



VILNIUS UNIVERSITY
LIFE SCIENCES CENTER

Aistė Židonytė

**Phylogeographic Analysis of Wuhan Mosquito Virus 6 and
Other RNA Viruses: Environmental and Host Correlates of
Spatial Diffusion**

Master's Thesis

Genetics study program

Thesis supervisor

Dr. Gytis Dudas

Thesis completed at

VU LSC Institute of Biotechnology

Vilnius 2026

Contents

Abbreviations	4
Introduction	5
1. Literature Review	7
1.1. RNA virus diversity, emergence and relevance in public health	7
1.1.1. <i>Orthomyxoviridae</i> virus family	7
1.2. Differences in transmission dynamics of arthropod-specific and dual-host RNA viruses	8
1.2.1. Arthropod-specific viruses	9
1.2.2. Arthropod-vertebrate infecting viruses	10
1.2.3. West Nile virus as a model for arbovirus spread analysis	11
1.3. Molecular evolution in epidemiology	12
1.3.1. Genomic epidemiology	12
1.3.2. Molecular clock – concepts and applications	13
1.3.3. Complications in molecular clock calculations	14
1.3.4. Molecular clock models in Bayesian modeling for epidemiology	15
1.4. Phylogeography – concepts and uses	16
1.4.1. Continuous phylogeography	18
1.4.2. Discrete phylogeographic analysis	18
2. Materials and Methods	20
2.1. Materials	20
2.1.1. Research subjects	20
2.1.2. Genomic data preparation	20
2.1.3. Raster dataset preparation for WuMV-6 continuous phylogeography in China	21
2.1.4. Predictor dataset preparation for discrete phylogeography	22
2.2. Methods	24
2.2.1. Continuous phylogeography of WuMV-6 in China	24
2.2.2. Impact of environmental factors on WuMV-6 spread in China	24
2.2.3. Discrete phylogeography of WuMV-6 and other RNA viruses	25
2.2.4. Evolutionary pressure analysis	27
2.2.5. Viral genome assembly for mixed lineage in pooled samples	28
2.2.6. Code source	28
3. Results	29
3.1. Reconstruction of WuMV-6 diffusion (continuous phylogeography)	29
3.2. Discrete phylogeography and environmental correlates of spatial diffusion of WuMV-6	30
3.3. Environmental and host correlates of spatial diffusion of other RNA viruses	33
3.4. Selective pressure on RNA virus proteins	34
3.5. Mixed viral lineage genome assembly and Shannon entropy calculation	35
4. Discussion	39
4.1. Rapid WuMV-6 migration	39
4.2. Correlates of WuMV-6 spatial spread	40
4.3. Selective pressure of WuMV-6 and other RNA virus' proteins	42

4.4. High Shannon entropy suggests mixed viral lineages in multiple WuMV-6 samples	44
4.5. Future directions	44
Conclusions	46
Acknowledgements	47
Author's Personal Contribution	48
Santrauka	49
Abstract	50
References	51
Appendices	67
Appendix A. Number of sequences per region for each virus analysis	68
Appendix B. Discrete phylogeography predictor matrices	69
Appendix C. WuMV-6 NP segment predictor colinearity	72
Appendix D. Discrete phylogeography GLM predictor statistics	74
Appendix E. WuMV-6 NP segment transitions between regions	76

Abbreviations

BF	–	Bayes Factor
CHIKV	–	Chikungunya Virus
CI	–	Credibility Interval
CxFV	–	<i>Culex</i> Flavivirus
DENV	–	Dengue Virus
dN/dS	–	Non-synonymous substitutions per non-synonymous site normalised by synonymous substitutions per synonymous site
ESS	–	Effective Sample Size
GLM	–	Generalized Linear Model
HKY	–	Hasegawa-Kishino-Yano substitution model
HPD	–	Highest Posterior Density
IAV	–	Influenza A Virus
ISV	–	Insect-Specific Virus
MCC	–	Maximum Clade Credibility
MCMC	–	Markov chain Monte Carlo
MRCA	–	Most Recent Common Ancestor
RdRp	–	RNA-dependent RNA polymerase
SIGV	–	Sigma Virus
UCLD	–	Uncorrelated Log-normal Relaxed Molecular Clock
USUV	–	Usutu Virus
WFBV	–	Wellfleet Bay Virus
WuMV-6	–	Wuhan Mosquito Virus 6
WNV	–	West Nile Virus
ZIKV	–	Zika Virus

Protein abbreviations:

IAV proteins: HA, NA, M, NEP, NP, PA, PB1, PB2	–	Hemagglutinin, Neuraminidase, Matrix, Nuclear Export protein, Nucleoprotein, Polymerase Acidic subunit, Polymerase Basic subunits (1 and 2)
WNV, CxFV proteins: C, E, M, NS	–	Capsid, Envelope, Membrane, Non-structural
WuMV-6 proteins: gp64, NP, PA, PB1, PB2	–	Glycoprotein, Nucleoprotein, Polymerase Acidic subunit, Polymerase Basic subunits (1 and 2)

Introduction

Wuhan Mosquito Virus 6 (WuMV-6) is a member of *Orthomyxoviridae* family and belongs to the genus *Quarantavirus*. Its genome consists of 8 RNA segments (Dudas and Batson, 2023). It was first identified through arthropod RNA sequencing in China more than 10 years ago, but has been detected during metatranscriptomic studies of mosquitoes on various continents since (Figure 1) (Li et al., 2015; Batson et al., 2021; Wang et al., 2021).

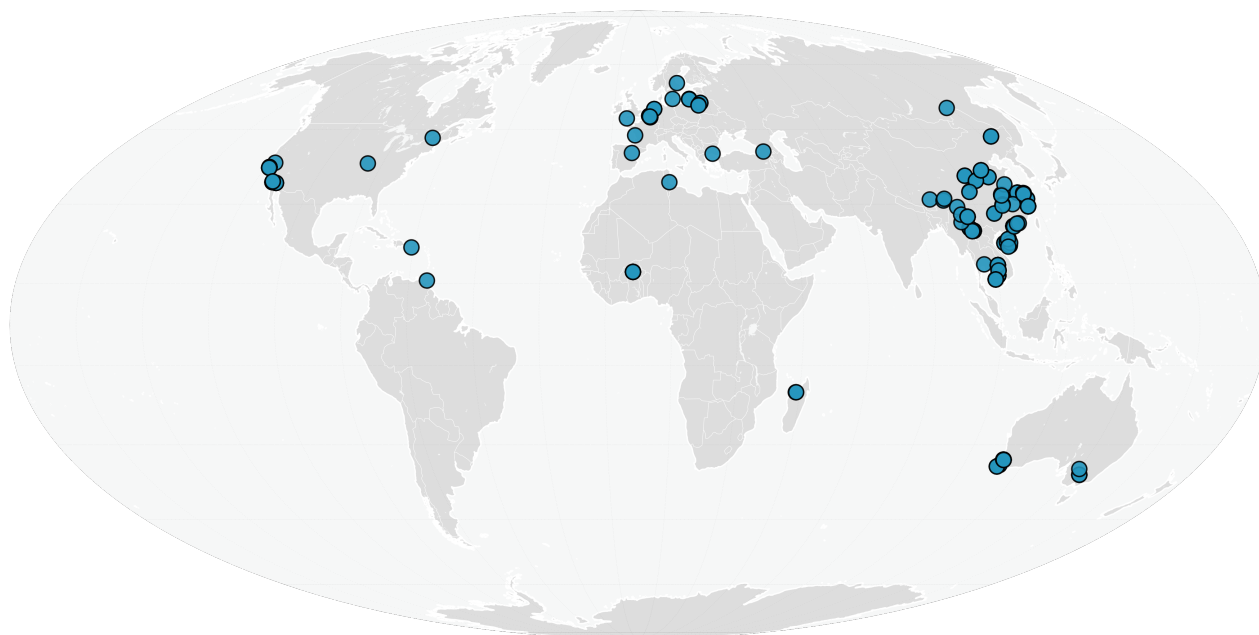


Figure 1. Locations where WuMV-6 was identified from metagenomic studies or PCR testing.

Unlike vertebrate-infecting orthomyxoviruses (influenza and isaviruses), which use class I membrane fusion proteins like haemagglutinin (HA) and related proteins, WuMV-6 and other arthropod-infecting orthomyxovirus genera (thogoto- and quarantaviruses) have a class III glycoprotein called gp64 (Garry and Garry, 2008). Within WuMV-6 populations this surface protein was found to evolve faster at the amino acid level compared to the rest of WuMV-6's proteome (Dudas and Batson, 2023). The highest values of dN/dS (ratio of non-synonymous substitutions per non-synonymous site and synonymous substitutions per synonymous site) concentrate around the surface protein's fusion loops, which directly interact with the host cell's surface (Podbilewicz, 2014; Dudas and Batson, 2023). Such rapid evolution of a surface protein resembles evolutionary pressure from vertebrate hosts with humoral immune systems (Dudas and Batson, 2023). Since mosquitoes do not have adaptive immune systems, the selective pressure they could theoretically exert is virtually undetectable compared to vertebrates (Woelk and Holmes, 2002; Obbard and Dudas, 2014; Kumar et al., 2018).

Rapid migration of WuMV-6 and fast surface protein evolution suggest the presence of a vertebrate host (Dudas and Batson, 2023). As related viruses have been discovered during outbreaks of disease in birds, which participate in periodic long distance migrations, and because of *Culex* mosquito preference for avian bloodmeals, we speculate that the vertebrate host of

WuMV-6 could be a bird (Takken and Verhulst, 2013).

Therefore, because the availability of enough genomic data from diverse geographic regions enables reliable molecular clock analyses, and a measurable temporal signal is present in the data, we can perform phylogeographic analyses to assess its' spread and try to identify the potential host of WuMV-6.

Hence, the aim of this thesis is to evaluate environmental variables that could be related to its spread, and assess if other RNA viruses with different (and known) hosts share the same correlates of spatial spread.

Thesis objectives:

1. Reconstruct the transmission pathways of WuMV-6 in China using continuous phylogeography methods.
2. Assess what environmental variables have an effect on the migration rates of WuMV-6 using discrete phylogeography methods.
3. Analyze other RNA viruses applying the same discrete phylogeography methods to compare if host differences play a role in migration patterns.
4. Compare selective pressure between surface and inner proteins of WuMV-6 and other RNA viruses.
5. Re-assemble WuMV-6 genomes accounting for genetic variability and the potential presence of mixed viral lineages within samples.

1. Literature Review

1.1. RNA virus diversity, emergence and relevance in public health

RNA viruses are found in abundance and great diversity in all types of ecosystems. They cause a wide range of human illnesses: from mild to severe respiratory infections, hemorrhagic fevers or encephalitis, and immunodeficiency syndrome to name a few (Bray, 2008; Zhang et al., 2018). A lot of pandemics and epidemics in the past two decades were caused by RNA viruses: SARS-CoV-2, influenza A virus, MERS-CoV, Ebola virus (Smith et al., 2009; Gire et al., 2014; Mohd et al., 2016; Batson et al., 2021; WHO, 2025). All these viruses emerged from zoonotic spillovers to the human population, therefore causing great challenges to public health, and placing RNA viruses as one of the most prominent emerging human pathogens (Woolhouse and Gowtage-Sequeria, 2005). Due to their vast diversity and tendency towards spillovers between species, RNA viruses also cause economic damage via infections of other species, including cattle, swine, poultry, and crops (Bray, 2008; Sastry and Zitter, 2014).

Despite advances in sequencing technology over the last decade that have decreased costs and increased throughput, novel viruses continue to be discovered every year (Rosenberg, 2015; Bassi et al., 2022). Recent efforts studying not only organisms with direct importance to humans that are already infected, but also a range of samples from seemingly healthy wildlife has increased understanding of virus host ranges, their distribution and diversity (Zhang et al., 2018). Just one study of aquatic environment samples from a harbour in China doubled the amount of RNA viruses known beforehand (Wolf et al., 2020).

1.1.1. *Orthomyxoviridae* virus family

The family of *Orthomyxoviridae* includes seven genera, all of which are enveloped RNA viruses with negative sense single-stranded segmented genomes: *Alphainfluenzavirus*, *Betainfluenzavirus*, *Deltainfluenzavirus*, *Gammainfluenzavirus*, *Isavirus*, *Thogotovirus*, and *Quarantavirus* (Ma, 2022). They include viruses causing familiar diseases: Influenza A, B, and C (flu) viruses are known to infect humans and other vertebrates, Influenza D virus – primarily infects pigs and cattle, with no human infections reported yet (Liang, 2023). Like Influenza viruses, Isaviruses are only found in vertebrates, namely – in fish, causing economic losses in Atlantic salmon farms (Rimstad and Markussen, 2020). Thogoto- and quarantaviruses are usually found in arthropods, but both have been found to cross the species barrier and infect humans and other vertebrates as well (Taylor et al., 1966; Kosoy et al., 2015).

Some other orthomyxoviruses were discovered because they caused disease in birds or other animals (Kessell et al., 2012; Hause et al., 2013; Allison et al., 2014). For example, Wellfleet Bay virus (WFBV), belonging to the same *Quarantavirus* genus as WuMV-6, has caused repeated mass die-offs in birds (Figure 2) (Allison et al., 2014). WFBV uses the same surface glycoprotein gp64 other arthropod-associated thogoto- and quarantaviruses (Allison et al., 2014). Due to its close relationship with tick-associated viruses, the recovery of its closest relative Quarantavirus

from ticks, and the presence of the same surface protein, WFBV is likely to be transmitted by ticks or other arthropods, similar to other members of this group (Allison et al., 2014; Cholleti et al., 2018).

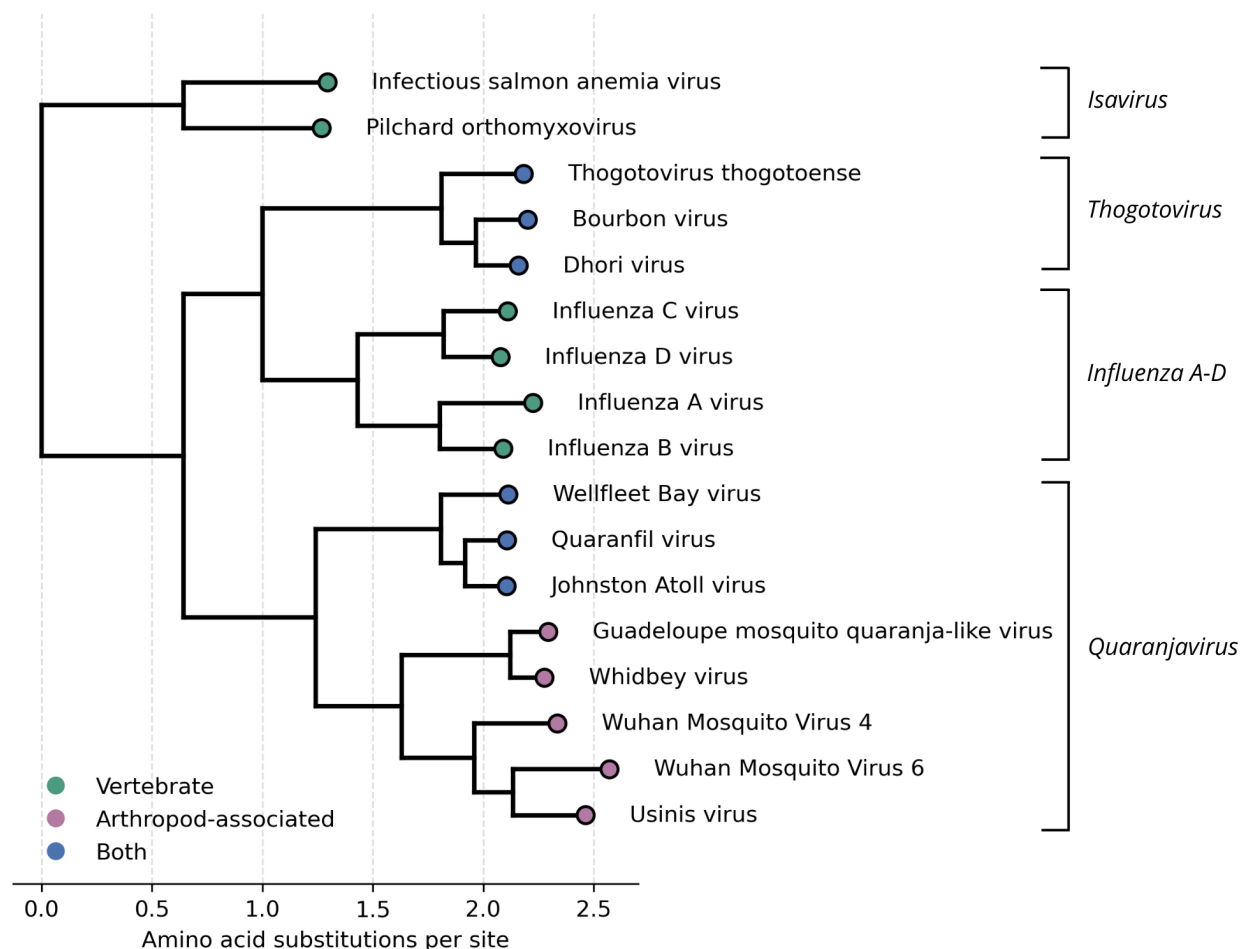


Figure 2. Maximum likelihood phylogenetic tree of PB1 segment of *Orthomyxoviridae* family viruses.

1.2. Differences in transmission dynamics of arthropod-specific and dual-host RNA viruses

RNA viruses vary in their rates of emergence, evolution, and spatial spread. These differences are influenced by multiple factors, including host range, host immune responses, the rate at which viruses invade new host populations or environments, virus replication speed, environmental determinants, and anthropogenic factors (Morley et al., 2015; Fitzsimmons et al., 2018; Moury and Desbiez, 2020; Alvarez-Munoz et al., 2021; Bisht and te Velthuis, 2022). In addition, viral transmission mechanisms also heavily influence the dynamics of their spread. Viruses can be transmitted through direct contact between infected hosts, vertical transmission to offspring, or indirectly via contaminated surfaces, or aerosols (Leung, 2021).

Since host range determines the species that are involved in the maintenance of the virus population, they are expected to impose distinct ecological and evolutionary constraints (Weaver

and Barrett, 2004). However, direct comparative studies of spatial transmission dynamics between insect-specific viruses (ISV) and viruses that spread via dual-host transmission cycles are limited.

1.2.1. Arthropod-specific viruses

Although the first ISV was isolated in 1975, 50 years later ISV's are still regarded as a relatively new research area (Stollar and Thomas, 1975; Wilkman et al., 2025). Since their first discovery, the majority of ISVs have been discovered in mosquitoes (Öhlund et al., 2019). Because it was shown that ISVs can influence arbovirus transmission in mosquitoes, recent studies in multiple countries have focused on identifying the core mosquito virome (the same viruses present among individuals of the same species) and characterizing the overall diversity of ISVs (Hobson-Peters et al., 2013; Konstantinidis et al., 2022; Olmo et al., 2023; De Coninck et al., 2024; Gangopadhyaya et al., 2024; Kacnik et al., 2025; Gómez-Palacio et al., 2025).

The spread of ISVs is limited by the movement of their hosts. Spread between individual animals can occur via vertical transmission (the infected insect transmits the virus directly to its offspring), transovarian transmission (the offspring is infected at the egg stage within the ovary) or via feeding (Blitvich and Firth, 2015). However, for most ISVs, transmission pathways have not been experimentally confirmed and have only been described theoretically, based on available knowledge of related viruses (Hollingsworth et al., 2023).

One way of studying virus transmission dynamics is genomic epidemiology, which studies sequencing data to reconstruct phylogenetic relations and disease transmission (more in section 1.4) (Hill et al., 2021). Typically epidemiological and molecular analyses require sufficiently large and well-sampled datasets that would have temporal and phylogenetic signal (Rambaut et al., 2016). Not all ISVs are sequenced in sufficient numbers or for long enough to yield useful datasets, and if temporal signal is weak or absent, molecular clock models can result in uncertain estimates (Webster et al., 2015). Given that ISV sequencing remains relatively limited compared to arboviruses with human public health relevance, epidemiological studies of insect specific viruses are limited.

Studying ISVs is relevant as it can also be used to study the ecology of hosts, particularly in mosquitoes, where population structure and dynamics are difficult to observe directly (Hollingsworth et al., 2023). Mosquito population structure is strongly shaped by land cover and human transportation networks, and species-specific ISVs are therefore expected to mirror this structuring (Schmidt et al., 2017; Hollingsworth et al., 2023). Incorporating transmission models into phylodynamic frameworks allows estimation of effective population sizes and transmission rates. For maternally inherited ISVs, viral prevalence and effective population size are proportional to that of the host (Hollingsworth et al., 2023). Therefore, changes in viral effective population size are expected to reflect changes in host population size (Hollingsworth et al., 2023). However, these inferences depend on sufficient viral genetic variation and sampling intensity.

1.2.2. Arthropod-vertebrate infecting viruses

Arthropod-borne viruses or arboviruses are defined as viruses that can replicate in both arthropods and vertebrates, and can transmit infections between these hosts (Shope and Meegan, 1997). They are usually transmitted by hematophagous insects during feeding (Wilkman et al., 2025). Multiple arboviruses have caused severe public health challenges in humans: West Nile virus (WNV), yellow fever virus, chikungunya virus (CHIKV), dengue virus (DENV), etc. (Gubler, 2001).

From an evolutionary perspective, phylogenetic analyses have shown that some ISVs are closely related to arboviruses that infect humans, despite being unable to replicate in vertebrate cells (Carvalho and Long, 2021). This relationship suggests a shared evolutionary history, and it has been proposed that some arboviruses may have evolved from ISVs by acquiring the ability to infect vertebrate hosts, potentially through blood-feeding vectors (Li et al., 2015). Arboviruses occur in multiple RNA virus families with different genome organisations and replication strategies, including families: *Orthomyxoviridae*, *Togaviridae*, *Flaviviridae*, *Reoviridae*, *Rhabdoviridae* and the order *Bunyavirales*. This broad taxonomic distribution suggests that the ability to infect vertebrate hosts and transmission via arthropod vectors has evolved independently multiple times (Hanley and Weaver, 2008).

Arbovirus epidemiological dynamics are complex and difficult to analyse due to multiple host transmission cycles. Many arboviruses circulate among livestock, wildlife (including migrating birds), humans, and their arthropod vectors such as ticks and mosquitoes (Morse, 1995; Hill et al., 2026). Because of their dependence on vector survival and reproduction, arbovirus transmission dynamics are strongly affected by climatic factors including temperature, humidity, precipitation (Dorigatti et al., 2025). Vector habitat availability, e.g. water bodies for mosquito breeding or vertebrate hosts for feeding, are also necessary for arbovirus spread. Since vector host preferences can vary during different life stages (particularly in ticks), it further complicates the dynamics of arbovirus transmission (Slowik and Lane, 2009).

Because arboviruses are strongly influenced by environmental conditions, climate change is expanding the geographic range of their vectors such as mosquitoes, ticks, and biting flies. This is also accompanied by extended transmission periods – from a few months to whole years (Paz, 2015; Leal Filho et al., 2025). Even regions previously considered at low risk for arbovirus infections are now reporting cases of CHIKV, DENV, Zika (ZIKV) viruses, and these diseases are expected to expand further into previously non-endemic, northern regions in the upcoming decades (Marcondes et al., 2017; Laverdeur et al., 2024). In addition to altering vector distributions, higher temperatures can accelerate the extrinsic incubation periods (the time it takes from mosquito getting a virus and becoming able to infect others) of some arboviruses, like CHIKV, ZIKV, DENV, WNV (Winokur et al., 2020; Hartner et al., 2025).

Increased human mobility and international trade are significant drivers of arbovirus spread due to both host and vector movement. Vertebrates can be brought to new areas either intentionally (e.g. through trade, pet transportation) or during their natural migration, potentially transporting pathogens to previously non-endemic regions (Pfeffer and Dobler, 2010). However,

the establishment of new arbovirus transmission cycles requires multiple conditions: sufficient arthropod vector and susceptible vertebrate host populations, and suitable climate conditions occurring simultaneously – making the probability of a successful arbovirus establishment in a new location quite low (Pfeffer and Dobler, 2010). Nevertheless, multiple examples of such emergences have been documented: CHIKV introduction into Italy by a viremic traveler, and the likely bird-mediated WNV introduction to the Americas, (Rezza et al., 2007; Kramer et al., 2008). Pathogens (not only arboviruses) may also be introduced through the international trade of live animals, which is not always accompanied by pathogen screening, as illustrated by the introduction of mpox into the United States via imported African rodents (CDC, 2003; Borsky et al., 2003).

Although by now there are more than 500 known arboviruses – the majority of them being transmitted by mosquitoes – out of which around 150 can cause human infections, only a few have commercially available vaccines: DENV, yellow fever and Japanese encephalitis viruses (Madewell, 2020). Therefore, vector population control remains one of the main prevention strategies (Achee et al., 2019). Control measures for mosquito-borne viruses target different life stages of the vector and may involve biological, chemical, or environmental approaches (Madewell, 2020). Currently, considerable attention is focused on biological control strategies, particularly the deployment of *Wolbachia* bacteria, which were shown to limit arbovirus replication in *Aedes* mosquitoes (Hurk et al., 2012; Ogunlade et al., 2021).

Taking all this into consideration, it is critically important to assess what influences the evolution and spread of arboviruses to be able to predict their potential emergence, further spread, and form prevention strategies.

1.2.3. West Nile virus as a model for arbovirus spread analysis

There are several well-characterised arboviruses, particularly those associated with significant human disease, with West Nile virus being one example. WNV belongs to the *Flaviviridae* family and is usually maintained in an enzootic transmission cycle, infecting birds and *Culex* species mosquitoes, while sometimes infecting humans and other vertebrates as dead-end hosts (Blitvich, 2008). WNV can be divided into multiple genetic lineages, broadly grouped into 2 major ones, but different studies separate less widespread lineages into up to 9 separate genetic lineages (Fall et al., 2017).

WNV was first isolated from the blood of a sick woman in Uganda (Smithburn et al., 1940). For about 50 years it was not considered as a major pathogen, as outbreaks were mild and rare (Blitvich, 2008). However, since the early 1990s, WNV outbreaks have increased in both frequency and severity, resulting in major epidemics in Africa, Asia, Australia, Europe, and, since 1999, in North and South America, where it spread dramatically from its introduction into New York in 1999 (Blitvich, 2008; Kramer et al., 2008). The Mediterranean Basin and temperate regions of North America now experience WNV as a seasonal recurring disease (Hayes et al., 2005). Due to its current public health relevance, a large amount of sequencing data is available for analysis, and it is possible to perform molecular clock analyses (May et al., 2011).

Multiple phylogeographic analyses of WNV were already conducted in order to find which factors are driving its spread and evolution. Phylogeographic studies have helped identify the likely pathways of global WNV spread, tracing its roots to Africa (May et al., 2011). More recent studies tested hypotheses about WNV spread in Americas and Europe, using discrete and continuous phylogeography models (more about phylogeographic models in Section 1.4) to reconstruct transmission dynamics and test potential drivers of its spread, and identified that agricultural activity and temperature were positively associated with the spread of WNV (Lu et al., 2024; Dellicour et al., 2020). Therefore, such approaches may provide a useful framework for analysing the spatial spread of other viruses, provided that sufficient genomic and epidemiological data are available.

1.3. Molecular evolution in epidemiology

1.3.1. Genomic epidemiology

Infectious disease surveillance can be enhanced by integrating genomic data, helping to identify relations between cases and transmissions. Without genomic data, epidemiological surveillance is largely limited to tracking case counts over time, while variant identification and antibiotic resistance testing would require additional targeted methods and resources (Tiwari et al., 2025). But with integration of pathogen genomic data, it becomes possible to detect distinct pathogen lineages and identify new introductions to populations (Figure 3) (Black and Dudas, 2024).

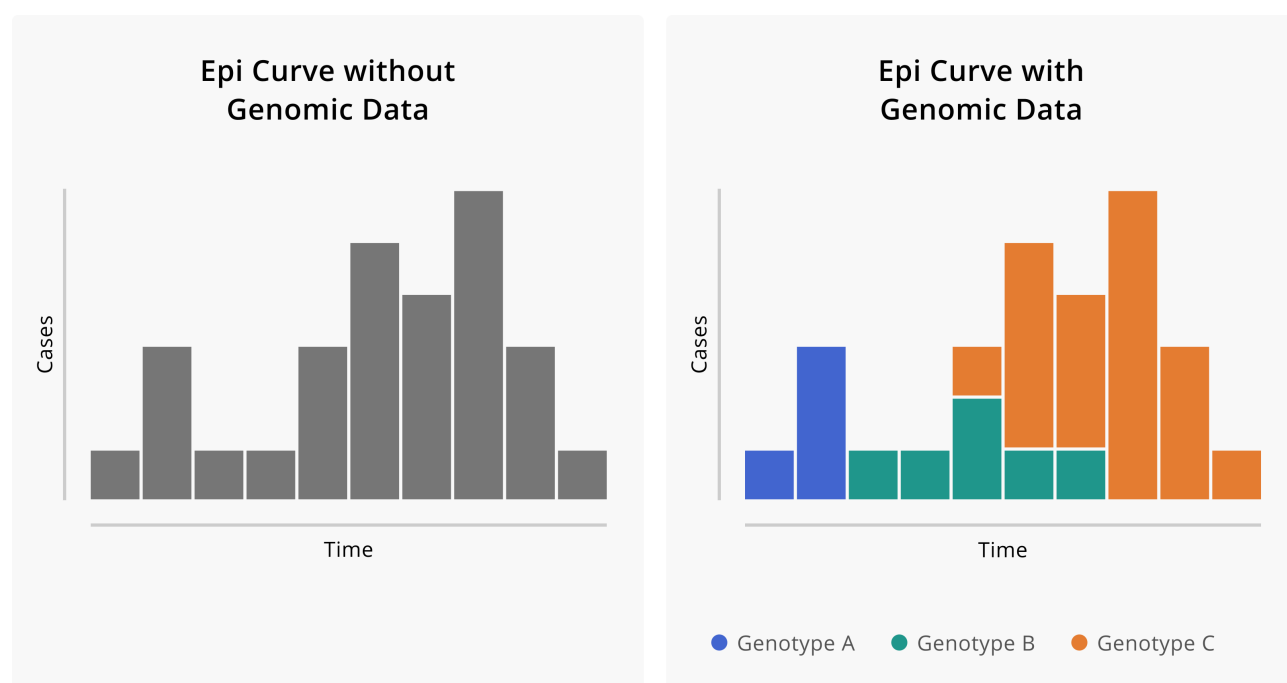


Figure 3. Epidemiologic curves without genomic data (left) and with the addition of genomic data (right). Black and Dudas, 2024

Before the availability of whole-genome-scale data, single markers were used for detecting and tracking outbreaks (Tümmler, 2020). Molecular epidemiology traditionally relied on

molecular biology approaches, primarily nucleic acid-based analyses, to investigate patterns of infectious disease transmission (Tümmler, 2020). It was mainly focused on studying single genes, microbial strains, mobile genetic elements, plasmids (Tümmler, 2020).

Previously, most outbreaks were analysed retrospectively, because preparation of samples for sequencing was long, and the process itself was costly (Tang et al., 2017). But now, the increasing availability of next-generation sequencing technologies, together with the progress of novel bioinformatics methods, has facilitated the development of genomic epidemiology – a relatively recent field that studies transmission of infectious diseases, pathogen evolution and their emergence using whole-genome data (Zhu et al., 2020; Hill et al., 2021). Genomic epidemiology studies can be conducted during ongoing outbreaks to provide nearly real-time insights that can inform infection control and prevention strategies for both bacterial and viral diseases (Halachev et al., 2014; Lane et al., 2021).

To fully make use of surveillance and genomic data over time, it is important not only to determine whether infections are linked, but also to infer the timing of transmission events and the origin of outbreaks (Menardo et al., 2019). Molecular clock analyses are essential for this purpose.

1.3.2. Molecular clock – concepts and applications

Genetic information can be characterized by how fast it accumulates mutations. If the speed of genetic changes remains approximately constant over time, the rate of the molecular clock can be estimated, reflecting the rate of sequence divergence through the accumulation of substitutions (Ho and Duchêne, 2014). It is determined by interactions of multiple factors: mutation types (neutral, advantageous or deleterious), population size, selection pressures (A. Drummond et al., 2003). Another related concept in molecular evolution is the mutation rate, which refers to the rate at which new mutations occur in the genome and is determined by the host or viral polymerase and repair systems. Although these concepts are closely related, they are not equivalent (A. Drummond et al., 2003).

Molecular clock can be estimated by analysing the correlation between the samples of interest and the root of the phylogenetic tree, which represents the common ancestor. By measuring how many mutations accumulate over time, a regression line through the correlation of root-to-tip distance and sample collection date can be drawn (Duchêne et al., 2016). If the samples are collected over a period of time, the slope of the regression line suggests the evolutionary rate of the dataset (Black and Dudas, 2024).

Because molecular clock can help estimate, how much time has passed between samples based on their genetic difference or vice versa, this approach is widely used. It can be used to generate temporally resolved phylogenetic trees, root-to-tip regression analyses can be used to identify outliers in the data, that do not follow the expected pattern of calculated molecular clock (Black and Dudas, 2024). Such outliers which might indicate sequencing errors, incorrectly annotated sampling dates, or rare latent infectious disease dynamics (Black and Dudas, 2024).

In practical applications, molecular clock analyses based on sequencing data can be used to estimate phylogenetic relations of pathogens, analyse their spread during outbreaks and monitor their genetic diversity for diagnostics and treatment (Hungnes et al., 2000). Recent molecular clock applications in outbreak epidemiology include SARS-CoV-2 non-human origin and variant evolution analyses, origins of recombinant swine Influenza A (H1N1) pandemic, aforementioned WNV introduction to the Americas, among others (Smith et al., 2009; Singh and Yi, 2021; Hill et al., 2022).

1.3.3. Complications in molecular clock calculations

RNA viruses are known to evolve faster than DNA viruses or cellular microorganisms, as their polymerases usually lack the 3' exonuclease proofreading activity (Steinhauer et al., 1992; Sanjuán and Domingo-Calap, 2016). Therefore, molecular clock rates differ substantially between DNA and RNA viruses. DNA viruses typically exhibit mutation rates on the order of 10^{-7} to 10^{-9} nucleotide substitutions per site per year, while RNA viruses usually show higher rates, ranging from 10^{-2} to 10^{-5} (Aiewsakun and Katzourakis, 2016). Such evolutionary rates for RNA viruses amount to approximately one mutation per genome per replication cycle (Sanjuán et al., 2010).

Accumulating mutations may increase or decrease virus fitness. According to the neutral theory of molecular evolution, the majority of changes are randomly fixed and are (nearly) neutral (Domingo et al., 1996). However, some mutations can be deleterious and are usually eliminated from the population. Therefore, such mutations are not sampled and can introduce bias in molecular clock rate estimates (Sanjuán et al., 2010). Bias can also be introduced when mutations accumulate rapidly, leading to substitution saturation, where multiple changes at the same site make the true number of substitution events indistinguishable and can reduce the reliability of evolutionary estimates (Holmes, 2003).

Multiple molecular processes occur during virus replication or virion assembly that can also complicate molecular clock analyses. For example, viruses with segmented genomes can undergo genome reassortment – a discrete form of recombination. It occurs when two or more virus strains or related viruses infect the same host cell. During virion assembly, genome segments from different lineages can be packed together, forming a new, hybrid virion (Brown and Yang, 2011). Genome reassortment can cause the emergence of new virus strains with changed pathogenicity, as seen in influenza A virus pandemics (Smith et al., 2009).

Because genome segments of segmented viruses can evolve at different rates, no single phylogenetic tree and molecular clock rate can describe the evolutionary history of such viruses (Schierup and Hein, 2000). Even small numbers of reassortment or recombination events during evolutionary history can complicate molecular clock analyses, as different genomic segments or regions may follow different phylogenetic tree topologies (Martin et al., 2011). Reliable inference of molecular clock depends on sufficient temporal signal – sufficient genetic change between samples over time to be able to infer the relationship between them (Rambaut et al., 2016).

1.3.4. Molecular clock models in Bayesian modeling for epidemiology

Bayesian modeling is based on Bayes' theorem for calculating conditional probabilities. In phylogenetics, prior knowledge about parameters of a model is updated with new information from observed (sequence or other) data to obtain phylogenetic trees that maximise the probability of topology, branch lengths and other included parameters, given sequence (and other) data (Bollback, 2002). One of the leading phylogenetic tools for molecular clock analyses is called BEAST (A. J. Drummond and Rambaut, 2007). BEAST uses the Markov chain Monte Carlo (MCMC) algorithm to stochastically sample from the posterior distribution of target parameters in probability space (A. J. Drummond and Rambaut, 2007). In recent BEAST versions, its functionality has been expanded beyond phylogenetic inference, allowing estimation of spatial dynamics by incorporation of discrete, continuous, or mixed traits, and supports a wide range of different evolutionary models (Baele et al., 2025).

Because the data used for analyses can be very different, determining which model fits the data best is important for accuracy and efficiency of calculations (Baele and Lemey, 2013). Multiple models for molecular clock estimation have been created to account for differences in datasets and goals for analyses.

The simplest model for estimating molecular clock rates is the strict clock, which assumes that the rate of evolution is constant across all lineages over time (Figure 4) (Ho and Duchêne, 2014). It is generally used where the variation in evolutionary rate is expected to be low (Brown and Yang, 2011). However, it is often unreliable, as rate variation between lineages is rarely homogeneous (Ho and Duchêne, 2014).

Multi-rate clocks can account for variation across phylogeny branches, but still assume that some of them will share the same rate (Bromham et al., 2018). The model will either group branches with the same assigned rates to clusters (local clocks) or scatter them across the phylogeny (discrete clocks) (Ho and Duchêne, 2014).

Variation across the phylogeny can be accommodated even further by using relaxed molecular clocks, where each branch can have its own unique rate (A. J. Drummond et al., 2006). In relaxed molecular clock models, substitution rates can be assumed to be independently and identically distributed in uncorrelated relaxed clocks or correlated between nearby branches of the tree in autocorrelated relaxed clocks (Thorne et al., 1998; A. J. Drummond et al., 2006; Bromham et al., 2018). Since rates naturally vary between species, lineages or genes, such models are biologically relevant (Ho and Duchêne, 2014).

Selecting a molecular clock model that fits the data and has the appropriate complexity is crucial, as misspecified models can produce inaccurate and misleading results (Ho and Duchêne, 2014; Worobey et al., 2014). Since overcomplicated models can have substantial computational demands, choosing a model that is as simple as possible yet takes into account the biological features of the data is necessary (Baele et al., 2012).

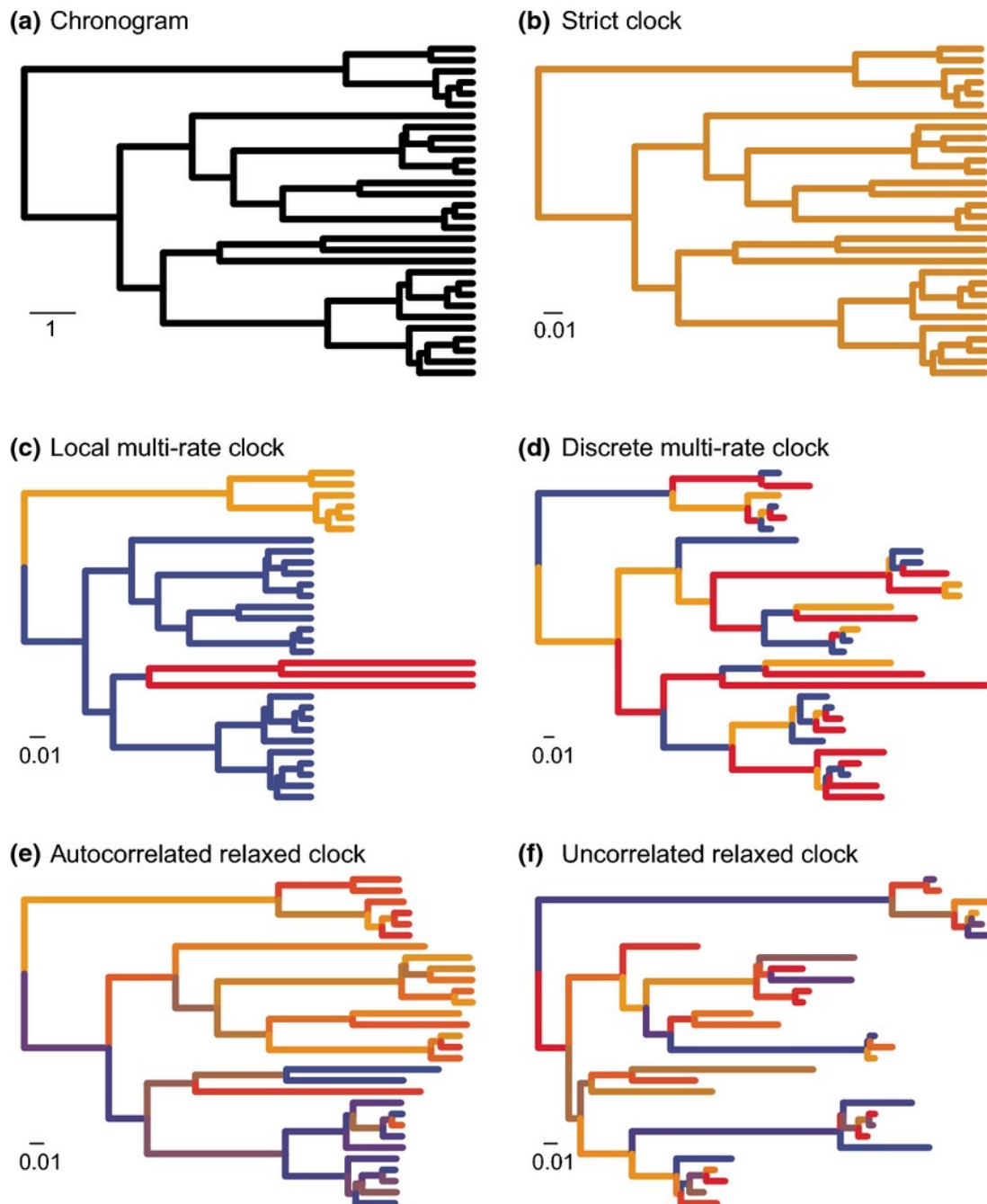


Figure 4. Differences between molecular clock models. Ho and Duchêne, 2014.

1.4. Phylogeography – concepts and uses

Phylogeography is a relatively new discipline, which focuses on studying genetic variation in a geographical context (J. Avise, 2009; Knowles, 2009). Since J. C. Avise et al., 1987 specified phylogeography as a term and highlighted the potential data yields of mitochondrial DNA phylogenies overlaid on geographic maps, the concept was widely adopted for a broader spectrum of findings, e.g. pathogen and host spatial dynamics, species conservation and their distribution (Bermingham and Moritz, 1998; J. Avise, 2009; Gutiérrez-García and Vázquez-Domínguez, 2011; Martinez et al., 2019; Nahata et al., 2021).

Phylogeographic studies can be conducted for one species (intraspecific) or comparative be-

tween numerous species (Gutiérrez-García and Vázquez-Domínguez, 2011). For both workflows the number and geographic diversity of samples is important for the quality and reliability of results, as sampling bias can strongly affect statistical inference (Maio et al., 2015). For example, if one or a few locations have sequenced many more genomes than others, overrepresentation of certain regions in datasets can occur (Guindon and De Maio, 2021). If some locations contribute disproportionately more sequences than others, this can bias the reconstruction of spatial spread. This was seen in SARS-CoV-2 analyses, where uneven sequencing efforts across countries and regions affected the reliability of phylogeographic inference and falsely marked better-sampled regions as sources of new variants (Hodcroft et al., 2021).

Bayesian phylogeography can be split into two main categories: discrete and continuous, depending on the type of geographic data that is included in the model (Figure 5) (Kalkauskas et al., 2021). Although both approaches rely on different assumptions and can give different inferences, comparing results across both frameworks is useful for assessing robustness (Nahata et al., 2021)

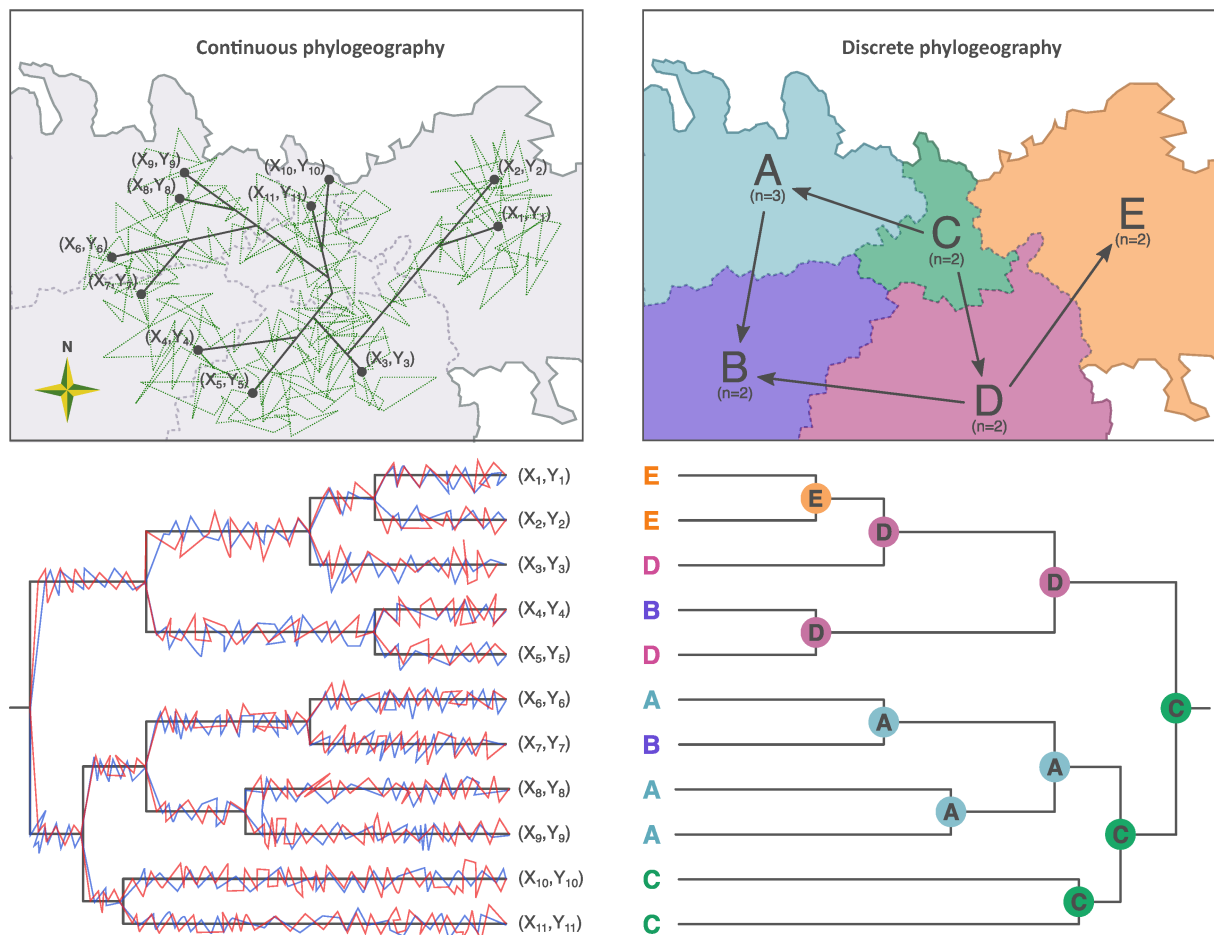


Figure 5. Comparison of continuous and discrete phylogeography approaches. Nahata et al., 2021.

1.4.1. Continuous phylogeography

Continuous phylogeographic inference can be performed using BEAST software, which can jointly reconstruct phylogenetic history and geographic spread using precise longitude and latitude coordinates of sampling (Dellicour et al., 2021). Early approaches used a Brownian diffusion model to reconstruct the spatial spread in a two-dimensional geographic space. In this framework, pathogen movement is modeled as consistent bivariate diffusion process over time (Lemey et al., 2010). Because a constant rate of spatial diffusion can be unrealistic for pathogen spread, more flexible models like the relaxed random walk, allowing diffusion rates to vary across branches, are used, creating a more biologically realistic reconstruction (Lemey et al., 2010; Dellicour et al., 2021).

Continuous phylogeography is used less often than discrete phylogeography although it can present a more precise reconstruction of the spatial movement history (Dellicour et al., 2021). Continuous phylogeography models are affected when sampling is uneven, resulting in less reliable calculations (Kalkauskas et al., 2021). Because of lack of consistent sequencing or information about precise sampling locations of publicly available data, this approach oftentimes cannot be properly used (Dellicour et al., 2021). In cases where sampling is inconsistent or only the country of sampling is known, discrete phylogeography can be a more appropriate approach (Dellicour et al., 2021).

1.4.2. Discrete phylogeographic analysis

Discrete phylogeographic approaches do not model the diffusion process but use sampling locations as discrete states, e.g. countries, cities or other administrative units. Although such information is often easier to acquire than precise sampling locations, it can provide an unrealistic or oversimplified reconstruction of the virus spread, as virus host movements usually do not follow geopolitical borders (Dellicour et al., 2018, 2021).

Discrete phlogeography uses a continuous-time Markov chain model, where transitions between locations are estimated similarly to nucleotide substitutions (Lemey et al., 2009; Layan et al., 2023). In this approach, the spread of each lineage is modeled as a trait evolving along the phylogenetic tree, allowing reconstruction of spatial transmission dynamics (Layan et al., 2023).

To analyse what factors can influence migration rates, discrete phylogeographic models can be extended to integrate generalized linear models (GLM) (Dellicour et al., 2018). Such integration allows to jointly calculate the spatial evolutionary history of virus movement and the potential factors (predictors) influencing the number of transitions between discrete states (Dellicour et al., 2018). GLMs can include various data (climate, geographic, demographic, economic) that can be associated with the virus movement (Magee et al., 2018).

By integrating GLMs, instead of estimating each migration rate as a free parameter, which would increase the amount of model parameters exponentially, predictor matrices – data tables with numerical values representing environmental variables – are included (Hong et al., 2021). Each predictor matrix has a “switch” that is either 1 or 0 (predictor is either included during

each MCMC step or not). If including a predictor makes the model fit the estimated phylogeny better, the predictor tends to be included, otherwise – tends to be left out, hence giving each predictor an inclusion probability (Lemey et al., 2014).

If the predictor is included, the model estimates of each predictor's contribution to the spatial diffusion process and calculates whether an association is positive (variable increasing the number of migrations between states) or negative (decreasing the number of migrations) (Langedijk et al., 2024). The magnitude of the effect is measured by the coefficient β , which represents if the predictor strongly influences the movements between locations, or if the effect size is low and its influence is negligible even when predictor is included in the model (Lemey et al., 2014). Because coefficients might not be included in each MCMC step, their effect size is relative, as it is only based on the results if the matrix is added to calculations.

Although GLM integration into discrete phylogeography can provide valuable biological insights, the model necessarily uses an abstract representation of geography. Because discrete locations are usually used as country- or county-level states, predictor values might not be precise due to local differences across large areas (Magee et al., 2018). Scarce and uneven sampling can also impact statistical inference, similarly to continuous phylogeography models (Maio et al., 2015; Kalkauskas et al., 2021).

2. Materials and Methods

2.1. Materials

2.1.1. Research subjects

The main research subject is Wuhan mosquito virus 6. Other RNA viruses were included in the analysis to compare how host and environmental factors influence viral spread. We included insect-specific viruses that infect *Culex* mosquitoes: *Culex pipiens* Iffa-like virus 2, *Phasivirus phasiense*, ISV infecting *Drosophila* fruit fly Sigma virus, arboviruses infecting insects and vertebrates (birds, humans): West Nile virus and Usutu virus, and avian-infecting Influenza A virus of subtype H5N1. All viruses have a sufficient number of publicly available genome sequences to conduct molecular clock and phylogeographic analyses.

2.1.2. Genomic data preparation

Viral genomes were obtained from publicly available databases: Bacterial and Viral Bioinformatics Resource Center (BV-BRC, <https://www.bv-brc.org/>), National Center for Biotechnology Information (NCBI, <https://www.ncbi.nlm.nih.gov/>), Science Data Bank (SciDB, <https://www.scidb.cn/en>).

Initial Wuhan Mosquito virus 6 sequences were obtained by screening published studies reporting virus detection and by searching Sequence Read Archives using PebbleScout (Shiriyev and Agarwala, 2024). Reads were assembled using a reference-based BWA pipeline. Bases with a Phred quality score >20 and a minimum coverage of 10 reads were retained. This step was done by Linas Gasparavičius (former MSc student from the lab).

West Nile Virus (WNV), Usutu Virus (USUV), and avian Influenza A (IAV) subtype H5N1 clade 2.3.4.4b sequences were obtained from the BV-BRC database. WNV dataset includes all available full genome sequences dated from 2000 to 2020, Usutu Virus – from 2000 to 2024. IAV dataset includes hemagglutinin segment sequences from 2020 to 2025.

Sigma virus (SIGV) and *Culex* flavivirus (CxFV) genomes were obtained from the NCBI database, excluding sequences from artificial infections. Since most of the available sequences of SIGV were partial genomes, sequences of at least 5 kb length were selected. The final dataset includes majority of sequences from Wilfert and Jiggins, 2014 study.

Since WNV, USUV and IAV had large datasets of more than 200 sequences, they were clustered by 95% similarity using CD-HIT (Fu et al., 2012). Some clusters had thousands of redundant sequences; therefore, to avoid oversampling and reduce computational load, we filtered clusters to take 2-3 random sequences per month per region. Because the IAV dataset was nearly three times larger than those of other viruses even after clustering and removal of redundant sequences, the dataset was randomly split into two subsets of 117 and 118 sequences for further analysis.

Culex pipiens Iffa-like virus 2 and *Phasivirus phasiense* genomes were assembled from sequence data retrieved through Logan-Search (<https://logan-search.org/>). This step was done

by Martynas Smičius (PhD student from the lab).

Sequence alignments were generated using MAFFT v7.526 global pairwise alignment and poorly aligned ends were trimmed (Kato and Standley, 2013). Initial maximum-likelihood phylogenetic trees were inferred using IQ-TREE v3.0.1 (Wong et al., 2025). TempEst, a software tool for the visualisation and analysis of temporally sampled sequence data, was used to identify sequences that represent outliers in root-to-tip regression analyses (Rambaut et al., 2016). Sequences that were distant from the regression line were removed from further analyses. Final numbers of sequences per cluster for each virus dataset and types of genome sequences used (segments or full genomes) are specified in Table 1.

Table 1. Summary of virus datasets, sampling structure, and genome characteristics.

Virus	Virus family	Host	Number of clusters used	Number of sequences per cluster	Number of regions per cluster	Genome structure used
Wuhan Mosquito Virus 6	<i>Orthomyxo-viridae</i>	<i>Culex</i> mosquitoes	3	127, 122, 118	27, 27, 28	NP, PB1, PA segments
<i>Culex pipiens</i> Iffa-like virus 2	<i>Iflaviridae</i>	<i>Culex</i> mosquitoes	1	42	12	Full genome
<i>Phasivirus phasiense</i>	<i>Phenuiviridae</i>	<i>Culex</i> mosquitoes	1	57	7	L segment
West Nile Virus	<i>Flaviviridae</i>	Birds, mosquitoes, humans	2	147, 103	33, 23	Full genome
Usutu Virus	<i>Flaviviridae</i>	Birds, mosquitoes, humans	1	73	13	Full genome
Influenza A Virus H5N1	<i>Orthomyxo-viridae</i>	Birds*	2	117, 118	33, 33	HA segment
Sigma Virus	<i>Rhabdoviridae</i>	<i>Drosophila</i> flies	1	113	10	Full genome

* Only IAV H5N1 infecting birds were included in the analysis.

Genomes for selective pressure analysis were prepared using the same filtering criteria. We used all protein coding gene sequences for WuMV-6, IAV H5N1, WNV, and CxFV. WNV genes were obtained using NC_009942.1 sequence annotations as a reference. CxFV reference was NC_008604.2, and gene annotations were accessed from Bittar et al., 2016 study.

2.1.3. Raster dataset preparation for WuMV-6 continuous phylogeography in China

Raster datasets representing geographic and human activity variables were used for continuous phylogeographic analyses. Data were obtained from publicly available databases. Global land-cover rasters for needle-leaf trees, broadleaf trees, shrublands, cultivated and managed

vegetation, and urban coverage were retrieved from the EarthEnv database (<https://www.earthenv.org>). Cropland coverage was obtained from EarthStat (<http://www.earthstat.org/>). Land elevation raster obtained from WorldClim (<https://www.worldclim.org>), evapotranspiration from FigShare (<https://figshare.com>). Human population density raster obtained from WorldPop (<https://hub.worldpop.org/>). Rasters were cropped to the extent of China borders and resampled to decrease their resolution to match with the lowest resolution of the rasters (0.0833 x 0.0833, or 5 arc minutes).

More about the usage of rasters for continuous phylogeography can be found in the methods section 2.2.1.

2.1.4. Predictor dataset preparation for discrete phylogeography

Data for discrete phylogeography predictors were obtained from publicly available databases (Table 2). We included predictors representing relevant climate, demographic, economic, and factors related to the movement of birds.

Climate-related predictors include mean annual temperature (°C), precipitation (mm), and seasonal index for both variables. Seasonal index was calculated by taking the difference between the highest and lowest monthly values (temperature and precipitation) for each region. Then regions were sorted from lowest to highest values and ranking index was assigned to each region.

Because bird movement can influence virus spread, predictors representing global bird migrations and the number of poultry farms were included. Poultry farms were assessed as the number of active farm business per region at the national or state level. The seasonal migrations of birds were calculated as overlaps between the areas of the bird migration routes and the sampling regions. If more than 50% of the region area overlaps with a bird flyway, that region gets assigned a value of 1, otherwise – gets a value of 0. The percentage of forest coverage was included because large areas of natural reserves can increase contact between wild birds and poultry, which is assumed to be related to outbreaks (Lu et al., 2017).

Variables representing the land area of each region, the size of the human population and the GDP of each region were included to assess whether human activity influences the movement of the virus. Predictors representing shared land borders between regions and distance between region centroids are included to infer if virus migration is faster between regions in closer proximity.

Predictor representing the number of sequences per region was included to infer if any regions are overrepresented in the analyses. Number sequences per region for each analysis are specified in Supplementary A1.

Predictor values were log-transformed when no negative values were present. All predictor values were standardized to have a mean value of 0 and a standard deviation of 1. Predictors were incorporated into the analysis within a generalized linear model framework as country-level or second administrative unit-level (e.g., state, province) matrices. Variables that differ between regions (e.g., mean annual temperature) were generated as separate origin and destination

Table 2. GLM predictors included in discrete phylogeography. “Log” indicates log transformation.

Predictor variable	Structure	Transformation	Temporal coverage	Data source	Comment
Mean annual temperature	Origin-destination	Standardized	1990–2020	https://weatherandclimate.com/	Not log-transformed due to negative values. Daily mean values.
Mean annual precipitation	Origin-destination	Log, standardized	1990–2020	https://weatherandclimate.com/	Daily mean values
Temperature seasonality index	Origin-destination	Log, standardized	1990–2020	https://weatherandclimate.com/	Difference between max and min monthly values
Precipitation seasonality index	Origin-destination	Log, standardized	1990–2020	https://weatherandclimate.com/	Difference between max and min monthly values
Distance between region centroids	Pairwise	Log, standardized	–	https://www.osmtoday.com/	Calculated from .geojson files
Shared land borders	Pairwise	Log, standardized	–	–	–
Number of bird farms per region	Origin-destination	Log, standardized	2025	https://www.poidata.io/report/	–
Region area size	Origin-destination	Log, standardized	–	–	Values in km ²
Forest coverage (%)	Origin-destination	Log, standardized	2020	https://www.globeforestwatch.org/map/	Excludes agricultural land
Human population size	Origin-destination	Log, standardized	2024	–	–
Gross domestic product	Origin-destination	Log, standardized	2024	https://www.worldometers.info/gdp/ , https://worldpopulationreview.com	–
Number of sequences per region	Origin-destination	Log, standardized	–	–	–
Overlap with bird migration flyways	Pairwise	Log, standardized	–	–	Included if >50% of flyway overlaps region

matrices to measure effect directionality, whereas pairwise predictors that are independent of movement direction (e.g., distances between region centroids) were represented by a single matrix (Supplementary B1, B2, B3).

All viral sequences were assigned to the region where they were sampled. Regions were defined at the country level, except for large countries with multiple samples, where second-level administrative units were used (specifically for China and the United States).

2.2. Methods

2.2.1. Continuous phylogeography of WuMV-6 in China

WuMV-6 diffusion dynamics in continuous space were inferred using BEAST v1.10.4 (Suchard et al., 2018). XML files were prepared using BEAUti v1.10.4. We used uncorrelated lognormal relaxed molecular clock (UCLD) that allows branches of the phylogenetic tree to have independent rates of evolution (A. J. Drummond et al., 2006). A Hasegawa-Kishino-Yano (HKY) substitution model with two partitions (1st + 2nd nucleotide positions vs. 3rd position) was used (Hasegawa et al., 1985). Phylogenetic trees were reconstructed using coalescent constant population size model, which assumes that the population remained constant through time (A. J. Drummond et al., 2002). The spatial spread was modelled using a relaxed random walk diffusion process (Lemey et al., 2010). Sampling locations were integrated as precise latitude and longitude coordinates, with a jitter window of 0.01 degrees to account for identical sampling locations and avoid improper posterior distribution (Lemey et al., 2010).

MCMC chains were run for 200 million steps, sampling every 20 000 steps and removing 10% burn-in. MCMC convergence was assessed using Tracer v1.7.2 (Rambaut et al., 2018). Results were combined from 5 independent runs using LogCombiner v10.5.0. Maximum Clade Credibility (MCC) Tree was generated with TreeAnnotator v10.5.0 from the BEAST software packages.

Example files, including the generated MCC tree (combined from five BEAST runs), the BEAST XML file, and the region KML files are available on GitHub: https://github.com/aisteid/WMV6-phylogeography/tree/main/continuous_phylogeography/data.

2.2.2. Impact of environmental factors on WuMV-6 spread in China

We used the R package SERAPHIM to assess the influence of environmental factors on viral diffusion rates (Dellicour et al., 2016). All environmental variables were represented as raster files that describe their spatial heterogeneity (raster preparation is detailed in Section 2.1.3). SERAPHIM estimates branch-specific environmental weights by extracting raster data along the inferred spatio-temporal paths of phylogenetic tree branches. These weights are then used to calculate if environmental variables facilitate or slow down the viral movement by testing correlation between environmental distances and branch dispersal velocities. Statistical support for these associations was evaluated using Bayes factor (BF) and considering values greater than 3 as significant.

A posterior sample of 100 phylogenetic trees from the BEAST phylogeographic analysis, together with the MCC tree, was used to extract spatio-temporal information for SERAPHIM analyses and compute association to environmental factors. SERAPHIM analysis was conducted by dr. Simon Dellicour (KU Leuven, Belgium).

2.2.3. Discrete phylogeography of WuMV-6 and other RNA viruses

For WuMV-6 and other negative-sense RNA viruses included in the analysis, discrete phylogeographic analyses were conducted using BEAST X (Baele et al., 2025). XML files were generated using BEAUti v10.5.0. UCLD molecular clock was used (A. J. Drummond et al., 2006). HKY substitution model with two partitions (1st + 2nd nucleotide positions vs. 3rd position) was used (Hasegawa et al., 1985). A flexible non-parametric Bayesian Skygrid tree prior with a cut-off of 50 years, which estimates the effective population size trajectories, was used (Gill et al., 2013). Sampling locations were integrated as discrete traits where regions were defined at the country level. For large countries with multiple samples (e.g. China, the United States), second-level administrative units were used as regions.

For WNV discrete phylogeographic analysis we applied a strict molecular clock and constant population size tree prior. An informative prior on the molecular clock rate was specified based on previous studies (Añez et al., 2013). Changes from previously described model were implemented because preliminary analyses with UCLD and Skygrid tree prior resulted in effective sample sizes (ESS) being <200 and poor MCMC mixing (i.e. inefficient exploration of parameter space).

Analysis of SIGV was conducted using a UCLD molecular clock and constant population size tree prior. Mean and standard deviation of the molecular clock rate were also included in the model as prior parameters (Wilfert and Jiggins, 2014). Skygrid tree prior was tested as well, but ESS values were low after combining 3 independent runs.

Three independent runs of 100 000 000 states were completed for each analysis. Generated log and tree files were combined using LogCombiner v10.5.0. Burn-in of 10% was removed. MCC tree was generated using TreeAnnotator v10.5.0.

We extended the discrete phylogeographic model by using GLMs. We simultaneously reconstruct the spatiotemporal history and assess the potential predictors influencing virus spatial spread (Lemey et al., 2014). Within this framework, each predictor is associated with a log-transformed effect size and an inclusion probability that shows how frequently the model includes the predictor in the calculations. Bayes factor supports were interpreted as significant for $BF > 3$ and strong for $BF > 10$, very strong if $BF > 30$ (Faria et al., 2013).

Predictors representing geographic, climate, demographic, and economic variables were included in the analysis (more about predictor preparation in Section 2.1.4). To account for predictor colinearity, we calculated their pairwise correlation and removed those that showed variance inflation factor above the value of 5 (Supplementary C1, C2). Predictors that were removed from each analysis are described in Table 3.

Table 3. Predictors and flyways excluded from each virus model, by reason for exclusion.

Virus	Removed due to correlation	Insufficient location data	Not statistically significant (not included by the model)
WuMV-6	Precipitation seasonal index (origin)	Central Asia, West African East African, Mississippi Americas, Pacific Americas	—
WuMV-6 (PA, PB1)	—	Central Asia, West African East African, Mississippi Americas, Pacific Americas	—
CxIfla-like2	—	Central Asia, Central Pacific, East Asian Australasian, West Asian East African, Mississippi Americas, Pacific Americas	—
<i>Phasivirus phasiense</i>	Temperature seasonal index (origin & dest.), mean annual precipitation (origin & dest.), precipitation seasonal index (origin & dest.), number of sequences per region (origin & dest.), human population size (origin & dest.), GDP (origin & dest.), number of bird farms per region (origin & dest.)	Atlantic Americas, Blacksea Mediterranean, Central Asia, Central Pacific, East Atlantic, West Asian East African, Mississippi Americas, Pacific Americas; no shared land borders	—
WNV cluster 1	Human population size (origin & dest.), GDP (origin & dest.)	Blacksea Mediterranean, Central Asia, Central Pacific, East Asian Australasian, West Asian East African	Mean annual temperature (origin & dest.)

Virus	Removed due to correlation	Not included due to lack of data (flyways)	Not included by the model
WNV cluster 2	—	Atlantic Americas, Blacksea Mediterranean, Central Asia, Central Pacific, East Asian Australasian, West Asian East African, Mississippi Americas, Pacific Americas	Mean annual temperature (origin), precipitation seasonal index (dest.)
USUV	—	Atlantic Americas, Central Asia, Central Pacific, East Asian Australasian, West Asian East African, Mississippi Americas, Pacific Americas	Blacksea Mediterranean
SIGV	GDP (origin & dest.)	All flyway predictors	Mean annual temperature (dest.)
IAV H5N1 cluster 1	Mean annual temperature (origin & dest.), mean annual precipitation (origin & dest.), human population size (origin & dest.)	Blacksea Mediterranean, West Asian East African	Mississippi Americas
IAV H5N1 cluster 2	Mean annual temperature (origin & dest.), mean annual precipitation (origin & dest.), human population size (origin & dest.)	Central Asia, Central Pacific, East Atlantic, Pacific Americas	Mississippi Americas

All used XML files and discrete trait files are available on Github: https://github.com/aisteid/MSc_Thesis_2026/tree/main/discrete_phyloge_BEAST_files

2.2.4. Evolutionary pressure analysis

To assess whether the surface protein of WuMV-6 virus evolves faster than its other proteins, in comparison with IAV, WNV, and CxFV (arbovirus in comparison to insect-specific virus), we performed evolutionary pressure analyses.

WuMV-6, IAV, WNV and CxFV genes were translated to amino acid sequences using Biopython (Cock et al., 2009). Amino acid phylogenetic trees were generated using IQTREE. Tree lengths were calculated using the *baltic* Python package.

Selective pressure on all protein coding genes separately was assessed by calculating dN/dS values using HyPhy fixed effects likelihood method which calculates nonsynonymous and synonymous substitution rates at each sequence position (Kosakovsky Pond and Frost, 2005). Based on dN/dS values, codons were assigned as undergoing purifying or neutral selection ($dN/dS \geq 1$) or under positive selection ($dN/dS > 1$).

2.2.5. Viral genome assembly for mixed lineage in pooled samples

Since a lot of mosquito metagenomic sequencing comes from pooled mosquito samples, we assessed if WuMV-6 samples have variable sites that could come from mixed lineages. We designed a pipeline to calculate variability at each genomic site and to assemble genomes while accounting for this sequence variability.

The pipeline downloads samples by accessing their SRA number or downloading fastp files directly from their source if SRA's are unavailable. Reads from downloaded files are trimmed using Fastp v1.0.1 (Chen, 2025). Trimmed reads were mapped to reference sequence using BWA-MEM2 v2.3 and SAMtools v1.22.1 (Vasimuddin et al., 2019; Danecek et al., 2021). Variant calling was done using BCFtools v1.22 (Danecek et al., 2021).

Variant information was extracted from the VCF files and site variability was calculated based on this data. Alleles with a minimum frequency of 0.9 were considered alternative allele (different from the reference) and were included in the assembly. Otherwise, if site is flagged as variable but alternative allele had frequency of less than 0.9, an ambiguous nucleotide is added to assembled sequence. If position is non-variable, the most frequent allele is taken. Positions with low read depth (less than 10 reads) are masked as ambiguous.

A report file is generated for each sample, including: 1) sites that differ from the reference sequence, 2) sites with low read depth, and 3) variable sites with multiple alleles. For each site, the following information was recorded: genome segment, position, read depth, reference nucleotide, called nucleotide, and the depth and frequency of each nucleotide.

Shannon Entropy is calculated for each genome site by assessing nucleotide frequencies from the VCF files. Then total entropy is calculated for each genome segment and written to a report file, which includes: genome segment, mean entropy and mean depth.

To check if mean entropy is correlated with mean read depth, Spearman correlation was calculated using Python Scipy package (Virtanen et al., 2020).

2.2.6. Code source

Code for generating GLM matrices for discrete phylogeography is available on Github: <https://github.com/aisteizid/WMV6-phylogeography>.

Code for mixed viral lineage genome assembly is available here: <https://github.com/aisteizid/mixed-viral-lineages-genome-assembly/tree/main>.

Code for other data visualization is available here: https://github.com/aisteizid/MSc_Thesis_2026. Most visualizations were made using Matplotlib and *baltic* Python packages (Hunter, 2007). Artificial intelligence was used to generate parts of the code.

3. Results

The study was aimed to reconstruct the spread of Wuhan Mosquito Virus 6 in China and globally and analyse, if environmental or host factors have influence on RNA virus spread.

3.1. Reconstruction of WuMV-6 diffusion (continuous phylogeography)

Genomic dataset for continuous phylogeography included 68 WuMV-6 NP segment sequences from China. By integrating sampling coordinates to Bayesian model we reconstructed the spatial spread pathways (Figure 6). WuMV-6 appears to have spread throughout the majority of the country in a span of around 10 years. Based on the available sampling data and reconstructed diffusion patterns, northern regions of China appear to have been colonized later than southern regions. The phylogeographic reconstruction suggests an origin in Yunnan province, from which the virus spread to the north and east.

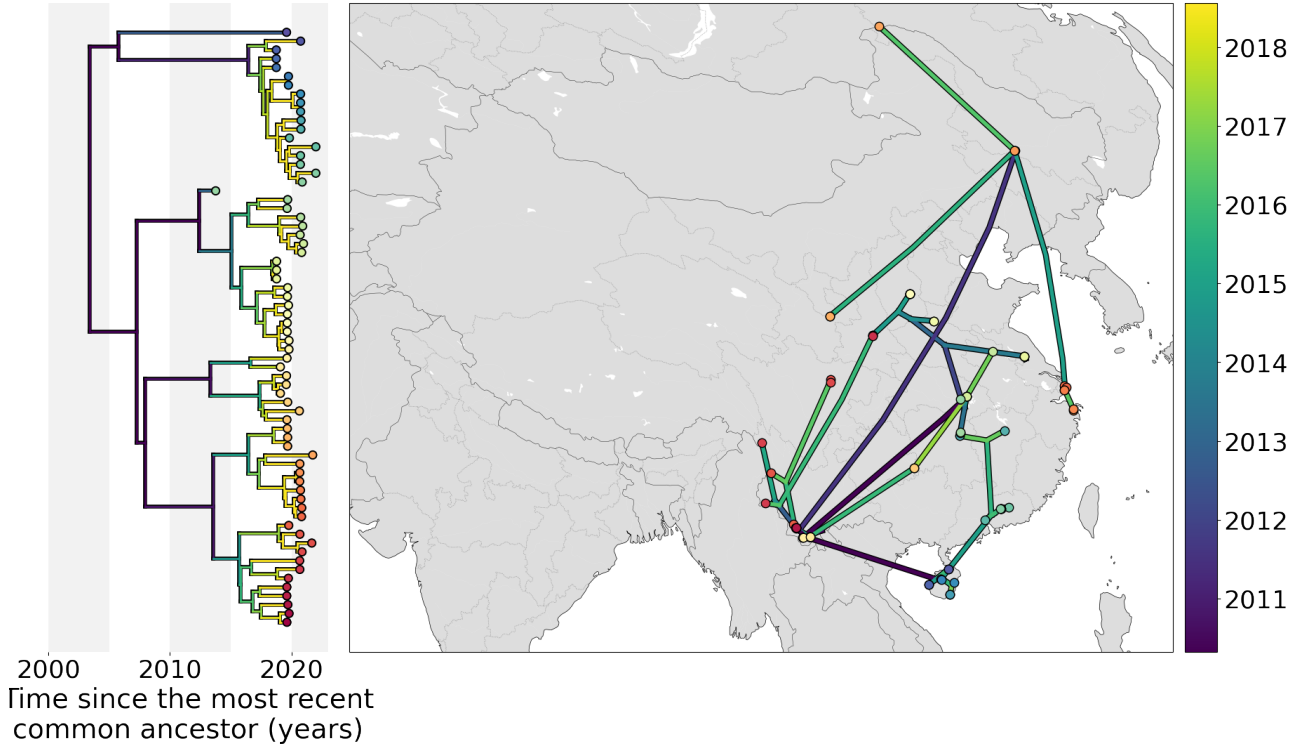


Figure 6. Bayesian phylogeny of WuMV-6 NP segment on the left and respective sampling locations connected by movement pathways on the map on the right. Tree branches and movement trajectories are coloured according to the midpoint time of each branch. Tips of the phylogenetic tree and samples on the map are colored according to the Y axis of the phylogenetic tree.

Using BEAST we estimated that the time to the most recent common ancestor (MRCA) for WuMV-6 NP segment was 2002.85 (95% highest posterior density (HPD) interval 1986.33, 2013.36). The mean nucleotide substitution rate was 1.28×10^{-3} substitutions per site per year (95% HPD 6.07×10^{-4} , 2.04×10^{-3} substitutions per site per year).

The subsequent analyses of the impact of environmental factors on the spread velocity of

WuMV-6 using SERAPHIM showed that out of the factors that were tested (land elevation, evapotranspiration, needleleaf, broadleaf, shrubland, cropland coverages, human activity coverage, human population density) none showed strong statistical support (Figure 7). Although Q statistic (“statistic estimating an association between dispersal durations and environmental distances”, SERAPHIM tutorial) had positive values, they were close to 0, while Bayes factor for each variable was less than 3.

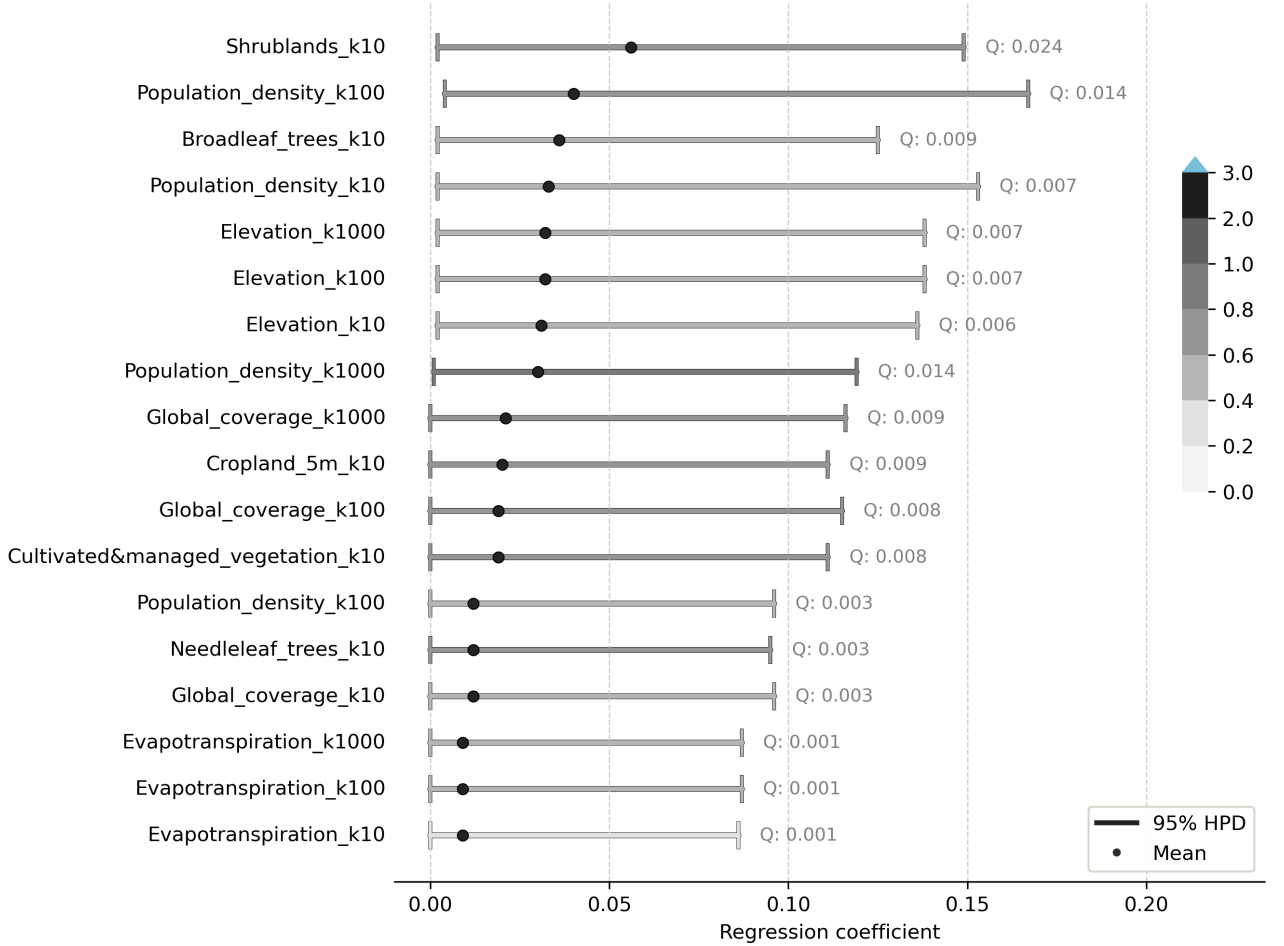


Figure 7. Environmental variables impact on WuMV-6 NP segment diffusion velocity. Mean effect sizes and their 95% HPD intervals and corresponding Q statistic and Bayes factor. Variables having $BF > 3$ are assigned color blue.

Continuous phylogeography was not applied to analyse other RNA viruses or other WuMV-6 diffusion locations due to inconsistent sampling or incorrect coordinates of the sampling location.

3.2. Discrete phylogeography and environmental correlates of spatial diffusion of WuMV-6

Discrete phylogeographic analysis of WuMV-6 NP genome segment was conducted using 127 sequences. These models calculate jumps between discrete states. It inferred migration events connecting Asia, Africa, Australia, Europe, and North America (Figure 8). Introductions to

4 continents as well as multiple provinces in China are estimated to have originated from the Yunnan province.

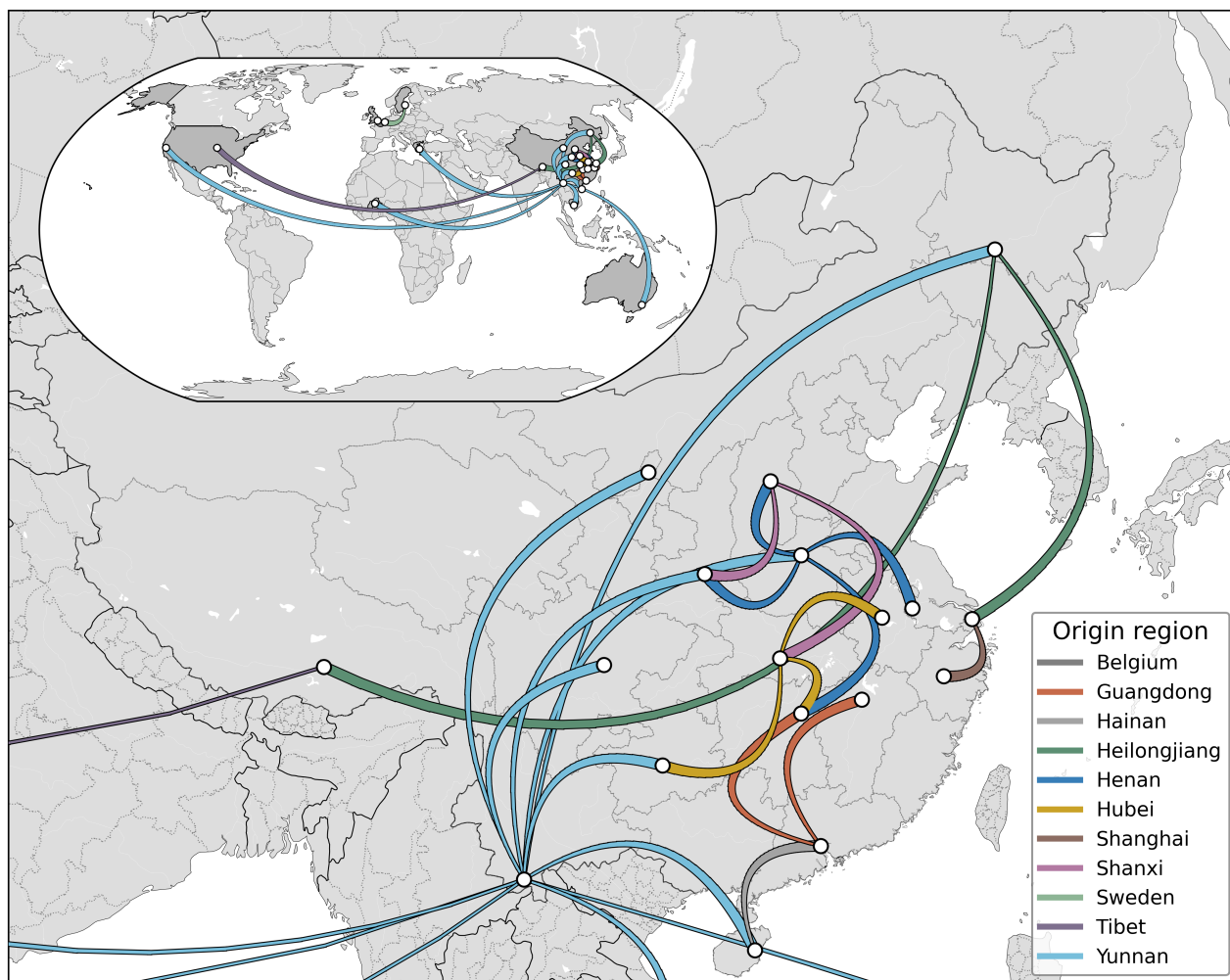


Figure 8. WuMV-6 migrations between countries and states/provinces. Curved lines indicate migration events. Lines are thin at the origin region and widen towards the destination region. Based on NP segment.

Phylogenetic analysis suggests that multiple jumps between distant regions happened in short periods of time: from China to the USA (California and Missouri), China to Australia, Greece, and Burkina Faso (Figure 9). All samples share a common ancestor dating to 1987.86 (95% HPD 1968.25, 2003.27). Although introductions to multiple continents are estimated to originate from China, there is no evidence that the global and Nordic European lineages would overlap in the time period since the most recent common ancestor estimated to have existed in the 1980s (Figure 9). However, the Nordic European lineage (including samples from Sweden, Belgium, and the United Kingdom) and the remaining global lineages show early divergence, with each group tracing back to separate ancestors in the early 2000s.

Molecular clock analyses using global WuMV-6 NP segment sequences infer similar clock rates as continuous phylogeographic analyses for samples from China, with mean rate at 1.64×10^{-3} substitutions per site per year (95% HPD 9.22×10^{-2} to 2.42×10^{-3} substitutions per site per year).

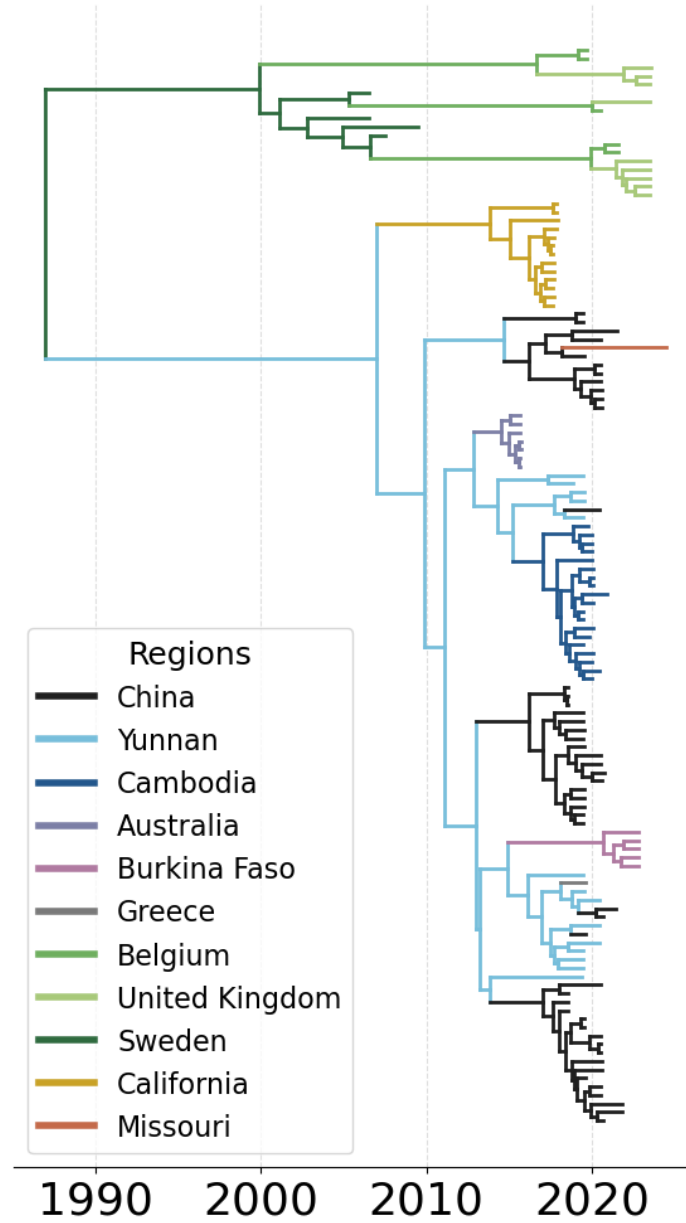


Figure 9. Time-scaled Bayesian phylogeny of the NP segment. The tree represents the MCC tree obtained by combining three independent BEAST runs. Branches are coloured according to inferred geographic transitions (region changes). Yunnan province is colored light blue, other China’s provinces are colored black.

We then analysed which environmental, climate, demographic and bird movement-related factors may be associated with the number of WuMV-6 migrations between sampling regions. Predictors were considered significant if $BF > 3$. Out of 25 predictors included in the analysis of WuMV-6 NP segment, 4 were evaluated as statistically significant, correlating to the higher number of migrations between sampling regions. East Atlantic bird migration flyway, regions sharing a land border, and the number of bird farms per region at the origin country showed a significant positive correlation, while distance between sampling region centroids showed negative association (Figure 10).

The effect of significant predictors (that have values other than 0 and 1) on the number of jumps between regions was evaluated. Viral jumps between regions were sorted according to

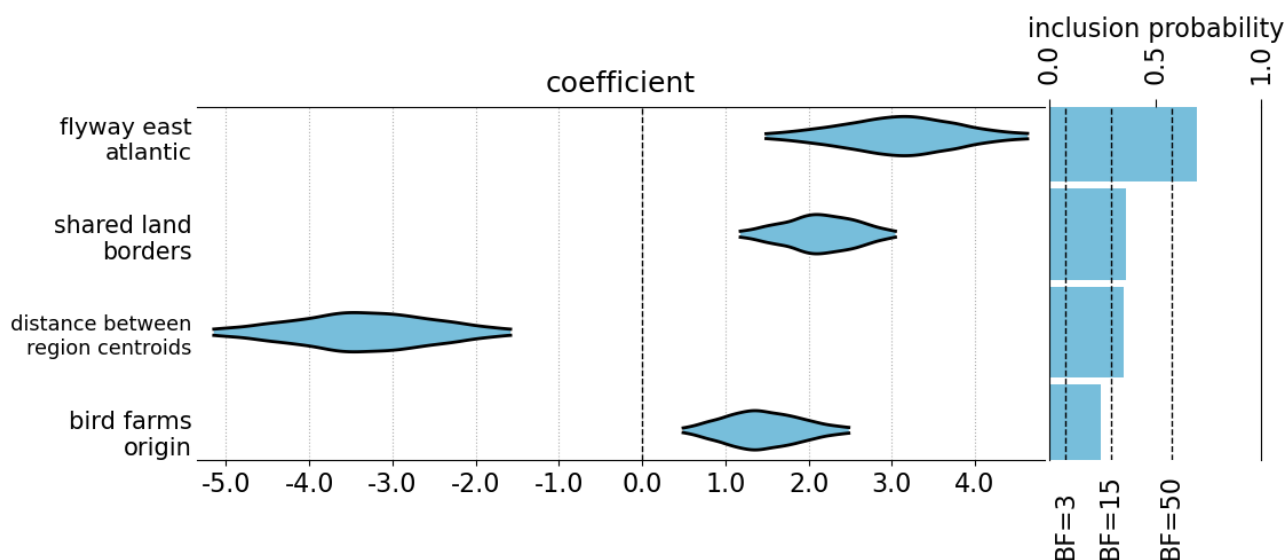


Figure 10. WuMV-6 NP segment 95% highest posterior density interval for migration predictor effect size with predictor inclusion probability and corresponding Bayes factor cut-offs.

farm predictor values (Supplementary E1). Although this predictor was inferred to be significant, the distribution of transition events was not consistent with predictor value magnitude.

Other analysed WuMV-6 segments also showed distance-based effects – both, PA and PB1 segment analyses inferred positive correlation between the number of migrations and regions sharing land borders, while larger distances between sampling region centroids have a negative effect. PA segment analysis also infers a positive effect for the forest area percentage of the region.

3.3. Environmental and host correlates of spatial diffusion of other RNA viruses

To evaluate if the same environmental factors are consistently evaluated as significant for viruses with different hosts, we applied the same discrete phylogeography with integrated GLMs framework to other RNA viruses. We included insect-specific viruses: *Culex pipiens* Ifla-like virus 2, *Phasivirus phasiense*, *Drosophila* Sigma virus, arboviruses: West Nile and Usutu viruses, and avian-infecting subtype of Influenza A virus H5N1 clade 2.3.4.4b. Results of all analyses (all included predictors and their coefficient effect sizes) are summarized in Figure 11. Inclusion probabilities, coefficient 95% HPD and BF of each predictor that were inferred to be significantly influencing virus migrations are provided in Supplementary D1.

Predictors that were inferred as significant in most analyses were representing distance-based variables. Regions sharing a land border were always inferred to have a positive effect size, while distance between region centroids have a negative effect. WNV cluster 1 analysis showed 15 significant predictors, although more than half of them had coefficient values close to 0.

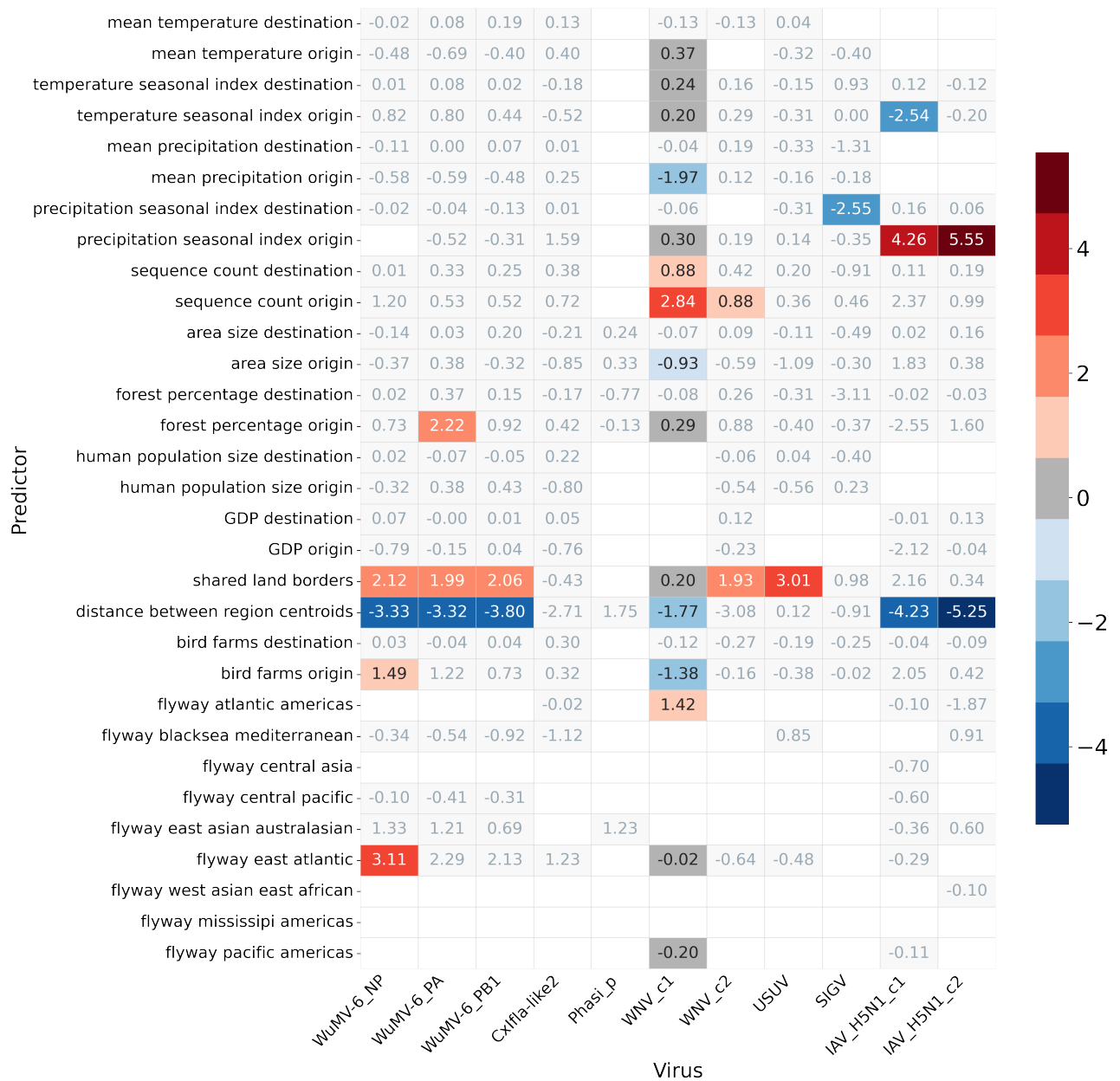


Figure 11. GLM predictor significance for each virus included in the analysis. If $BF > 3$, cells are colored according to the direction of the effect (red for positive, blue for negative). If $BF < 3$, cells are colored light grey. If predictors were not included in the analysis due to correlation or lack of data, cells are white.

3.4. Selective pressure on RNA virus proteins

To assess whether the WuMV-6 surface protein gp64 evolves faster than the other viral proteins, and compared to protein evolutionary rates in the avian IAV H5N1, we calculated the selective pressure and amino acid tree length for each protein-coding gene individually. To have an additional arbovirus-ISV comparison, WNV and CxFV were included in the analysis.

Phylogenetic tree lengths were assessed as amino acid substitutions per site over all tree branches (Figure 12). WuMV-6 has a large difference between the surface protein and the remaining proteins in terms of tree length, gp64 having 0.49 amino acid substitutions per site. This value is nearly twofold higher than the nucleoprotein, RNA-dependent RNA polymerase

(RdRp) subunits, hypothetical 1 and 2. Hypothetical 3 has a much higher tree length of approximately 1.62 amino acid substitutions per site. IAV showed a similar pattern, where surface protein has values higher than RdRp subunits, NP and matrix protein. In comparison, both flaviviruses (WNV and CxFV) showed relatively similar amino acid substitution levels across most proteins, with the exception of CxFV capsid protein, which had a substantially higher amino acid evolution rate.

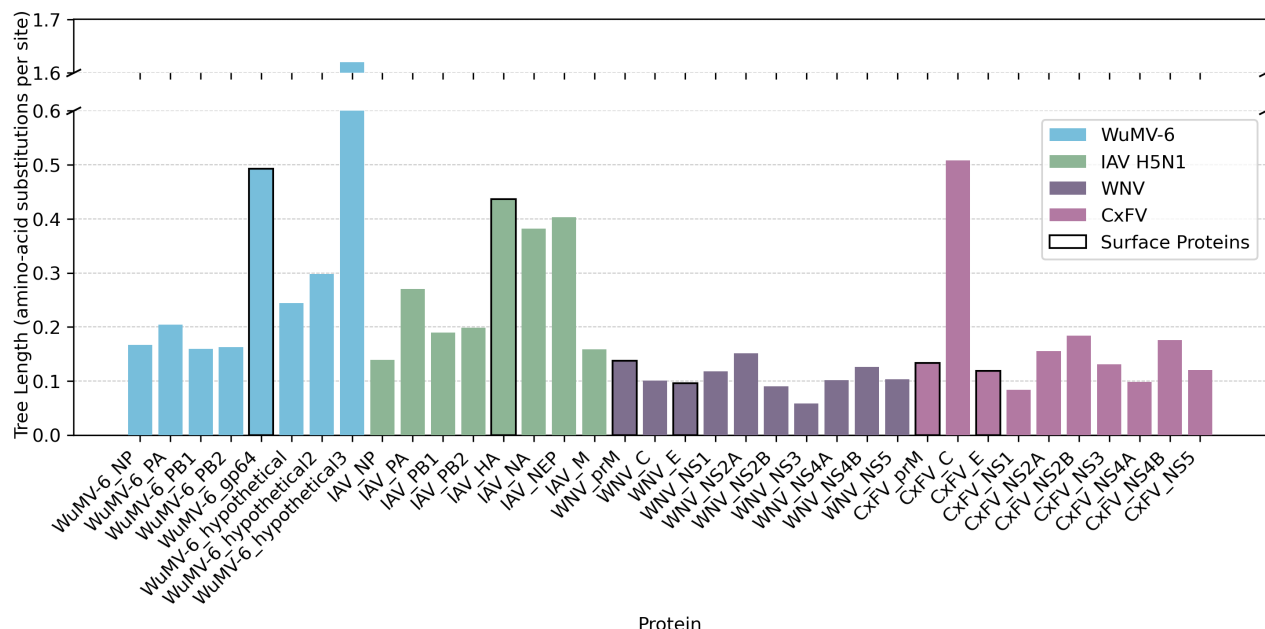


Figure 12. Maximum-likelihood tree lengths of WuMV-6, IAV H5N1, WNV and CxFV proteins. Surface proteins are highlighted with black edgecolor.

dN/dS values were calculated for each gene and genome sites were inferred as undergoing positive/diversifying selection if $dN/dS > 1$, and under neutral or purifying selection if $dN/dS < 1$ (Figure 13). WuMV-6 gp64 and hypothetical proteins 1, 2 and 3 had largest amounts of genome sites under diversifying selection. Similarly, IAV surface protein, PA and NEP had largest numbers of sites under diversifying selection. Flaviviruses do not have a consistent pattern, as WNV shows NS4B and CxFV – C and NS2B with the most sites under positive selection.

To compare the excess of non-synonymous mutations over synonymous mutations between surface and inner proteins, a Mann-Whitney U non-parametric test was performed. However, none of the analysed viruses had statistically significant difference (WuMV-6 p-value = 1.00, IAV p-value = 1.00, WNV p-value = 1.00, CxFV p-value = 0.89).

3.5. Mixed viral lineage genome assembly and Shannon entropy calculation

WuMV-6 genomes were re-assembled to assess samples that could contain mixed viral lineages. Mean entropy per sample and mean entropy per segment for each sample were calculated (Figure 14). A mean entropy per sample threshold for defining highly variable samples was

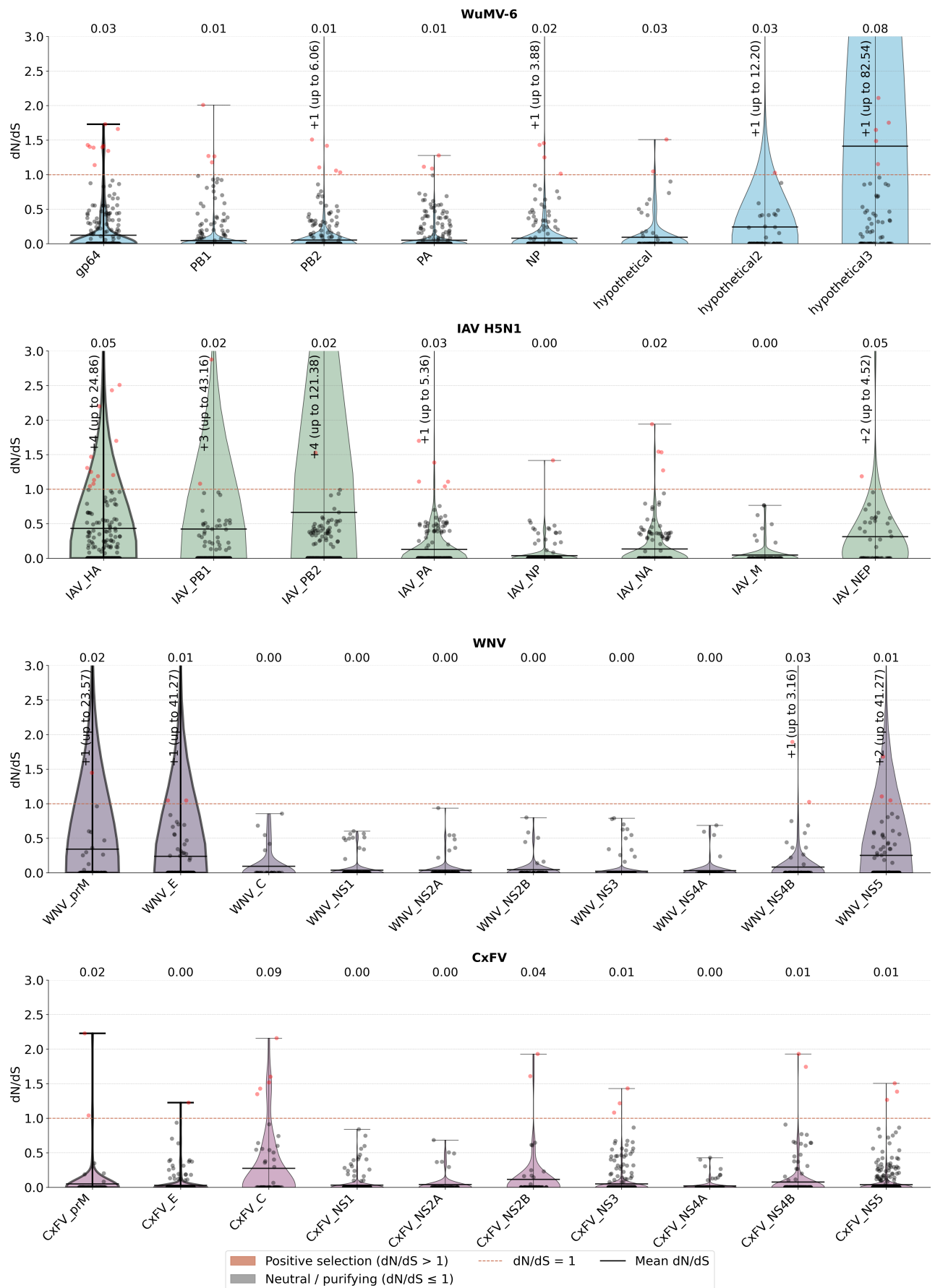


Figure 13. dN/dS values of RNA virus proteins. Values above the violin plots indicate the amount of sites under positive selection. Surface proteins are highlighted with thicker black edgecolor.

empirically chosen to be 0.1. It was based on assessment of the relationship between mean entropy per sample and sequencing depth. Out of 122 samples that were assembled, 48 samples (39%) were above the set entropy threshold.

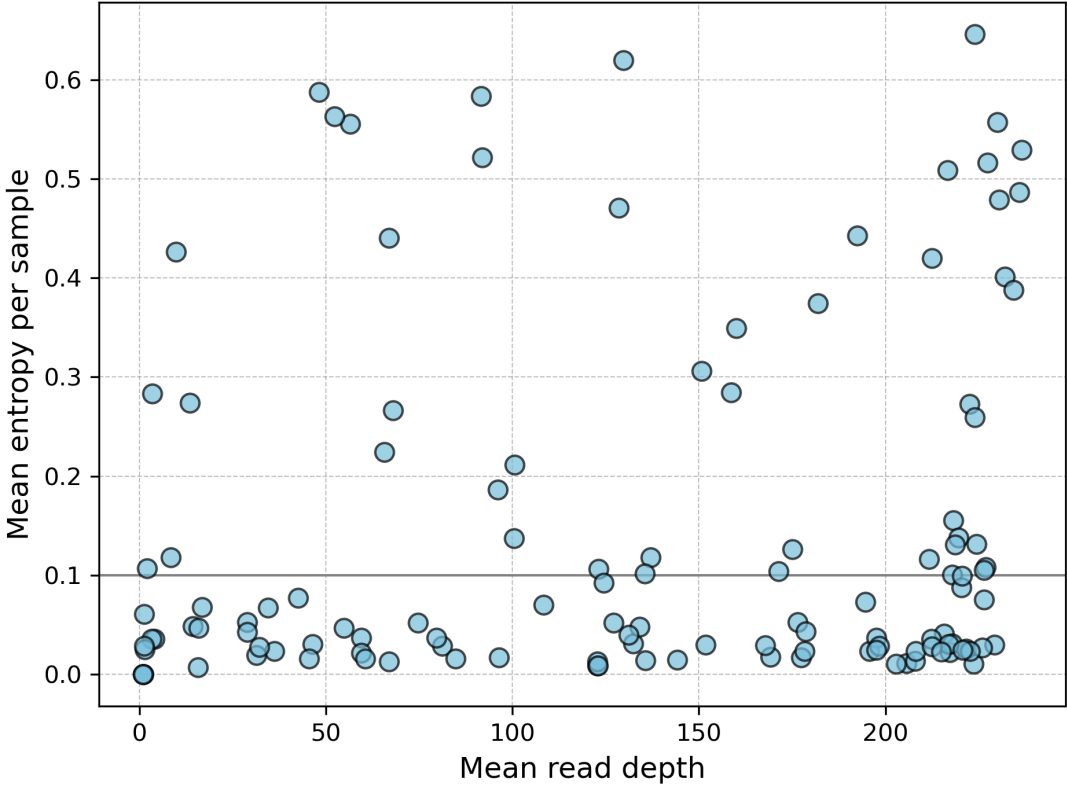


Figure 14. Mean read depth vs mean sample entropy. Each point represents a sample. Grey line at $y = 0.1$ represents the set entropy threshold.

Spearman correlation between mean sample entropy and mean read depth shows weak positive correlation (correlation coefficient = 0.19, pvalue = 0.03).

Per-segment entropy for each sample was also assessed (Figure 15). A tendency of samples having similar mean Shannon entropy across all genome segments is seen. Some samples that were used in the initial analysis did not pass the quality threshold for this genome assembly.

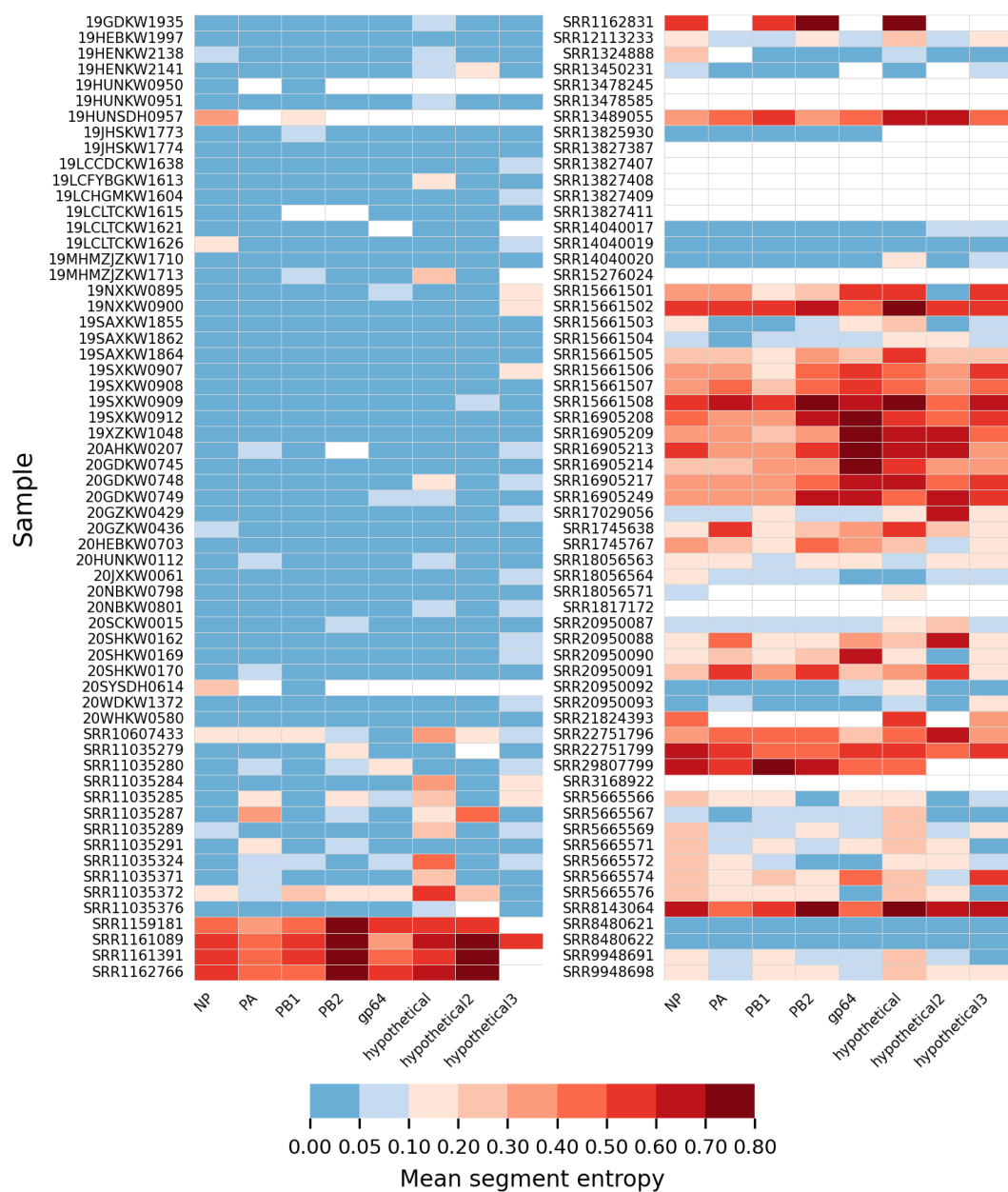


Figure 15. Per-segment (columns) Shannon entropy of each sample (rows). White cells represent segments that had low read depth (below 10 reads).

4. Discussion

This is the first study to analyse the correlates of WuMV-6's spatial diffusion using phylogeography. To compare if other RNA viruses: ISVs or arboviruses, share similar environmental predictors influencing their spatial spread, we extended the study by applying the discrete phylogeographic workflow to both: well-characterised model viruses with known correlates of diffusion (e.g. WNV, IAV), and less studied insect viruses (e.g. Phasivirus phasiense, *Culex pipiens* Ifla-like virus 2).

Initial molecular clock analyses of WuMV-6 conducted by Dudas and Batson, 2023 and used a dataset of 27 WuMV-6 genomes. We now expand the study and use a dataset that is nearly 5 times larger.

4.1. Rapid WuMV-6 migration

Both, continuous and discrete phylogeography results estimate that the origin of WuMV-6 spread is likely to be Yunnan province, China. Virus moved from Yunnan throughout the whole country and had introductions to other continents: Europe, Australia, Africa, North America. However, half of the sequences used in WuMV-6 NP discrete phylogeography come from China, most of them – from the Yunnan province, therefore, it is likely overrepresented in the global analysis. Because China has good sampling distribution, it's local spatial reconstruction (movements between provinces) is likely more reliable.

Previous study of WuMV-6 showed that the virus is unusually mobile, as reassortment network showed recent contacts between distant lineages (e.g. between China and the US, the US and Australia) (Dudas and Batson, 2023). Fast global movement is supported by discrete phylogeography analysis, where samples from distant countries are estimated to share a common ancestor in the past 10-15 years (China and the United States, China and Burkina Faso, China and Australia). The current dataset suggests that there were two independent introductions from Asia to North America. First introduction originated from Yunnan province and jumped to California (MRCA dating around 2010) and the second introduction – from Tibet to Missouri (around 2018). However, this reconstruction is biased by available sampling locations. To assess the precise origin of WuMV-6 jumps, samples from more Asian countries would be needed.

Molecular clock analysis of WuMV-6 shows that the global lineage, excluding samples from the Northern Europe, share a common ancestor in the 2000s. Previous findings estimated ancestor of the global lineage to be a few years older (Dudas and Batson, 2023). However, our analysis estimates that the MRCA of all available WuMV-6 NP samples dates to 1987, while Dudas and Batson, 2023 inferred a much older root, dating to 1950s. This is likely due to inclusion of newly available samples, which may represent the previously unsampled diversity, resulting in a more recent estimate of the tree root (Linder et al., 2005).

Consistently with findings of Dudas and Batson, 2023, a distinct lineage from European countries around the Baltic sea is identified. Although WuMV-6 is seen to migrate globally very rapidly, this lineage appears to have remained separate from the rest of the global lineage

and pushes back the estimated date to the MRCA of all samples. This unusual population structure might be influenced by *Wolbachia* bacteria, which are known to be able to regulate ISV replication in mosquitoes, specifically decrease replication or transmission of RNA arboviruses (Bian et al., 2010; Altinli et al., 2020). *Wolbachia* in *Culex* mosquitoes has been described to have 5 main haplotypes distributed globally, out of which one is found distinctly in Europe (Figure 16) (Dumas et al., 2013). A previous study by Altinli et al., 2020 showed that mosquitoes infected with different *Wolbachia* strains differed in how likely they were to be infected with *Culex pipiens* densovirus (ssDNA). Although one strain was associated with higher viral prevalence, *Wolbachia* was not identified as the major force shaping its diversity. Therefore, even if not being the main factor, *Wolbachia* might play a part in WuMV-6 distinct population structure, separating two major virus lineages.

4.2. Correlates of WuMV-6 spatial spread

We analysed whether environmental conditions, climate, human activity, and bird movement-related factors influence the spatial spread of WuMV-6. We performed continuous phylogeographic analyses with additional environmental variable assessments, and discrete phylogeography analyses with integrated GLMs, representing predictors which might be associated to virus spread.

Although continuous phylogeography and SERAPHIM calculations did not show any significant variables influencing WuMV-6 spatial diffusion, discrete phylogeography showed significant effect of distance-based predictors, with regions sharing land borders having a positive effect, while larger distances between regions – negative. Geographic distance is often found to be a significant predictor for vertebrate or arbovirus spread, and it is inferred as significant for both arboviruses included in this study (Gass et al., 2023). However, phylogeographic studies that would focus on ISVs and drivers of their spread are limited. None of the ISVs included in this study (CxFV, CxIf1a-like2, SIGV) showed distance to be a significant predictor. This may be due to smaller dataset sizes and more limited geographic sampling, which could reduce the ability to reliably assess the significance of distance and other predictors.

The number of bird farms in the origin country was assessed as positively correlated for WuMV-6, negatively – for WNV cluster 1. However, distribution of predictor and transition event values for WuMV-6 were not consistent. Most of the signal is coming from China's provinces, as they have larger numbers of poultry farms (Supplementary E1). While the majority of transitions between regions are also observed in China, they are not distributed in the same pattern. Therefore, the number of bird farms does not seem to be the main factor influencing WuMV-6 spread.

The East Atlantic bird migration flyway is inferred as a strongly positive predictor for WuMV-6. However, the signal for this predictor is most likely only coming from the Nordic European samples (including Sweden, United Kingdom, Belgium), whose distribution is very similar to the flyway path. While the birds migrating via the East Atlantic flyway could be influencing WuMV-6 spread and separating the Nordic-European and global lineages, other

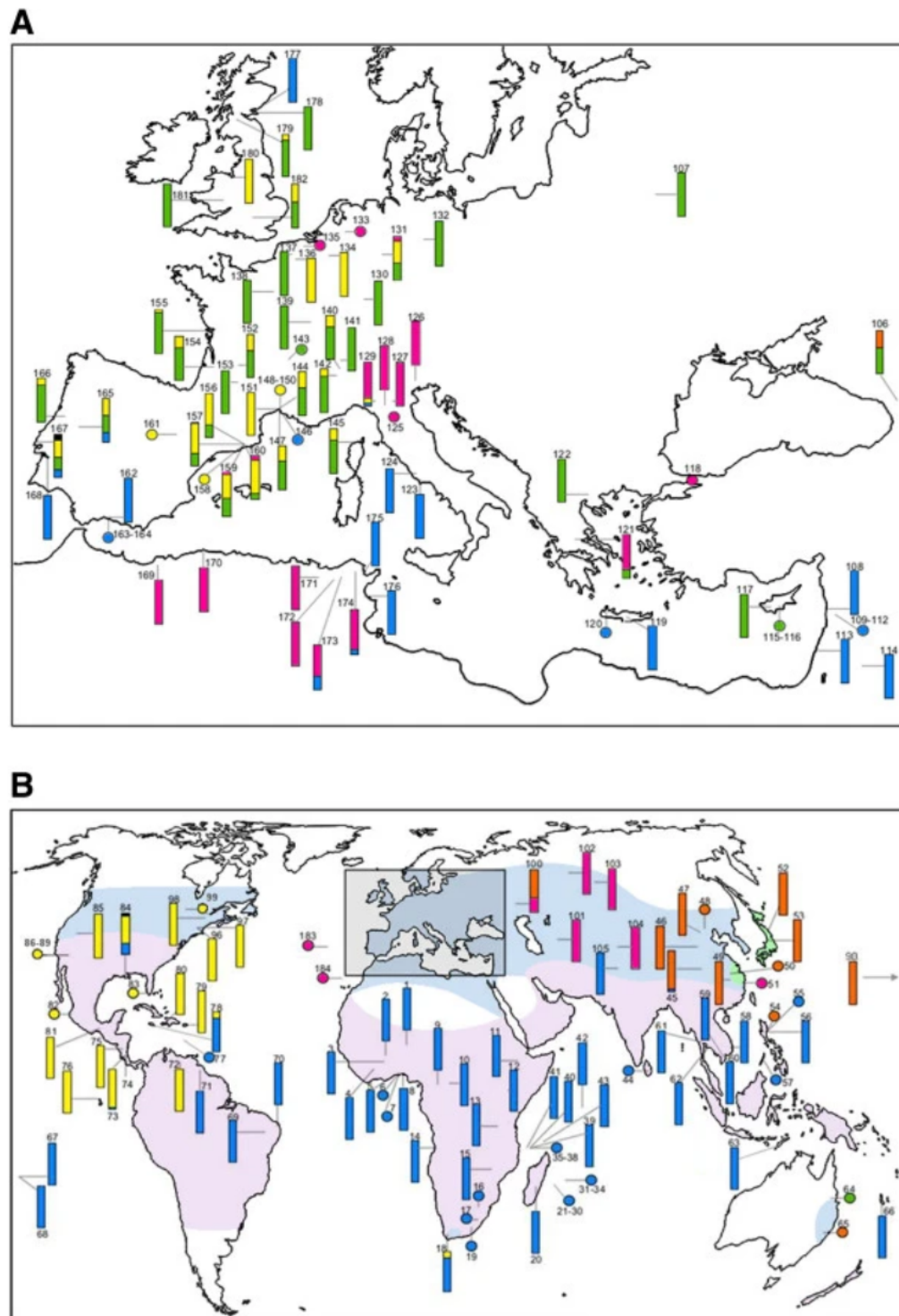


Figure 16. Geographic distribution of *Wolbachia* population structure in mosquitoes. Bars represent natural population, dots – laboratory strains. Colored by haplotypes. Dumas et al., 2013

flyways which include sampling regions do not appear to be significant.

Arboviruses that are known to be transmitted by birds, e.g. WNV, have been assessed to have bird migration flyways as significant predictors in previous studies (Lu et al., 2024). In this study, WNV cluster 1 also found multiple bird migration-representing predictors to be significant. Contrary to WNV cluster 1, cluster 2 did not have many significant predictors.

However, both clusters of WNV analyses showed that the number of sequences included to the model was a significant factor. These datasets were prepared by taking 1-2 sequences per region per year, so regions that are sampled more consistently are likely influencing the result. Therefore, because both clusters had different sampling regions included in the model, GLMs that were included and inferred as significant are also different.

Similarly to bird farm predictor, forest percentage is inferred as significant only for WuMV-6 PA segment. Its value distribution did not correspond to the distribution of WuMV-6 transition between regions numbers (Supplementary E2). Since the sampling regions of the PA segment are very similar to those of other WuMV-6 segments, with only Tibet, the UK and Missouri absent compared to the NP segment dataset, reassortment events affecting inferred migration patterns between regions may also influence correlations with the tested predictors as well.

While arboviruses have been shown to preferentially circulate in regions with higher temperatures (e.g. Getah virus, WNV) and in this study temperature and precipitation-related predictors were found to be significant in WNV and IAV, WuMV-6 and CxFV did not show similar associations (J. Zhao et al., 2022; Lu et al., 2024).

As this is the first phylogeographic study of WuMV-6, and high-quality data and previous research on insect-specific viruses remain limited, comparisons with previous studies are largely restricted to arbovirus research. Like arboviruses, WuMV-6 consistently shows that geographic distance-based predictors correlate with the number of transitions between regions. Similarly to the avian-transmitted WNV, WuMV-6 has correlation with bird movement-related predictors. However, as these predictors do not explain all variation, there might be another variable we have not considered that is correlated with both – avian densities and WuMV-6 (and inherently *Culex* mosquito) migrations.

4.3. Selective pressure of WuMV-6 and other RNA virus' proteins

Selective pressure analysis indicates that most of WuMV-6 genome is under purifying selection. Consistently with findings of Dudas and Batson, 2023, gp64 is evolving faster than most of its proteome. gp64 amino acid tree length (amino acid substitutions per site) is nearly twice as high as other WuMV-6's proteins, except hypothetical 3, which has unusually high value of 1.6 substitutions per site. Such high variation might be caused by having mixed viral lineages and calling many alternative alleles during consensus calling. Mean dN/dS values of gp64 sites is similar to other WuMV-6 proteins and varies below 0.3 (except hypothetical 3). However, the amount of gp64 sites that are under positive selection is higher than that of conservative proteins: RdRp subunits, nucleoprotein. This suggests that gp64 accumulates mutations fast, yet diversifying selection is occurring at specific, non-random sites, as indicated by low averaged dN/dS values showing mostly purifying selection (Spielman and Wilke, 2015).

IAV tree length results show that proteins that are on the virion surface (HA, NA) and nuclear export protein (NEP) accumulate mutations more rapidly than the other IAV proteins. It is typical for IAV to have higher tree lengths of surface proteins than inner proteins, as they undergo antigenic drift (Wei et al., 2014). We also inferred slightly higher dN/dS values of

IAV RdRp basic subunits (yet lower tree length) than that of the surface protein, although HA has both: high tree length and high dN/dS values. Inferred dN/dS value of NEP protein (0.35) is higher than usually inferred in the literature, where it varies between 0.08-0.20 for avian and/or human infecting lineages of IAV (Wei and Li, 2018; Yao et al., 2024; Zheng et al., 2025). Because the dataset consists of IAV samples from multiple bird species, this might suggest that the specific lineage is undergoing host jumps which drive host-specific adaptations in subunits of RdRp and the surface protein.

WNV results differ from what is usually presented in the literature. Although capsid protein is usually estimated to have the highest dN/dS values across WNV genome, while membrane and envelope proteins have the lowest values, our results show the opposite (Di Giallonardo et al., 2015; Bialosuknia et al., 2019). Because the analysed dataset consisted of sequences clustered at 95% similarity, part of the observed evolutionary signal may reflect mutations that are not maintained in the broader viral population over time. As the amino acid evolution of these proteins does not appear to be substantially faster compared to the rest of the genome, it would be informative to additionally assess how many amino acid changes are unique to individual sequences versus shared among multiple sequences to differentiate between mildly deleterious circulating variation and more persistent variation.

CxFV selective pressure results indicate that most of the genome is under purifying selection with low dN/dS values and only a few individual sites where dN/dS is above 1. Previous studies of insect flaviviruses showed similar results (Morais et al., 2021). Only the capsid protein shows higher number of sites under positive selection and has larger values of amino acid substitution per site. This suggests that the protein has less functional constraint and is susceptible to more changes. More rapid capsid protein evolution compared to the rest of the genome is also seen in previous *Flaviviridae* arbovirus and insect flavivirus studies (Di Giallonardo et al., 2015; Morais et al., 2021). Overall, both analysed flaviviruses show similar rates of amino acid evolution.

Both orthomyxoviruses and CxFV show elevated evolutionary rates of surface proteins, however the difference in rates between the surface protein and internal proteins is non-significant. While surface protein evolution at amino acid level does not prove that WuMV-6 is able to infect a vertebrate host by itself, it suggests that it might have another unknown factor influencing its evolution. Previous experimental studies using pseudoviruses by dr. Milda Norkienė demonstrated that WuMV-6 gp64-pseudotyped lentiviruses are unable to enter vertebrate or non-mosquito insect cells (A549 (human lung carcinoma), Sf9 (moth ovary), SH-SY5Y (human neuroblastoma), U87 (human brain), DF-1 (chicken fibroblast), PBMCs (chicken peripheral blood mononuclear cells), LLC1 (mouse lung carcinoma)). Hence, because the amino acid divergence of WuMV-6 proteins was generally similar to that observed in avian IAV H5N1 surface proteins (with the exception of hypothetical protein 3), it is unlikely that the observed genomic variation in WuMV-6 is solely explained by the presence of mixed viral lineages within samples.

4.4. High Shannon entropy suggests mixed viral lineages in multiple WuMV-6 samples

Because WuMV-6 shows high evolutionary rates, we accounted for potential within-sample variability by analysing if samples contain mixed viral lineages. Since part of the samples we used for the study came from pooled mosquito sequencing, some samples might have multiple distinct WuMV-6 lineages. Mixed lineages occurring in the same sample could create a mosaic between multiple genotypes if they are present in the same sequencing sample without being able to distinguish which nucleotide belongs to which lineage. This would increase the number of single nucleotide polymorphisms and falsely infer high sequence diversity in samples with higher read depth, which would allow to infer more variation. Low sample read depth might also influence the calculation of entropy, as lower depth can cause biased diversity estimates (L. Zhao and Illingworth, 2019).

Nearly 40% of per-sample Shannon entropy values were above the set threshold. This suggests that WuMV-6 lineages are quite diverse, as next-generation sequencing errors can only account for up to 1 % of false variability, depending on the platform (Cheng et al., 2023). However, the association between mean entropy per-sample and mean read depth is weak – this might be an indication that lineage diversity is not explained by the quality of sequencing samples (read depth). Comparison of Shannon entropy between samples and genome segments showed that entropy values were similar among segments within the same sample. However, the hypothetical 3 segment more frequently exhibited higher entropy values compared with other WuMV-6 segments, corresponding to its higher dN/dS and amino acid tree lengths. Per-segment mean Shannon entropy analysis also suggests that some samples have higher number of alternative alleles, hence – higher entropy values, suggesting mixed viral lineages might be present in those samples.

By removing samples with mean per-sample Shannon entropy above 0.1 we would remove the majority of highly variable samples. Molecular clock and phylogeographic analyses should be repeated with reassembled sequences to compare if rapid evolution can be explained by sample quality and mixed viral lineages.

4.5. Future directions

To expand on this topic further, the first step would be to redo all phylogeographic analyses with reassembled WuMV-6 genomes that pass the Shannon entropy threshold. It would also be informative to include the Gamma-distributed rate heterogeneity parameter to BEAST models to account for unequal evolution rates across all sites. Reassembling CxFV genomes and repeating selective pressure analysis would also be interesting, as other ISVs might have the same genome assembly problem with mixed viral lineages as WuMV-6.

Obtaining more WuMV-6 samples from more countries across different continents would also be beneficial, as phylogeographic analyses are heavily influenced by uneven sampling.

To assess the connection between WuMV-6 and *Wolbachia* bacteria distribution, it would be

interesting to assess the available mosquito metatranscriptomic samples and infer if *Wolbachia* strains can be distinguished from the samples to compare it with WuMV-6 Nordic European and global lineages distribution. Although metagenomic sequencing would be more efficient for bacterial analysis, metatranscriptomic studies were shown to be able to identify strain-level variation for fungi, although it remains challenging to detect microorganisms more precisely than at the species-level (Marcelino et al., 2019).

Conclusions

The following conclusions can be drawn:

- WuMV-6 spatial spread reconstruction using continuous phylogeography methods suggest the virus origin in Yunnan province, China, and the most recent common ancestor dating to 2002.85.
- Discrete phylogeography analysis shows consistent distinct distance-based effects – positive associations between WuMV-6 migration and sharing land borders and negative associations with more distant region centroids, which is consistent with arboviruses infecting vertebrate hosts. As with avian-infecting WNV, WuMV-6 has positive associations with bird movement-representing predictors, yet it is not inferred to be the main factor for WuMV-6 spread.
- Evolutionary analyses of WuMV-6 and other RNA viruses do not show significant differences between surface protein and inner protein amino acid evolution rates. However, WuMV-6 surface protein gp64 amino acid substitution per site rate remains higher than the rest of the proteome with the exception of hypothetical 3, which has higher Shannon entropy values comparing to other genome segments.
- WuMV-6 genome assembly using developed pipeline for mixed viral lineage assembly shows high nucleotide diversity (mean Shannon entropy > 0.1) in 39% of used samples.

Acknowledgements

I want to sincerely thank my supervisor dr. Gytis Dudas for mentorship, continuous support, inspiration, and for providing opportunities that extended far beyond the scope of this work.

I also want to thank:

- Linas Gasparavičius and Martynas Smičius for assembling initial WuMV-6 genomes;
- Lander de Coninck for sharing unpublished WuMV-6 sequences from Belgium and Burkina Faso;
- Jacco Boon and Drew Karp for sharing unpublished WuMV-6 sequences from Missouri;
- Simon Dellicour for help with SERAPHIM analysis;
- Miglė Gabrielaitė for advice on the development of WuMV-6 genome assembly pipeline;
- Barney Potter for advice on discrete phylogeography approaches;
- Information Technology Research Center of Vilnius University for providing high performance computing resources.

Author's Personal Contribution

Data analysis and visualization, writing (original draft) was done by the author.
Project conceptualization, draft revision was done by dr. Gytis Dudas.

VILNIAUS UNIVERSITETAS
GYVYBĖS MOKSLŲ CENTRAS

Aistė Židonytė

Magistro baigiamasis darbas

Uhano uodų viruso 6 ir kitų RNR virusų filogeografinė analizė: aplinkos ir šeimininko veiksniai, susiję su plitimu erdvėje

SANTRAUKA

Uhano uodų virusas 6 (WuMV-6) – tai *Orthomyxoviridae* šeimai priklausančias RNR virusas, giminingas gripo virusams. Nors iki dabar WuMV-6 buvo rastas infekuojantis tik uodus, tačiau jis identifikuojamas tyrimuose įvairiose pasaulio šalyse. Ankstesnis tyrimas parodė, kad WuMV-6 paviršiaus baltymas aminorūgščių lygmenyje evoliucionuoja greičiau nei kiti jo baltymai. Jis taip pat paplito globaliai per trumpą laiko tarpą. Tai indikuoja, kad globalus WuMV-6 plitimas gali būti susijęs su mobiliu stuburiniu šeimininku.

Šio tyrimo tikslas buvo atlikti WuMV-6 ir kitų vabzdžiams specifinių virusų bei arbovirusų filogeografinės analizės siekiant identifikuoti, kokie veiksniai koreliuoja su jų plitimu. Pritaikydami tęstines ir diskrečias filogeografinės analizės identifikavome, kad su atstumu susiję veiksniai daro reikšmingą įtaką migracijų tarp regionų skaičiui. Taip pat identifiukuota, kad paukščių ūkių skaičius ir didesnė miškų dalis regione teigiamai koreliuoja su didesniu migracijos mastu, tačiau tai nepaaiškina visos variacijos.

Kaip ir ankstesnių paukščių gripo (IAV) H5N1 tyrimų atveju, WuMV-6 atrankos spaudimo analizė rodo, kad jo paviršiaus baltymas gp64 aminorūgščių lygmeniu evoliucionuoja greičiau nei kiti baltymai. Tačiau WuMV-6 baltymų, įskaitant gp64, dN/dS vertės buvo panašios į tas, kurios buvo nustatytos likusioje genomo dalyje, ir į IAV baltymų dN/dS vertes. Palyginus *Flaviviridae* šeimos arboviruso ir vabzdžiams būdingo viruso baltymus nustatyta, kad jų evoliucijos greičiai yra panašūs, nors *Culex pipiens* flaviviruso kapsidės baltymo aminorūgščių skirtumai ir dN/dS santykis – didesni.

Siekiant atsižvelgti į WuMV-6 genomų įvairovę, buvo sukurta mišrių virusų linijų genomų surinkimo sistema. 39% iš naujo surinktų WuMV-6 genomų viršijo nustatytą *Shannon* entropijos ribą, indikuojant, kad mėginio kokybė buvo prasta arba jame buvo daugiau nei viena viruso linija.

VILNIUS UNIVERSITY
LIFE SCIENCES CENTER

Aistė Židonytė

Master's thesis

Phylogeographic Analysis of Wuhan Mosquito Virus 6 and Other RNA Viruses: Environmental and Host Correlates of Spatial Diffusion

ABSTRACT

Wuhan mosquito virus 6 (WuMV-6) is an RNA virus that belongs to the *Orthomyxoviridae* family and is related to viruses like Influenza. Although by now WuMV-6 has only been found to infect mosquitoes, it has been identified in multiple studies all over the world. Previous study has found that WuMV-6 surface protein evolves faster at amino acid level than the rest of its' proteins. It has also spread globally in a short period of time. This suggests that the global spread might be facilitated by a mobile vertebrate host.

The goal of this study was to conduct phylogeographic analysis of WuMV-6 and other insect-specific viruses and arboviruses to infer what variables are influencing its' spread. By applying continuous and discrete phylogeography methods, we identified that distance-based predictors significantly influence the number of transitions between sampling regions. It was also inferred that the number of bird farms and the percentage higher percentage of forest area in the region positively correlates to more migrations, but does not explain the whole variation.

Similarly to previous studies of avian Influenza (IAV) H5N1, selective pressure analysis of WuMV-6 suggests that its surface protein gp64 is evolving faster at amino acid level than other proteins. However, the dN/dS values of WuMV-6 proteins, including gp64, were comparable to those observed across the rest of the genome and to those of IAV proteins. Comparison of arboviral and insect-specific virus proteins within the *Flaviviridae* family showed generally similar evolutionary rates, although the *Culex pipiens* flavivirus capsid protein exhibited elevated amino acid divergence and higher dN/dS values. However, overall comparison between virus surface and inner proteins did not show significant difference.

To address the variability in WuMV-6 genomes, pipeline for mixed viral lineage genome assembly was developed. 39% of reassembled WuMV-6 genomes were above the set Shannon entropy threshold, suggesting either low sample quality or presence of more than one viral lineage in the sample.

References

- Achee, N. L., Grieco, J. P., Vatandoost, H., Seixas, G., Pinto, J., Ching-NG, L., Martins, A. J., Juntarajumnong, W., Corbel, V., Gouagna, C., David, J.-P., Logan, J. G., Orsborne, J., Marois, E., Devine, G. J., & Vontas, J. (2019). Alternative strategies for mosquito-borne arbovirus control. *PLOS Neglected Tropical Diseases*, *13*(1), e0006822. <https://doi.org/10.1371/journal.pntd.0006822>
- Aiewsakun, P., & Katzourakis, A. (2016). Time-dependent rate phenomenon in viruses. *Journal of Virology*, *90*(16), 7184–7195. <https://doi.org/10.1128/jvi.00593-16>
- Allison, A. B., Ballard, J. R., Tesh, R. B., Brown, J. D., Ruder, M. G., Keel, M. K., Munk, B. A., Mickley, R. M., Gibbs, S. E. J., Travassos da Rosa, A. P. A., Ellis, J. C., Ip, H. S., Shearn-Bochsler, V. I., Rogers, M. B., Ghedin, E., Holmes, E. C., Parrish, C. R., & Dwyer, C. (2014). Cyclic avian mass mortality in the northeastern united states is associated with a novel orthomyxovirus. *Journal of Virology*, *89*(2), 1389–1403. <https://doi.org/10.1128/jvi.02019-14>
- Altinli, M., Lequime, S., Atyame, C., Justy, F., Weill, M., & Sicard, M. (2020). Wolbachia modulates prevalence and viral load of culex pipiens densovirus in natural populations [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.15609>]. *Molecular Ecology*, *29*(20), 4000–4013. <https://doi.org/10.1111/mec.15609>
- Alvarez-Munoz, S., Upegui-Porras, N., Gomez, A. P., & Ramirez-Nieto, G. (2021). Key factors that enable the pandemic potential of RNA viruses and inter-species transmission: A systematic review. *Viruses*, *13*(4). <https://doi.org/10.3390/v13040537>
- Añez, G., Grinev, A., Chancey, C., Ball, C., Akolkar, N., Land, K. J., Winkelman, V., Stramer, S. L., Kramer, L. D., & Rios, M. (2013). Evolutionary dynamics of west nile virus in the united states, 1999–2011: Phylogeny, selection pressure and evolutionary time-scale analysis. *PLOS Neglected Tropical Diseases*, *7*(5), e2245. <https://doi.org/10.1371/journal.pntd.0002245>
- Avise, J. (2009). Phylogeography: Retrospect and prospect. *Journal of Biogeography - J BIOGEOGR*, *36*. <https://doi.org/10.1111/j.1365-2699.2008.02032.x>
- Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., Reeb, C. A., & Saunders, N. C. (1987). Intraspecific phylogeography: The mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, *18*, 489–522. Retrieved March 24, 2026, from <https://www.jstor.org/stable/2097141>
- Baele, G., Ji, X., Hassler, G. W., McCrone, J. T., Shao, Y., Zhang, Z., Holbrook, A. J., Lemey, P., Drummond, A. J., Rambaut, A., & Suchard, M. A. (2025). BEAST x for bayesian phylogenetic, phylogeographic and phylodynamic inference. *Nature Methods*, *22*(8), 1653–1656. <https://doi.org/10.1038/s41592-025-02751-x>
- Baele, G., & Lemey, P. (2013). Bayesian evolutionary model testing in the phylogenomics era: Matching model complexity with computational efficiency. *Bioinformatics*, *29*(16), 1970–1979. <https://doi.org/10.1093/bioinformatics/btt340>

- Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M. A., & Alekseyenko, A. V. (2012). Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular Biology and Evolution*, 29(9), 2157–2167. <https://doi.org/10.1093/molbev/mss084>
- Bassi, C., Guerriero, P., Pierantoni, M., Callegari, E., & Sabbioni, S. (2022). Novel virus identification through metagenomics: A systematic review. *Life*, 12(12). <https://doi.org/10.3390/life12122048>
- Batson, J., Dudas, G., Haas-Stapleton, E., Kistler, A. L., Li, L. M., Logan, P., Ratnasiri, K., & Retallack, H. (2021). Single mosquito metatranscriptomics identifies vectors, emerging pathogens and reservoirs in one assay. *eLife*, 10, e68353. <https://doi.org/10.7554/eLife.68353>
- Bermingham, E., & Moritz, C. (1998). Comparative phylogeography: Concepts and applications. *Molecular Ecology*, 7(4), 367–369. <https://doi.org/10.1046/j.1365-294x.1998.00424.x>
- Bialosuknia, S. M., Tan, Y., Zink, S. D., Koetzner, C. A., Maffei, J. G., Halpin, R. A., Mueller, E. A., Novotny, M., Shilts, M., Fedorova, N. B., Amedeo, P., Das, S. R., Pickett, B., Kramer, L. D., & Ciota, A. T. (2019). Evolutionary dynamics and molecular epidemiology of west nile virus in new york state: 1999–2015. *Virus Evolution*, 5(2), vez020. <https://doi.org/10.1093/ve/vez020>
- Bian, G., Xu, Y., Lu, P., Xie, Y., & Xi, Z. (2010). The endosymbiotic bacterium wolbachia induces resistance to dengue virus in aedes aegypti. *PLOS Pathogens*, 6(4), e1000833. <https://doi.org/10.1371/journal.ppat.1000833>
- Bisht, K., & te Velhuis, A. J. W. (2022). Decoding the role of temperature in RNA virus infections. *mBio*, 13(5), e02021–22. <https://doi.org/10.1128/mbio.02021-22>
- Bittar, C., Machado, D. C., Vedovello, D., Ullmann, L. S., Rahal, P., Araújo Junior, J. P., & Nogueira, M. L. (2016). Genome sequencing and genetic characterization of culex flavivirus (CxFV) provides new information about its genotypes. *Virology Journal*, 13, 158. <https://doi.org/10.1186/s12985-016-0614-3>
- Black, A., & Dudas, G. (2024, March 18). *The applied genomic epidemiology handbook: A practical guide to leveraging pathogen genomic data in public health*. Chapman; Hall/CRC. <https://doi.org/10.1201/9781003409809>
- Blitvich, B. J. (2008). Transmission dynamics and changing epidemiology of west nile virus. *Animal Health Research Reviews*, 9(1), 71–86. <https://doi.org/10.1017/S1466252307001430>
- Blitvich, B. J., & Firth, A. E. (2015). Insect-specific flaviviruses: A systematic review of their discovery, host range, mode of transmission, superinfection exclusion potential and genomic organization. *Viruses*, 7(4), 1927–1959. <https://doi.org/10.3390/v7041927>
- Bollback, J. P. (2002). Bayesian model adequacy and choice in phylogenetics. *Molecular Biology and Evolution*, 19(7), 1171–1180. <https://doi.org/10.1093/oxfordjournals.molbev.a004175>

- Borsky, S., Hennighausen, H., Leiter, A., & Williges, K. (2003). CITES and the zoonotic disease content in international wildlife trade. *Environmental and Resource Economics*, 76(4), 1001–1017. <https://doi.org/10.1007/s10640-020-00456-7>
- Bray, M. (2008). Highly pathogenic RNA viral infections: Challenges for antiviral research. *Antiviral Research*, 78(1), 1–8. <https://doi.org/10.1016/j.antiviral.2007.12.007>
- Bromham, L., Duchêne, S., Hua, X., Ritchie, A. M., Duchêne, D. A., & Ho, S. Y. W. (2018). Bayesian molecular dating: Opening up the black box. *Biological Reviews*, 93(2), 1165–1191. <https://doi.org/10.1111/brv.12390>
- Brown, R. P., & Yang, Z. (2011). Rate variation and estimation of divergence times using strict and relaxed clocks. *BMC Evolutionary Biology*, 11(1), 271. <https://doi.org/10.1186/1471-2148-11-271>
- Carvalho, V. L., & Long, M. T. (2021). Insect-specific viruses: An overview and their relationship to arboviruses of concern to humans and animals. *Virology*, 557, 34–43. <https://doi.org/10.1016/j.virol.2021.01.007>
- CDC, C. f. D. C. (2003). *Update: Multistate outbreak of monkeypox — illinois, indiana, kansas, missouri, ohio, and wisconsin, 2003*. Retrieved March 10, 2026, from <https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5227a5.htm>
- Chen, S. (2025). Fastp 1.0: An ultra-fast all-round tool for FASTQ data quality control and preprocessing. *iMeta*, 4(5), e70078. <https://doi.org/10.1002/imt2.70078>
- Cheng, C., Fei, Z., & Xiao, P. (2023). Methods to improve the accuracy of next-generation sequencing. *Frontiers in Bioengineering and Biotechnology*, 11. <https://doi.org/10.3389/fbioe.2023.982111>
- Cholleti, H., Hayer, J., Mulandane, F. C., Falk, K., Fafetine, J., Berg, M., & Blomström, A.-L. (2018). Viral metagenomics reveals the presence of highly divergent quaranjavirus in rhipicephalus ticks from mozambique. *Infection Ecology & Epidemiology*, 8(1), 1478585. <https://doi.org/10.1080/20008686.2018.1478585>
- Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11), 1422–1423. <https://doi.org/10.1093/bioinformatics/btp163>
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2), giab008. <https://doi.org/10.1093/gigascience/giab008>
- De Coninck, L., Soto, A., Wang, L., De Wolf, K., Smits, N., Deblauwe, I., Mbigba Donfack, K. C., Müller, R., Delang, L., & Matthijnsens, J. (2024). Lack of abundant core virome in culex mosquitoes from a temperate climate region despite a mosquito species-specific virome. *mSystems*, 9(6), e00012–24. <https://doi.org/10.1128/msystems.00012-24>

- Dellicour, S., Gill, M. S., Faria, N. R., Rambaut, A., Pybus, O. G., Suchard, M. A., & Lemey, P. (2021). Relax, keep walking — a practical guide to continuous phylogeographic inference with BEAST. *Molecular Biology and Evolution*, 38(8), 3486–3493. <https://doi.org/10.1093/molbev/msab031>
- Dellicour, S., Lequime, S., Vrancken, B., Gill, M. S., Bastide, P., Gangavarapu, K., Matteson, N. L., Tan, Y., du Plessis, L., Fisher, A. A., Nelson, M. I., Gilbert, M., Suchard, M. A., Andersen, K. G., Grubaugh, N. D., Pybus, O. G., & Lemey, P. (2020). Epidemiological hypothesis testing using a phylogeographic and phylodynamic framework. *Nature Communications*, 11(1), 5620. <https://doi.org/10.1038/s41467-020-19122-z>
- Dellicour, S., Rose, R., Faria, N. R., Lemey, P., & Pybus, O. G. (2016). SERAPHIM: Studying environmental rasters and phylogenetically informed movements. *Bioinformatics*, 32(20), 3204–3206. <https://doi.org/10.1093/bioinformatics/btw384>
- Dellicour, S., Vrancken, B., Trovão, N. S., Fargette, D., & Lemey, P. (2018). On the importance of negative controls in viral landscape phylogeography. *Virus Evolution*, 4(2). <https://doi.org/10.1093/ve/vey023>
- Di Giallonardo, F., Geoghegan, J. L., Docherty, D. E., McLean, R. G., Zody, M. C., Qu, J., Yang, X., Birren, B. W., Malboeuf, C. M., Newman, R. M., Ip, H. S., & Holmes, E. C. (2015). Fluid spatial dynamics of west nile virus in the united states: Rapid spread in a permissive host environment. *Journal of Virology*, 90(2), 862–872. <https://doi.org/10.1128/jvi.02305-15>
- Domingo, E., Escarmís, C., Sevilla, N., Moya, A., Elena, S. F., Quer, J., Novella, I. S., & Holland, J. J. (1996). Basic concepts in RNA virus evolution. *The FASEB Journal*, 10(8), 859–864. <https://doi.org/10.1096/fasebj.10.8.8666162>
- Dorigatti, I., Gaythorpe, K. A. M., Cox, V. M., Windram, F. A., & Cator, L. (2025). Priorities for modelling arbovirus transmission under climate change. *Trends in Molecular Medicine*, 31(10), 885–894. <https://doi.org/10.1016/j.molmed.2025.05.010>
- Drummond, A., Pybus, O. G., & Rambaut, A. (2003). Inference of viral evolutionary rates from molecular sequences. 54, 331–358. [https://doi.org/10.1016/S0065-308X\(03\)54008-8](https://doi.org/10.1016/S0065-308X(03)54008-8)
- 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., Nicholls, G. K., Rodrigo, A. G., & Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, 161(3), 1307–1320. <https://doi.org/10.1093/genetics/161.3.1307>
- Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(1), 214. <https://doi.org/10.1186/1471-2148-7-214>
- Duchêne, S., Geoghegan, J. L., Holmes, E. C., & Ho, S. Y. (2016). Estimating evolutionary rates using time-structured data: A general comparison of phylogenetic methods. *Bioinformatics*, 32(22), 3375–3379. <https://doi.org/10.1093/bioinformatics/btw421>

- Dudas, G., & Batson, J. (2023). Accumulated metagenomic studies reveal recent migration, whole genome evolution, and undiscovered diversity of orthomyxoviruses. *Journal of Virology*, 97(10), e01056–23. <https://doi.org/10.1128/jvi.01056-23>
- Dumas, E., Atyame, C. M., Milesi, P., Fonseca, D. M., Shaikevich, E. V., Unal, S., Makoundou, P., Weill, M., & Duron, O. (2013). Population structure of wolbachia and cytoplasmic introgression in a complex of mosquito species. *BMC Evolutionary Biology*, 13(1), 181. <https://doi.org/10.1186/1471-2148-13-181>
- Fall, G., Paola, N. D., Faye, M., Dia, M., Freire, C. C. d. M., Loucoubar, C., Zanotto, P. M. d. A., Faye, O., & Sall, A. A. (2017). Biological and phylogenetic characteristics of west african lineages of west nile virus. *PLOS Neglected Tropical Diseases*, 11(11), e0006078. <https://doi.org/10.1371/journal.pntd.0006078>
- Faria, N. R., Suchard, M. A., Rambaut, A., Streicker, D. G., & Lemey, P. (2013). Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1614), 20120196. <https://doi.org/10.1098/rstb.2012.0196>
- Fitzsimmons, W. J., Woods, R. J., McCrone, J. T., Woodman, A., Arnold, J. J., Yennawar, M., Evans, R., Cameron, C. E., & Lauring, A. S. (2018). A speed–fidelity trade-off determines the mutation rate and virulence of an RNA virus. *PLOS Biology*, 16(6), e2006459. <https://doi.org/10.1371/journal.pbio.2006459>
- Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23), 3150–3152. <https://doi.org/10.1093/bioinformatics/bts565>
- Gangopadhyaya, A., Lole, K., Ghuge, O., Ramdasi, A., Kamble, A., Roy, D., Thakar, S., Nath, A., Sudeep, A. B., & Cherian, S. (2024). Metagenomic analysis of viromes of aedes mosquitoes across india. *Viruses*, 16(1). <https://doi.org/10.3390/v16010109>
- Garry, C. E., & Garry, R. F. (2008). Proteomics computational analyses suggest that baculovirus GP64 superfamily proteins are class III penetrenes. *Virology Journal*, 5(1), 28. <https://doi.org/10.1186/1743-422X-5-28>
- Gass, J. D., Hill, N. J., Damodaran, L., Naumova, E. N., Nutter, F. B., & Runstadler, J. A. (2023). Ecogeographic drivers of the spatial spread of highly pathogenic avian influenza outbreaks in europe and the united states, 2016–early 2022. *International Journal of Environmental Research and Public Health*, 20(11), 6030. <https://doi.org/10.3390/ijerph20116030>
- 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), 713–724. <https://doi.org/10.1093/molbev/mss265>
- Gire, S. K., Goba, A., Andersen, K. G., Sealfon, R. S. G., Park, D. J., Kanneh, L., Jalloh, S., Momoh, M., Fullah, M., Dudas, G., Wohl, S., Moses, L. M., Yozwiak, N. L., Winnicki, S., Matranga, C. B., Malboeuf, C. M., Qu, J., Gladden, A. D., Schaffner, S. F., ... Sabeti,

- P. C. (2014). Genomic surveillance elucidates ebola virus origin and transmission during the 2014 outbreak. *Science*, *345*(6202), 1369–1372. <https://doi.org/10.1126/science.1259657>
- Gómez-Palacio, A., Junca, H., Vivero-Gomez, R. J., Suaza, J., Moreno-Herrera, C. X., Cadavid-Restrepo, G., Pieper, D. H., & Uribe, S. (2025). Metagenomic profiling of the insect-specific virome in non-urban mosquitoes (culicidae: Culicinae) from colombia's northern inter-andean valleys. *PLOS ONE*, *20*(9), e0331552. <https://doi.org/10.1371/journal.pone.0331552>
- Gubler, D. J. (2001). Human arbovirus infections worldwide. *Annals of the New York Academy of Sciences*, *951*(1), 13–24. <https://doi.org/10.1111/j.1749-6632.2001.tb02681.x>
- Guindon, S., & De Maio, N. (2021). Accounting for spatial sampling patterns in bayesian phylogeography. *Proceedings of the National Academy of Sciences*, *118*(52), e2105273118. <https://doi.org/10.1073/pnas.2105273118>
- Gutiérrez-García, T. A., & Vázquez-Domínguez, E. (2011). Comparative phylogeography: Designing studies while surviving the process. *BioScience*, *61*(11), 857–868. <https://doi.org/10.1525/bio.2011.61.11.5>
- Hayes, E. B., Komar, N., Nasci, R. S., Montgomery, S. P., O'Leary, D. R., & Campbell, G. L. (2005). Epidemiology and transmission dynamics of west nile virus disease. *Emerging Infectious Diseases*, *11*(8), 1167–1173. <https://doi.org/10.3201/eid1108.050289a>
- Halachev, M. R., Chan, J. Z.-M., Constantinidou, C. I., Cumley, N., Bradley, C., Smith-Banks, M., Oppenheim, B., & Pallen, M. J. (2014). Genomic epidemiology of a protracted hospital outbreak caused by multidrug-resistant acinetobacter baumannii in birmingham, england. *Genome Medicine*, *6*(11), 70. <https://doi.org/10.1186/s13073-014-0070-x>
- Hanley, K. A., & Weaver, S. C. (2008, January 1). Arbovirus evolution. In *Origin and evolution of viruses* (pp. 351–391). Academic Press. <https://doi.org/10.1016/B978-0-12-374153-0.00016-3>
- Hartner, A.-M., Herath, T., Gomez, L. R. E., Körber, N., Pfeil, J., Escudié, J.-B., Wieler, L., & Irrgang, C. (2025, August 8). The temperature sensitivity of arboviral disease extrinsic incubation periods: A systematic review [ISSN: 2692-8205 Pages: 2025.08.08.669272 Section: New Results]. <https://doi.org/10.1101/2025.08.08.669272>
- Hasegawa, M., Kishino, H., & Yano, T.-a. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, *22*(2), 160–174. <https://doi.org/10.1007/BF02101694>
- Hause, B. M., Ducatez, M., Collin, E. A., Ran, Z., Liu, R., Sheng, Z., Armien, A., Kaplan, B., Chakravarty, S., Hoppe, A. D., Webby, R. J., Simonson, R. R., & Li, F. (2013). Isolation of a novel swine influenza virus from oklahoma in 2011 which is distantly related to human influenza c viruses. *PLOS Pathogens*, *9*(2), e1003176. <https://doi.org/10.1371/journal.ppat.1003176>

- Hill, V., Dellicour, S., Giovanetti, M., & Grubaugh, N. D. (2026). Phylogenetic insights into the transmission dynamics of arthropod-borne viruses. *Nature Reviews Genetics*, 27(1), 47–61. <https://doi.org/10.1038/s41576-025-00854-x>
- Hill, V., Du Plessis, L., Peacock, T. P., Aggarwal, D., Colquhoun, R., Carabelli, A. M., Ellaby, N., Gallagher, E., Groves, N., Jackson, B., McCrone, J. T., O'Toole, Á., Price, A., Sanderson, T., Scher, E., Southgate, J., Volz, E., Barclay, W. S., Barrett, J. C., ... The COVID-19 Genomics UK (COG-UK) Consortium. (2022). The origins and molecular evolution of SARS-CoV-2 lineage b.1.1.7 in the UK. *Virus Evolution*, 8(2), veac080. <https://doi.org/10.1093/ve/veac080>
- Hill, V., Ruis, C., Bajaj, S., Pybus, O. G., & Kraemer, M. U. G. (2021). Progress and challenges in virus genomic epidemiology. *Trends in Parasitology*, 37(12), 1038–1049. <https://doi.org/10.1016/j.pt.2021.08.007>
- Ho, S. Y. W., & Duchêne, S. (2014). Molecular-clock methods for estimating evolutionary rates and timescales [_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.12953>]. *Molecular Ecology*, 23(24), 5947–5965. <https://doi.org/10.1111/mec.12953>
- Hobson-Peters, J., Yam, A. W. Y., Lu, J. W. F., Setoh, Y. X., May, F. J., Kurucz, N., Walsh, S., Prow, N. A., Davis, S. S., Weir, R., Melville, L., Hunt, N., Webb, R. I., Blitvich, B. J., Whelan, P., & Hall, R. A. (2013). A new insect-specific flavivirus from northern australia suppresses replication of west nile virus and murray valley encephalitis virus in co-infected mosquito cells. *PLOS ONE*, 8(2), e56534. <https://doi.org/10.1371/journal.pone.0056534>
- Hodcroft, E. B., De Maio, N., Lanfear, R., MacCannell, D. R., Minh, B. Q., Schmidt, H. A., Stamatakis, A., Goldman, N., & Dessimoz, C. (2021). Want to track pandemic variants faster? fix the bioinformatics bottleneck [Bandiera_abtest: a Cg_type: Comment Subject_term: Public health, Bioinformatics, SARS-CoV-2, Genomics]. *Nature*, 591(7848), 30–33. <https://doi.org/10.1038/d41586-021-00525-x>
- Hollingsworth, B. D., Grubaugh, N. D., Lazzaro, B. P., & Murdock, C. C. (2023). Leveraging insect-specific viruses to elucidate mosquito population structure and dynamics. *PLOS Pathogens*, 19(8), e1011588. <https://doi.org/10.1371/journal.ppat.1011588>
- Holmes, E. C. (2003). Molecular clocks and the puzzle of RNA virus origins. *Journal of Virology*, 77(7), 3893–3897. <https://doi.org/10.1128/jvi.77.7.3893-3897.2003>
- Hong, S. L., Lemey, P., Suchard, M. A., & Baele, G. (2021). Bayesian phylogeographic analysis incorporating predictors and individual travel histories in BEAST. *Current protocols*, 1(4), e98. <https://doi.org/10.1002/cpz1.98>
- Hungnes, O., Jonassen, T. ø., Jonassen, C. M., & Grinde, B. (2000). Molecular epidemiology of viral infections. *APMIS*, 108(2), 81–97. <https://doi.org/10.1034/j.1600-0463.2000.d01-31.x>
- Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>

- Hurk, A. F. v. d., Hall-Mendelin, S., Pyke, A. T., Frentiu, F. D., McElroy, K., Day, A., Higgs, S., & O'Neill, S. L. (2012). Impact of wolbachia on infection with chikungunya and yellow fever viruses in the mosquito vector *aedes aegypti*. *PLOS Neglected Tropical Diseases*, 6(11), e1892. <https://doi.org/10.1371/journal.pntd.0001892>
- Yao, Q., Liu, J., Liu, H., Zhou, Y., Huo, M., Li, Y., Gao, Y., & Ge, Y. (2024). One-health challenge in h9n2 avian influenza: Novel human-avian reassortment virus in guangdong province, china [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1155/2024/9913934>]. *Transboundary and Emerging Diseases*, 2024(1), 9913934. <https://doi.org/10.1155/2024/9913934>
- Kacnik, S., MacIntyre, C., Guarido, M., & Venter, M. (2025). Identification of insect-specific viruses in mosquitoes collected from wildlife and rural areas in north-eastern parts of south africa using a metagenomic RNA sequencing approach. *One Health Outlook*, 7(1), 62. <https://doi.org/10.1186/s42522-025-00185-1>
- Kalkauskas, A., Perron, U., Sun, Y., Goldman, N., Baele, G., Guindon, S., & Maio, N. D. (2021). Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk. *PLOS Computational Biology*, 17(1), e1008561. <https://doi.org/10.1371/journal.pcbi.1008561>
- Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772–780. <https://doi.org/10.1093/molbev/mst010>
- Kessell, A., Hyatt, A., Lehmann, D., Shan, S., Crameri, S., Holmes, C., Marsh, G., Williams, C., Tachedjian, M., Yu, M., Bingham, J., Payne, J., Lowther, S., Wang, J., Wang, L.-F., & Smith, I. (2012). Cygnet river virus, a novel orthomyxovirus from ducks, australia. *Emerging Infectious Diseases*, 18(12), 2044–2046. <https://doi.org/10.3201/eid1812.120500>
- Knowles, L. L. (2009). Statistical phylogeography. *Annual Review of Ecology, Evolution, and Systematics*, 40(1), 593–612. <https://doi.org/10.1146/annurev.ecolsys.38.091206.095702>
- Konstantinidis, K., Dovrolis, N., Kouvela, A., Kassela, K., Rosa Freitas, M. G., Nearchou, A., de Courcy Williams, M., Veletza, S., & Karakasiliotis, I. (2022). Defining virus-carrier networks that shape the composition of the mosquito core virome of a local ecosystem. *Virus Evolution*, 8(1), veac036. <https://doi.org/10.1093/ve/veac036>
- Kosakovsky Pong, S. L., & Frost, S. D. W. (2005). Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution*, 22(5), 1208–1222. <https://doi.org/10.1093/molbev/msi105>
- Kosoy, O. I., Lambert, A. J., Hawkinson, D. J., Pastula, D. M., Goldsmith, C. S., Hunt, D. C., & Staples, J. E. (2015). Novel thogotovirus associated with febrile illness and death, united states, 2014 - volume 21, number 5—may 2015 - emerging infectious diseases journal - CDC. <https://doi.org/10.3201/eid2105.150150>

- Kramer, L. D., Styer, L. M., & Ebel, G. D. (2008). A global perspective on the epidemiology of west nile virus. *Annual Review of Entomology*, 53, 61–81. <https://doi.org/10.1146/annurev.ento.53.103106.093258>
- Kumar, A., Srivastava, P., Sirisena, P., Dubey, S. K., Kumar, R., Shrinet, J., & Sunil, S. (2018). Mosquito innate immunity. *Insects*, 9(3). <https://doi.org/10.3390/insects9030095>
- Layan, M., Müller, N. F., Dellicour, S., De Maio, N., Bourhy, H., Cauchemez, S., & Baele, G. (2023). Impact and mitigation of sampling bias to determine viral spread: Evaluating discrete phylogeography through CTMC modeling and structured coalescent model approximations. *Virus Evolution*, 9(1), vead010. <https://doi.org/10.1093/ve/vead010>
- Lane, C. R., Sherry, N. L., Porter, A. F., Duchene, S., Horan, K., Andersson, P., Wilmot, M., Turner, A., Dougall, S., Johnson, S. A., Sait, M., Silva, A. G. d., Ballard, S. A., Hoang, T., Stinear, T. P., Caly, L., Sintchenko, V., Graham, R., McMahon, J., ... Howden, B. P. (2021). Genomics-informed responses in the elimination of COVID-19 in victoria, australia: An observational, genomic epidemiological study. *The Lancet Public Health*, 6(8), e547–e556. [https://doi.org/10.1016/S2468-2667\(21\)00133-X](https://doi.org/10.1016/S2468-2667(21)00133-X)
- Langedijk, A. C., Vrancken, B., Lebbink, R. J., Wilkins, D., Kelly, E. J., Baraldi, E., Mascareñas de Los Santos, A. H., Danilenko, D. M., Choi, E. H., Palomino, M. A., Chi, H., Keller, C., Cohen, R., Papenburg, J., Pernica, J., Greenough, A., Richmond, P., Martínón-Torres, F., Heikkinen, T., ... Bont, L. J. (2024). The genomic evolutionary dynamics and global circulation patterns of respiratory syncytial virus. *Nature Communications*, 15(1), 3083. <https://doi.org/10.1038/s41467-024-47118-6>
- Laverdeur, J., Desmecht, D., Hayette, M.-P., & Darcis, G. (2024). Dengue and chikungunya: Future threats for northern europe? *Frontiers in Epidemiology*, 4. <https://doi.org/10.3389/fepid.2024.1342723>
- Leal Filho, W., Nagy, G. J., Gbaguidi, G. J., Paz, S., Dinis, M. A. P., Luetz, J. M., & Sharifi, A. (2025). The role of climatic changes in the emergence and re-emergence of infectious diseases: Bibliometric analysis and literature-supported studies on zoonoses. *One Health Outlook*, 7(1), 12. <https://doi.org/10.1186/s42522-024-00127-3>
- 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), e1003932. <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*, 5(9), e1000520. <https://doi.org/10.1371/journal.pcbi.1000520>
- Lemey, P., Rambaut, A., Welch, J. J., & Suchard, M. A. (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology and Evolution*, 27(8), 1877–1885. <https://doi.org/10.1093/molbev/msq067>

- Leung, N. H. L. (2021). Transmissibility and transmission of respiratory viruses. *Nature Reviews Microbiology*, 19(8), 528–545. <https://doi.org/10.1038/s41579-021-00535-6>
- Li, C.-X., Shi, M., Tian, J.-H., Lin, X.-D., Kang, Y.-J., Chen, L.-J., Qin, X.-C., Xu, J., Holmes, E. C., & Zhang, Y.-Z. (2015). Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses (S. P. Goff, Ed.). *eLife*, 4, e05378. <https://doi.org/10.7554/eLife.05378>
- Liang, Y. (2023). Pathogenicity and virulence of influenza. *Virulence*, 14(1), 2223057. <https://doi.org/10.1080/21505594.2023.2223057>
- Linder, H. P., Hardy, C. R., & Rutschmann, F. (2005). Taxon sampling effects in molecular clock dating: An example from the african restionaceae. *Molecular Phylogenetics and Evolution*, 35(3), 569–582. <https://doi.org/10.1016/j.ympev.2004.12.006>
- Lu, L., Leigh Brown, A. J., & Lycett, S. J. (2017). Quantifying predictors for the spatial diffusion of avian influenza virus in china. *BMC Evolutionary Biology*, 17(1), 16. <https://doi.org/10.1186/s12862-016-0845-3>
- Lu, L., Zhang, F., Munnink, B. B. O., Munger, E., Sikkema, R. S., Pappa, S., Tsioka, K., Sinigaglia, A., Molin, E. D., Shih, B. B., Günther, A., Pohlmann, A., Ziegler, U., Beer, M., Taylor, R. A., Bartumeus, F., Woolhouse, M., Aarestrup, F. M., Barzon, L., ... Koopmans, M. P. G. (2024). West nile virus spread in europe: Phylogeographic pattern analysis and key drivers. *PLOS Pathogens*, 20(1), e1011880. <https://doi.org/10.1371/journal.ppat.1011880>
- Ma, W. (2022). Orthomyxoviridae [Section: 57]. In *Veterinary microbiology* (pp. 573–588). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781119650836.ch57>
- Madewell, Z. J. (2020). Arboviruses and their vectors. *Southern Medical Journal*, 113(10), 520–523. <https://doi.org/10.14423/SMJ.00000000000001152>
- Magee, D., Taylor, J. E., & Scotch, M. (2018). The effects of sampling location and predictor point estimate certainty on posterior support in bayesian phylogeographic generalized linear models. *Scientific Reports*, 8(1), 5905. <https://doi.org/10.1038/s41598-018-24264-8>
- May, F. J., Davis, C. T., Tesh, R. B., & Barrett, A. D. T. (2011). Phylogeography of west nile virus: From the cradle of evolution in africa to eurasia, australia, and the americas. *Journal of Virology*, 85(6), 2964–2974. <https://doi.org/10.1128/jvi.01963-10>
- Maio, N. D., Wu, C.-H., O'Reilly, K. M., & Wilson, D. (2015). New routes to phylogeography: A bayesian structured coalescent approximation. *PLOS Genetics*, 11(8), e1005421. <https://doi.org/10.1371/journal.pgen.1005421>
- Marcelino, V. R., Irinyi, L., Eden, J.-S., Meyer, W., Holmes, E. C., & Sorrell, T. C. (2019). Metatranscriptomics as a tool to identify fungal species and subspecies in mixed communities – a proof of concept under laboratory conditions. *IMA Fungus*, 10(1), 12. <https://doi.org/10.1186/s43008-019-0012-8>

- Marcondes, C. B., Contigiani, M., & Gleiser, R. M. (2017). Emergent and reemergent arboviruses in south america and the caribbean: Why so many and why now? *Journal of Medical Entomology*, 54(3), 509–532. <https://doi.org/10.1093/jme/tjw209>
- Martin, D. P., Lemey, P., & Posada, D. (2011). *Analysing recombination in nucleotide sequences - MARTIN - 2011 - molecular ecology resources - wiley online library*. Retrieved March 22, 2026, from <https://onlinelibrary.wiley.com/doi/full/10.1111/j.1755-0998.2011.03026.x>
- Martinez, P. A., Zurano, J. P., Molina, W. F., & Bidau, C. J. (2019). Applications and implications of phylogeography for canid conservation. *Mastozoología Neotropical*. Retrieved March 24, 2026, from <https://mn.sarem.org.ar/article/applications-and-implications-of-phylogeography-for-canid-conservation/>
- Menardo, F., Duchêne, S., Brites, D., & Gagneux, S. (2019). The molecular clock of mycobacterium tuberculosis. *PLOS Pathogens*, 15(9), e1008067. <https://doi.org/10.1371/journal.ppat.1008067>
- Mohd, H. A., Al-Tawfiq, J. A., & Memish, Z. A. (2016). Middle east respiratory syndrome coronavirus (MERS-CoV) origin and animal reservoir. *Virology Journal*, 13(1), 87. <https://doi.org/10.1186/s12985-016-0544-0>
- Morais, P., Trovão, N. S., Abecasis, A. B., & Parreira, R. (2021). Genetic lineage characterization and spatiotemporal dynamics of classical insect-specific flaviviruses: Outcomes and limitations. *Virus Research*, 303, 198507. <https://doi.org/10.1016/j.virusres.2021.198507>
- Morley, V. J., Mendiola, S. Y., & Turner, P. E. (2015). Rate of novel host invasion affects adaptability of evolving RNA virus lineages. *Proceedings of the Royal Society B: Biological Sciences*, 282(1813), 20150801. <https://doi.org/10.1098/rspb.2015.0801>
- Morse, S. S. (1995). Factors in the emergence of infectious diseases. *Emerging Infectious Diseases*, 1(1), 7–15. <https://doi.org/10.3201/eid0101.950102>
- Moury, B., & Desbiez, C. (2020). Host range evolution of potyviruses: A global phylogenetic analysis. *Viruses*, 12(1). <https://doi.org/10.3390/v12010111>
- Nahata, K. D., Bollen, N., Gill, M. S., Layan, M., Bourhy, H., Dellicour, S., & Baele, G. (2021). On the use of phylogeographic inference to infer the dispersal history of rabies virus: A review study. *Viruses*, 13(8). <https://doi.org/10.3390/v13081628>
- Obbard, D. J., & Dudas, G. (2014). The genetics of host–virus coevolution in invertebrates. *Current Opinion in Virology*, 8, 73–78. <https://doi.org/10.1016/j.coviro.2014.07.002>
- Ogunlade, S. T., Meehan, M. T., Adekunle, A. I., Rojas, D. P., Adegboye, O. A., & McBryde, E. S. (2021). A review: Aedes-borne arboviral infections, controls and wolbachia-based strategies. *Vaccines*, 9(1). <https://doi.org/10.3390/vaccines9010032>
- Öhlund, P., Lundén, H., & Blomström, A.-L. (2019). Insect-specific virus evolution and potential effects on vector competence. *Virus Genes*, 55(2), 127–137. <https://doi.org/10.1007/s11262-018-01629-9>
- Olmo, R. P., Tadjro, Y. M. H., Aguiar, E. R. G. R., de Almeida, J. P. P., Ferreira, F. V., Armache, J. N., de Faria, I. J. S., Ferreira, A. G. A., Amadou, S. C. G., Silva, A. T. S.,

- de Souza, K. P. R., Vilela, A. P. P., Babarit, A., Tan, C. H., Diallo, M., Gaye, A., Paupy, C., Obame-Nkoghe, J., Visser, T. M., ... Marques, J. T. (2023). Mosquito vector competence for dengue is modulated by insect-specific viruses. *Nature Microbiology*, 8(1), 135–149. <https://doi.org/10.1038/s41564-022-01289-4>
- Paz, S. (2015). Climate change impacts on west nile virus transmission in a global context. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1665), 20130561. <https://doi.org/10.1098/rstb.2013.0561>
- Pfeffer, M., & Dobler, G. (2010). Emergence of zoonotic arboviruses by animal trade and migration. *Parasites & Vectors*, 3(1), 35. <https://doi.org/10.1186/1756-3305-3-35>
- Podbilewicz, B. (2014). Virus and cell fusion mechanisms. *Annual Review of Cell and Developmental Biology*, 30, 111–139. <https://doi.org/10.1146/annurev-cellbio-101512-122422>
- Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior summarization in bayesian phylogenetics using tracer 1.7. *Systematic Biology*, 67(5), 901–904. <https://doi.org/10.1093/sysbio/syy032>
- 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>
- Rezza, G., Nicoletti, L., Angelini, R., Romi, R., Finarelli, A. C., Panning, M., Cordioli, P., Fortuna, C., Boros, S., Magurano, F., Silvi, G., Angelini, P., Dottori, M., Ciufolini, M. G., Majori, G. C., & Cassone, A. (2007). Infection with chikungunya virus in Italy: An outbreak in a temperate region. *The Lancet*, 370(9602), 1840–1846. [https://doi.org/10.1016/S0140-6736\(07\)61779-6](https://doi.org/10.1016/S0140-6736(07)61779-6)
- Rimstad, E., & Markussen, T. (2020). Infectious salmon anaemia virus—molecular biology and pathogenesis of the infection. *Journal of Applied Microbiology*, 129(1), 85–97. <https://doi.org/10.1111/jam.14567>
- Rosenberg, R. (2015). Detecting the emergence of novel, zoonotic viruses pathogenic to humans. *Cellular and Molecular Life Sciences*, 72(6), 1115–1125. <https://doi.org/10.1007/s00018-014-1785-y>
- Sanjuán, R., & Domingo-Calap, P. (2016). Mechanisms of viral mutation. *Cellular and Molecular Life Sciences*, 73(23), 4433–4448. <https://doi.org/10.1007/s00018-016-2299-6>
- Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M., & Belshaw, R. (2010). Viral mutation rates. *Journal of Virology*, 84(19), 9733–9748. <https://doi.org/10.1128/jvi.00694-10>
- Sastry, K. S., & Zitter, T. A. (2014). Management of virus and viroid diseases of crops in the tropics. In K. S. Sastry & T. A. Zitter (Eds.), *Plant virus and viroid diseases in the tropics: Volume 2: Epidemiology and management* (pp. 149–480). Springer Netherlands. https://doi.org/10.1007/978-94-007-7820-7_2
- Schierup, M. H., & Hein, J. (2000). Recombination and the molecular clock. *Molecular Biology and Evolution*, 17(10), 1578–1579. <https://doi.org/10.1093/oxfordjournals.molbev.a026256>

- Schmidt, T. L., Rašić, G., Zhang, D., Zheng, X., Xi, Z., & Hoffmann, A. A. (2017). Genome-wide SNPs reveal the drivers of gene flow in an urban population of the asian tiger mosquito, *aedes albopictus*. *PLOS Neglected Tropical Diseases*, 11(10), e0006009. <https://doi.org/10.1371/journal.pntd.0006009>
- Shiryeve, S. A., & Agarwala, R. (2024). Indexing and searching petabase-scale nucleotide resources. *Nature Methods*, 21(6), 994–1002. <https://doi.org/10.1038/s41592-024-02280-z>
- Shope, R. E., & Meegan, J. M. (1997). Arboviruses. In A. S. Evans & R. A. Kaslow (Eds.), *Viral infections of humans: Epidemiology and control* (pp. 151–183). Springer US. https://doi.org/10.1007/978-1-4899-0036-4_6
- Singh, D., & Yi, S. V. (2021). On the origin and evolution of SARS-CoV-2. *Experimental & Molecular Medicine*, 53(4), 537–547. <https://doi.org/10.1038/s12276-021-00604-z>
- Slowik, T. J., & Lane, R. S. (2009). Feeding preferences of the immature stages of three western north american ixodid ticks (acari) for avian, reptilian, or rodent hosts. *Journal of Medical Entomology*, 46(1), 115–122. <https://doi.org/10.1603/033.046.0115>
- Smith, G. J. D., Vijaykrishna, D., Bahl, J., Lycett, S. J., Worobey, M., Pybus, O. G., Ma, S. K., Cheung, C. L., Raghwani, J., Bhatt, S., Peiris, J. S. M., Guan, Y., & Rambaut, A. (2009). Origins and evolutionary genomics of the 2009 swine-origin h1n1 influenza a epidemic. *Nature*, 459(7250), 1122–1125. <https://doi.org/10.1038/nature08182>
- Smithburn, K. C., Hughes, T. P., Burke, A. W., & Paul, J. H. (1940). A neurotropic virus isolated from the blood of a native of uganda. *The American Journal of Tropical Medicine and Hygiene*, 1(4), 471–492. <https://doi.org/10.4269/ajtmh.1940.s1-20.471>
- Spielman, S. J., & Wilke, C. O. (2015). The relationship between dN/dS and scaled selection coefficients. *Molecular Biology and Evolution*, 32(4), 1097–1108. <https://doi.org/10.1093/molbev/msv003>
- Steinhauer, D. A., Domingo, E., & Holland, J. J. (1992). Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase. *Gene*, 122(2), 281–288. [https://doi.org/10.1016/0378-1119\(92\)90216-C](https://doi.org/10.1016/0378-1119(92)90216-C)
- Stollar, V., & Thomas, V. L. (1975). An agent in the *Aedes aegypti* cell line (peleg) which causes fusion of *Aedes albopictus* cells. *Virology*, 64(2), 367–377. [https://doi.org/10.1016/0042-6822\(75\)90113-0](https://doi.org/10.1016/0042-6822(75)90113-0)
- 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>
- Taylor, R. M., Hurlbut, H. S., Work, T. H., Kingston, J. R., & Hoogstraal, H. (1966). Arboviruses isolated from argas ticks in egypt: Quarantil, chenuda, and nyamanini. *The American Journal of Tropical Medicine and Hygiene*, 15(1), 76–86. <https://doi.org/10.4269/ajtmh.1966.15.76>
- Takken, W., & Verhulst, N. O. (2013). Host preferences of blood-feeding mosquitoes. *Annual Review of Entomology*, 58, 433–453. <https://doi.org/10.1146/annurev-ento-120811-153618>

- Tang, P., Croxen, M. A., Hasan, M. R., Hsiao, W. W. L., & Hoang, L. M. (2017). Infection control in the new age of genomic epidemiology. *American Journal of Infection Control*, 45(2), 170–179. <https://doi.org/10.1016/j.ajic.2016.05.015>
- Thorne, J. L., Kishino, H., & Painter, I. S. (1998). Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution*, 15(12), 1647–1657. <https://doi.org/10.1093/oxfordjournals.molbev.a025892>
- Tiwari, S., Dhakal, T., Kim, B.-J., Jang, G. S., & Oh, Y. (2025). Genomics in epidemiology and disease surveillance: An exploratory analysis. *Life*, 15(12), 1848. <https://doi.org/10.3390/life15121848>
- Tümmler, B. (2020). Molecular epidemiology in current times [_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1462-2920.15238>]. *Environmental Microbiology*, 22(12), 4909–4918. <https://doi.org/10.1111/1462-2920.15238>
- Vasimuddin, M., Misra, S., Li, H., & Aluru, S. (2019). Efficient architecture-aware acceleration of BWA-MEM for multicore systems [ISSN: 1530-2075]. *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 314–324. <https://doi.org/10.1109/IPDPS.2019.00041>
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3), 261–272. <https://doi.org/10.1038/s41592-019-0686-2>
- Wang, L., Rosales Rosas, A. L., De Coninck, L., Shi, C., Bouckaert, J., Matthijssens, J., & Delang, L. (2021). Establishment of culex modestus in belgium and a glance into the virome of belgian mosquito species. *mSphere*, 6(2), 10.1128/msphere.01229–20. <https://doi.org/10.1128/msphere.01229-20>
- Weaver, S. C., & Barrett, A. D. T. (2004). Transmission cycles, host range, evolution and emergence of arboviral disease. *Nature Reviews. Microbiology*, 2(10), 789–801. <https://doi.org/10.1038/nrmicro1006>
- Webster, C. L., Waldron, F. M., Robertson, S., Crowson, D., Ferrari, G., Quintana, J. F., Brouqui, J.-M., Bayne, E. H., Longdon, B., Buck, A. H., Lazzaro, B. P., Akorli, J., Haddrill, P. R., & Obbard, D. J. (2015). The discovery, distribution, and evolution of viruses associated with drosophila melanogaster. *PLOS Biology*, 13(7), e1002210. <https://doi.org/10.1371/journal.pbio.1002210>
- Wei, K., & Li, Y. (2018). Global genetic variation and transmission dynamics of h9n2 avian influenza virus [_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/tbed.12733>]. *Transboundary and Emerging Diseases*, 65(2), 504–517. <https://doi.org/10.1111/tbed.12733>
- Wei, K., Chen, Y., Lin, Y., & Pan, Y. (2014). Genetic dynamic analysis of the influenza a h5n1 NS1 gene in china. *PLOS ONE*, 9(7), e101384. <https://doi.org/10.1371/journal.pone.0101384>

- WHO. (2025). *Independent assessment of the origins of SARS-CoV-2, developed by the scientific advisory group for the origins of novel pathogens (SAGO)*. Retrieved February 15, 2026, from <https://www.who.int/publications/m/item/independent-assessment-of-the-origins-of-sars-cov-2-from-the-scientific-advisory-group-for-the-origins-of-novel-pathogens>
- Wilfert, L., & Jiggins, F. M. (2014). Flies on the move: An inherited virus mirrors drosophila melanogaster's elusive ecology and demography. *Molecular Ecology*, 23(8), 2093–2104. <https://doi.org/10.1111/mec.12709>
- Wilkman, L., Kaarle, E., Luande, V. N., Lantto, R., Evander, M., & Lwande, O. W. (2025). Insect-specific viruses: Transmission dynamics and biological control strategies against arboviruses. *Frontiers in Microbiology*, 16. <https://doi.org/10.3389/fmicb.2025.1624662>
- Winokur, O. C., Main, B. J., Nicholson, J., & Barker, C. M. (2020). Impact of temperature on the extrinsic incubation period of zika virus in aedes aegypti. *PLoS Neglected Tropical Diseases*, 14(3), e0008047. <https://doi.org/10.1371/journal.pntd.0008047>
- Woelk, C. H., & Holmes, E. C. (2002). Reduced positive selection in vector-borne RNA viruses. *Molecular Biology and Evolution*, 19(12), 2333–2336. <https://doi.org/10.1093/oxfordjournals.molbev.a004059>
- Wolf, Y. I., Silas, S., Wang, Y., Wu, S., Bocek, M., Kazlauskas, D., Krupovic, M., Fire, A., Dolja, V. V., & Koonin, E. V. (2020). Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nature Microbiology*, 5(10), 1262–1270. <https://doi.org/10.1038/s41564-020-0755-4>
- Wong, T. K. F., Ly-Trong, N., Ren, H., Baños, H., Roger, A. J., Susko, E., Bielew, C., Maio, N. D., Goldman, N., Hahn, M. W., Huttley, G., Lanfear, R., & Minh, B. Q. (2025). IQ-TREE 3: Phylogenomic inference software using complex evolutionary models. Retrieved February 15, 2026, from <https://ecoevorxiv.org/repository/view/8916/>
- Woolhouse, M. E., & Gowtage-Sequeria, S. (2005). Host range and emerging and reemerging pathogens. *Emerging Infectious Diseases*, 11(12), 1842–1847. <https://doi.org/10.3201/eid1112.050997>
- Worobey, M., Han, G.-Z., & Rambaut, A. (2014). A synchronized global sweep of the internal genes of modern avian influenza virus. *Nature*, 508(7495), 254–257. <https://doi.org/10.1038/nature13016>
- Zhang, Y.-Z., Shi, M., & Holmes, E. C. (2018). Using metagenomics to characterize an expanding virosphere. *Cell*, 172(6), 1168–1172. <https://doi.org/10.1016/j.cell.2018.02.043>
- Zhao, J., Dellicour, S., Yan, Z., Veit, M., Gill, M. S., He, W.-T., Zhai, X., Ji, X., Suchard, M. A., Lemey, P., & Su, S. (2022). Early genomic surveillance and phylogeographic analysis of getah virus, a reemerging arbovirus, in livestock in china. *Journal of Virology*, 97(1), e01091–22. <https://doi.org/10.1128/jvi.01091-22>
- Zhao, L., & Illingworth, C. J. R. (2019). Measurements of intrahost viral diversity require an unbiased diversity metric. *Virus Evolution*, 5(1), vey041. <https://doi.org/10.1093/ve/vey041>

- Zheng, X., Cao, F., Yu, Y., Yu, X., Zhou, Y., Cheng, S., Qiu, X., Ao, L., Yang, X., Sun, Z., & Li, J. (2025). The ongoing epidemics of seasonal influenza a(h3n2) in hangzhou, china, and its viral genetic diversity. *Viruses*, 17(4), 526. <https://doi.org/10.3390/v17040526>
- Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G. F., & Tan, W. (2020). A novel coronavirus from patients with pneumonia in china, 2019. *New England Journal of Medicine*, 382(8), 727–733. <https://doi.org/10.1056/NEJMoa2001017>

Appendices

Appendix A. Number of sequences per region for each virus analysis

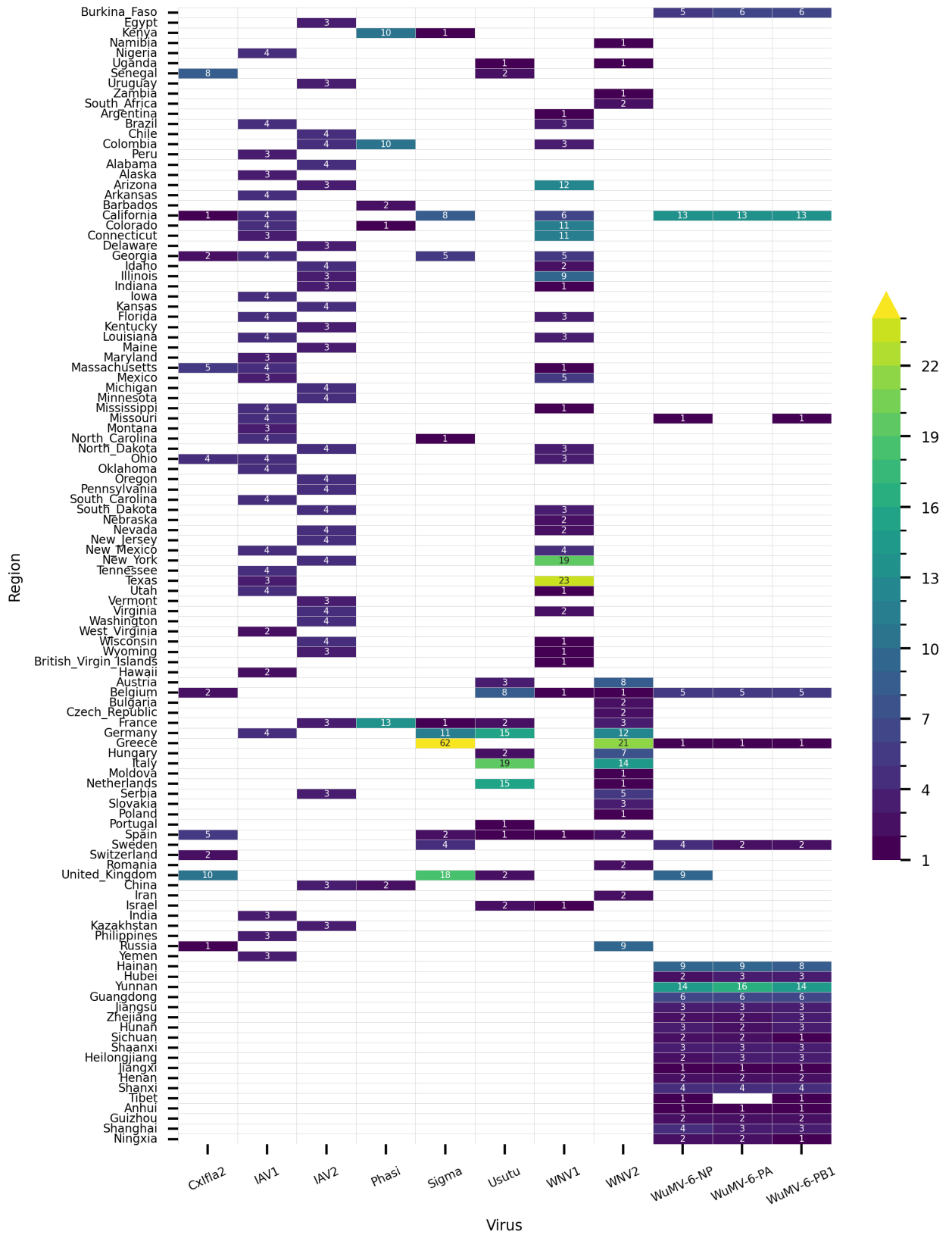


Figure A1. Sequence counts per region for each discrete phylogeographic analysis.

Appendix B. Discrete phylogeography predictor matrices

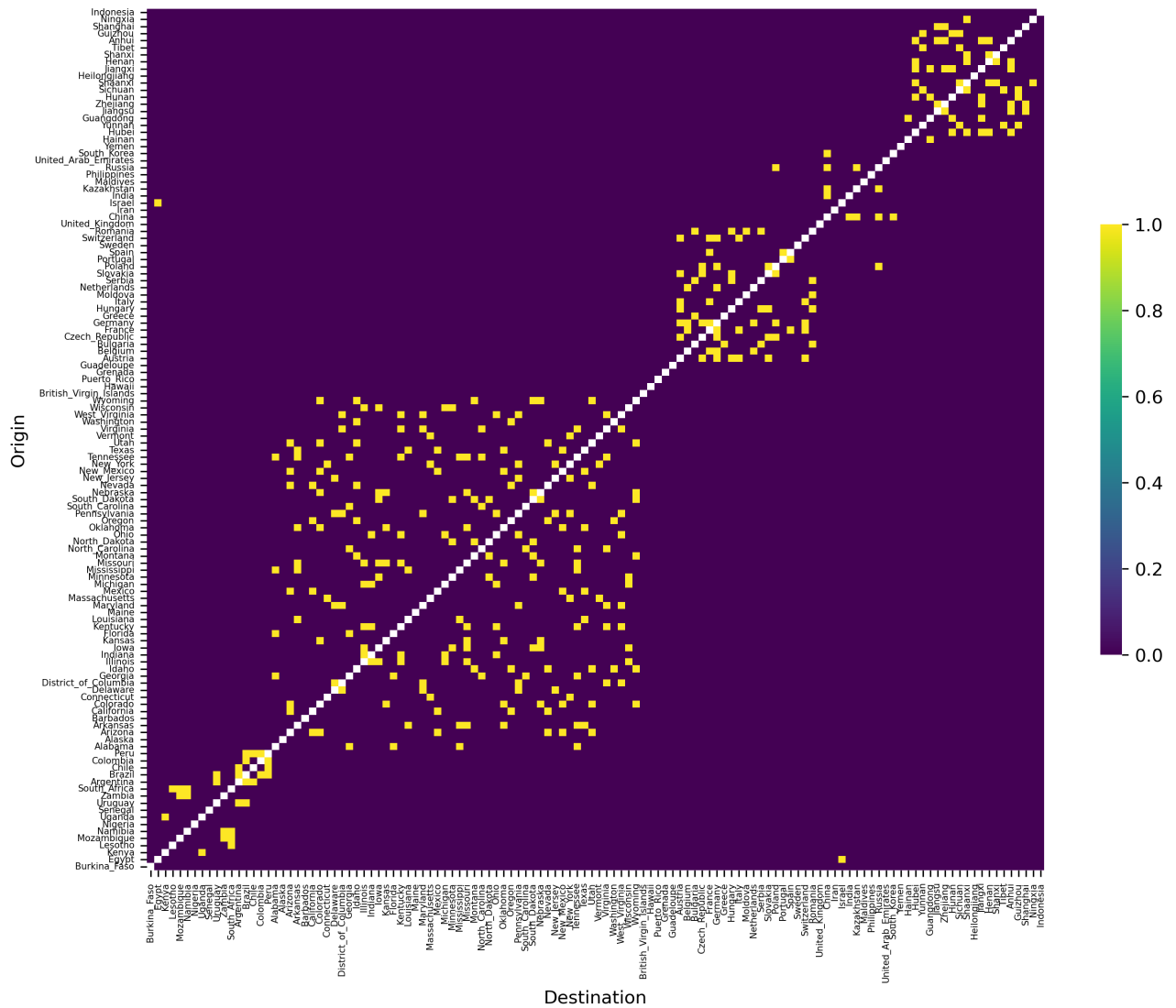


Figure B1. Shared land borders predictor used for discrete phylogeography. If regions share a land border, matrix has a value of 1, if it does not – 0.

Figure B2. Predictor representing distance between region centroids.

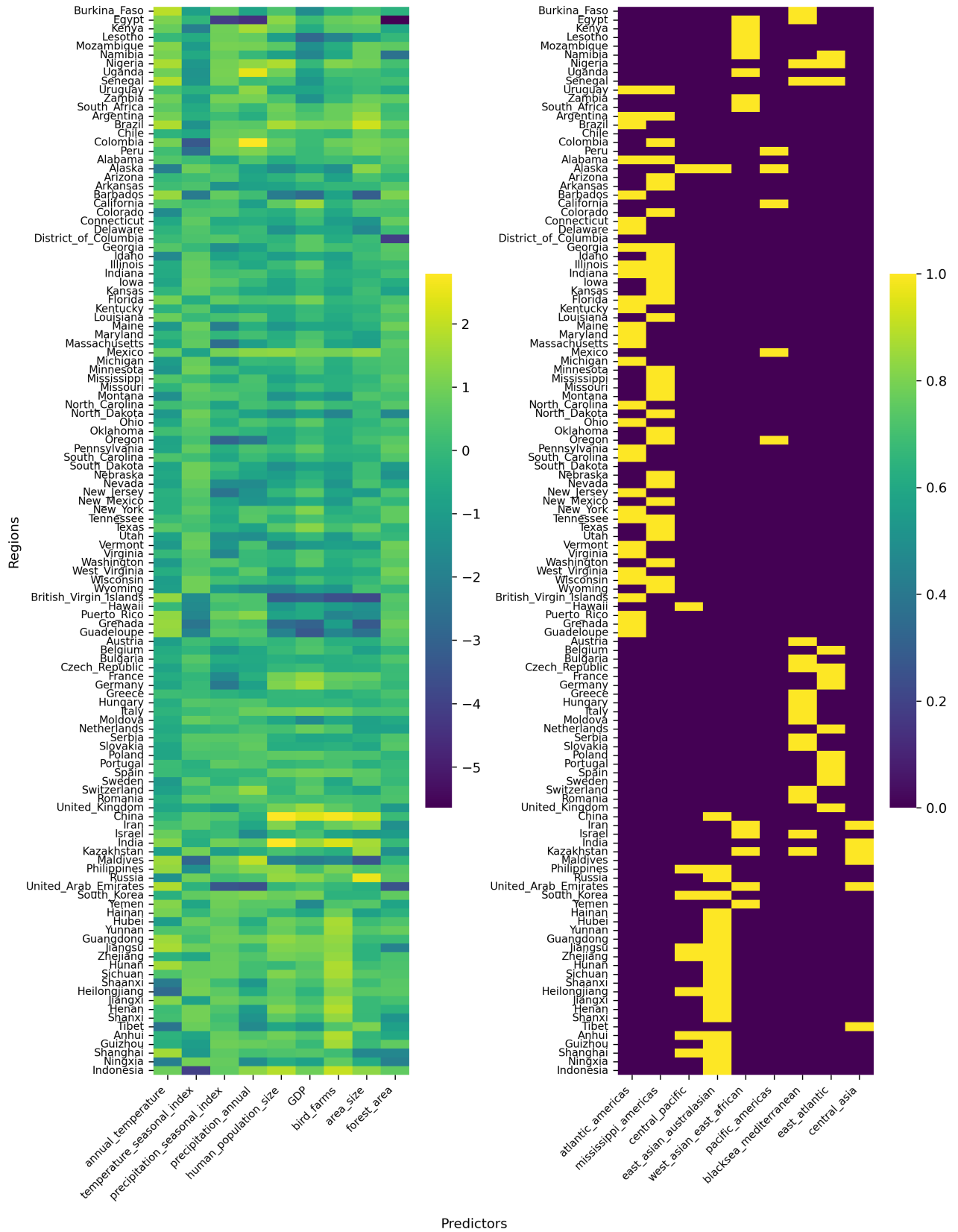


Figure B3. Predictor values for origin/destination predictors (left) and bird migration flyway overlap with sampling regions (right). If region is in the bird migration flyway path, it has a value of 1, otherwise - 0.

Appendix C. WuMV-6 NP segment predictor colinearity

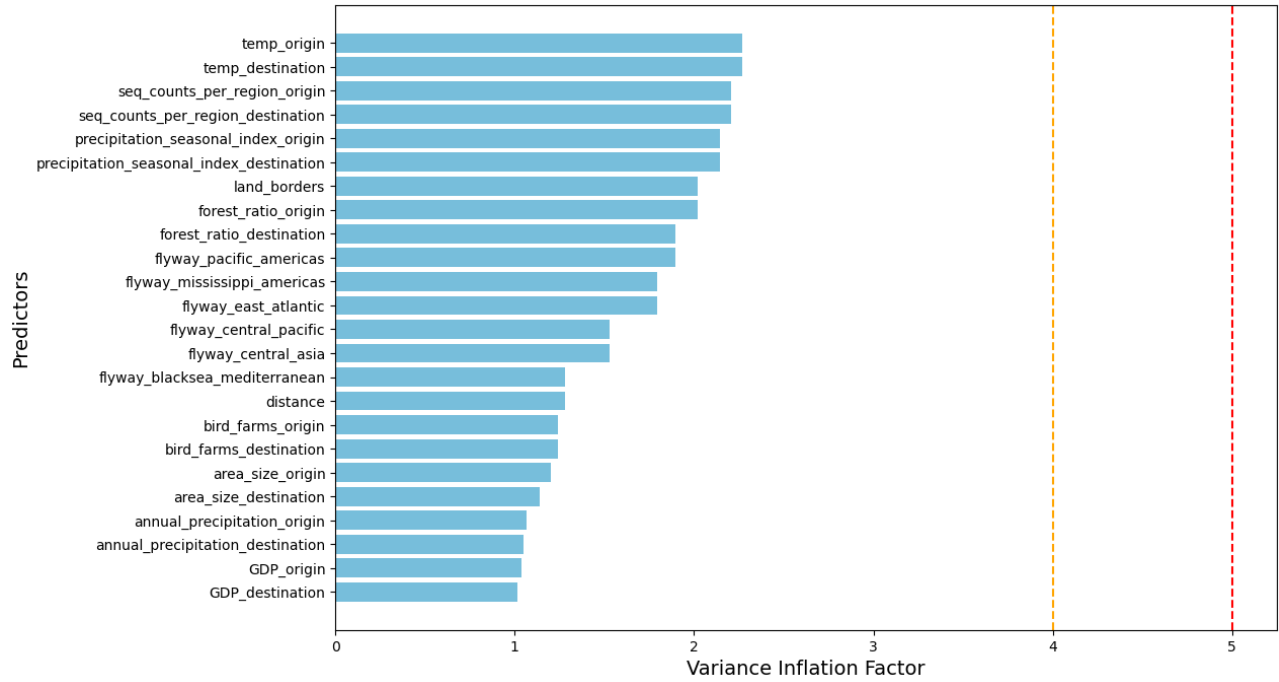


Figure C1. Variance inflation factor of WuMV-6 NP segment discrete phylogeography predictors after removing correlated predictors.

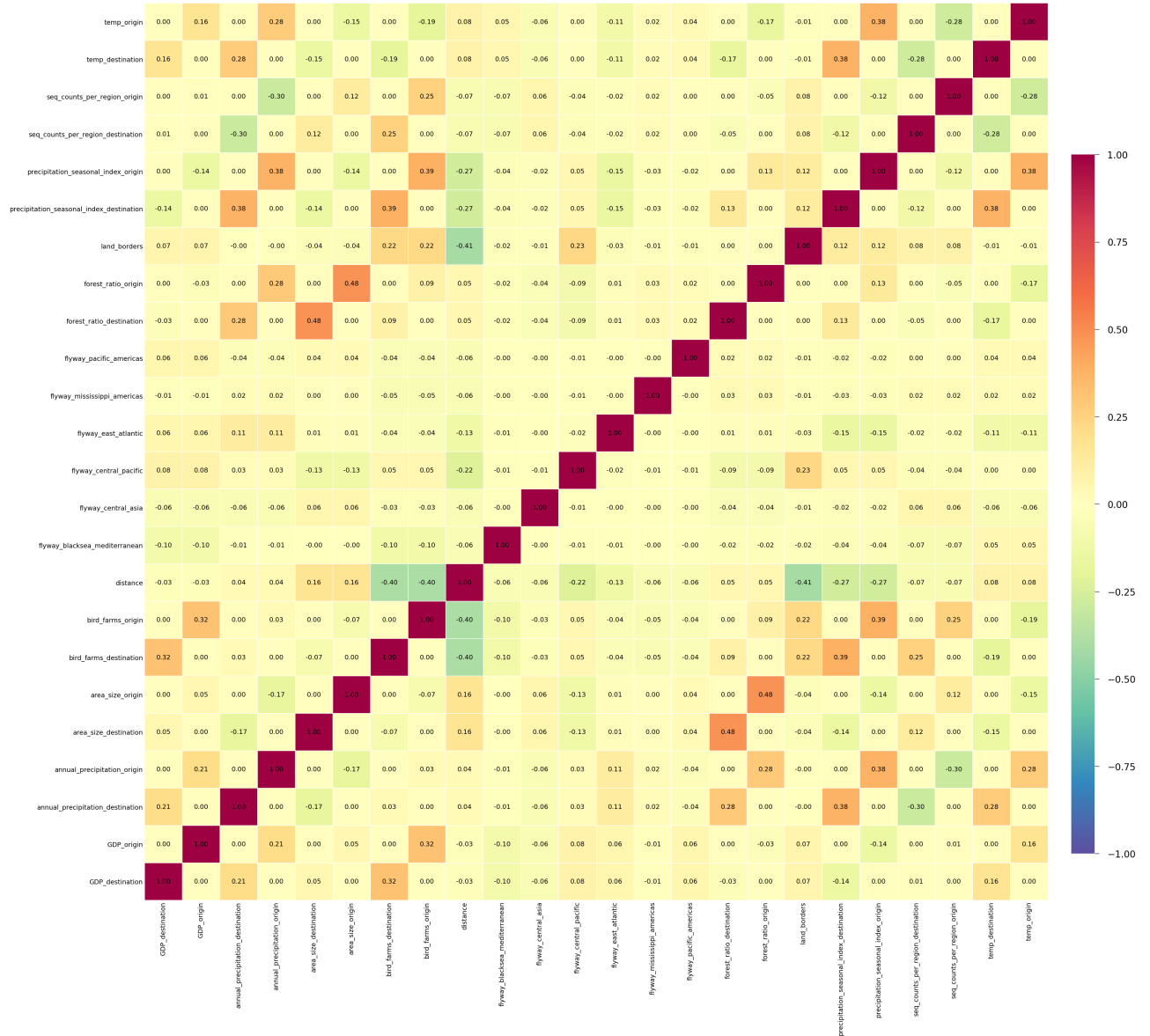


Figure C2. Pairwise Spearman correlations between all predictors after removing correlated predictors. Color indicates magnitude and direction of the correlation (red - positive correlation, blue - negative).

Appendix D. Discrete phylogeography GLM predictor statistics**Table D1.** GLM predictor statistics: Mean coefficient, 95% highest posterior density credible interval (CI), Bayes Factor and inclusion probability. Only predictors, that have $BF > 3$ are included.

Sample	Predictor	Mean Coeffi- cient	95% CI	BF	Inclusion probabil- ity
WuMV-6_NP	flyway east atlantic	3.11	1.52, 4.63	>50.00	0.69
	shared land borders	2.13	1.18, 3.09	21.01	0.36
	distance between region centroids	-3.33	-5.16, -1.60	19.99	0.35
	bird farms origin	1.49	0.49, 2.51	11.72	0.24
WuMV-6_PA	distance between region centroids	-3.32	-5.00, -1.55	30.98	0.46
	shared land borders	1.99	1.16, 2.79	28.32	0.43
	forest percentage origin	2.22	0.44, 4.51	3.57	0.09
WuMV-6_PB1	distance between region centroids	-3.8	-5.54, -2.12	>50.00	0.87
	shared land borders	2.06	0.92, 2.98	7.46	0.17
CxIfla2					
WNV_c1	sequence count origin	2.84	1.23, 4.74	>50.00	1.00
	sequence count destina- tion	0.88	0.64, 1.11	>50.00	1.00
	bird farms origin	-1.38	-3.18, 0.13	41.25	0.57
	flyway atlantic americas	1.42	0.06, 2.86	40.17	0.56
	mean precipitation ori- gin	-1.97	-4.03, 0.06	35.01	0.53
	distance between region centroids	-1.77	-3.96, 0.46	26.57	0.46
	area size origin	-0.93	-2.40, 0.27	15.59	0.33
	flyway east atlantic	-0.02	-3.90, 3.80	13.63	0.30
	flyway pacific americas	-0.20	-3.71, 3.46	12.84	0.29
	forest percentage origin	0.31	-1.71, 2.36	7.36	0.19
	precipitation seasonal index origin	0.30	-1.49, 2.36	6.75	0.18
	temperature seasonal in- dex origin	0.20	-1.68, 2.07	6.67	0.18
	mean temperature origin	0.37	-1.25, 1.98	4.39	0.12

Sample	Predictor	Mean Coeffi- cient	95% CI	BF	Inclusion probabil- ity
WNV_c2	flyway mississippi ameri- cas	0.46	-0.52, 1.45	4.19	0.12
	temperature seasonal in- dex destination	0.24	-0.08, 0.55	3.69	0.11
	shared land borders	0.20	-0.57, 0.99	3.61	0.10
	shared land borders	1.93	1.13, 2.73	>50.00	0.84
	sequence count origin	0.88	0.38, 1.45	14.41	0.31
USUV	shared land borders	3.01	2.07, 4.09	>50.00	1.00
SIGV	precipitation seasonal index destination	-2.55	-4.36, -0.87	18.34	0.41
IAV_c1	distance between region centroids	-4.23	-5.78, -2.79	>50.00	0.68
	precipitation seasonal index origin	4.26	1.30, 7.11	>50.00	0.67
	temperature seasonal in- dex origin	-2.54	-3.98, -1.44	8.25	0.20
IAV_c2	precipitation seasonal index origin	5.54	3.68, 7.68	>50.00	1.00
	distance between region centroids	-5.25	-7.25, -3.21	>50.00	0.99

Appendix E. WuMV-6 NP segment transitions between regions

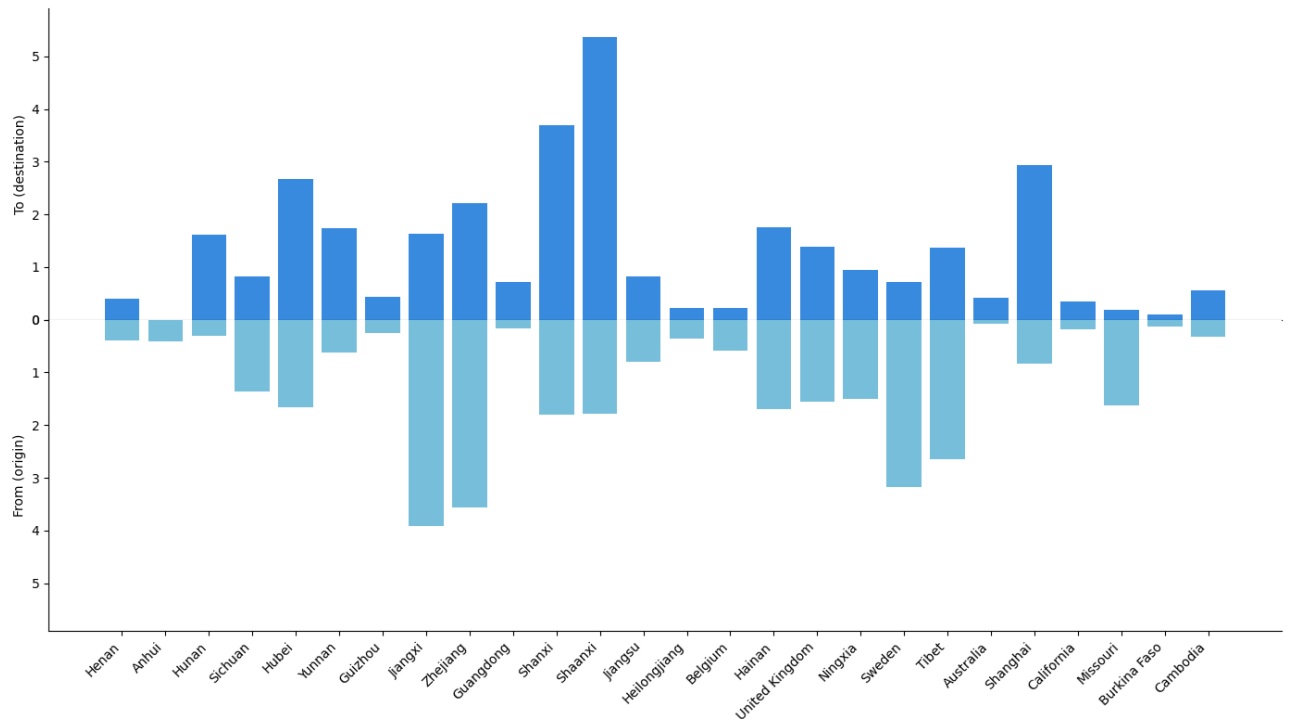


Figure E1. Number of WuMV-6 transitions from and to regions sorted by predictor values of the number of bird farms per region (origin).

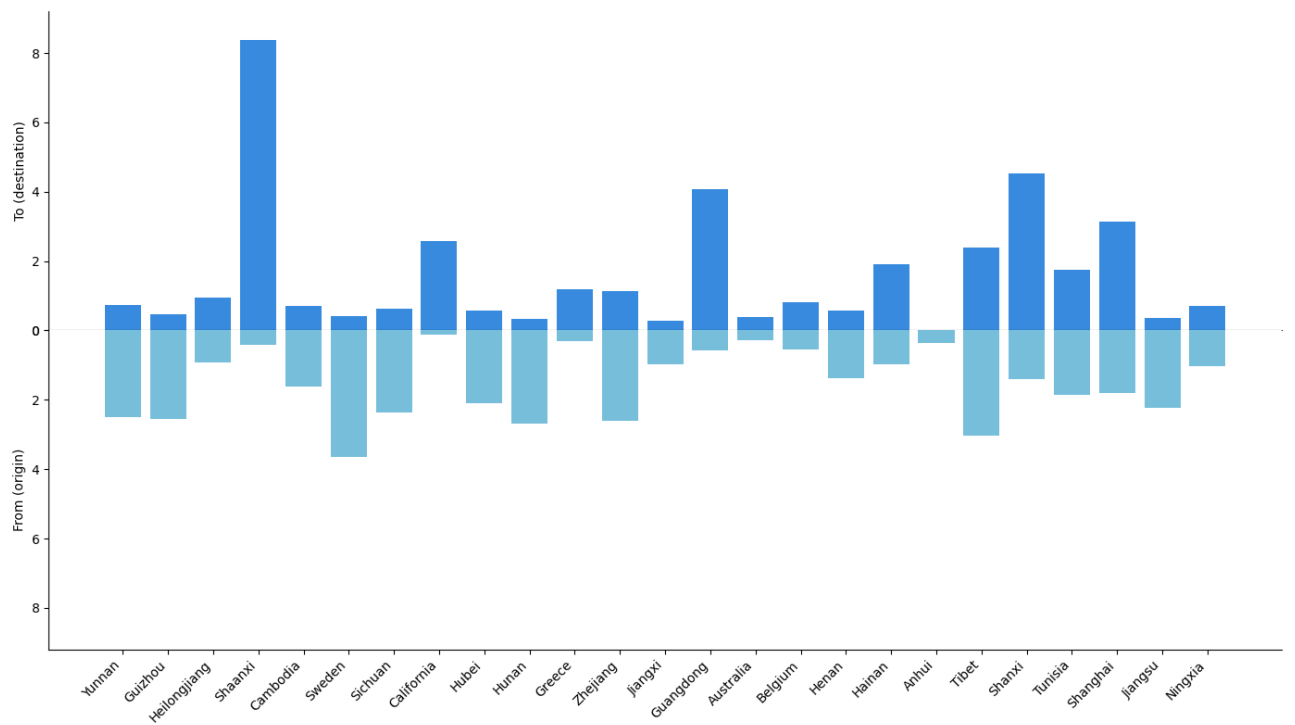


Figure E2. Number of WuMV-6 transitions from and to regions sorted by predictor values of the forest percentage per region (origin).