nature communications

Article

https://doi.org/10.1038/s41467-024-53076-w

Untangling lineage introductions, persistence and transmission drivers of HP-PRRSV sublineage 8.7

Received: 23 October 2023

Accepted: 27 September 2024

Published online: 13 October 2024

Check for updates

Yankuo Sun[1,2,3,4,9], Jiabao Xing[1,9], Samuel L. Hong[5,9], Nena Bollen[5,9], Sijia Xu[1], Yue Li[1], Jianhao Zhong[1], Xiaopeng Gao[1], Dihua Zhu[1], Jing Liu[1], Lang Gong[1], Lei Zhou[6], Tongqing An[7], Mang Shi[8], Heng Wang[1,2,3,4,10] ✉, Guy Baele[5,10] ✉ & Guihong Zhang[1,2,3,4,10] ✉

Despite a rapid expansion of Porcine reproductive and respiratory syndrome virus (PRRSV) sublineage 8.7 over recent years, very little is known about the patterns of virus evolution, dispersal, and the factors influencing this dispersal. Relying on a national PRRSV surveillance project established over 20 years ago, we expand the available genomic data of sublineage 8.7 from China. We perform independent interlineage and intralineage recombination analyses for the entire study period, which showed a heterogeneous recombination pattern. A series of Bayesian phylogeographic analyses uncover the role of Guangdong as an important infection hub within Asia. The spatial spread of PRRSV is highly linked with a composite of human activities and the heterogeneous provincial distribution of the swine industry, largely propelled by the smaller-scale Chinese rural farming systems in the past years. We sequence all four available modified live vaccines (MLVs) and perform genomic analyses with publicly available data, of which our results suggest a key "leaky" period spanning 2011–2017 with two concurrent amino acid mutations in ORF1a 957 and ORF2 250. Overall, our study provides an in-depth overview of the evolution, transmission dynamics, and potential leaky status of HP-PRRS MLVs, providing critical insights into new MLV development.

Porcine reproductive and respiratory syndrome (PRRS) is one of the most devastating diseases currently affecting the global pork industry, especially in China and the United States. PRRS has caused significant economic loss over the last decades[1]. The etiologic agent, porcine reproductive and respiratory syndrome virus (PRRSV), is a single positive-strand enveloped RNA virus, with a genome size of 15kbs, containing at least ten open reading frames (ORFs): ORF1a, ORF1b, ORF2a, ORF2b, ORF3, ORF4, ORF5, ORF6, and ORF7. PRRSV belongs to

[1]Key Laboratory of Zoonosis Prevention and Control of Guangdong Province, College of Veterinary Medicine, South China Agricultural University, Guangzhou 510642, China. [2]Guangdong Laboratory for Lingnan Modern Agriculture, Guangzhou 510642, China. [3]National Engineering Research Center for Breeding Swine Industry, South China Agricultural University, Guangzhou 510642, China. [4]Maoming Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Maoming 525000, China. [5]Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Leuven, Belgium. [6]Key Laboratory of Animal Epidemiology of the Ministry of Agriculture and Rural Affairs, College of Veterinary Medicine, China Agricultural University, Beijing 100193, People's Republic of China. [7]State Key Laboratory for Animal Disease Control and Prevention, Harbin Veterinary Research Institute, Chinese Academy of Agricultural Sciences, Harbin, China. [8]School of Medicine, Shenzhen campus of Sun Yat-sen University, Sun Yat-sen University, Shenzhen, China. [9]These authors contributed equally: Yankuo Sun, Jiabao Xing, Samuel L. Hong, and Nena Bollen. [10]These authors jointly supervised this work: Heng Wang, Guy Baele, Guihong Zhang. ✉e-mail: wangheng2009@scau.edu.cn; guy.baele@kuleuven.be; guihongzh@scau.edu.cn

the order *Nidovirales* and the family *Arteriviridae*, emerging almost simultaneously as two species (*Betaarterivirus suid* 1 and *Betaarterivirus suid* 2), and with almost 50%–70% nucleotide homogeneity[2]. *Betaarterivirus suid* 2 can be further divided into 9 separate lineages based on the ORF5 gene[3]. To date, lineage 1, lineage 3, lineage 5, and sublineage 8.7 have circulated in the mainland of China[4,5].

PRRSV lineage 8 strains have a long evolutionary history on a global scale. In 1995, Iowa saw a large number of spontaneous abortions and deaths in pregnant sows, characterized as an "abortion storm", which spilled over to the entire United States. Lineage 8 PRRSVs constituted a large proportion of the emerging strains during this event[6]. Subsequently, sublineage 8.7, specifically the CH-1a cluster, was detected in China and later lead to the widespread viral dissemination on farms which lacked strict biosafety measures. Since then, sublineage 8.7 has been frequently detected and has established itself as endemic in China. In 2006, the outbreak of a highly pathogenic PRRSV (HP-PRRSV) strain with much higher virulence was reported, and resulted in more economic losses[7]. Evolutionary analyses suggested that these highly pathogenic HP-PRRSV isolates belonged to sublineage 8.7[7]. In following years, recombination events and the adaptive evolution of HP-PRRSV further added to the complexity of the genetic diversity of this sublineage[3,8]. Given the increased harm from HP-PRRSV, three modified live vaccines (MLV) with the attenuated strains JXA1-R, HuN4-F112 and TJM-F92 were rapidly licensed for emergency use. These vaccines have been widely used in China until recently[9–11], when another HP-PRRSV vaccine, GDr180, was licensed in 2015. Due to the lack of 3′–5′ exonuclease proofreading during replication, this MLV is characterized by low replication fidelity and high mutability, which raises the risk of reversion to virulence. In the last 25 years since the first PRRSV MLV was licensed, many clinical investigations have shown the potential of virulence revision of HP-PRRS MLVs[12–14]. For example, Jiang et al. isolated three field strains and found that these had the highest nucleotide similarity to the HP-PRRSV-derived vaccine strain JXA1-R. These strains were able to cause high fever and mortality in the inoculated pigs, indicating the reversion to fatal virulence[15]. Another study proved the ability to regain virulence through an in vivo reverse passage test[16]. A final example deployed an intranasal inoculation experiment where the vaccine JXwn06-P80 regains its fatal virulence at the 9th in vivo passage. JXwn06-P80 also regains fatal virulence through the reverse passage in porcine alveolar macrophages (PAMs)[17]. However, although several in vivo and in vitro experiments have confirmed the potential that HP-PRRSV-derived MLV vaccines can regain their virulence, a study to comprehensively assess the "leaky" status of HP-PRRS MLVs – i.e., the history of reversion to virulence of each HP-PRRSV MLVs–was still lacking[10,15,17].

When confronted with low diversity and potential issues related to sampling bias, evolutionary reconstructions that shed light on the spatiotemporal evolution of viruses may greatly benefit from integrating additional sources of information. Bayesian phylodynamic approaches are particularly adept for this purpose. Furthermore, phylogeographic methods have been extended to take advantage of human transportation data as proxies of population-level connectivity between locations. This approach has been utilized in a wide range of applications, including the identification of the key drivers of Ebola virus spread in West Africa and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) Omicron BA.1 in the United Kingdom[18,19]. Additionally, recent studies that employ Bayesian phylogeographic inference with a GLM extension effectively demonstrated the role of anthropogenic activities (i.e., swine trade) in the transmission of porcine epidemic diarrhea virus (PEDV) and PRRSV[20,21]. As a result of our national epidemiological surveillance project of PRRSV, we have observed that China is experiencing a key genotypic shift from lineage 8 to lineage 1[4,22]. However, we still failed to investigate how sublineage 8.7 spread globally and locally.

Here we assembled an extended genomic dataset regarding PRRSV-2 sublineage 8.7, including 242 novel sublineage 8.7 ORF5 sequences and 42 new complete genomes collected in China between 2005 and 2022. Importantly, we also sequenced all available MLV vaccines derived from sublineage 8.7 HP-PRRSV. With this genomic dataset, we performed a series of genomic analyses to answer important questions on the emergence and spread of sublineage 8.7, including: (1) How did sublineage 8.7 emerge and spread in China? (2) Which factors affect the spread of the virus in China? (3) What are the recombination dynamics (if any)? (4) Did the vaccine strains contribute to the persistence of sublineage 8.7 in China? Answering these questions allows us to fill key knowledge gaps concerning the evolution and spread of PRRSV.

Results

Generation of large-scale dataset

Here we quantified the distribution according to genotypic and spatiotemporal information in our genomic database (Fig. 1). From the timeline, we observed that from 1994 to 1999, only a very few sequences from the the lineage 8 paraphyletic cluster were obtained, whereas an extremely pronounced surge of infections was observed from 2000 to 2005. Subsequently, lineage 8 strains have undergone rapid population expansion from 2006 to 2016, most of which were clustered into sublineage 8.7 and located in China, with most sequences in the paraphyletic cluster detected in USA (90.1%). We have also illustrated the distribution of samples at the provincial level in China, since there is an unprecedented proportion of sublineage 8.7 in China, to better characterize the geographic distribution of lineage 8 isolates (Fig. 1). Our dataset included sequences from all provinces in China, and southern China (Guangdong) played a key role in terms of overall contribution.

Sublineage 8.7 underwent a geographically centralized spread in Asia

The phylogenetic estimation of sublineage 8.7 indicated that after a short period in which classical CH-1a-like cluster were mostly present, the HP-PRRSV cluster gained prominence, with major genetic distance. In the classical CH-1a-like cluster, we were able to identify that sublineage 8.7 cluster has spread substantially in eastern China (i.e., Zhejiang, Fujian, Shandong, and Jiangsu), Southern China (i.e., Hubei), Central China (i.e., Henan), and Northern China (i.e., Beijing) in the early stage (Fig. 2). Also note that, with the transition to the HP-PRRSV cluster as the dominant cluster, South China harbored more transmission, typically in Guangdong. Afterwards, HP-PRRSV developed into an endemic cluster and was found in over 30 provinces and autonomous regions of China and several other Asian countries. Dominating phylogenetic branches seem to be related to the Guangdong backbone, indicative of a potential origin of these epizootics (Fig. 2). Although Chinese strains have yet to develop into geographically specific clades, isolates located in Guangdong were distributed in all the main branches, indicating that Guangdong played a crucial role in the dispersal of the virus to other locations.

Bayesian phylogeographic reconstruction and drivers of sublineage 8.7 spread within China

From our time-resolved maximum clade credibility (MCC) tree, we estimate the emergence and origin of PRRSV sublineage 8.7 HP-PRRSV in China to be around 1987 [95% HPD interval = 1976–1996] in Guangdong province (Fig. 3A). The dispersal pattern obtained from our analysis suggests that the spread of this sublineage is broadly characterized by a source-sink dynamic, with Guangdong province acting as the major source of viral lineages to the remaining provinces, and Henan, Shandong, and Jiangsu acting as minor amplifying hubs (Fig. 3B, C). Coupled with a strong support (BF > 100) for distance as driver of spread (Fig. 4), which indicates that transmissions occurred

https://doi.org/10.1038/s41467-024-53076-w

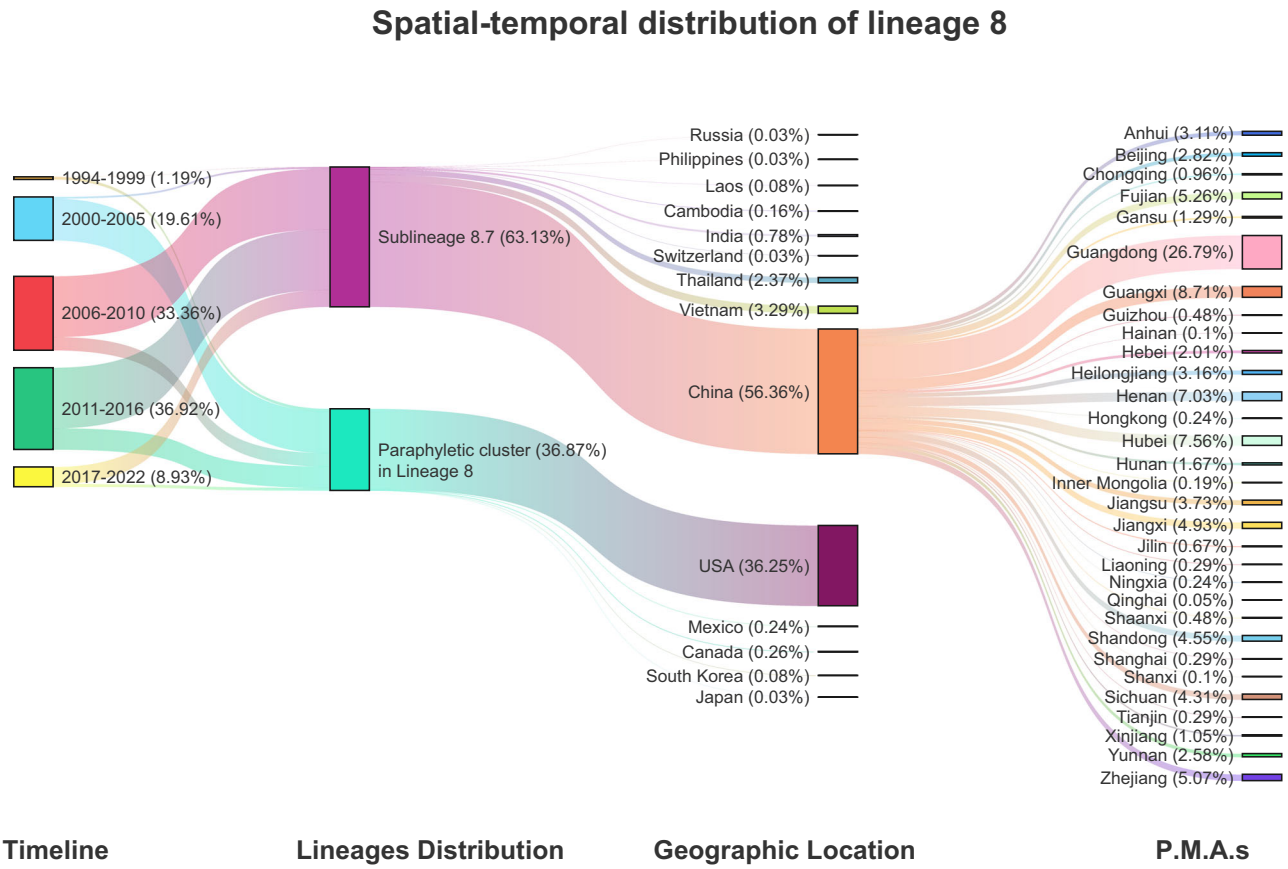

Fig. 1 | Global spatiotemporal distribution of lineage 8 sequences. Global spatiotemporal proportion and lineage distribution of 3708 lineage 8 sequences.

more frequently between nearby provinces, we observe that the spread follows a gravity-like pattern, with the number of introduction events from Guangdong decreasing as the distance between the locations becomes larger (Fig. 3B). Any measure of the scale of swine farming will be negatively correlated, which we term as an 'anti-gravity' effect. This suggests that regions with larger industrial swine farming operations may have better biosecurity measures in place, reducing the likelihood of viral introductions despite their larger population sizes.

Another key goal of this study was to assess and quantify the contribution of various factors that influenced the geographical dissemination of sublineage 8.7 in China. We considered and incorporated ecological, anthropocentric, economic, and geographical variables at the provincial level, which may impact the process of viral spread using a discrete phylogeographic generalized linear model (GLM) approach. Figure 4 shows the posterior estimates of the inclusion probabilities and conditional effect sizes of the log-transformed covariates to quantify the contribution of predictor variables to the among-province lineage transition rates. Only predictors with a Bayes Factor above 3 are displayed (for posterior estimates for all predictors, see Supplementary Fig. 3). Besides geographical distance, we see that five other covariates are strongly supported (BF > 100) in the model: *per capita* pork sold by rural residents (destination), gross population (origin and destination), breeding stock (destination) and "from Guangdong". The high support and large effect size of this "from Guangdong" predictor further support the source-sink gravity pattern previously described.

The overall significant correlated covariates seem to indicate that human-related activities might be influencing the spread of PRRSV (Fig. 4). Specifically, this is corroborated by the support for the distance predictor as closer proximity facilitates interprovincial activities. Additionally, in recent years, rural-scale pig farms have generally operated on a smaller scale, and their biosecurity measures for PRRSV prevention and control have been less emphasized. Consequently, rural swine farming becomes a more probable factor for the dispersion of the pathogen to other proximate regions through activities such as the transport and sale of pork (Fig. 4). Moreover, our analysis shows strong support for "breeding stock" at the destination as a predictor of PRRSV spread with a negative effect size. We note that to properly interpret this predictor, we must consider the methodology behind selecting the covariates considered in our model. Our covariate correlation analysis revealed that this predictor is part of a highly correlated ($r > 0.97$) cluster of covariates that includes "breeding stock", "pork pigs slaughtered", and "pork production" (Supplementary Fig. S7). Although these covariates measure distinct quantities, the choice of which covariate to include as a potential predictor in the model has no effect in our interpretation. However, the high cross-correlation between them steers us to broaden our interpretation beyond the narrow scope measured by each variable. Instead, we consider them collectively as proxies for a latent variable that captures the overall scale of industrial swine production within a given province. As a result, the negative effect size of "breeding stock" should be interpreted as a protective effect of industrial pork production on the spread of PRRSV.

Furthermore, the inverse relationship between rural production and breeding stock can be intuitively explained once we consider them as components of an unobserved covariate measuring the ratio of rural to industrial pork production in a province. This inverse relationship further suggests that dispersal of HP-PRRSV occurs more frequently into provinces where the ratio of rural-to-industrial swine farming is higher. As for population size, an increase in population size inherently promotes pork consumption, which may lead to a higher intensity of

https://doi.org/10.1038/s41467-024-53076-w

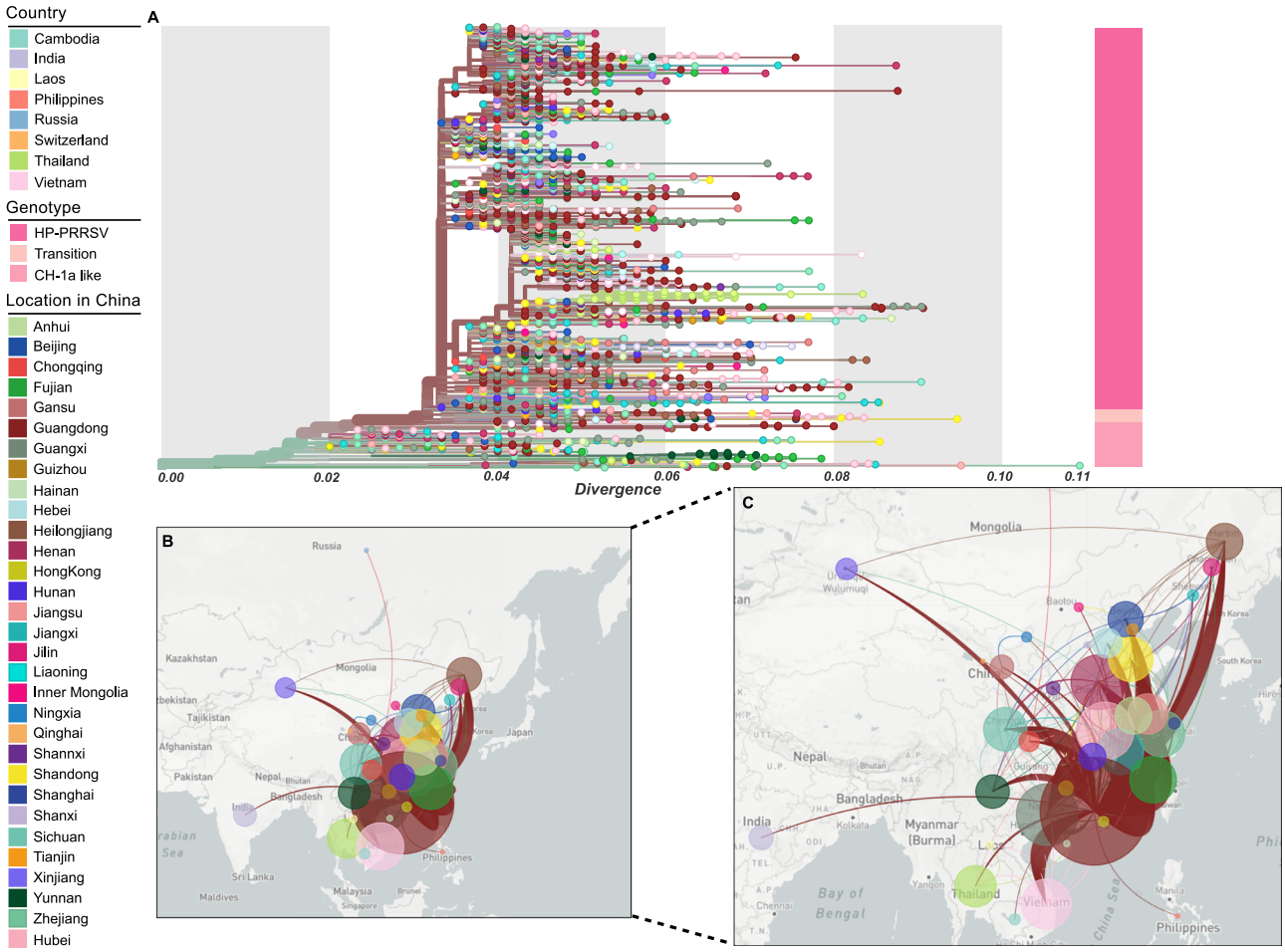

Fig. 2 | PRRSV sublineage 8.7 global phylogeographic reconstruction. **A** Maximum-likelihood phylogeny of global sublineage 8.7 strains until 2022 with countries annotated on the ancestral nodes and branches. Lineage 8 transition periods were attached in the right panel with different genotypes (deep pink: HP-PRRSV isolates, pink: transition isolates from classical CH-1a cluster to HP-PRRSV cluster). **B** Phylogeographic reconstruction of PRRSV sublineage 8.7. The line thickness signifies the captured spatial transmission routes. The colors of the circle and relevant routes correspond to the color of the ancestral nodes. **C** Zoom-in map represents the detailed transmission routes in Southeast Asia. The map showed in Panels B and C were generated using a custom-built Nextstrain pipeline.

pig trade between provinces. This relationship was also corroborated in a PEDV phylogeographic study[21].

Taking these aspects into consideration, we hypothesize that the spatial spread of PRRSV is highly correlated with integrated human activities and the provincially heterogeneous distribution of the swine industry. Our results suggest that interprovincial spread is primarily sourced from Guangdong province and is driven by overloaded rural small-scale farming and commerce in China in the past several years.

Intralineage and interlineage recombination investigation present divergent landscape of recombinant preference

We employed two different approaches to identify potential recombination events. First, we constructed a phylogenetic network to detect the distribution of inter- and intra-lineage recombination (Fig. 5A). Using the pairwise homology index of the neighbor-net method, we identified an significant recombination signal ($p < 0.001$). Secondly, a more detailed investigation of inter- and intra-lineage recombination revealed a divergent recombination pattern.

We tracked the recombinant history of lineage 8, taking into account recombination with other lineages as well as intralineage recombination and the temporal distribution of recombination events. Regarding interlineage recombination, the first recombinant event can be traced back to 2007, with fewer recombination events detected between 2007 and 2013. However, since 2014, the number of interlineage recombinant events increases exponentially, with lineage 1 and lineage 3 contributing frequently as minor parents, particularly during 2014-2018 (Fig. 5B, D and E). Since 2010, ORF1ab and ORF3-ORF5 were the regions that code for the structural protein which saw the most recombination. These regions encode GP3, GP4, GP5, as well as a series of non-structural proteins (nsp) (Fig. 5E). Specific to ORF1a is that frequent recombination events were detected in the nsp2 region (40.0%), with most of these events associated with lineage 1 (91.7%) in 2014-2016. In ORF1b, most events were found on the nsp12 region (38.5%), contributed by lineage 1, 3, and 5. The number of recombination events in structural regions was about 55% higher than in non-structural regions, although the genomic length of non-structural regions was relatively longer than that of structural regions.

Intralineage recombination events were detected more frequently compared with interlineage recombination. As for interlineage recombination, the first event of intralineage recombination dates back to 2007, with fewer events between 2008 and 2009. However, we observed a surge in recombination events between 2010 and 2015, followed by significant fluctuations in the frequency of these events. Unlike interlineage recombination, the genomic region with the most intralineage recombination was found in the nsp region, i.e., ORF1ab. Events in ORF1a showed a relatively uniform distribution regardless of the genomic length of a specific region. Comparatively, nsp9 was detected with higher frequency in ORF1b. Besides ORF1ab, ORF4 also exhibited a relatively high frequency among structural protein regions. Overall, both the interlineage recombination likelihood and the

https://doi.org/10.1038/s41467-024-53076-w

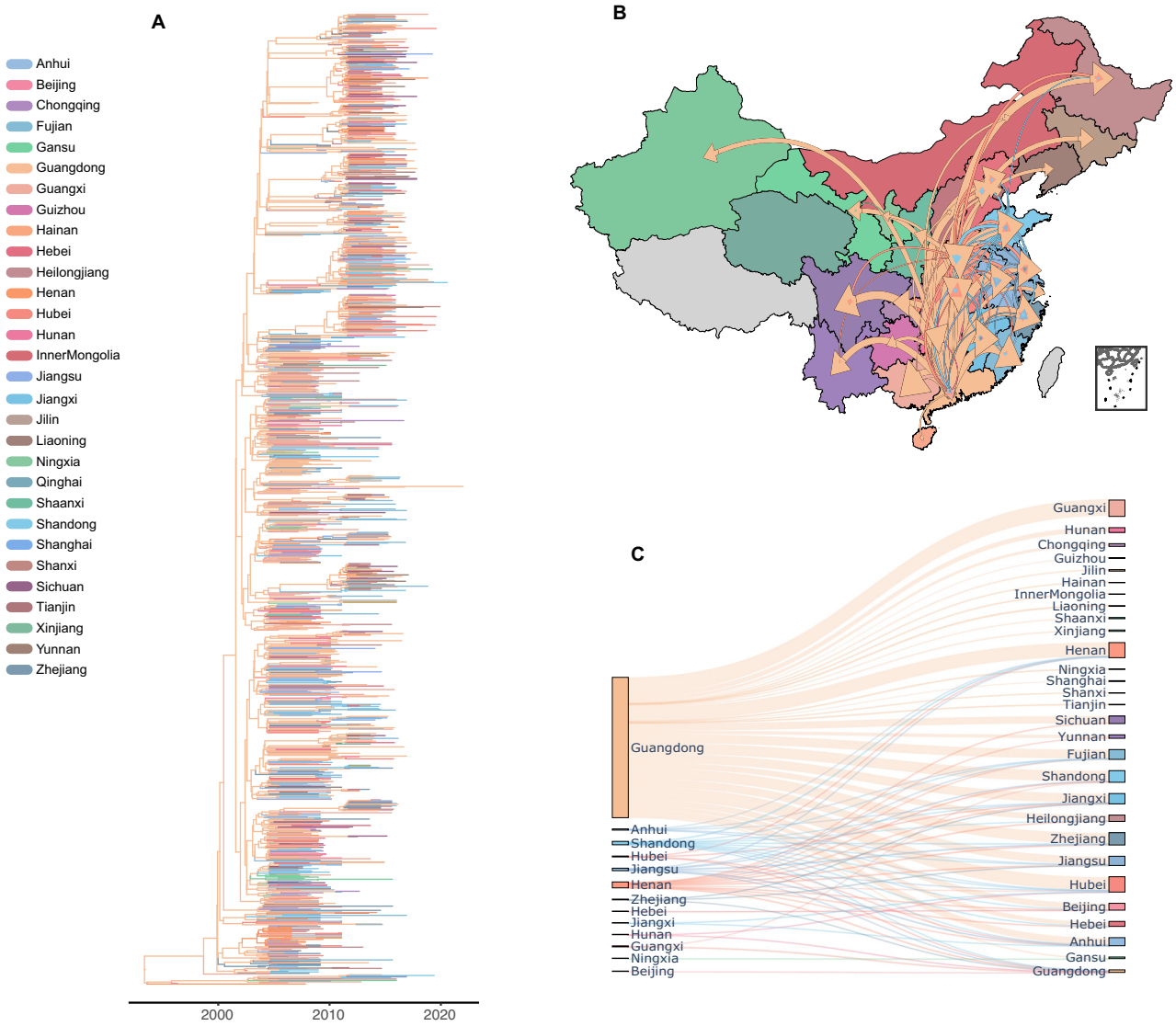

Fig. 3 | PRRSV sublineage 8.7 phylogeographic reconstruction within China. **A** Maximum clade credibility tree with ancestral nodes and branches colored according to estimated (province) location, depicting the spread of PRRSV within China. **B** Spatial spread of PRRSV in China based on the posterior expectations of Markov jumps. In this plot, the colors of migration link of each location correspond to the source location. The thickness of migration link correspond to the values of Markov jumps. **C** Sankey plots summarizing Markov jump estimates for the transition between provinces. The plots show the relative number of transitions between origin (top) and destination (bottom) locations. Note that locations may both be origin locations (in the left row) and destination locations (in the right row), and there is no temporal order for the transitions involved.

genomic hotspot region differed with the region in intralineage 8. Specifically, intralineage recombination occurred earlier and more frequently compared to interlineage recombination (interlineage: 2014-2016 and intralineage: 2010-2015). Considering there are five approved lineage 8 MLVs to market in China until now and they share a higher administration rate compared with MLVs of other lineages, we speculate that the heterogeneous frequency between interlineage and intralineage recombination may be related to the lineage 8 MLV administration in China[23].

Genomic insights of HP-PRRSV Modified Live Vaccines reversion

Although several in vivo and in vitro experiments have confirmed the potential for HP-PRRSV-derived MLV vaccines to regain their virulence, we still lacked evidence to assess the "leaky" status of HP-PRRSV MLVs. Note that genomic evidence of the MLV vaccine-derived clinical sequences suggests it has been widely used in China in past decades[24]. Therefore, we sequenced all HP-PRRSV-related vaccines approved for clinical use in China to obtain complete genomes for the four approved HP-PRRS MLVs (i.e., JXA1-R, HuN4-F112, GDr180, and TJM-F92). Then, using several phylogenetic approaches as well as a temporal analysis, we characterize the specific molecular marker for clinical vaccine-homogeneous strains.

We first estimated an ML phylogeny using our complete genome dataset as well as the vaccine strains to identify monophyletic clusters corresponding to each vaccine strain, i.e., vaccine clusters (Supplementary Fig. S11). Using ClusterPicker, a total of 41 clinical strains associated with vaccines were selected with a fairly robust threshold of homogeneity (bootstrap value: 85, genetic distance: 97%). Specifically, we constructed a haplotype using the nsp9 gene (encoding RdRp) to identify the homogeneous relationship between field isolates and vaccine strains. Except for the GDr180 cluster, all clinical strains fell into the ancestral node of homogenous vaccine strains, suggesting that these field strains were likely to be homogeneous with corresponding MLV vaccines (Fig. 6). To elucidate further, within the JXA1-R haplotype relationship, a consecutive series of passage viruses, specifically from JXA1/P10 to JXA1/P70, were observed progressively

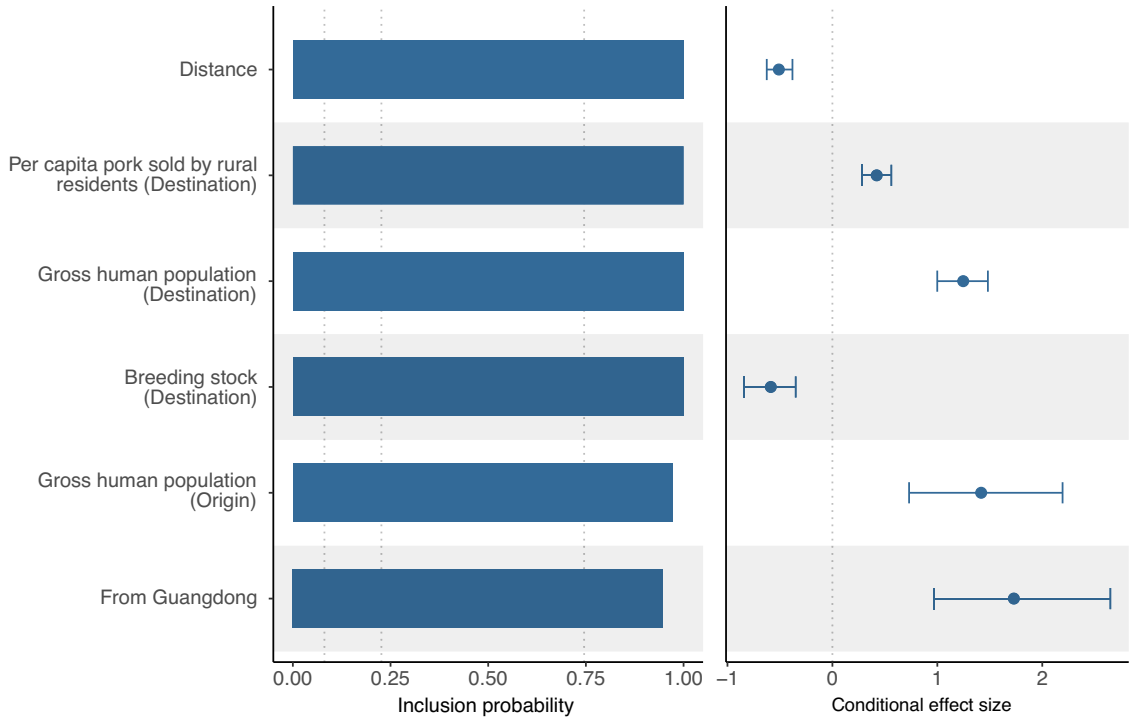

Fig. 4 | The support and contribution of PRRSV diffusion predictors among 30 Chinese provinces. Support for each predictor is represented by an inclusion probability that is estimated as the posterior expectation for the indicator variable associated with each predictor. Indicator expectations corresponding to Bayes factor support values of 3, 20, and 150 are represented by a dotted vertical line in this bar plot. Here we only showed the predictors which had BF values >3. The contribution of each predictor is represented by the mean and credible intervals of the GLM coefficients on a log scale conditional on the predictor being included in the model. The support and contribution of all predictors was included in Supplementary Fig. S10.

converging towards a progeny node, denoted as JXA1-R. Several field variants were subsequently noted to diverge from the aforementioned JXA1-R node, signifying a plausible homogeneous lineage relationship between the ancestral sequences and their progeny (Fig. 6A). Of particular interest is the NT2/2015 node, a reported reversion case from JXA1-R, which speciated at the terminal branch of the haplotype[15]. Furthermore, MLV vaccines TJM-F92 and HuN4-F112 were also embedded in a key position, which indicates a key hub of viral dissemination among TJM F92 and HuN4-F112 related field strains. As a counterexample, GDr180 - the latest approved vaccine in 2015 - with a smaller market share, had a less homogeneous relationship with field strains. All strains in this cluster were embedded at the terminal of the haplotype, which suggested a less likely homogeneous relationship (Fig. 6A). We further analyzed the cumulative time series cases of the homogenous strains (Fig. 6B). In the JXA1, TJM and HuN4 clusters, yearly reported cases of field strains remained zero until the corresponding vaccines were approved for clinical use in 2011. Specifically, since the MLVs (including JXA1-R, TJM-F92, and HuN4-F112) were widely used, each vaccine cluster has increased remarkably for a period of 6 years (from 2011 to 2017), during which the new lineage (lineage 1) was introduced into Asia. The sharp rise of the number of clinical strain cases during six continuous years reflects the clinical impact of MLV vaccination. Cases declined since 2017, as the dominant lineage changed from lineage 8 to lineage 1 in China. In the GDr180 cluster however, we observed an abnormal surge spanning 2006 – 2009, during which GDr180 was not yet developed. In fact, GDr180 was not in use until 2015. This, combined with its low market coverage further corroborates that GDr180 is less likely to reverse (within our current surveillance dataset).

Furthermore, we identified the potential amino acid markers associated with MLV reversion. We applied the following criteria to identify sites of interest: (i) the amino acid site was substituted between the parental strain and the corresponding MLV strain (for example, JXA1 and JXA1-R); (ii) the amino acid site was consistently mutated in the field strains (at least 50% of cases); (iii) the amino acid mutation site in the field strains is consistent with the one in the MLV strains. In light of our previous analysis, the GDr180 cluster was excluded entirely. Using these criteria, we identified 35 concurrent amino acid mutations for the TJM-F92 cluster isolates, specifically in ORF1ab, ORF3, and ORF5 (Supplementary Fig. S12); for JXA1-R we found 32 concurrent amino acid mutations distributed among ORF1ab, ORF2, ORF3, ORF4, and ORF5 (Supplementary Fig. S13); for HuN4-F112 cluster isolates we found 13 concurrent amino acid mutations distributed among ORF1ab, ORF2, and ORF5 (Supplementary Fig. S14). We identified that the JXA1-R and HuN4-F112 clusters shared an identical amino acid substitution (JXA1:F250S、HuN4: T250I) in ORF2. Similarly, both the JXA1-R and TJM-F92 clusters shared the T225A mutation in ORF3 and an identical amino acid substitution in ORF1a (JXA1-R: E957G TJM-F92:T957S).

Discussion

Despite a rapid increase in the number of sublineage 8.7 infections in Asian countries over recent years, very little was known about the dynamics of PRRSV emergence and spread. Relying on a national long-term PRRSV surveillance project, we collected over 6000 suspected positive samples to sequence and obtained 242 new ORF5 sequences and 42 complete genome sequences belonging to sublineage 8.7; these data spanned approximately two decades. We integrated these novel sequences with publicly available genomic data in order to form a large collection of available PRRSV sublineage 8.7 sequences. Our goal was to explore how sublineage 8.7 emerged, evolved, was transmitted, and recombined (intra- and interlineage) in the nearly two decades since its emergence in 2006[4,5,25,26].

Since several HP-PRRS MLVs were hastily approved for use on an emergency basis in China, and given that few studies focused on the potential impact of these vaccinations, we sequenced all HP-PRRS

https://doi.org/10.1038/s41467-024-53076-w

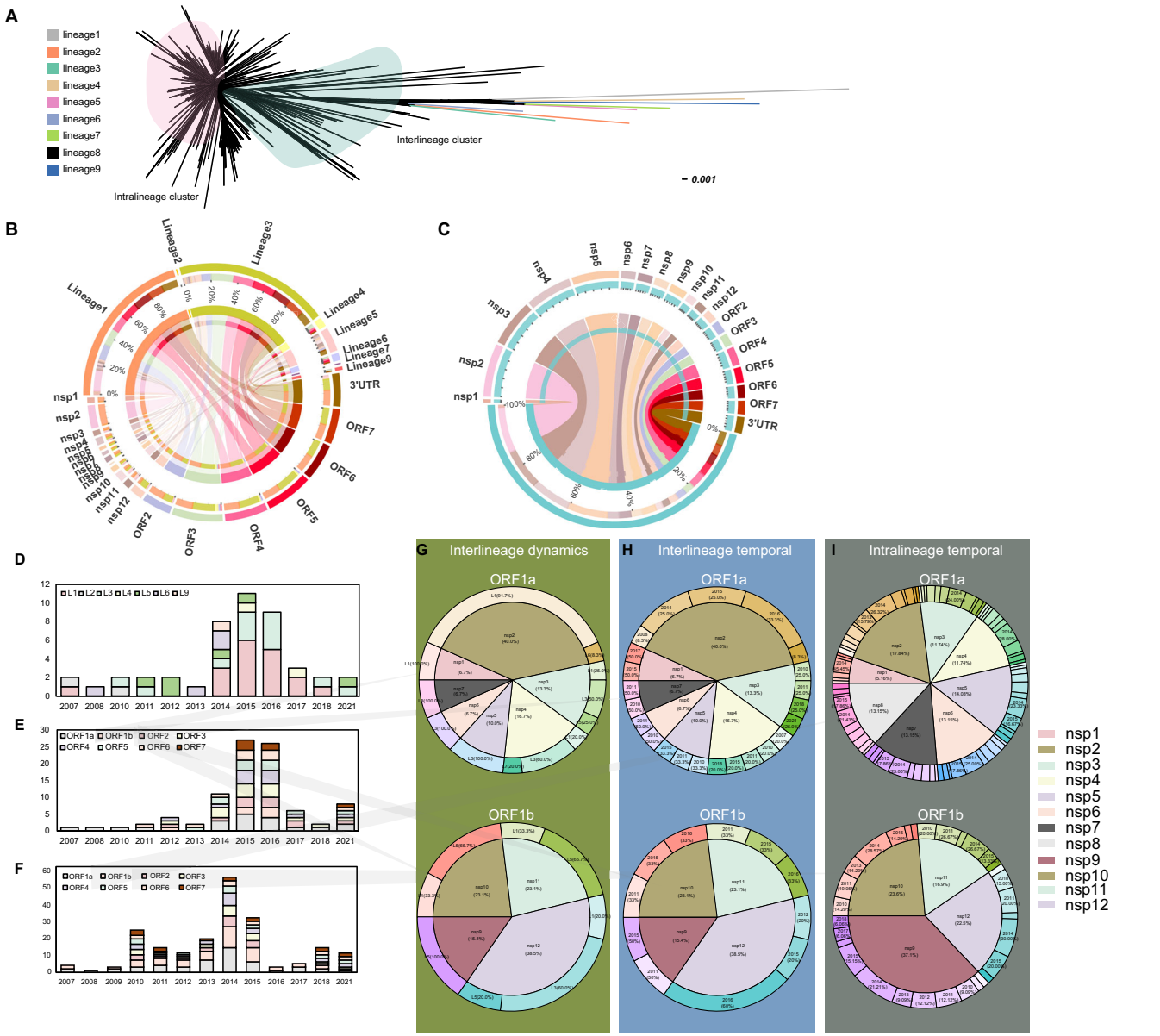

Fig. 5 | Recombination landscape. A Phylogenetic network of full-length genomes of lineage 8, using the SplitsTree5 software with the Kimura 2-parameter model. Isolates in the red shaded region corresponding to the intralineage recombination within lineage 8, the green shaded region corresponds to the interlineage recombination with lineage8, with a statistically significant difference using Phi Test ($p < 0.0001$). **B** Overview of interlineage recombination patterns. The relative size of linkages from the upper part to the lower part correlates to the recombination frequency of each lineage as minor parent specific to the recombined region. For example, lineage 1 is more likely recombined as minor parent in ORF4-ORF7. **C** Overview of intralineage recombination patterns. The relative thickness of curve in the upper part correlates to the recombination frequency of each region as minor parent. For example, most intra-lineage recombination events result in a new doner in non-structural region. **D** Cumulative number of interlineage recombination events per year with color corresponding to different lineages. **E** Cumulative number of interlineage recombination events per year with color corresponding to different regions. **F** Cumulative number of intralineage recombination events per year with color corresponding to different regions. **G** Cumulative proportion of interlineage recombination events relating to each lineage in specific region of ORF1a (top panel) and ORF1b (bottom). **H** Cumulative proportion of interlineage recombination events relating to each year in specific region of ORF1a (top panel) and ORF1b (bottom). **I** Cumulative proportion of intralineage recombination events relating to each year in ORF1a (top panel) and ORF1b (bottom).

MLVs to analyze their clinical impact. We found strong evidence that HP-PRRS MLVs were "leaky", which may have restored the virulence of PRRSV, based on our multivariate analysis.

In this study, we investigated the spatiotemporal dispersal patterns of sublineage 8.7 using a CTMC-based discrete phylogeographic analysis with covariates. We identified the importance of rural swine activities and provincial distance as contributing factors to the spatial spread of sublineage 8.7. The CTMC model has previously been shown to be sensitive to sampling bias, which is a common concern in phylogeographic analyses. Besides CTMC, two approximations of the structured coalescent model are also widely used for this purpose, as they are theoretically better at handling sampling bias: the Bayesian structured coalescent approximation (BASTA)[27] and the marginal approximation of the structured coalescent (MASCOT)[28]. All three inference methods are potentially affected by geographic sampling bias, but their performance varies depending on the degree of sampling bias in the data[29] Specifically, while the reconstructed spatiotemporal histories were impacted by sampling bias for the three approaches, BASTA and MASCOT reconstructions were shown to also be biased when employing unbiased samples. In contrast, increasing

https://doi.org/10.1038/s41467-024-53076-w

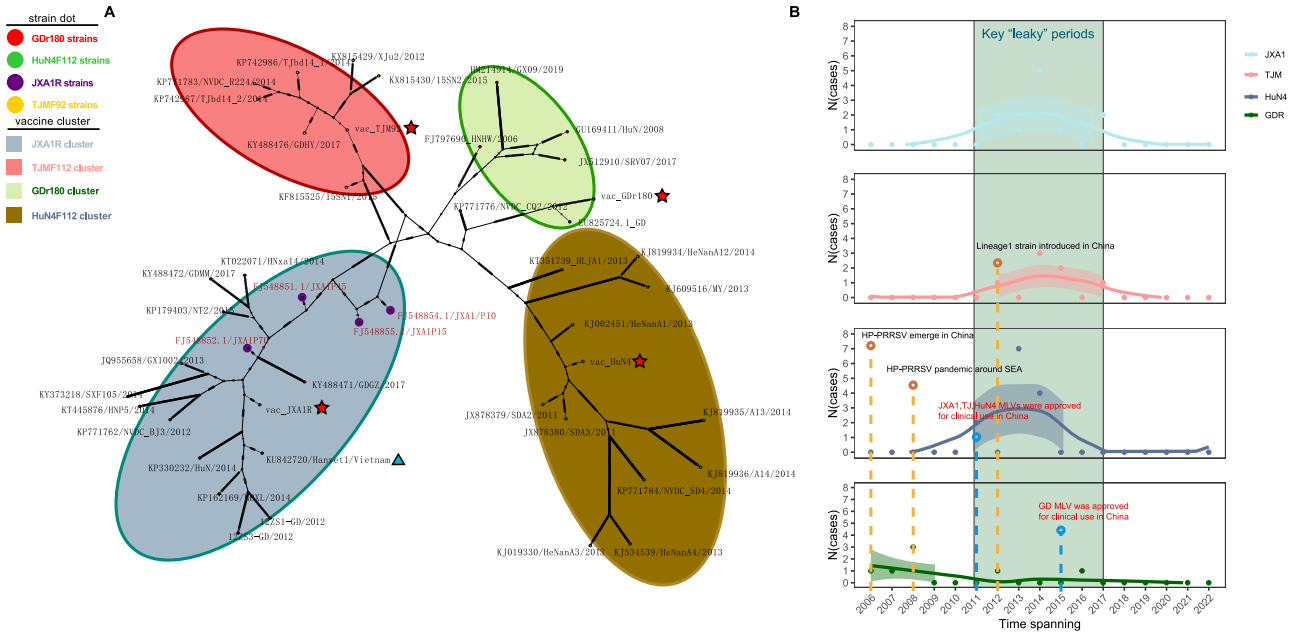

Fig. 6 | Vaccine homogeneous analysis. A TCS haplotype network reconstruction, with nodes colored by MLV clusters. MLV vaccines were annotated with a red star, with a single Vietnam isolate annotated with a triangle, indicative of the potential implication of MLV-related isolate transmission in the South-east. **B** Temporal homogeneous analysis of field strains in each vaccine cluster, respectively. The shaded area represents the 95% confidence interval of the fitted values using Poisson parameterization estimation.

the number of analyzed genomes led to more robust estimates at low sampling bias for the CTMC model. Alternative sampling strategies that maximize the spatiotemporal coverage greatly improved the inference at intermediate sampling bias for the CTMC model, and to a lesser extent for BASTA and MASCOT. Further, despite the theoretical advantages of these structured coalescent models, their current implementations are unable to scale to datasets as large as the one in our study (1371 sequences and 30 locations). These are strong arguments in favor of our CTMC approach as the most suitable for the spatiotemporal diffusion analysis of sublineage 8.7 in China.

To correctly accommodate for computational demands while accounting for sampling bias, we implemented a subsampling strategy meant to downsize our dataset to computationally manageable numbers and assessed the degree of bias by comparing our sample sizes with PRRSV incidence data (for details of the surveillance work see Supplementary Information). The results showed that our sampling was in fact representative of the distribution of PRRSV in China, however the high correlation between some of our predictors presented further challenges. Although the spike-and-slab prior we use promotes sparsity in the included predictors and partially accounts for multicollinearity in the covariates, our sensitivity analysis showed that pairwise correlations of up to 0.80 resulted in convergence issues (see Supplementary information for more detail). Thus, the application of shrinkage priors in phylogeography may provide a better approach for GLM analyses. Taking all of this into consideration, we believe that the spatial reconstruction and drivers of spread we identified are robust to sampling bias and representative of the true spatial spread history of the pathogen.

We performed a Nextstrain analysis of the sublineage 8.7 clusters. Although several branches were detected in the USA and Russia, nearly the whole phylogenetic trunk was located within Asia, suggesting sporadic transmission events from China to other countries and without any outbreak events identified in other regions. In addition, both the classical sublineage 8.7 cluster and the transition cluster – containing sequences that exhibit many mutations prior to the emergence of sublineage 8.7 - exhibited longer branch lengths, indicative of genetic divergence from the more recent HP-PRRSV cluster (Fig. 2). This finding suggests that the virus was under greater host innate immune pressure and underwent adaptative evolution during the early invasion period. This observation is reminiscent of a study suggesting that the ongoing convergence of SARS-CoV-2 lineages includes multiple mutations that encourage the existence of diverse virus lineages during host immune recognition[30]. Regarding the dispersal history, the results of the Nextstrain analysis allowed us to hypothesize on how PRRSV sublineage 8.7 may be maintained in strict transmission foci. The dissemination pattern of sublineage 8.7 points to an interconnected network of Asian regions. South China serves as an important reservoir of PRRSV, from which the virus spreads not only to the rest of China but also to other neighboring Asian countries such as Vietnam and Thailand. Thailand and Vietnam possibly act as secondary infection hubs to neighboring countries (Laos and Cambodia).

The role of Guangdong as the epicenter of the infection was further corroborated by our GLM analyses and by Chinese surveillance data (Figs. 3, 4 and Supplementary Fig. S8). In our Bayesian discrete phylogeographic analysis, we accurately estimated the early transmission from Guangdong to nearby provinces (e.g., Guangxi) and to provinces in central China, such as Henan and Hubei, with strong Markov jump support. Similarly, He et al. also identified Guangdong as the epicenter of another important porcine virus (PEDV) using Bayesian discrete phylogeographic analysis[21]. This study also successfully linked the trade and consumption of pork with the spread of PEDV in China using a GLM extension. In our GLM model, we found strong support for provincial distance as well as demographic factors such as human population size at origin and pork sale in rural areas. We estimated this difference may be attributable to the host infectivity heterogeneity (PEDV: piglets; PRRSV: boar and pregnant sow) and different transmission capabilities between PRRSV and PEDV. Makau et al. similarly implemented a discrete-space phylogeographic GLM study to explore factors associated with variability in between-sector diffusion rates of PRRSV lineage 1 in the United States[20]. ovement of growing pigs (as opposed to movement of weaned pigs coming straight from breeding farms) was found to be more associated with PRRSV dispersal. In our study, our phylogeographic GLM suggested that spread of HP-PRRSV is more associated with rural farming activity

and that an increasing amount of breeding stock serves as a deterring effect to the dispersal of PRRSV. We speculate that more stringent biosafety practices in breeding farms (compared to growing farms) are likely prohibitive to the circulation of PRRSV in China. We assume this difference is attributable to differences in pig breeding systems between China and the United States, which deserves further analysis to explore how to better prevent and control PRRSV introductions and dispersal in different countries with heterogeneous farming systems.

Recombination occurs as a result of virulence enhancement, host shifting, and adaptability strengthening. PRRSV recombination is significant and pervasive in that it largely enhances genetic diversities and reduces the cross-protection of vaccines. In this study, we analyzed the intra- and interlineage recombination of PRRSV lineage 8, taking into account its temporal dynamics, and found a principal recombination wave spanning from 2014 to 2016. It is commonly accepted that frequent homogeneous RNA viral recombination is the result of random template conversion during replication and is thought to be deployed by the "copy-choice" mechanism of RdRp. Although a high level of both intra- and interlineage events were found, interlineage recombination was more targeted to the structural protein regions (GP3-GP5), whereas intralineage recombination was more concentrated on nonstructural protein regions (ORF1a), with a breakpoint at nsp2-nsp5. This mainly involved antagonizing host innate immune systems such as deubiquitin, IFN antagonist and membrane modification[1]. The significant differences among the number of inter- and intralineage recombinations may be due to the flush vaccination of lineage 8 MLVs. Until now, lineage 8 possessed the largest amounts of approved MLV vaccines of PRRSV in China. Since all PRRS MLV were able to continue to replicate after administration, the "copy-choice" characteristic of RNA polymerase offered the possibility to recombine with field strains within the host. Currently, China possesses only L5 lineage vaccines, derived from the VR2332 lineage, as well as L8 lineage vaccine strains. The use of L8 lineage vaccines significantly outweighs that of L5 lineage vaccines, thereby elevating the probability of genetic recombination occurring. We note that understanding the intricate interplay of vaccines and field strains is a delicate undertaking, and we should be conservative in our conclusions.

Our study is the first to explore how the HP-PRRSV MLVs are likely to affect immunized herds in the field in China. Using multiple phylogenetic reconstructions and recombination elimination, we have identified four MLV groups. We inspected the temporal signal of potential descendants within each group. The JXA1-R, TJM-F92, and HuN4-F112 groups supported our hypothesis, which relied on the premise that the time at which vaccines were approved predates the prevalence of the vaccine-associated field isolates. However, in the case of the GDr180 cluster, we did not find this pattern and found no temporal link between GDr180 and the field isolates. We hypothesized that it is partly due to GDr180 being the latest MLV vaccine to be approved (2015), and as such, it has been used with relatively low frequency. Among the other three HP-PRRS MLV vaccines, JXA1-R is the most frequently administered and the one which was mandatory before 2017 in China. JXA1-R was also the vaccine that was associated with the most field strains. Note that the JXA1-R-homogeneity strain, KU842720/Hanvet1/Vietnam, was detected before the approval of the JXA1-R vaccine in Vietnam and was thus likely imported from abroad. This illustrates the importance of continuous monitoring and of quarantine procedures in the context of cross-regional livestock trading. Although multiple approaches such as infectious clones and challenge experiments have been attempted previously, the commonality of these results allows us to draw conclusions only for specific cases related to reversion sites. There is little knowledge surrounding amino acid markers from MLV supported by comprehensive clinical whole genome data. Our results showed several common amino acid substitution positions on a whole-genome scale, which may be associated with HP-PRRS MLV reversion markers, although specific molecular markers varied with different vaccine clusters. These results should prove helpful when it comes to studying potential vaccine reversion cases, potential vaccine escape cases and other potentially problematic variants.

Our study also has certain limitations. Although we conducted a comprehensive investigation into potential cases/sequences of HP-PRRSV MLV reversion, supported by detailed epidemiological and pathogenetic data, as well as time-dependent phylogenetic evidence linking MLV reversion to field strains associated with MLVs, experimental evidence regarding which non-synonymous mutations significantly affect the pathogenic phenotype of MLVs remains elusive. Nonetheless, our study has produced a comprehensive atlas of non-synonymous mutations across the full genome involved in MLV reversion. In the future, leveraging reverse genetic systems and animal-challenging assays will allow us to screen this atlas and identify specific mutations that alter the pathogenicity of MLVs, thus facilitating the design of next-generation vaccines.

In summary, we constructed the largest possible dataset to reconstruct sublineage 8.7 spatial dynamics, assessed the implication of its associated ecological, demographic and geographic variables as well as swine-farming practices. We also provided evidence regarding the potential leaky status of HP-PRRS MLVs. As the NADC30-like and NADC34-like lineages within PPRSV-2 propagate and evolve into predominant strains within the global epidemic, there is a crucial need to extend research to these novel lineages to prevent pandemic like HP-PRRSV. Our study potentially provides crucial insights and reference for future research in these novel lineages.

Methods

Dataset generation

Our national surveillance project on PRRSV focused on suspected PRRSV-positive farms in China and included nearly all provinces with pig-production activities. Over 6000 clinical specimens (i.e., whole blood, spleens, and lungs) were collected from 2005 to 2022 to sequence the ORF5 of PRRSV. Specimens were ground by a freezing grinder (JXFSTPRP-CLN-48, Shanghai Jingxin Industrial Development Co., Ltd., China) and the viral genomes were extracted by RNA fast200 kit (Fastagen, Shanghai, China) following the instructions of the manufacturer. Collectively, 242 ORF5 sequences and 42 complete genomes belonging to sublineage 8.7 were obtained from mainland China. Furthermore, we downloaded all *Betaarterivirus suid 2* ORF5 and complete genome sequences until 2022 (specifically up to March 2022) from the GenBank database. The ORF5 dataset was filtered to exclude sequences that: (1) didn't include a collection date or location data (location data were included when available at the provincial level for Chinese isolates), (2) vaccine strains, (3) unverified sequences, (4) the virus was serially passaged in cells, (5) ambiguous and deleted residues. The resulting database of ORF5 sequences was aligned using the MAFFTv7 algorithm, manually truncating all the nucleotide sites except those in ORF5 in MEGA7 software[31,32]. For the complete-genomes database, we used MAFFT to align the sequences and then removed ambiguous regions using the TrimAL 1.4 algorithm[33].

Subsequently, multiple rounds of maximum-likelihood analyses were run in order to screen the final lineage 8 database using IQ-TREE 2[34]. Briefly, contextual reference sequence of each lineages were combined with our global ORF5 datasets to reconstruct the ML phylogeny (L1: NADC30, L2: XW008, L3: GM2, L4: EDRD-1, L5: VR-2332, L6: P129, L7: SP, L8: CH-1a, HP-PRRSV: JXA1, L9: MN30100). Then the lineage 8 cluster was selected from this tree. In total, 3708 ORF5 sequences belonging to lineage 8 were selected. Furthermore, 2340 ORF5 sequences (including 242 sequences from our lab) were identified as sublineage 8.7 and extracted from the global lineage 8 phylogeny to generate the final sublineage 8.7 ORF5 database. Of note, as the phylogeographic analysis of sublineage 8.7 is our main interest, we included the geographical information of viruses in China at the

provincial level. Using this approach, 341 complete-genome sequences were identified as sublineage 8.7 and extracted from the global lineage 8 phylogeny to generate the final sublineage 8.7 complete-genome database.

Phylogenetic, phylogeographic, and phylodynamic analysis

Nextstrain workflow. To reconstruct the evolutionary relationship of all the sequences in our dataset, we first utilized RDP4 to detect recombination events in our dataset. Multiple detection methods were tested including RDP[35], Chimaera[36], 3SEQ[37], GENECONV[38], and MaxChi[39]. Furthermore, BootScan[40] and SiScan[41] were employed as secondary detection with a highest acceptable *p*-value threshold of 0.05. Other parameters were left at their default setting. A sequence was excluded when three or more methods identified it as recombinant[42].

Following the Nextstrain pipeline, we performed a maximum likelihood analyses to infer the ancestral nodes, the phylogeny and the dispersal history of sublineage 8.7, using the built-in python framework TreeTime[43]. The goal is to estimate the time and involved locations whenever a transmission event took place. To be more specific, sample node colors indicate the ancestral state (in this case, the location) and shifts are drawn as links between demes on the map[44]. We first employed the *align* augur command[31] to match sequences into a qualified layout. Next, we employed the tree building augur command with built-in algorithm IQ-TREE 2[45] with a general time-reversible (GTR) model to build the preliminary maximum-likelihood tree without any time and ancestral node annotation, which was determined through the automated model selection procedure (i.e., ModelFinder) in IQ-TREE 2. This "raw" tree was then used as input for TreeTime via augur to infer a time-resolved phylogenetic tree[43]. Then, we matched all the location-data to the raw tree via a call to augur. Finally, we employed TreeTime again to jointly estimate the phylogeny and the ancestral locations at all of its internal nodes.

Subsampling strategy. The magnitude of the sublineage 8.7 dataset computationally prohibits performing a fully Bayesian phylodynamic analysis. Hence, we deployed a subsampling strategy that was previously used in an HIV study to enable robust phylogeographic analyses[46]. The subsampling consisted of removing sequences such that monophyletic clusters that entirely consist of samples from the same geographical units are represented by a single sample. This is justified by the province-level geographic resolution of our data. Monophyletic clusters consisting solely of sequences from the same province do not bring any additional information on the between-province diffusion process we aim to infer. We discarded all but at least one randomly selected sequence per monophyletic cluster to provide a systematic way to reduce the initial dataset. In practice, this was done by first estimating a maximum-likelihood tree using FastTree 2.1[47] and removing outlier sequences by performing a root-to-tip regression using TempEst[42]. We then constructed a new tree without the outliers by parsing the tree using the 'ape' R package to identify the state-specific clusters and removing the redundant sequences from their corresponding clades[48]. Since the main objective of this step was to reduce the number of sequences, we did not take into account branch support values when selecting the clusters from which we subsampled. This also allowed us to avoid setting arbitrary threshold values in the clustering step. The resulting dataset consists of 1371 sequences from the original database of sublineage 8.7 (1782 sequences), which made a fully Bayesian phylodynamic inference approach possible (Supplementary Data 1-2).

Bayesian discrete trait phylogeographic GLM analysis. Our next goal was to run a generalized linear model (GLM) extension of the discrete phylogeographic model to determine which factors were associated with viral spread between locations[49]. For this, we considered all possible explanatory predictors that were collected by the Chinese Bureau of Statistics and the Chinese Center for Animal Disease Control and Prevention. Predictors included climatological, ecological, physical (e.g., altitude), and anthropogenic factors (e.g., gross population). For non-pairwise predictor, we collected a total of 17 province-specific potential covariates for the spatial diffusion process of PRRSV (Supplementary Data 3). In addition to non-pairwise predictors, we accounted for the distance between pairs of provincial centroids as a predictor of geographic distance as a pairwise predictor (Supplementary Data 4)[49]. For all non-pairwise predictors, a separate origin and destination predictor was included. This brought the initial total number of predictors to 35. For a detailed description of each of these predictors we refer to the Supplementary information. Often, these types of analyses include a sample size predictor as a sanity check against sampling bias. We avoided the inclusion of this sample size predictor given that the number of samples present for each province is highly correlated to the incidence of lineage 8 in each location ($r = 0.95$). As such, we considered the sampling to be representative of the underlying HP-PRRSV circulating in the country. In addition, we performed a linear regression between the number of cases and the sequences included and performed a Spearman cross-correlation check between each predictor and the residuals (Supplementary Figs. S8–S9) and included the residuals of this linear regression as a predictor in our GLM (Supplementary Fig. S10). Analysis showed that no predictor was significantly correlated with the regression residuals and as such assured that sampling bias would not be a concern in our phylogeographic reconstruction (Supplementary Fig. S9). Further, this analysis revealed the presence of highly correlated covariates (Supplementary Fig. S2–S3). As a next step, we systematically removed covariates so that the pairwise Pearson correlation coefficients between all predictors were <0.80. This brought the final number of covariates considered in our model to 24. For a detailed explanation of this step, we refer to the "Correlation analysis of GLM covariates" section of the Supplementary information. A preliminary analysis using this GLM setup showed an overwhelming "out of Guangdong" source-sink dynamic which, coupled with the fact that many of the extreme values in the covariates come from Guangdong province, led us to include a binary "from Guangdong" predictor to more reliably ascertain the independent contribution of the remaining predictors in our analysis. Such an inference has been deployed in a previous analysis to assess the effect of London as a transmission hub in the spread of SARS-CoV-2 in the United Kingdom[18].

In order to decrease computation time and to deal with the large number of locations in our dataset, we split the estimation of the phylogeny and the dispersal history into two separate analyses. First, we performed a purely phylogenetic analysis without geographical information to obtain an empirical distribution of 1000 time-calibrated phylogenies. We subsequently conditioned on this distribution and performed a discrete trait phylogeographical reconstruction under the GLM formulation to reconstruct the geographic spread of the virus and identify drivers of spread. We performed both analyses using BEAST v1.10 using the BEAGLE library v4 to improve computational performance[50,51].

We generated the empirical tree distribution using the following model specifications: a HKY + Γ_4 substitution model[52,53], a skygrid coalescent prior[54], and an uncorrelated relaxed clock with an underlying lognormal distribution for rate heterogeneity[55]. Furthermore, we made use of a Hamiltonian Monte Carlo transition kernel to achieve efficient sampling of the skygrid model parameters[56], and inferred a preliminary time-calibrated phylogeny using IQ-TREE 2[45] + TreeTime[43] as a starting tree to minimize burn-in. The Markov chain Monte Carlo analysis was run for 10 ^ 9 iterations and convergence and mixing of all parameters were assessed using Tracer v1.7[57].

To reconstruct the process of spatial dispersal, we modeled the transition rates between (discrete) locations through a continuous-time Markov chain (CTMC) approach[49]. The GLM parameterization we

used models the log-transformed transition rates as a log-linear function of the previously mentioned predictors and is able to estimate the effect sizes of each covariate along with their inclusion probability through the use of a spike-and-slab prior[43]. We further generate realizations of this CTMC to estimate the number of Markov jumps between locations[58]. We ran this conditioned phylogeographical analysis for 10^7 iterations and assessed convergence and mixing as previously described. Posterior summarization of the trees was done using TreeAnnotator[50].

Recombination analysis

By merging the complete-genome dataset of sublineage 8.7 and the reference sequences of each lineage (Supplementary Data 5), we constructed a complete-genome dataset that allows us to evaluate the recombination history of lineage 8 (including inter- and intra- recombination). We characterized the recombination history of interlineage and intralineage 8 PRRSV following two independent approaches. Firstly, we assessed the overview of interlineage and intralineage recombination events of total lineage in SplitsTree5. We visualized the splits with the EqualAngle method using 1000 bootstrap replicates. The remaining parameters were kept at default[59]. Secondly, we calculated the frequency of recombination regions to understand the recombination heterogeneity through time. For interlineage recombination characterization, we used RDP4 to detect the recombination events in our dataset with the different lineage reference strains respectively using the methods described before. As for intralineage recombination, we incorporated all the sublineage 8.7 strains to do a full exploratory recombination scan using the same methods. Each event was further examined using Simplot v3.5.1 as a robustness check. To diminish the off-spring spread of a single recombination event, we perform a deduplication of each recombination event by selecting a unique breakpoint. The events with repeated breakpoints were excluded for reducing repetition. Detailed information of inter- and intra-recombination events was curated in Supplementary Data 6–7.

Analysis of the relationship of vaccine strains and field isolates

Regarding the HP-PRRSV vaccines used in China, there have been four legally approved vaccines, including JXA1-R, TJM-F92, HuN4-F112, and GDr180 since the emergence of HP-PRRSV in 2006. We have procured a copy of each vaccine and obtained the full genomes using metatranscriptome as in a previous project[25]. We constructed four libraries of vaccine sequences then sequenced on the MGISEQ-200 RS sequencer platform with a pair-end length of 150 bp. Then we trimmed the adapter of short reads by Trimmomatic[60] and removed all low-quality reads (QC < 20). The refined reads were then assembled by MEGAHIT[61]. The assembled contigs were mapped with their database number using DIAMOND. Samtools[62] and iVar[63] were finally run to obtain consensus sequences with a criteria of sequencing depth >100 and a minimum threshold 10 times, or to be written with N.

Together with the lineage 8 complete genome database, we have aligned the sequences using MAFFT7, then constructed the maximum likelihood tree using IQ-TREE 2 based on the best-fit nucleotide substitution model GTR + F + I + Γ_4 (according to the Bayesian Information Criterion and 1000 bootstrap replicates). We subsequently used ClusterPicker[64] (bootstrap threshold: 85%, genetic similarity threshold: 97%) to select four clusters related to each MLV vaccine (Supplementary Data 8–10). For each selected cluster, we further estimated haplotype using RNA-dependent RNA polymerase (RdRp) to reflect its homogeneous relationship. After our homogeneous estimation, we further questioned if these clinical sequences show clinical pathogenicity. To prove its clinical impact, we accordingly curated a new table (i.e., Supplementary Data 11) to present the detailed clinical pathogenicity data of each sequence (case), which suggests that most of the clinical cases were proven to be highly pathogenic in the clinic by publication retrieval. We analyzed the concurrent amino acid mutation motifs existing in each clinical sequence of a single cluster, however distinct from parental strains, aiming to characterize the specific molecular marker for clinical vaccine-homogeneous strains using R v4.1.3 (ggtree and ggmsa packages)[65]. Specifically, concurrent amino acid mutation sites were defined following the criteria: over half of the clinical sequences showed identical mutation with vaccine strain but distinct from vaccine-derived parental strain. For example: in TJM-92 cluster, the clinical strains and vaccine strain (TJM-92) shared identical mutation of H257Y in ORF1a against TJM strain.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The ORF5 sequences, complete genomes, and four vaccine sequences generated in this study have been deposited in the GenBank with the accession number of PQ271638-PQ271879 (ORF5 sequences), PQ325306-PQ325309 (vaccine sequences), and PQ325310-PQ325355 (whole genome sequences). Source Data for ecological covariates, recombination breakpoints, and subsampling have been provided in Supplementary Data. Source data are provided with this paper.

Code availability

BEAST XML files and log files used for this work are publicly available on GitHub: https://github.com/jobsxing1228/lineage8.

References

1. Lunney, J. K. et al. Porcine reproductive and respiratory syndrome virus (PRRSV): pathogenesis and interaction with the immune system. *Annu Rev. Anim. Biosci.* **4**, 129–154 (2016).
2. Walker, P. J. et al. Changes to virus taxonomy and to the international code of virus classification and nomenclature ratified by the international committee on taxonomy of viruses (2021). *Arch. Virol.* **166**, 2633–2648 (2021).
3. Shi, M. et al. Phylogeny-based evolutionary, demographical, and geographical dissection of North American type 2 porcine reproductive and respiratory syndrome viruses. *J. Virol.* **84**, 8700–8711 (2010).
4. Sun, Y. K. et al. Insights into the evolutionary history and epidemiological characteristics of the emerging lineage 1 porcine reproductive and respiratory syndrome viruses in China. *Transbound. Emerg. Dis.* **67**, 2630–2641 (2020).
5. Sun, Y. K. et al. Phylogeography, phylodynamics and the recent outbreak of lineage 3 porcine reproductive and respiratory syndrome viruses in China. *Transbound. Emerg. Dis.* **66**, 2152–2162 (2019).
6. Key, K. F. et al. Genetic variation and phylogenetic analyses of the ORF5 gene of acute porcine reproductive and respiratory syndrome virus isolates. *Vet. Microbiol.* **83**, 249–263 (2001).
7. Tian, K. et al. Emergence of fatal PRRSV variants: unparalleled outbreaks of atypical PRRS in China and molecular dissection of the unique hallmark. *PLoS ONE* **2**, e526 (2007).
8. Shi, M. et al. Recombination is associated with an outbreak of novel highly pathogenic porcine reproductive and respiratory syndrome viruses in China. *J. Virol.* **87**, 10904–10907 (2013).
9. Leng, X. et al. Evaluation of the efficacy of an attenuated live vaccine against highly pathogenic porcine reproductive and respiratory syndrome virus in young pigs. *Clin. Vaccin. Immunol.* **19**, 1199–1206 (2012).
10. Tian, Z. J. et al. An attenuated live vaccine based on highly pathogenic porcine reproductive and respiratory syndrome virus (HP-PRRSV) protects piglets against HP-PRRS. *Vet. Microbiol.* **138**, 34–40 (2009).
11. Han, W. et al. Molecular mutations associated with the in vitro passage of virulent porcine reproductive and respiratory syndrome virus. *Virus Genes* **38**, 276–284 (2009).

12. Li, B. et al. Recombination in vaccine and circulating strains of porcine reproductive and respiratory syndrome viruses. *Emerg. Infect. Dis.* **15**, 2032–2035 (2009).
13. Lu, W. H. et al. Re-emerging of porcine respiratory and reproductive syndrome virus (lineage 3) and increased pathogenicity after genomic recombination with vaccine variant. *Vet. Microbiol.* **175**, 332–340 (2015).
14. Zhao H. et al. 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. *Virol. J.* **14**, 67 (2017).
15. Jiang, Y. F. et al. Characterization of three porcine reproductive and respiratory syndrome virus isolates from a single swine farm bearing strong homology to a vaccine strain. *Vet. Microbiol.* **179**, 242–249 (2015).
16. Liu, P. et al. High reversion potential of a cell-adapted vaccine candidate against highly pathogenic porcine reproductive and respiratory syndrome. *Vet. Microbiol.* **227**, 133–142 (2018).
17. Wang, J. et al. Attenuated porcine reproductive and respiratory syndrome virus regains its fatal virulence by serial passaging in pigs or porcine alveolar macrophages to increase its adaptation to target cells. *Microbiol. Spectr.* **10**, e0308422 (2022).
18. Tsui, J. L. et al. Genomic assessment of invasion dynamics of SARS-CoV-2 Omicron BA.1. *Science* **381**, 336–343 (2023).
19. Dudas, G. et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* **544**, 309–315 (2017).
20. Makau, D. N. et al. Integrating animal movements with phylogeography to model the spread of PRRSV in the USA. *Virus Evol.* **7**, veab060 (2021).
21. He W. T., et al. Phylogeography reveals sociation between swine trade and the spread of Porcine Epidemic Diarrhea Virus in China and across the World. *Mol Biol Evol.* **39**, msab364 (2022).
22. Yu F., et al. Phylogenetics, genomic recombination, and NSP2 polymorphic patterns of porcine reproductive and respiratory syndrome virus in China and the United States in 2014–2018. *J Virol.* **28**, 94 (2020).
23. Han, J. et al. Pathogenesis and control of the Chinese highly pathogenic porcine reproductive and respiratory syndrome virus. *Vet. Microbiol.* **209**, 30–47 (2017).
24. Zhou L., Ge X., Yang H. Porcine reproductive and respiratory syndrome modified live virus vaccine: a "leaky" vaccine with debatable efficacy and safety. *Vaccines (Basel).* **9**, 9 (2021).
25. Xing J. B., et al. Whole genome sequencing of clinical specimens reveals the genomic diversity of porcine reproductive and respiratory syndrome viruses emerging in China. *Transbound. Emerg. Dis.* **69**, e2530-e2540 (2022).
26. Sun, Y. K. et al. Emergence of novel recombination lineage 3 of porcine reproductive and respiratory syndrome viruses in Southern China. *Transbound. Emerg. Dis.* **66**, 578–587 (2019).
27. De Maio, N. et al. New routes to phylogeography: a bayesian structured coalescent approximation. *PLoS Genet* **11**, e1005421 (2015).
28. Muller, N. F., Rasmussen, D. A. & Stadler, T. The structured coalescent and its approximations. *Mol. Biol. Evol.* **34**, 2970–2981 (2017).
29. Layan, M. et al. Impact and mitigation of sampling bias to determine viral spread: evaluating discrete phylogeography through CTMC modeling and structured coalescent model approximations. *Virus Evol.* **9**, vead010 (2023).
30. Martin, D. P. et al. The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell* **184**, 5189–5200.e7 (2021).
31. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
32. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).
33. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
34. Nguyen, L. T. et al. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
35. Martin, D. & Rybicki, E. RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16**, 562–563 (2000).
36. Arenas, M. & Posada, D. The effect of recombination on the reconstruction of ancestral sequences. *Genetics* **184**, 1133–1139 (2010).
37. Boni, M. F., Posada, D. & Feldman, M. W. An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics* **176**, 1035–1047 (2007).
38. Padidam, M., Sawyer, S. & Fauquet, C. M. Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**, 218–225 (1999).
39. Wiuf, C., Christensen, T. & Hein, J. A simulation study of the reliability of recombination detection methods. *Mol. Biol. Evol.* **18**, 1929–1939 (2001).
40. Salminen, M. O. et al. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS Res Hum. Retrovirus.***11**, 1423–1425 (1995).
41. Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16**, 573–582 (2000).
42. Rambaut, A. et al. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016).
43. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018).
44. Hadfield, J. et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinforma. (Oxf., Engl.)* **34**, 4121–4123 (2018).
45. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
46. Hong S. L., et al. In search of covariates of HIV-1 subtype B spread in the United States-a cautionary tale of large-scale Bayesian phylogeography. *Viruses*. **5**, 12 (2020).
47. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
48. Paradis E., Schliep K. J. B. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. **35**, 526–528 (2019).
49. Lemey, P. et al. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS Pathog.* **10**, e1003932 (2014).
50. Suchard, M. A. et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).
51. Ayres, D. L. et al. BEAGLE 3: Improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. *Syst. Biol.* **68**, 1052–1061 (2019).
52. Yang, Z. A space-time process model for the evolution of DNA sequences. *Genetics* **139**, 993–1005 (1995).
53. Tavaré SJLoMfLS. Some probabilistic and statistical problems on the analysis of DNA sequence. *Life Sci.* **17**, 57 (1986).
54. Gill, M. S. et al. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.* **30**, 713–724 (2013).
55. Drummond, A. J. et al. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).
56. Baele G. et al. Hamiltonian Monte Carlo sampling to estimate past population dynamics using the skygrid coalescent model in a Bayesian phylogenetics framework. *Wellcome Open Res.* **5**, 53 (2020).
57. Rambaut, A. et al. Posterior summarization in Bayesian phylogenetics using tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).

58. Minin, V. N. & Suchard, M. A. Fast, accurate and simulation-free stochastic mapping. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 3985–3995 (2008).
59. Huson, D. H. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68–73 (1998).
60. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
61. Li, D. et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
62. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
63. Grubaugh, N. D. et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 8 (2019).
64. Lycett S., Hodcroft E., Ragonnet-Cronin M. *Cluster Picker and Cluster Matcher: A Tool Combination for the Phylogenetic Analysis of Clusters of Nucleotide Sequences (in java)*. https://www.research.ed.ac.uk (2013).
65. Yu, G. et al. Two methods for mapping and visualizing associated data on phylogeny using Ggtree. *Mol. Biol. Evol.* **35**, 3041–3043 (2018).

Acknowledgements

Y.S., J.X., S.X., Y.L., H.W., and G.Z. acknowledge support from the National Natural Science Foundation of China [grant number 32102704], Key-Area Research and Development Program of Guangdong Province [grant number 2019B020211003] and China Agriculture Research System of MOF and MARA. GB and SLH acknowledge support from the Research Foundation—Flanders ("Fonds voor Wetenschappelijk Onderzoek—Vlaanderen," G0E1420N) and from the DURABLE EU4Health project 02/2023-01/2027, which is co-funded by the European Union (call EU4H-2021-PJ4) under Grant Agreement No. 101102733. Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. GB and NB acknowledge support from the Research Foundation—Flanders ("Fonds voor Wetenschappelijk Onderzoek—Vlaanderen," G098321N). We would like to gratefully thank all the researchers and laboratories for their generous genomic uploaded data in NCBI that we have used in this study. Furthermore, we would thank Xiaoqin Xu, Yuli Luo, and Qian Kuang for the large-scale sampling and routine monitoring relying on national surveillance of PRRSV in China.

Author contributions

G.Z. Y.S. and G.B. conceived the research. Y.S. H.W. and J.X. generated sequence data along with substantial help from S.X. Y.L. J.Z. X.G. D.Z. J.L. and G.L. J.X. G.Z. drafted the manuscript with the substantial help of S.L.H. N.B. and G.B. J.X., Y.S. S.L.H., and G.B. performed data analyses along with help from L.Z. M.S. and T.A. Y.S. H.W. G.B. and G.Z. supervised this work. Additionally, H.W. G.B. and G.Z. contributed equally to this study. All authors approved the final version of the manuscript and accept responsibility for the data therein.

Competing interests

The authors declare no competing interests.

Ethics statement

Our sampling procedures were approved by the Animal Ethics Committee of South China Agricultural University and conducted under the guidance of the South China Agricultural University Institutional Animal Care and Use Committee (SCAU-AEC-2022A010).

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-53076-w.

Correspondence and requests for materials should be addressed to Heng Wang, Guy Baele or Guihong Zhang.

Peer review information *Nature Communications* thanks Kimberly VanderWaal, and the other, anonymous, reviewer for their contribution to the peer review of this work. A peer review file is available.