. A., Lambert, M. ève, & D'Allaire, S. (2012). Porcine reproductive and respiratory syndrome virus diversity of Eastern Canada swine herds in a large sequence dataset reveals two hypervariable regions under positive selection. *Infection, Genetics and Evolution*. https://doi.org/10.1016/j.meegid.2012.03.015

Delputte, P. L., & Nauwynck, H. J. (2004). Porcine Arterivirus Infection of Alveolar Macrophages Is Mediated by Sialic Acid on the Virus. *Journal of Virology*. https://doi.org/10.1128/jvi.78.15.8094-8101.2004

Didelot, X., Fraser, C., Gardy, J., Colijn, C., & Malik, H. (2017). Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msw275

Do, H. Q., Trinh, D. T., Nguyen, T. L., Vu, T. T. H., Than, D. D., Van Lo, T., Yeom, M., Song, D., Choe, S. E., An, D. J., & Le, V. P. (2016). Molecular evolution of type 2 porcine reproductive and respiratory syndrome viruses circulating in Vietnam from 2007 to 2015. *BMC Veterinary Research*. https://doi.org/10.1186/s12917-016-0885-3

Domingo, E., Sheldon, J., & Perales, C. (2012). Viral Quasispecies Evolution. *Microbiology and Molecular Biology Reviews*. https://doi.org/10.1128/mmbr.05023-11

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., Mcclean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1). https://doi.org/10.1111/j.1600-0587.2012.07348.x

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., & Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology*, *4*(5), e88. https://doi.org/10.1371/journal.pbio.0040088

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*. https://doi.org/10.1186/1471-2148-7-214

Du, L., Wang, H., Liu, F., Wei, Z., Weng, C., Tang, J., & Feng, W. H. (2021). NSP2 Is Important for Highly Pathogenic Porcine Reproductive and Respiratory Syndrome Virus to Trigger High Fever-Related COX-2-PGE2 Pathway in Pigs. *Frontiers in Immunology*, *12*. https://doi.org/10.3389/fimmu.2021.657071

Duault, H., Durand, B., & Canini, L. (2022). Methods Combining Genomic and Epidemiological Data in the Reconstruction of Transmission Trees: A Systematic Review. *Pathogens*, *11*(2). https://doi.org/10.3390/pathogens11020252

Duffy, S. (2018). Why are RNA virus mutation rates so damn high? *PLoS Biology*, *16*, 1–6. https://doi.org/10.1371/journal.pbio.3000003

Eclercy, J., Renson, P., Lebret, A., Hirchaud, E., Normand, V., Andraud, M., Paboeuf, F., Blanchard, Y., Rose, N., & Bourry, O. (2019). A field recombinant strain derived from two type 1 porcine reproductive and respiratory syndrome virus (PRRSV-1) modified live vaccines shows increased viremia and transmission in spf pigs. *Viruses*. https://doi.org/10.3390/v11030296

Economic Research Service & USDA. (2022). *Hogs: Annual and cumulative year-to-date U.S. trade (head)*. https://www.ers.usda.gov/webdocs/DataFiles/81475/Hog_YearlyFull.xlsx?v=7597.1

Evangelista, P. F., & Beskow, D. (2019). Geospatial point density. *R Journal*. https://doi.org/10.32614/RJ-2018-061

Faaberg, K. S., Hocker, J. D., Erdman, M. M., Harris, D. L. H., Nelson, E. A., Torremorell, M., & Plagemann, P. G. W. (2006). Neutralizing antibody responses of pigs infected with natural GP5 N-glycan mutants of porcine reproductive and respiratory syndrome virus. *Viral Immunology*. https://doi.org/10.1089/vim.2006.19.294

Fang, L., Jiang, Y., Xiao, S., Niu, C., Zhang, H., & Chen, H. (2006). Enhanced immunogenicity of the modified GP5 of porcine reproductive and respiratory syndrome virus. *Virus Genes*, *32*(1), 5–11. https://doi.org/10.1007/s11262-005-5839-y

Fang, Y., & Snijder, E. J. (2010). The PRRSV replicase: Exploring the multifunctionality of an intriguing set of nonstructural proteins. *Virus Research*. https://doi.org/10.1016/j.virusres.2010.07.030

Faria, N. R., Suchard, M. A., Rambaut, A., Streicker, D. G., & Lemey, P. (2013). Simultaneously reconstructing viral crossspecies transmission history and identifying the underlying constraints. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1614). https://doi.org/10.1098/rstb.2012.0196

Felsenstein, J. (2009). PHYLIP - Phylogeny Inference Package, Version 3.69. (Seattle, WA: University of Washington). *The American Naturalist*, *171*(6).

Firestone, S. M., Hayama, Y., Bradhurst, R., Yamamoto, T., Tsutsui, T., & Stevenson, M. A. (2019). Reconstructing foot-and-mouth disease outbreaks: A methods comparison of transmission network models. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-41103-6

Frias-De-Diego, A., Jara, M., Pecoraro, B. M., & Crisci, E. (2021). Whole Genome or Single Genes? A Phylodynamic and Bibliometric Analysis of PRRSV. *Frontiers in Veterinary Science*. https://doi.org/10.3389/fvets.2021.658512

Gagnon, C. A., & Dea, S. (1998). Differentiation between Porcine Reproductive and Respiratory Syndrome Virus Isolates by Restriction Fragment Length Polymorphism of Their ORFs 6 and 7 Genes. *Canadian Journal of Veterinary Research*, *62*(2).

Gagnon, C. A., Lalonde, C., & Provost, C. (2021). Porcine reproductive and respiratory syndrome virus whole-genome sequencing efficacy with field clinical samples using a poly(A)-tail viral genome purification method. *Journal of Veterinary Diagnostic Investigation*, *33*(2), 216–226. https://doi.org/10.1177/1040638720952411

Gail, M. H., Lubin, J. H., & Rubinstein, L. V. (1981). Likelihood calculations for matched case-control studies and survival studies with tied death times. *Biometrika*, *68*(3). https://doi.org/10.1093/biomet/68.3.703

Galvis, J. A., Corzo, C. A., Prada, J. M., & Machado, G. (2021). Modelling the transmission and vaccination strategy for porcine reproductive and respiratory syndrome

virus. *Transboundary and Emerging Diseases*. https://doi.org/10.1111/tbed.14007

Gavel, Y., & Heijne, G. V. (1990). Sequence differences between glycosylated and non-glycosylated asn-x-thr/ser acceptor sites: Implications for protein engineering. *Protein Engineering, Design and Selection*, *3*(5). https://doi.org/10.1093/protein/3.5.433

Geoghegan, J. L., Senior, A. M., Giallonardo, F. D., & Holmes, E. C. (2016). Virological factors that increase the transmissibility of emerging human viruses. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(15). https://doi.org/10.1073/pnas.1521582113

Gibbs, M. J., Armstrong, J. S., & Gibbs, A. J. (2000). Sister-scanning: A Monte Carlo procedure for assessing signals in rebombinant sequences. *Bioinformatics*, *16*(7), 573–582. https://doi.org/10.1093/bioinformatics/16.7.573

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B., & Suchard, M. A. (2013). Improving bayesian population dynamics inference: A coalescent-based model for multiple loci. *Molecular Biology and Evolution*, *30*(3). https://doi.org/10.1093/molbev/mss265

Ginestet, C. (2011). ggplot2: Elegant Graphics for Data Analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *174*(1). https://doi.org/10.1111/j.1467-985x.2010.00676_9.x

Giovani Trevisan & Daniel Linhares. (2022, September 20). *Swine Disease Detection Dashboards*. https://fieldepi.research.cvm.iastate.edu/domestic-swine-disease-monitoring-program/

Goldstein, R. A. (2013). Population Size Dependence of Fitness Effect Distribution and Substitution Rate Probed by Biophysical Model of Protein Thermostability. *Genome Biology and Evolution*, *5*(9), 1584–1593. https://doi.org/10.1093/gbe/evt110

Gorbalenya, A. E., Enjuanes, L., Ziebuhr, J., & Snijder, E. J. (2006). Nidovirales: Evolving the largest RNA virus genome. *Virus Research*. https://doi.org/10.1016/j.virusres.2006.01.017

Gregori, J., Perales, C., Rodriguez-Frias, F., Esteban, J. I., Quer, J., & Domingo, E. (2016). Viral quasispecies complexity measures. *Virology*. https://doi.org/10.1016/j.virol.2016.03.017

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. A., & Holmes, E. C. (2004). Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science*, *303*(5656). https://doi.org/10.1126/science.1090727

Guo, Z., Chen, X. X., Li, R., Qiao, S., & Zhang, G. (2018). The prevalent status and genetic diversity of porcine reproductive and respiratory syndrome virus in China: A molecular epidemiological perspective. *Virology Journal*. https://doi.org/10.1186/s12985-017-0910-6

Gutiérrez, A. H., Martin, W. D., Bailey-Kellogg, C., Terry, F., Moise, L., & De Groot, A. S. (2015). Development and validation of an epitope prediction tool for swine (PigMatrix) based on the pocket profile method. *BMC Bioinformatics*, *16*(1), 290. https://doi.org/10.1186/s12859-015-0724-8

Haddad, D., John, S. E., Mohammad, A., Hammad, M. M., Hebbar, P., Channanath, A., Nizam, R., Al-Qabandi, S., Madhoun, A. A., Alshukry, A., Ali, H., Thanaraj, T. A., & Al-Mulla, F. (2021). SARS-CoV-2: Possible recombination and emergence of potentially more virulent strains. *PLoS ONE*, *16*(5), e0251368. https://doi.org/10.1371/journal.pone.0251368

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., & Neher, R. A. (2018). NextStrain: Real-time tracking of pathogen evolution. *Bioinformatics*, *34*(23), 4121–4123. https://doi.org/10.1093/bioinformatics/bty407

Haley, M. M. (2004). Market Integration in the North American hog industries. In *Electronic Outlook Report from the Economic Research Service* (Issue LDP-M-125-0).

Han, J., Wang, Y., & Faaberg, K. S. (2006). Complete genome analysis of RFLP 184 isolates of porcine reproductive and respiratory syndrome virus. *Virus Research*, *122*(1–2). https://doi.org/10.1016/j.virusres.2006.06.003

Hanada, K., Suzuki, Y., Nakane, T., Hirose, O., & Gojobori, T. (2005). The origin and evolution of porcine reproductive and respiratory syndrome viruses. *Molecular*

Biology and Evolution, *22*(4), 1024–1031. https://doi.org/10.1093/molbev/msi089

Hao, X., Lu, Z., Kuang, W., Sun, P., Fu, Y., Wu, L., Zhao, Q., Bao, H., Fu, Y., Cao, Y., Li, P., Bai, X., Li, D., & Liu, Z. (2011). Polymorphic genetic characterization of the ORF7 gene of porcine reproductive and respiratory syndrome virus (PRRSV) in China. *Virology Journal*. https://doi.org/10.1186/1743-422X-8-73

Harris, D. L. (1992). Multiple site production. *Proceedings of the Southeast Swine Practitioner Conference*, 1–19.

Hatherell, H. A., Didelot, X., Pollock, S. L., Tang, P., Crisan, A., Johnston, J. C., Colijn, C., & Gardy, J. L. (2016). Declaring a tuberculosis outbreak over with genomic epidemiology. *Microbial Genomics*. https://doi.org/10.1099/mgen.0.000060

Hayati, M., Biller, P., & Colijn, C. (2020). Predicting the short-term success of human influenza virus variants with machine learning. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1924). https://doi.org/10.1098/rspb.2020.0319

Hijmans, R. J., Williams, E., & Vennes, C. (2019). Geosphere: Spherical Trigonometry. R package version 1.5-10. *Package Geosphere*.

Holtkamp, D. J., Kliebenstein, J. B., Neumann, E. J., Zimmerman, J. J., Rotto, H. F., Yoder, T. K., Wang, C., Yeske, P. E., Mowrer, C. L., & Haley, C. A. (2013). Assessment of the economic impact of porcine reproductive and respiratory syndrome virus on United States pork producers. *Journal of Swine Health and Production*, *21*, 72–84. https://doi.org/10.31274/ans_air-180814-28

Holtkamp, D. J., Polson, D. D., Torremorell, M., Morrison, B., Classen, D. M., Becton, L., Henry, S., Rodibaugh, M. T., Rowland, R. R., Snelson, H., Straw, B., Yeske, P., & Zimmerman, J. (2011). Terminology for classifying swine herds by porcine reproductive and respiratory syndrome virus status. *Tierarztliche Praxis Ausgabe G: Grosstiere - Nutztiere*. https://doi.org/10.1055/s-0038-1624624

Hu, H., Li, X., Zhang, Z., Shuai, J., Chen, N., Liu, G., & Fang, W. (2009). Porcine reproductive and respiratory syndrome viruses predominant in southeastern China from 2004 to 2007 were from a common source and underwent further divergence. *Archives of Virology*, *154*(3). https://doi.org/10.1007/s00705-009-0316-x

Huddleston, J., Barnes, J. R., Rowe, T., Xu, X., Kondor, R., Wentworth, D. E., Whittaker, L., Ermetal, B., Daniels, R. S., McCauley, J. W., Fujisaki, S. I., Nakamura, K., Kishida, N., Watanabe, S., Hasegawa, H., Barr, I., Subbarao, K., Barrat-Charlaix, P., Neher, R. A., & Bedford, T. (2020). Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza A/H3N2 evolution. *eLife*, *9*. https://doi.org/10.7554/ELIFE.60067

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., & Morris, M. (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v024.i03

Jara, M., Rasmussen, D. A., Corzo, C. A., & Machado, G. (2021). Porcine reproductive and respiratory syndrome virus dissemination across pig production systems in the

United States. *Transboundary and Emerging Diseases*. https://doi.org/10.1111/tbed.13728

Johnson, W., Roof, M., Vaughn, E., Christopher-Hennings, J., Johnson, C. R., & Murtaugh, M. P. (2004). Pathogenic and humoral immune responses to porcine reproductive and respiratory syndrome virus (PRRSV) are related to viral load in acute infection. *Veterinary Immunology and Immunopathology*, *102*(3). https://doi.org/10.1016/j.vetimm.2004.09.010

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kalkauskas, A., Perron, U., Sun, Y., Goldman, N., Baele, G., Guindon, S., & De Maio, N. (2021). Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk. *PLoS Computational Biology*, *17*(1). https://doi.org/10.1371/JOURNAL.PCBI.1008561

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*. https://doi.org/10.1038/nmeth.4285

Kapur, V., Elam, M. R., Pawlovich, T. M., & Murtaugh, M. P. (1996). Genetic variation in porcine reproductive and respiratory syndrome virus isolates in the midwestern United States. *Journal of General Virology*. https://doi.org/10.1099/0022-1317-77-6-1271

Katoh, K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Keffaber, K. K. (1989). Reproduction failure of unknown etiology. *Am. Assoc. Swine Pract. Newsl.*, *1*, 1–9.

Key, N., & McBride, W. D. (2011). The Changing Economics of U.S. Hog Production. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1084881

Kikuti, M., Geary, E., Picasso-Risso, C., Medrano, M., & Corzo, C. (2021). *Updated epidemiological curve of cases associated with the new Lineage 1C RFLP1-4-4 variant*. Morrison Swine Health Monitoring Project; Department of Veterinary Population Medicine, College of Veterinary Medicine, University of Minnesota.

Kikuti, M., Paploski, I. A. D., Pamornchainavakul, N., Picasso-Risso, C., Schwartz, M., Yeske, P., Leuwerke, B., Bruner, L., Murray, D., Roggow, B. D., Thomas, P., Feldmann, L., Allerson, M., Hensch, M., Bauman, T., Sexton, B., Rovira, A., VanderWaal, K., & Corzo, C. A. (2021). Emergence of a New Lineage 1C Variant of Porcine Reproductive and Respiratory Syndrome Virus 2 in the United States. *Frontiers in Veterinary Science*, *8*. https://doi.org/10.3389/fvets.2021.752938

Kikuti, M., Sanhueza, J., Vilalta, C., Paploski, I. A. D., VanderWaal, K., & Corzo, C. A. (2021). Porcine reproductive and respiratory syndrome virus 2 (PRRSV-2) genetic diversity and occurrence of wild type and vaccine-like strains in the United States swine industry. *PLOS ONE*, *16*(11), e0259531. https://doi.org/10.1371/journal.pone.0259531

Kim, W. I., Kim, J. J., Cha, S. H., Wu, W. H., Cooper, V., Evans, R., Choi, E. J., & Yoon, K. J. (2013). Significance of genetic variation of PRRSV ORF5 in virus neutralization and molecular determinants corresponding to cross neutralization among PRRS viruses. *Veterinary Microbiology*, *162*(1). https://doi.org/10.1016/j.vetmic.2012.08.005

Kinsley, A. C., Perez, A. M., Craft, M. E., & Vanderwaal, K. L. (2019). Characterization of swine movements in the United States and implications for disease control. *Preventive Veterinary Medicine*. https://doi.org/10.1016/j.prevetmed.2019.01.001

Kwon, B., Ansari, I. H., Pattnaik, A. K., & Osorio, F. A. (2008). Identification of virulence determinants of porcine reproductive and respiratory syndrome virus through construction of chimeric clones. *Virology*. https://doi.org/10.1016/j.virol.2008.07.030

Lam, H. M., Ratmann, O., & Boni, M. F. (2018). Improved Algorithmic Complexity for the 3SEQ Recombination Detection Algorithm. *Molecular Biology and Evolution*, *35*(1), 247–251. https://doi.org/10.1093/molbev/msx263

Larochelle, R., D'Allaire, S., & Magar, R. (2003). Molecular epidemiology of porcine reproductive and respiratory syndrome virus (PRRSV) in Québec. *Virus Research*. https://doi.org/10.1016/S0168-1702(03)00168-0

Lauring, A. S., & Andino, R. (2010). Quasispecies theory and the behavior of RNA viruses. *PLoS Pathogens*. https://doi.org/10.1371/journal.ppat.1001005

Lee, K., Polson, D., Lowe, E., Main, R., Holtkamp, D., & Martínez-López, B. (2017). Unraveling the contact patterns and network structure of pig shipments in the United States and its association with porcine reproductive and respiratory syndrome virus (PRRSV) outbreaks. *Preventive Veterinary Medicine*. https://doi.org/10.1016/j.prevetmed.2017.02.001

Lemey, P., Rambaut, A., Bedford, T., Faria, N., Bielejec, F., Baele, G., Russell, C. A., Smith, D. J., Pybus, O. G., Brockmann, D., & Suchard, M. A. (2014). Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2. *PLoS Pathogens*, *10*(2). https://doi.org/10.1371/journal.ppat.1003932

Lemey, P., Rambaut, A., Drummond, A. J., & Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1000520

Lemoine, F., Domelevo Entfellner, J. B., Wilkinson, E., Correia, D., Dávila Felipe, M., De Oliveira, T., & Gascuel, O. (2018). Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature*. https://doi.org/10.1038/s41586-018-0043-0

Li, H., & Roossinck, M. J. (2004). Genetic Bottlenecks Reduce Population Variation in an Experimental RNA Virus Population. *Journal of Virology*, *78*(19). https://doi.org/10.1128/jvi.78.19.10582-10587.2004

Li, J., & Murtaugh, M. P. (2012). Dissociation of porcine reproductive and respiratory syndrome virus neutralization from antibodies specific to major envelope protein surface epitopes. *Virology*. https://doi.org/10.1016/j.virol.2012.08.026

Li, J., Wang, S., Li, C., Wang, C., Liu, Y., Wang, G., He, X., Hu, L., Liu, Y., Cui, M., Bi, C., Shao, Z., Wang, X., Xiong, T., Cai, X., Huang, L., & Weng, C. (2017). Secondary Haemophilus parasuis infection enhances highly pathogenic porcine reproductive and respiratory syndrome virus (HP-PRRSV) infection-mediated inflammatory responses. *Veterinary Microbiology*, *204*. https://doi.org/10.1016/j.vetmic.2017.03.035

Li, X., Galliher-Beckley, A., Pappan, L., Trible, B., Kerrigan, M., Beck, A., Hesse, R., Blecha, F., Nietfeld, J. C., Rowland, R. R., & Shi, J. (2014). Comparison of host immune responses to homologous and heterologous type II porcine reproductive and respiratory syndrome virus (PRRSV) challenge in vaccinated and unvaccinated pigs. *BioMed Research International*, *2014*. https://doi.org/10.1155/2014/416727

Li, Y., Zhou, L., Zhang, J., Ge, X., Zhou, R., Zheng, H., Geng, G., Guo, X., & Yang, H. (2014). Nsp9 and Nsp10 Contribute to the Fatal Virulence of Highly Pathogenic Porcine

Reproductive and Respiratory Syndrome Virus Emerging in China. *PLoS Pathogens*. https://doi.org/10.1371/journal.ppat.1004216

Libin, P. J. K., Deforche, K., Abecasis, A. B., & Theys, K. (2019). VIRULIGN: Fast codon-correct alignment and annotation of viral genomes. *Bioinformatics*, *35*(10). https://doi.org/10.1093/bioinformatics/bty851

Linhares, D. C. L., Cano, J. P., Torremorell, M., & Morrison, R. B. (2014). Comparison of time to PRRSv-stability and production losses between two exposure programs to control PRRSv in sow herds. *Preventive Veterinary Medicine*. https://doi.org/10.1016/j.prevetmed.2014.05.010

Liu, J., Xu, Y., Lin, Z., Fan, J., Dai, A., Deng, X., Mao, W., Huang, X., Yang, X., & Wei, C. (2021). Epidemiology investigation of PRRSV discharged by faecal and genetic variation of ORF5. *Transboundary and Emerging Diseases*, *68*(4), 2334–2344. https://doi.org/10.1111/TBED.13894

Liu, P., Song, Y., Colijn, C., & MacPherson, A. (2022). The impact of sampling bias on viral phylogeographic reconstruction. *PLOS Global Public Health*, *2*(9), e0000577-.

Liu Pengyu AND Song, Y. A. N. D. C. C. A. N. D. M. A. (2022). The impact of sampling bias on viral phylogeographic reconstruction. *PLOS Global Public Health*, *2*(9), 1–19. https://doi.org/10.1371/journal.pgph.0000577

Liu, Q. H., Ajelli, M., Aleta, A., Merler, S., Moreno, Y., & Vespignani, A. (2018). Measurability of the epidemic reproduction number in data-driven contact

networks. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1811115115

Loving, C. L., Osorio, F. A., Murtaugh, M. P., & Zuckermann, F. A. (2015). Innate and adaptive immunity against Porcine Reproductive and Respiratory Syndrome Virus. *Veterinary Immunology and Immunopathology*, *167*(1–2). https://doi.org/10.1016/j.vetimm.2015.07.003

Lunney, J. K., Benfield, D. A., & Rowland, R. R. R. (2010). Porcine reproductive and respiratory syndrome virus: An update on an emerging and re-emerging viral disease of swine. *Virus Research*. https://doi.org/10.1016/j.virusres.2010.10.009

Lunney, J. K., Fang, Y., Ladinig, A., Chen, N., Li, Y., Rowland, B., & Renukaradhya, G. J. (2016). Porcine Reproductive and Respiratory Syndrome Virus (PRRSV): Pathogenesis and Interaction with the Immune System. *Annual Review of Animal Biosciences*, *4*(1), 129–154. https://doi.org/10.1146/annurev-animal-022114-111025

MacLean, O. A., Lytras, S., Weaver, S., Singer, J. B., Boni, M. F., Lemey, P., Kosakovsky Pond, S. L., & Robertson, D. L. (2021). Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. *PLoS Biology*, *19*(3). https://doi.org/10.1371/journal.pbio.3001115

Makau, D. N., Alkhamis, M. A., Paploski, I. a. D., Corzo, C. A., Lycett, S., & VanderWaal, K. (2021). Integrating animal movements with phylogeography to model the spread of PRRSV in the USA. *Virus Evolution*. https://doi.org/10.1093/ve/veab060

Makau, D. N., Paploski, I. A. D., Corzo, C. A., & VanderWaal, K. (2022). Dynamic network connectivity influences the spread of a sub-lineage of porcine reproductive and respiratory syndrome virus. *Transboundary and Emerging Diseases*, *69*(2). https://doi.org/10.1111/tbed.14016

Makau, D. N., Paploski, I. A. D., & VanderWaal, K. (2021). Temporal stability of swine movement networks in the U.S. *Preventive Veterinary Medicine*. https://doi.org/10.1016/j.prevetmed.2021.105369

Malgarin, C. M., Moser, F., Pasternak, J. A., Hamonic, G., Detmer, S. E., MacPhee, D. J., & Harding, J. C. S. (2021). Fetal hypoxia and apoptosis following maternal porcine reproductive and respiratory syndrome virus (PRRSV) infection. *BMC Veterinary Research*, *17*(1). https://doi.org/10.1186/s12917-021-02883-0

Mardassi, H., Mounir, S., & Dea, S. (1995). Molecular analysis of the ORFs 3 to 7 of porcine reproductive and respiratory syndrome virus, Québec reference strain. *Archives of Virology*, *140*(8). https://doi.org/10.1007/BF01322667

Marshall, R. D. (1974). The nature and metabolism of the carbohydrate peptide linkages of glycoproteins. *Biochemical Society Symposia*.

Martin, D. P., Murrell, B., Golden, M., Khoosal, A., & Muhire, B. (2015). RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution*, *1*(1). https://doi.org/10.1093/ve/vev003

Martin, D. P., Posada, D., Crandall, K. A., & Williamson, C. (2005). A Modified Bootscan Algorithm for Automated Identification of Recombinant Sequences and

Recombination Breakpoints. *AIDS Research and Human Retroviruses*, *21*(1), 98–102. https://doi.org/10.1089/aid.2005.21.98

Martin, D. P., Varsani, A., Roumagnac, P., Botha, G., Maslamoney, S., Schwab, T., Kelz, Z., Kumar, V., & Murrell, B. (2021). RDP5: A computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evolution*, *7*(1). https://doi.org/10.1093/ve/veaa087

Martin, D., & Rybicki, E. (2000). RDP: Detection of recombination amongst aligned sequences. *Bioinformatics*, *16*(6), 562–563. https://doi.org/10.1093/bioinformatics/16.6.562

Martínez-Lobo, F. J., Díez-Fuertes, F., Simarro, I., Castro, J. M., & Prieto, C. (2011). Porcine Reproductive and Respiratory Syndrome Virus isolates differ in their susceptibility to neutralization. *Vaccine*. https://doi.org/10.1016/j.vaccine.2011.07.076

Martin-Valls, G. E., Kvisgaard, L. K., Tello, M., Darwich, L., Cortey, M., Burgara-Estrella, A. J., Hernandez, J., Larsen, L. E., & Mateu, E. (2014). Analysis of ORF5 and Full-Length Genome Sequences of Porcine Reproductive and Respiratory Syndrome Virus Isolates of Genotypes 1 and 2 Retrieved Worldwide Provides Evidence that Recombination Is a Common Phenomenon and May Produce Mosaic Isolates. *Journal of Virology*. https://doi.org/10.1128/jvi.02858-13

McCrone, J. T., & Lauring, A. S. (2018). Genetic bottlenecks in intraspecies virus transmission. *Current Opinion in Virology*, *28*. https://doi.org/10.1016/j.coviro.2017.10.008

Meier, W. A., Galeota, J., Osorio, F. A., Husmann, R. J., Schnitzlein, W. M., & Zuckermann, F. A. (2003). Gradual development of the interferon-γ response of swine to porcine reproductive and respiratory syndrome virus infection or vaccination. *Virology*. https://doi.org/10.1016/S0042-6822(03)00009-6

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. *Journal of General Virology*, *76*(12). https://doi.org/10.1099/0022-1317-76-12-3181

Meulenberg, J. J. M., Hulst, M. M., De Meijer, E. J., Moonen, P. L. J. M., Den Besten, A., De Kluyver, E. P., Wensvoort, G., & Moormann, R. J. M. (1993). Lelystad virus, the causative agent of porcine epidemic abortion and respiratory syndrome (PEARS), is related to LDV and EAV. *Virology*. https://doi.org/10.1006/viro.1993.1008

Minin, V. N., Bloomquist, E. W., & Suchard, M. A. (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular Biology and Evolution*, *25*(7), 1459–1471. https://doi.org/10.1093/molbev/msn090

Montaner-Tarbes, S., del Portillo, H. A., Montoya, M., & Fraile, L. (2019). Key gaps in the knowledge of the porcine respiratory reproductive syndrome virus (PRRSV). *Frontiers in Veterinary Science*. https://doi.org/10.3389/fvets.2019.00038

Morrison, R. B. (2015, April 24). *PRRS RFLP 1-7-4 Summary*.

MSHMP History | College of Veterinary Medicine—University of Minnesota. (n.d.). Retrieved January 28, 2021, from https://vetmed.umn.edu/centers-programs/swine-program/outreach-leman-mshmp/mshmp-history

Müller, N. F., Rasmussen, D. A., & Stadler, T. (2017). The structured coalescent and its approximations. *Molecular Biology and Evolution*, *34*(11). https://doi.org/10.1093/molbev/msx186

Müller, N. F., Rasmussen, D., & Stadler, T. (2018). MASCOT: Parameter and state inference under the marginal structured coalescent approximation. *Bioinformatics*, *34*(22). https://doi.org/10.1093/bioinformatics/bty406

Murtaugh, M. P. (2012). Use and interpretation of sequencing in PRRSV control programs. *2012 Allen D. Leman Swine Conference: Disease Diagnostics*, *39*, 49–55. https://conservancy.umn.edu/bitstream/handle/11299/139321/Murtaugh.pdf?sequence=1

Murtaugh, M. P., & Genzow, M. (2011). Immunological solutions for treatment and prevention of porcine reproductive and respiratory syndrome (PRRS). *Vaccine*. https://doi.org/10.1016/j.vaccine.2011.09.013

Murtaugh, M. P., Stadejek, T., Abrahante, J. E., Lam, T. T. Y., & Leung, F. C. C. (2010). The ever-expanding diversity of porcine reproductive and respiratory syndrome virus. *Virus Research*. https://doi.org/10.1016/j.virusres.2010.08.015

Nan, Y., Wu, C., Gu, G., Sun, W., Zhang, Y. J., & Zhou, E. M. (2017). Improved vaccine against PRRSV: Current Progress and future perspective. *Frontiers in Microbiology*. https://doi.org/10.3389/fmicb.2017.01635

Nathues, C., Perler, L., Bruhn, S., Suter, D., Eichhorn, L., Hofmann, M., Nathues, H., Baechlein, C., Ritzmann, M., Palzer, A., Grossmann, K., Schüpbach-Regula, G., & Thür, B. (2016). An Outbreak of Porcine Reproductive and Respiratory Syndrome Virus in Switzerland Following Import of Boar Semen. *Transboundary and Emerging Diseases*. https://doi.org/10.1111/tbed.12262

Nathues, H., Alarcon, P., Rushton, J., Jolie, R., Fiebig, K., Jimenez, M., Geurts, V., & Nathues, C. (2017). Cost of porcine reproductive and respiratory syndrome virus at individual farm level – An economic disease model. *Preventive Veterinary Medicine*, *142*. https://doi.org/10.1016/j.prevetmed.2017.04.006

National Agricultural Statistics Service. (2019). *2017 Census of Agriculture* (pp. 419–429).

Neher, R. A., Russell, C. A., & Shraiman, B. I. (2014a). Predicting evolution from the shape of genealogical trees. *eLife*. https://doi.org/10.7554/eLife.03568

Neher, R. A., Russell, C. A., & Shraiman, B. I. (2014b). Predicting evolution from the shape of genealogical trees. *eLife*, *3*. https://doi.org/10.7554/eLife.03568

Neher, R. A., Russell, C. A., & Shraiman, B. I. (2014c). Predicting evolution from the shape of genealogical trees. *eLife*, *3*. https://doi.org/10.7554/eLife.03568

Nelson, M. I., Culhane, M. R., Trovão, N. S., Patnayak, D. P., Halpin, R. A., Lin, X., Shilts, M. H., Das, S. R., & Detmer, S. E. (2017). The emergence and evolution of influenza A (H1α) viruses in swine in Canada and the United States. *Journal of General Virology*, *98*(11). https://doi.org/10.1099/jgv.0.000924

Neumann, E. J., Kliebenstein, J. B., Johnson, C. D., Mabry, J. W., Bush, E. J., Seitzinger, A. H., Green, A. L., & Zimmerman, J. J. (2005). Assessment of the economic impact of porcine reproductive and respiratory syndrome on swine production in the United States. *Journal of the American Veterinary Medical Association*, *227*(3). https://doi.org/10.2460/javma.2005.227.385

Nishiura, H., & Chowell, G. (2009). The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends. In *Mathematical and Statistical Estimation Approaches in Epidemiology*. https://doi.org/10.1007/978-90-481-2313-1_5

Nodelijk, G., De Jong, M. C. M., Van Nes, A., Vernooy, J. C. M., Van Leengoed, L. A. M. G., Pol, J. M. A., & Verheijden, J. H. M. (2000). Introduction, persistence and fade-out of porcine reproductive and respiratory syndrome virus in a Dutch breeding herd: A mathematical analysis. *Epidemiology and Infection*. https://doi.org/10.1017/S0950268899003246

Obermeyer, F., Jankowiak, M., Barkas, N., Schaffner, S. F., Pyle, J. D., Yurkovetskiy, L., Bosso, M., Park, D. J., Babadi, M., MacInnis, B. L., Luban, J., Sabeti, P. C., & Lemieux, J. E. (2022). Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science*, *376*(6599), 1327–1332. https://doi.org/10.1126/science.abm1208

Ostrowski, M., Galeota, J. A., Jar, A. M., Platt, K. B., Osorio, F. A., & Lopez, O. J. (2002). Identification of Neutralizing and Nonneutralizing Epitopes in the Porcine Reproductive and Respiratory Syndrome Virus GP5 Ectodomain. *Journal of Virology*. https://doi.org/10.1128/jvi.76.9.4241-4250.2002

Otake, S., Dee, S., Corzo, C., Oliveira, S., & Deen, J. (2010). Long-distance airborne transport of infectious PRRSV and Mycoplasma hyopneumoniae from a swine population infected with multiple viral variants. *Veterinary Microbiology*. https://doi.org/10.1016/j.vetmic.2010.03.028

Padidam, M., Sawyer, S., & Fauquet, C. M. (1999). Possible emergence of new geminiviruses by frequent recombination. *Virology*, *265*(2), 218–225. https://doi.org/10.1006/viro.1999.0056

Pamornchainavakul, N., Kikuti, M., Paploski, I. A. D., Makau, D. N., Rovira, A., Corzo, C. A., & VanderWaal, K. (2022). Measuring How Recombination Re-shapes the Evolutionary History of PRRSV-2: A Genome-Based Phylodynamic Analysis of the Emergence of a Novel PRRSV-2 Variant. *Frontiers in Veterinary Science*, *9*. https://doi.org/10.3389/fvets.2022.846904

Pamornchainavakul, N., Makau, D. N., Paploski, I. A. D., Corzo, C. A., & VanderWaal, K. (2023). Unveiling invisible farm-to-farm PRRSV -2 transmission links and routes through transmission tree and network analysis. *Evolutionary Applications*, eva.13596. https://doi.org/10.1111/eva.13596

Pamornchainavakul, N., Paploski, I. A. D., Makau, D. N., Kikuti, M., Rovira, A., Lycett, S., Corzo, C. A., & VanderWaal, K. (2023). Mapping the Dynamics of Contemporary PRRSV-2 Evolution and Its Emergence and Spreading Hotspots in the U.S. Using Phylogeography. *Pathogens*, *12*(5). https://doi.org/10.3390/pathogens12050740

Paploski, I. A. D., Corzo, C., Rovira, A., Murtaugh, M. P., Sanhueza, J. M., Vilalta, C., Schroeder, D. C., & VanderWaal, K. (2019). Temporal Dynamics of Co-circulating Lineages of Porcine Reproductive and Respiratory Syndrome Virus. *Frontiers in Microbiology*, *10,2486*. https://doi.org/10.3389/fmicb.2019.02486

Paploski, I. A. D., Makau, D. N., Pamornchainavakul, N., Baker, J. P., Schroeder, D., Rovira, A., & VanderWaal, K. (2022). Potential Novel N-Glycosylation Patterns Associated with the Emergence of New Genetic Variants of PRRSV-2 in the U.S. *Vaccines*, *10*(12). https://doi.org/10.3390/vaccines10122021

Paploski, I. A. D., Pamornchainavakul, N., Makau, D. N., Rovira, A., Corzo, C. A., Schroeder, D. C., Cheeran, M. C. J., Doeschl-Wilson, A., Kao, R. R., Lycett, S., & Vanderwaal, K. (2021). Phylogenetic structure and sequential dominance of sub-lineages of prrsv type-2 lineage 1 in the United States. *Vaccines*, *9(6),608*. https://doi.org/10.3390/vaccines9060608

Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Passafaro, T. L., Fernandes, A. F. A., Valente, B. D., Williams, N. H., & Rosa, G. J. M. (2020). Network analysis of swine movements in a multi-site pig production system in Iowa, USA. *Preventive Veterinary Medicine*. https://doi.org/10.1016/j.prevetmed.2019.104856

Pasternak, A. O., Spaan, W. J. M., & Snijder, E. J. (2006). Nidovirus transcription: How to make sense...? *Journal of General Virology*. https://doi.org/10.1099/vir.0.81611-0

Paules, C. I., McDermott, A. B., & Fauci, A. S. (2019). Immunity to Influenza: Catching a Moving Target To Improve Vaccine Design. *The Journal of Immunology*, *202*(2). https://doi.org/10.4049/jimmunol.1890025

Pearl, J. (1982). *REVEREND BAYES ON INFERENCE ENGINES: A DISTRIBUTED HIERARCHICAL APPROACH.* https://doi.org/10.1145/3501714.3501727

Pileri, E., & Mateu, E. (2016). Review on the transmission porcine reproductive and respiratory syndrome virus between pigs and farms and impact on vaccination. *Veterinary Research*. https://doi.org/10.1186/s13567-016-0391-4

Pirzadeh, B., & Dea, S. (1997). Monoclonal antibodies to the ORF5 product of porcine reproductive and respiratory syndrome virus define linear neutralizing

determinants. *Journal of General Virology*, *78*(8), 1867–1873. https://doi.org/10.1099/0022-1317-78-8-1867

Pirzadeh, B., Gagnon, C. A., & Dea, S. (1998). Genomic and Antigenic Variations of Porcine Reproductive and Respiratory Syndrome Virus Major Envelope GP5 Glycoprotein. *Canadian Journal of Veterinary Research*, *62*(3).

Poon, A. F. Y., Walker, L. W., Murray, H., McCloskey, R. M., Harrigan, P. R., & Liang, R. H. (2013). Mapping the shapes of phylogenetic trees from human and zoonotic RNA viruses. *PLoS ONE*, *8*(11). https://doi.org/10.1371/journal.pone.0078122

Popescu, L. N., Trible, B. R., Chen, N., & Rowland, R. R. R. (2017). GP5 of porcine reproductive and respiratory syndrome virus (PRRSV) as a target for homologous and broadly neutralizing antibodies. *Veterinary Microbiology*. https://doi.org/10.1016/j.vetmic.2017.04.016

Porcine reproductive and respiratory syndrome: OIE - World Organisation for Animal Health. (n.d.). Retrieved February 24, 2021, from https://www.oie.int/en/animal-health-in-the-world/animal-diseases/Porcine-reproductive-and-respiratory-syndrome/

Posada, D., & Crandall, K. A. (2001). Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(24), 13757–13762. https://doi.org/10.1073/pnas.241370698

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2—Approximately maximum-likelihood trees for large alignments. *PLoS ONE*, *5*(3). https://doi.org/10.1371/journal.pone.0009490

Proctor, J., Wolf, I., Brodsky, D., Cortes, L. M., Frias-De-Diego, A., Almond, G. W., Crisci, E., Negrão Watanabe, T. T., Hammer, J. M., & Käser, T. (2022). Heterologous vaccine immunogenicity, efficacy, and immune correlates of protection of a modified-live virus porcine reproductive and respiratory syndrome virus vaccine. *Frontiers in Microbiology*, *13*. https://doi.org/10.3389/fmicb.2022.977796

R Core Team. (2019). R: A language and environment for statistical computing. In *R Foundation for Statistical Computing* [Computer software]. https://www.r-project.org/

Ragonnet-Cronin, M., Hodcroft, E., Hué, S., Fearnhill, E., Delpech, V., Brown, A. J. L., & Lycett, S. (2013). Automated analysis of phylogenetic clusters. *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-14-317

Rahe, M. C., & Murtaugh, M. P. (2017). Mechanisms of adaptive immunity to porcine reproductive and respiratory syndrome virus. *Viruses*, *9*(6). https://doi.org/10.3390/v9060148

Rambaut, A. (2018). *FigTree v. 1.4.4* [Computer software]. http://tree.bio.ed.ac.uk/software/figtree/

Rambaut, A., Lam, T. T., Max Carvalho, L., & Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution*, *2*(1), vew007. https://doi.org/10.1093/ve/vew007

Ramirez, A., Whitney, D., & Bickett–Weddle, D. (2011). Swine Industry Manual. FAD PReP: Foreign Animal Disease Preparedness & Response Plan / National Animal Health Emergency Management System. *United States Department of Agriculture, Animal and Plant Health Inspection Service*.

Reid, S., & Tibshirani, R. (2014). Regularization paths for conditional logistic regression: The clogitL1 package. *Journal of Statistical Software*, *58*(12). https://doi.org/10.18637/jss.v058.i12

Roca, M., Gimeno, M., Bruguera, S., Segalés, J., Díaz, I., Galindo-Cardiel, I. J., Martínez, E., Darwich, L., Fang, Y., Maldonado, J., March, R., & Mateu, E. (2012). Effects of challenge with a virulent genotype II strain of porcine reproductive and respiratory syndrome virus on piglets vaccinated with an attenuated genotype I strain vaccine. *Veterinary Journal*, *193*(1). https://doi.org/10.1016/j.tvjl.2011.11.019

Rodriguez, M. J., Sarraseca, J., Garcia, J., Sanz, A., Plana-Durán, J., & Casal, J. I. (1997). Epitope mapping of the nucleocapsid protein of European and North American isolates of porcine reproductive and respiratory syndrome virus. *Journal of General Virology*, *78*(9). https://doi.org/10.1099/0022-1317-78-9-2269

Rose, N., Renson, P., Andraud, M., Paboeuf, F., Le Potier, M. F., & Bourry, O. (2015). Porcine reproductive and respiratory syndrome virus (PRRSv) modified-live vaccine reduces virus transmission in experimental conditions. *Vaccine*. https://doi.org/10.1016/j.vaccine.2015.03.040

Rosendal, T., Dewey, C., Friendship, R., Wootton, S., Young, B., & Poljak, Z. (2014). Spatial and temporal patterns of porcine reproductive and respiratory syndrome virus (PRRSV) genotypes in Ontario, Canada, 2004-2007. *BMC Veterinary Research*. https://doi.org/10.1186/1746-6148-10-83

Ruedas-Torres, I., Rodríguez-Gómez, I. M., Sánchez-Carvajal, J. M., Larenas-Muñoz, F., Pallarés, F. J., Carrasco, L., & Gómez-Laguna, J. (2021). The jigsaw of PRRSV virulence. *Veterinary Microbiology*, *260,109168*. https://doi.org/10.1016/j.vetmic.2021.109168

Saberi, A., Gulyaeva, A. A., Brubacher, J. L., Newmark, P. A., & Gorbalenya, A. E. (2018). A planarian nidovirus expands the limits of RNA genome size. *PLoS Pathogens*, *14*(11), e1007314. https://doi.org/10.1371/journal.ppat.1007314

Sagulenko, P., Puller, V., & Neher, R. A. (2018). TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution*. https://doi.org/10.1093/ve/vex042

Sanhueza, J. M., Vilalta, C., Corzo, C., & Arruda, A. G. (2019). Factors affecting Porcine Reproductive and Respiratory Syndrome virus time-to-stability in breeding herds in the Midwestern United States. *Transboundary and Emerging Diseases*. https://doi.org/10.1111/tbed.13091

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., & Crowley, J. (2022). *GGally: Extension to "ggplot2."* https://ggobi.github.io/ggally/, https://github.com/ggobi/ggally

Schneider, W. L., & Roossinck, M. J. (2001). Genetic Diversity in RNA Virus Quasispecies Is Controlled by Host-Virus Interactions. *Journal of Virology*. https://doi.org/10.1128/jvi.75.14.6566-6571.2001

Schroeder, D. C., Odogwu, N. M., Kevill, J., Yang, M., Krishna, V. D., Kikuti, M., Pamornchainavakul, N., Vilalta, C., Sanhueza, J., Corzo, C. A., Rovira, A., Dee, S., Nelson, E., Singrey, A., Zhitnitskiy, P., Balestreri, C., Makau, D. N., Paploski, I. A. D., Cheeran, M. C.-J., … Torremorell, M. (2021). Phylogenetically Distinct Near-Complete Genome Sequences of Porcine Reproductive and Respiratory Syndrome Virus Type 2 Variants from Four Distinct Disease Outbreaks at U.S. Swine Farms over the Past 6 Years. *Microbiology Resource Announcements*, *10,33*. https://doi.org/10.1128/mra.00260-21

Scotch, M., & Mei, C. (2013). Phylogeography of swine influenza H3N2 in the United States: Translational public health for zoonotic disease surveillance. *Infection, Genetics and Evolution*, *13*(1). https://doi.org/10.1016/j.meegid.2012.09.015

Sellman, S., Beck-Johnson, L. M., Hallman, C., Miller, R. S., Owers Bonner, K. A., Portacci, K., Webb, C. T., & Lindström, T. (2022). Modeling nation-wide U.S. swine movement networks at the resolution of the individual premises. *Epidemics*, *41*, 100636. https://doi.org/10.1016/j.epidem.2022.100636

Shi, M., Holmes, E. C., Brar, M. S., & Leung, F. C.-C. (2013). Recombination Is Associated with an Outbreak of Novel Highly Pathogenic Porcine Reproductive and Respiratory Syndrome Viruses in China. *Journal of Virology*. https://doi.org/10.1128/jvi.01270-13

Shi, M., Lam, T. T. Y., Hon, C. C., Hui, R. K. H., Faaberg, K. S., Wennblom, T., Murtaugh, M. P., Stadejek, T., & Leung, F. C. C. (2010). Molecular epidemiology of PRRSV: A phylogenetic perspective. *Virus Research*, *154*(1–2). https://doi.org/10.1016/j.virusres.2010.08.014

Shi, M., Lam, T. T.-Y., Hon, C.-C., Murtaugh, M. P., Davies, P. R., Hui, R. K.-H., Li, J., Wong, L. T.-W., Yip, C.-W., Jiang, J.-W., & Leung, F. C.-C. (2010). Phylogeny-Based Evolutionary, Demographical, and Geographical Dissection of North American Type 2 Porcine Reproductive and Respiratory Syndrome Viruses. *Journal of Virology*, *84*(17), 8700–8711. https://doi.org/10.1128/jvi.02551-09

Shi, M., Lemey, P., Singh Brar, M., Suchard, M. A., Murtaugh, M. P., Carman, S., D'Allaire, S., Delisle, B., Lambert, M. È., Gagnon, C. A., Ge, L., Qu, Y., Yoo, D., Holmes, E. C., & Chi-Ching Leung, F. (2013). The spread of type 2 porcine reproductive and

respiratory syndrome virus (prrsv) in North America: A phylogeographic approach. *Virology*. https://doi.org/10.1016/j.virol.2013.08.028

Shields, D. A., & Mathews, K. H. (2003). Interstate livestock movements. *Non*.

Simon-Loriere, E., & Holmes, E. C. (2011). Why do RNA viruses recombine? *Nature Reviews Microbiology*, *9*, 617–626. https://doi.org/10.1038/nrmicro2614

Smith, J. M. (1992). Analyzing the mosaic structure of genes. *Journal of Molecular Evolution*, *34*(2), 126–129. https://doi.org/10.1007/BF00182389

Smith, M. D., Wertheim, J. O., Weaver, S., Murrell, B., Scheffler, K., & Kosakovsky Pond, S. L. (2015). Less is more: An adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Molecular Biology and Evolution*, *32*(5). https://doi.org/10.1093/molbev/msv022

Snijder, E. J., Kikkert, M., & Fang, Y. (2013). Arterivirus molecular biology and pathogenesis. *Journal of General Virology*. https://doi.org/10.1099/vir.0.056341-0

Song, J., Gao, P., Kong, C., Zhou, L., Ge, X., Guo, X., Han, J., & Yang, H. (2019). The nsp2 Hypervariable Region of Porcine Reproductive and Respiratory Syndrome Virus Strain JXwn06 Is Associated with Viral Cellular Tropism to Primary Porcine Alveolar Macrophages. *Journal of Virology*, *93*(24), e01436-19. https://doi.org/10.1128/jvi.01436-19

Spielman, S. J., Weaver, S., Shank, S. D., Magalis, B. R., Li, M., & Kosakovsky Pond, S. L. (2019). Evolution of viral genomes: Interplay between selection, recombination,

and other forces. In *Methods in Molecular Biology* (Vol. 1910). https://doi.org/10.1007/978-1-4939-9074-0_14

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btu033

Storgaard, T., Oleksiewicz, M., & Bøtner, A. (1999). Examination of the selective pressures on a live PRRS vaccine virus. *Archives of Virology*. https://doi.org/10.1007/s007050050652

Streicker, D. G., Turmelle, A. S., Vonhof, M. J., Kuzmin, I. V., McCracken, G. F., & Rupprecht, C. E. (2010). Host Phylogeny Constrains Cross-Species Emergence and Establishment of Rabies Virus in Bats. *Science*, *329*(5992), 676–679. https://doi.org/10.1126/science.1188836

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., & Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, *4*(1), vey016. https://doi.org/10.1093/ve/vey016

Swinton, J. (2023). *Vennerable: Venn and Euler area-proportional diagrams*. https://github.com/js229/Vennerable

Talevich, E., Invergo, B. M., Cock, P. J. A., & Chapman, B. A. (2012). Bio.Phylo: A unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics*, *13*(1). https://doi.org/10.1186/1471-2105-13-209

Thakur, K. K., Revie, C. W., Hurnik, D., Poljak, Z., & Sanchez, J. (2015). Simulation of between-farm transmission of porcine reproductive and respiratory syndrome virus in Ontario, Canada using the North American Animal Disease Spread Model. *Preventive Veterinary Medicine*. https://doi.org/10.1016/j.prevetmed.2015.01.006

Thanawongnuwech, R., Brown, G. B., Halbur, P. G., Roth, J. A., Royer, R. L., & Thacker, B. J. (2000). Pathogenesis of Porcine Reproductive and Respiratory Syndrome Virus-induced Increase in Susceptibility to Streptococcus suis Infection. *Veterinary Pathology*. https://doi.org/10.1354/vp.37-2-143

Thanawongnuwech, R., Thacker, B., Halbur, P., & Thacker, E. L. (2004). Increased production of proinflammatory cytokines following infection with porcine reproductive and respiratory syndrome virus and Mycoplasma hyopneumoniae. *Clinical and Diagnostic Laboratory Immunology*, *11*(5). https://doi.org/10.1128/CDLI.11.5.901-908.2004

The Morrison Swine Health Monitoring Project. (2022, September 21). *PRRS CUMULATIVE INCIDENCE*. https://vetmed.umn.edu/centers-programs/swine-program/outreach-leman-mshmp/mshmp/mshmp-prrs-figures

The OIE AD HOC group on porcine reproductive respiratory syndrome. (2008). *PRRS : the disease , its diagnosis , prevention and control*. Group.

The Swine Health Information Center. (2022, September 22). *The Swine Health Information Center's Rapid Response Program*. https://www.swinehealth.org/rapid-response-to-emerging-disease-program/

Therneau, T. M. (2021). survival: A Package for Survival Analysis in R. *R Package Version 2.38*.

Tian, K., Yu, X., Zhao, T., Feng, Y., Cao, Z., Wang, C., Hu, Y., Chen, X., Hu, D., Tian, X., Liu, D., Zhang, S., Deng, X., Ding, Y., Yang, L., Zhang, Y., Xiao, H., Qiao, M., Wang, B., … Gao, G. F. (2007). Emergence of Fatal PRRSV Variants: Unparalleled Outbreaks of Atypical PRRS in China and Molecular Dissection of the Unique Hallmark. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0000526

Tokach, M. D., Goodband, B. D., & O'Quinn, T. G. (2016). Performance-enhancing technologies in swine production. *Animal Frontiers*, *6*(4). https://doi.org/10.2527/af.2016-0039

Trevisan, G., Linhares, L. C. M., Crim, B., Dubey, P., Schwartz, K. J., Burrough, E. R., Wang, C., Main, R. G., Sundberg, P., Thurn, M., Lages, P. T. F., Corzo, C. A., Torrison, J., Henningson, J., Herrman, E., Hanzlicek, G. A., Raghavan, R., Marthaler, D., Greseth, J., … Linhares, D. C. L. (2020). Prediction of seasonal patterns of porcine reproductive and respiratory syndrome virus RNA detection in the U.S. swine industry. *Journal of Veterinary Diagnostic Investigation*, *32*(3). https://doi.org/10.1177/1040638720912406

Trifinopoulos, J., Nguyen, L. T., von Haeseler, A., & Minh, B. Q. (2016). W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research*, *44*(W1), W232–W235. https://doi.org/10.1093/nar/gkw256

Valdes-Donoso, P., Alvarez, J., Jarvis, L. S., Morrison, R. B., & Perez, A. M. (2018). Production losses from an endemic animal disease: Porcine reproductive and respiratory syndrome (PRRS) in selected Midwest US Sow Farms. *Frontiers in Veterinary Science*, *5*(MAY). https://doi.org/10.3389/fvets.2018.00102

Valdes-Donoso, P., & Jarvis, L. S. (2022). Combining epidemiology and economics to assess control of a viral endemic animal disease: Porcine Reproductive and Respiratory Syndrome (PRRS). *PLOS ONE*, *17*(9), e0274382-.

Van Breedam, W., Van Gorp, H., Zhang, J. Q., Crocker, P. R., Delputte, P. L., & Nauwynck, H. J. (2010). The M/GP5 glycoprotein complex of porcine reproductive and respiratory syndrome virus binds the sialoadhesin receptor in a sialic acid-dependent manner. *PLoS Pathogens*, *6*(1), e1000730. https://doi.org/10.1371/journal.ppat.1000730

van Geelen, A. G. M., Anderson, T. K., Lager, K. M., Das, P. B., Otis, N. J., Montiel, N. A., Miller, L. C., Kulshreshtha, V., Buckley, A. C., Brockmeier, S. L., Zhang, J., Gauger, P. C., Harmon, K. M., & Faaberg, K. S. (2018). Porcine reproductive and respiratory disease virus: Evolution and recombination yields distinct ORF5 RFLP 1-7-4 viruses with individual pathogenicity. *Virology*, *513*, 168–179. https://doi.org/10.1016/j.virol.2017.10.002

Van Rossum, G., Drake, F. L., Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., … Oliphant, T. E. (2009). Python 3 Reference Manual. In *Nature* (Vol. 585, Issue 7825).

VanderWaal, K., & Deen, J. (2018). Global trends in infectious diseases of swine. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1806068115

VanderWaal, K., Paploski, I. A. D., Makau, D. N., & Corzo, C. A. (2020). Contrasting animal movement and spatial connectivity networks in shaping transmission pathways of a genetically diverse virus. *Preventive Veterinary Medicine*, *178*. https://doi.org/10.1016/j.prevetmed.2020.104977

Vignuzzi, M., & Andino, R. (2012). Closing the gap: The challenges in converging theoretical, computational, experimental and real-life studies in virus evolution. *Current Opinion in Virology*. https://doi.org/10.1016/j.coviro.2012.09.008

Volz, E. M., Koelle, K., & Bedford, T. (2013). Viral Phylodynamics. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1002947

Vu, H. L. X., Kwon, B., Yoon, K.-J., Laegreid, W. W., Pattnaik, A. K., & Osorio, F. A. (2011). Immune Evasion of Porcine Reproductive and Respiratory Syndrome Virus through Glycan Shielding Involves both Glycoprotein 5 as Well as Glycoprotein 3. *Journal of Virology*. https://doi.org/10.1128/jvi.00189-11

Walker, P. J., Siddell, S. G., Lefkowitz, E. J., Mushegian, A. R., Adriaenssens, E. M., Alfenas-Zerbini, P., Davison, A. J., Dempsey, D. M., Dutilh, B. E., García, M. L., Harrach, B., Harrison, R. L., Hendrickson, R. C., Junglen, S., Knowles, N. J., Krupovic, M., Kuhn, J. H., Lambert, A. J., Łobocka, M., … Zerbini, F. M. (2021). Changes to virus taxonomy and to the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2021). *Archives of Virology*, *166*, 2633–2648. https://doi.org/10.1007/s00705-021-05156-1

Wasik, B. R., & Turner, P. E. (2013). On the biological success of viruses. *Annual Review of Microbiology*, *67*. https://doi.org/10.1146/annurev-micro-090110-102833

Wei, C. J., Crank, M. C., Shiver, J., Graham, B. S., Mascola, J. R., & Nabel, G. J. (2020). Next-generation influenza vaccines: Opportunities and challenges. *Nature Reviews Drug Discovery*, *19*(4). https://doi.org/10.1038/s41573-019-0056-x

Wesley, R. D., Mengeling, W. L., Lager, K. M., Clouser, D. F., Landgraf, J. G., & Frey, M. L. (1998). 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. *Journal of Veterinary Diagnostic Investigation*, *10*(2), 140–144. https://doi.org/10.1177/104063879801000204

Whiting, T. L. (2008). Special welfare concerns in countries dependent on live animal trade: The real foreign animal disease emergency for Canada. *Journal of Applied Animal Welfare Science*, *11*(2). https://doi.org/10.1080/10888700801926008

Wissink, E. H. J., Kroese, M. V., van Wijk, H. A. R., Rijsewijk, F. A. M., Meulenberg, J. J. M., & Rottier, P. J. M. (2005). Envelope Protein Requirements for the Assembly of Infectious Virions of Porcine Reproductive and Respiratory Syndrome Virus. *Journal of Virology*. https://doi.org/10.1128/jvi.79.19.12495-12506.2005

Wissink, E. H. J., van Wijk, H. A. R., Kroese, M. V., Weiland, E., Meulenberg, J. J. M., Rottier, P. J. M., & van Rijn, P. A. (2003). The major envelope protein, GP5, of a European porcine reproductive and respiratory syndrome virus contains a neutralization epitope in its N-terminal ectodomain. *Journal of General Virology*. https://doi.org/10.1099/vir.0.18957-0

Yang, L., Frey, M. L., Yoon, K.-J., Zimmerman, J. J., & Platt, K. B. (2000). Categorization of North American porcine reproductive and respiratory syndrome viruses: Epitopic profiles of the N, M, GP5 and GP3 proteins and susceptibility to neutralization. *Archives of Virology*, *145*(8), 1599–1619. https://doi.org/10.1007/s007050070079

Yang, Z., Nielsen, R., Goldman, N., & Pedersen, A. M. K. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*. https://doi.org/10.1093/genetics/155.1.431

Yang, Z., & Nielsent, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution*, *19*(6). https://doi.org/10.1093/oxfordjournals.molbev.a004148

Ye, Y., Si, Z. H., Moore, J. P., & Sodroski, J. (2000). Association of Structural Changes in the V2 and V3 Loops of the gp120 Envelope Glycoprotein with Acquisition of Neutralization Resistance in a Simian-Human Immunodeficiency Virus Passaged In Vivo. *Journal of Virology*. https://doi.org/10.1128/jvi.74.24.11955-11962.2000

Yeske, P. E., Betlach, A., Evelsizer, R. W., & Hammer, J. M. (2021). Evaluation of shedding and effect on pig performance of Prevacent® PRRS vaccine in commercial conditions. *52nd American Association of Swine Veterinarians Annual Meeting 2021*, 203–207.

Yeske, P., & Murtaugh, M. (2008). Epidemiology of a new PRRS virus isolate and Outbreak. In T. Smith (Ed.), *Proceedings of the Allen D. Leman Swine Conference* (pp. 11–15). Proceedings of the Allen D. Leman Swine Conference.

Yim-im, W., Anderson, T. K., Paploski, I. A. D., VanderWaal, K., Gauger, P., Krueger, K., Shi, M., Main, R., & Zhang, J. (2023). Refining PRRSV-2 genetic classification based on global ORF5 sequences and investigation of their geographic distributions and temporal changes. *Microbiology Spectrum*, e02916-23. https://doi.org/10.1128/spectrum.02916-23

Yoshii, M., Okinaga, T., Miyazaki, A., Kato, K., Ikeda, H., & Tsunemitsu, H. (2008). Genetic polymorphism of the nsp2 gene in North American type-Porcine reproductive and respiratory syndrome virus. *Archives of Virology*, *153*(7), 1323–1334. https://doi.org/10.1007/s00705-008-0098-6

Yu, F., Yan, Y., Shi, M., Liu, H.-Z., Zhang, H.-L., Yang, Y.-B., Huang, X.-Y., Gauger, P. C., Zhang, J., Zhang, Y.-H., Tong, G.-Z., Tian, Z.-J., Chen, J.-J., Cai, X.-H., Liu, D., Li, G., & An, T.-Q. (2020). Phylogenetics, Genomic Recombination, and NSP2 Polymorphic Patterns of Porcine Reproductive and Respiratory Syndrome Virus in China and the United States in 2014–2018. *Journal of Virology*, *94*(6), e01813-19. https://doi.org/10.1128/jvi.01813-19

Yu, G. (2021). *seqcombo: Visualization Tool for Sequence Recombination and Reassortment* [Computer software]. https://doi.org/10.18129/B9.bioc.seqcombo

Zhang, J., Zheng, Y., Xia, X. Q., Chen, Q., Bade, S. A., Yoon, K. J., Harmon, K. M., Gauger, P. C., Main, R. G., & Li, G. (2017). High-throughput whole genome sequencing of Porcine reproductive and respiratory syndrome virus from cell culture materials and clinical specimens using next-generation sequencing technology. *Journal of Veterinary Diagnostic Investigation*, *29*(1), 41–50. https://doi.org/10.1177/1040638716673404

Zhang, M., Gaschen, B., Blay, W., Foley, B., Haigwood, N., Kuiken, C., & Korber, B. (2004). Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin. *Glycobiology*. https://doi.org/10.1093/glycob/cwh106

Zhang, Q., Jun, S. R., Leuze, M., Ussery, D., & Nookaew, I. (2017). Viral phylogenomics using an alignment-free method: A three-step approach to determine optimal length of k-mer. *Scientific Reports*. https://doi.org/10.1038/srep40712

Zhao, H., Han, Q., Zhang, L., Zhang, Z., Wu, Y., Shen, H., & Jiang, P. (2017). Emergence of mosaic recombinant strains potentially associated with vaccine JXA1-R and predominant circulating strains of porcine reproductive and respiratory syndrome virus in different provinces of China. *Virology Journal*, *14*(1), 67. https://doi.org/10.1186/s12985-017-0735-3

Zhao, K., Gao, J.-C., Xiong, J.-Y., Guo, J.-C., Yang, Y.-B., Jiang, C.-G., Tang, Y.-D., Tian, Z.-J., Cai, X.-H., Tong, G.-Z., & An, T.-Q. (2018). Two Residues in NSP9 Contribute to the Enhanced Replication and Pathogenicity of Highly Pathogenic Porcine Reproductive and Respiratory Syndrome Virus. *Journal of Virology*, *92*(7), e02209-17. https://doi.org/10.1128/jvi.02209-17

Zhou, X., Shen, X. X., Hittinger, C. T., & Rokas, A. (2018). Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. *Molecular Biology and Evolution*, *35*(2). https://doi.org/10.1093/molbev/msx302

ProQuest Number: 30989469

INFORMATION TO ALL USERS
The quality and completeness of this reproduction is dependent on the quality
and completeness of the copy made available to ProQuest.

Distributed by ProQuest LLC (2024).
Copyright of the Dissertation is held by the Author unless otherwise noted.

This work may be used in accordance with the terms of the Creative Commons license
or other rights statement, as indicated in the copyright statement or in the metadata
associated with this work. Unless otherwise specified in the copyright statement
or the metadata, all rights are reserved by the copyright holder.

This work is protected against unauthorized copying under Title 17,
United States Code and other applicable copyright laws.

Microform Edition where available © ProQuest LLC. No reproduction or digitization
of the Microform Edition is authorized without permission of ProQuest LLC.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346 USA