Durrant, Rowan (2025) *Genetic analysis of the rabies virus.* PhD thesis.

https://theses.gla.ac.uk/84927/

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses
https://theses.gla.ac.uk/
research-enlighten@glasgow.ac.uk

Genetic Analysis of the Rabies Virus

Rowan Durrant MRes, BSc

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

School of Biodiversity, One Health and Veterinary Medicine

College of Medical, Veterinary and Life Sciences

University of Glasgow

OCTOBER 2024

Abstract

Rabies is a fatal disease caused by a negative-strand RNA virus with a genome size of approximately 12 kilobases. Rabies kills an estimated 60,000 people per year, most of whom would have been bitten by a rabid domestic dog. In recent years whole genome sequencing of the rabies virus has become more accessible through the development of portable sequencing technologies and inexpensive protocols, which has led to an increase in the amount of publicly available genomic data, and in the capacity to rapidly acquire sequence data from new rabies outbreaks. In this thesis I aim to use rabies genomic and genetic data to investigate how rabies evolves, and how to best use this data to meet the global goal of achieving zero human rabies deaths by 2030. I developed a simulation framework consisting of an existing branching process epidemiological model and a novel mutation model to generate synthetic rabies sequences associated with known underlying transmission dynamics. In chapter 2 I used this framework to investigate whether the lack of temporal signal required to conduct Bayesian phylogenetic analyses on rabies sequence/added datasets could be due to rabies' variable incubation period lengths. I found that at substitution rates comparable to rabies', it is not possible to distinguish root-to-tip divergence plots for synthetic genomes generated using a per-unit time or per-generation model of substitution; it is possible, however, at rates more representative of other RNA viruses, due to distinctive "ridges" that form under the per-generation model after unusually long or short incubation periods. I conclude that rabies' slow evolution is more likely to be the cause of the lack of temporal signal than its variable incubation periods, but that thinking about evolution on a per-generation scale could be useful in certain contexts. Existing methods of estimating outbreak sizes, such as serological surveys and randomised testing, are unsuitable for estimating rabies outbreak sizes due to the fatality of the virus and the testing method respectively. In chapter 3, using the same simulation framework as in chapter 2, I developed a novel method of estimating outbreak sizes from phylogenetic trees which is simple, computationally inexpensive and takes advantage of the genomic data already usually gathered as part of the outbreak surveillance. I apply this method to a new outbreak of rabies in the Romblon province of the Philippines, confirming that there has been widespread undetected transmission, but that the outbreak surveillance was perhaps more effective at detecting cases than is usual for a rabies outbreak. In chapter 4 I used publicly available rabies sequence data to investigate to what extent codon usage was biased between different host-species-specific minor clades, and whether these differences were evidence of adaptation by the virus to the host. I found that while there was little evidence of the virus adapting its codon usage specifically to new host species, differences exist in RABV's CpG content which suggest that bat- and carnivore-associated rabies clades are under differing levels of selection pressure from the host immune system on CpG dinucleotides. Together these findings demonstrate that genomic data is a valuable resource that can be used to inform outbreak responses and tell us about how rabies evolves and interacts with its wide range of hosts.

Contents

List of Tables

List of Figures

Acknowledgements

I feel very lucky that my PhD experience has been relaxed and enjoyable all the way to submission day, and for that I have many people to thank.

Firstly I must thank my supervisors Katie Hampson and Christina Cobbold for putting up with me for the past three and a half years. They gave me freedom to explore the aspects of infectious disease research that I thought were cool, and encouraged me to develop my vague ideas into pieces of work that other people would actually find interesting. Their mentorship has made me a better researcher by far.

Jonathan Dushoff deserves a massive thank you for adopting me as an unofficial supervisee and for hosting me at McMaster earlier this year. Those two weeks were quite pivotal to my PhD and probably to my career as a whole, and I'm never going to look at train carriage numbers the same way again.

I must also thank Dan Haydon for our useful discussions and his ideas surrounding the formula in chapter 3 and for lending me some non-rabies data to play with, and Denise Marston for our discussions surrounding chapter 4.

My fellow SBOHVM-ers have helped me out on all things rabies, virology and phylogenetics, and have been excellent company in office 308 and beyond over the last three years; these people include, but are certainly not limited to, Kat, Kirstyn, Essel, Hollie, Matt, Danni, Camille, Elaine and Martha.

Miguel Lurgi and Konstans Wells have continued to help me grow long after their roles as my MRes supervisors had officially ended, and I hope I've made them proud and can one day pay their kindness forward.

It has been a massive privilege to work with everyone mentioned so far, and I am very grateful that there are people out there who seem to be genuinely invested in my growth as a researcher.

Finally, moving away from the academic world, I'd like to thank the Glasgow Uni Mountaineering Club Oldies and the Non-Volant Oranges for keeping me sane for the last few years. I've seen more of Scotland than I ever would have by myself and had some genuinely life-changing experiences on trips with the ex-GUMC crowd (even the sugar-free Pabbay custard was character building). The Oranges have been very tolerant of me annoying them about organising a zoom call every once in a while and have kept me supplied with high quality memes. Love you all.

Declaration

I declare that, except where explicit reference is made to the contribution of others in the “Thesis aims & organisation” section of the Introduction chapter, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Abbreviations

- AIC - Akaike information criterion
- CAI - Codon adaptation index
- CNS - Central nervous system
- CpG - Cytosine-phosphate-guanine
- ENC - Effective number of codons
- GC3 - Guanine or cytosine in third codon position
- MCMC - Markov chain Monte Carlo
- NCBI - National Center for Biotechnology Information
- N_e - Effective population size
- PCA - Principal component analysis
- PEP - Post-exposure prophylaxis
- R_0 - Basic reproduction number
- RABV - Rabies virus
- R_e - Effective reproduction number
- RNA - Ribonucleic acid
- RSCU - Relative synonymous codon usage
- SNP - Single nucleotide polymorphism
- WHO - World Health Organization
- ZAP - Zinc-finger antiviral protein

Chapter 1

Introduction

1.1 Evolution, molecular clocks and phylogenetics

1.1.1 Neutral theory and the molecular clock

When comparing the number of amino acid substitutions in the different haemoglobin chains of various mammals and the species' divergence times, Zuckerkandl and Pauling (1965) observed that genetic divergence increased linearly with time. Their observation was considered incompatible with the contemporary view that natural selection was the main driver of evolution, and that most mutations either had an advantageous or deleterious effect. If this view was true and most mutations did have a fitness effect, the increase in divergence would be unlikely to be so constant over time due to the unpredictability in changes in evolutionary pressures from the environment.

This observation resulted in their introduction of the term "molecular clock", and contributed to the development of the neutral theory of molecular evolution: the idea that most mutations have no fitness effect, so whether they persist within the population or not is random, leading to genetic drift (Kimura 1968). Kimura further developed this idea, stating that the variation in substitution rate (the rate at which substitutions become fixed in the population) observed between genes was likely due to the functional constraint of a gene. As the “rigidity” of the structure and function of a protein decreased, the evolutionary rate would approach the synonymous mutation rate, based on the assumption that synonymous mutations were always neutral and therefore equal to the mutation rate (Kimura 1977, 1979).

1.1.2 Divergence time estimation and phylogenetic tree construction

The assumption that the substitution rate remains constant through time allows the divergence time to be estimated if the substitution rate is known. These methods first construct a phylogenetic tree of the available sequences, and then use the branch lengths of the tree and the substitution rate to estimate the divergence time at each node. Various tree-building algorithms were developed around this time, including the neighbour joining (Saitou and Nei 1987), maximum parsimony (Farris 1970) and maximum likelihood methods (Felsenstein 1981). Maximum likelihood tree building uses evolutionary models to capture different patterns of mutation; for example, the JC69 model assumes that all nucleotide substitutions have an equal probability of occurring (Jukes and Cantor 1969), whereas the HKY model assumes that transition and transversion substitutions occur with different probabilities (Hasegawa et al. 1985), and the GTR model assumes that all nucleotide substitutions occur at different rates (Tavaré 1986).

While many early studies surrounding the molecular clock were conducted using eukaryote genes (such as Sarich and Wilson (1967) using relative levels of cross-reactivity between albumins and antiserums instead of amino acid or genetic sequences to estimate the divergence time between humans and other primates), these methods were soon also applied to viral sequences. Gojobori et al. (1990) observed the same linear relationship between genetic divergence and time in viral genes, albeit at a much higher rate than their hosts, and concluded that the neutral theory was also applicable to viruses. They then used the assumption of a molecular clock to estimate the divergence times of hepatitis viruses and other hepadnaviruses (much more recently than their respective host species' divergence times) and HIVs and SIVs (150-200 years ago). Li et al. (1988) also used this divergence time estimation to date the emergence of HIV to before 1960, and Buonagurio et al. (1986), Saitou and Nei (1986) and Gorman et al. (1990) applied this method to influenza A sequences, all using maximum parsimony trees.

These methods were of particular relevance to RNA viruses as their fast rate of evolution meant that the substitution rate could be measured from a few years' worth of sampling. This contrasts the required "fossil data" (estimates of the divergence time of at least two clades in the dataset, perhaps acquired from fossil evidence) when all sequences are acquired in a timeframe that is relatively short compared to the timeframe over which divergence has occurred, such as in modern eukaryotes (Hasegawa et al. 1985). RNA viruses, like their DNA counterparts, require their host's translation machinery to produce their proteins, but have a much higher evolutionary rate due to their error prone mode of replication. A study of a range of RNA virus species found that there was evolutionary rate variation between these lineages outwith the constraints of a molecular clock (Jenkins et al. 2002). Furthermore, Jenkins et al. found that while evidence for a molecular clock was stronger when only investigating synonymous substitutions, there was still sufficient variation to suggest that these substitutions were not truly neutral.

1.1.3 Alternative clock models and Bayesian phylogenetics

One of the most important developments was the realisation that Kimura's "strict" clock too simple; in reality, there is variation in the substitution rate between lineages. Initially, this resulted in genes that failed tests of strict clock-likeness being removed from analyses, until methods that could handle non-clock-likeness were developed (Sanderson 1997; Thorne et al. 1998). These methods relied on the evolutionary rates being autocorrelated (more closely related lineages are likely to have similar evolutionary rates).

These methods also required that the tree topology be pre-specified, which could lead to inaccurate results if the topology was determined under the assumption of a global molecular clock. Drummond et al. (2006) developed relaxed clock models for use with the Bayesian phylogenetics software BEAST (Drummond and Rambaut 2007), which allowed the tree topology itself to be determined under the assumption of between-lineage rate variation. BEAST also incorporated the use of sequences sampled at different times to aid in estimating the evolutionary rate, which is particularly relevant to RNA viruses (Drummond et al. 2003b). The ability to construct trees scaled in time rather than substitutions per site distinguished BEAST from other Markov-chain Monte Carlo-based phylogenetic methods such as MrBayes (Huelsenbeck and Ronquist 2001). Later developments to the BEAST software included phylogeographic inference and models of phenotypic trait evolution (Drummond et al. 2012).

1.1.4 Codon usage

One argument against the neutral theory of evolution is that synonymous mutations in some cases can, in fact, have a fitness effect, unrelated to the functional constraint of the resulting protein. This can be due to biases in codon use, where certain codons are used preferentially to encode an amino acid over others, or selection against dinucleotides such as CpG and UpA.

The extent of codon usage bias varies between groups of viruses with different host life history strategies; viruses spread by aerosols have stronger codon usage biases, whereas vector-borne viruses have weaker biases (Jenkins and Holmes 2003), and positive-strand RNA viruses with a narrow host range have stronger biases than those with a broad host range (Tian et al. 2018). Codon usage biases are mainly driven by nucleotide and dinucleotide usage in human RNA viruses (Jenkins and Holmes 2003), particularly due to biases against CpG and UpA dinucleotides, but this fails to explain all variation observed in codon usage. Another possible explanation for these biases is that more closely matching the host's codon usage can improve translational efficiency of viral proteins (Haas et al. 1996), although there is poor evidence that negative-strand RNA viruses' codon usage mimics their host species (Rima 2015). One possible reason for this could be that optimising translational efficiency may not actually be the best strategy for the virus; suboptimal codon usage can be used as a way to delay glycoprotein production in lentiviruses and herpesviruses (Shin et al. 2015).

A significant selective bias against CpG dinucleotides has been noted to occur in almost all RNA viruses (Karlin et al. 1994), which was suggested to be due to a previously unknown host immune mechanism (Greenbaum et al. 2009), as vertebrate genomes are also CpG deficient due to the high mutation rate of methylated cytosine (which exists mainly in CpG dinucleotides in vertebrates) to thymine (Sved and Bird 1990). We now know this host immune mechanism is likely to be the zinc-finger antiviral protein (ZAP; Takata et al. 2017), which selectively binds to CpG-rich RNA sequences. While the exact mechanism of the bias against UpA is currently unknown, ZAP may also be involved here (Goonawardane et al. 2021).

Codon and dinucleotide usage biases in RNA viruses have particular relevance to vaccine development. Deoptimising codon usage has been suggested as a way to develop live attenuated viruses (Baker et al. 2015), and has successfully resulted in effective vaccines for a variety of viruses (Konopka-Anstadt et al. 2020; Lorenzo et al. 2022; Broadbent et al. 2016; Cheng et al. 2015). Conversely, altering the codon usage of the herpesvirus glycoprotein to make it more translationally efficient induces an antibody response which could assist in its usage as a vaccine vector (Shin et al. 2015).

1.2 Rabies

Rabies is an infectious disease of the nervous system which is estimated to cause approximately 60,000 human deaths per year (Hampson et al. 2015). The majority of these deaths occur in Africa and Asia, and 99% result from a bite from a rabid domestic dog. Rabies can also infect a range of other mammal species, including cats, bats, mustelids, raccoons and livestock (Hanlon et al. 2003), although herbivores and humans are usually dead-end hosts. Symptoms of rabies in humans commonly include hydrophobia, seizures, and behavioural changes (ibid.). The disease can present as either “furious” or “paralytic” rabies (Ghosh et al. 2009), where patients either become agitated and aggressive, or become anorexic, depressed and paralysed, eventually leading to coma; in either case, rabies is virtually always fatal. Rabies, especially in its paralytic form, is often initially misdiagnosed in humans as other neurological conditions, such as Guillain-Barré syndrome (Surve et al. 2021).

Rabies is caused by a negative sense, single-stranded RNA virus in the genus Lyssavirus. While other Lyssaviruses can also cause rabies-like disease (Shope 1982), this thesis focuses on the rabies virus, commonly abbreviated to RABV. After being transferred into the wound of a bite victim via the saliva of a rabid animal, the virus enters the nervous system by binding to a range of host cell receptors (Lentz et al. 1982; Tuffereau et al. 1998; Thoulouze et al. 1998) with the glycoprotein, which is the virus' only surface protein. The virus then proceeds to travel along the nerves at a rate of 50 to 100 millimetres per day (Tsiang et al. 1991). As it travels along the nerves towards the central nervous system (Charlton and Casey 1979; Tsiang 1979), the bite victim may experience pain or itching at the healed bite site (Jackson 2011), but in general no symptoms are displayed during this incubation period. While bite victims are generally assumed to not become infectious until a few days before or after they become symptomatic, detectable levels of the virus have been found in the saliva of infected dogs nearly two weeks before symptoms were observed (Fekadu et al. 1982), although there is some scepticism surrounding this finding (Hanlon et al. 2003). This asymptomatic, non-infectious phase of the infection can routinely last multiple weeks, and in some cases over a year (Johnson et al. 2008; Lakhanpal and Sharma 1985). In many diseases where the infecting event is almost impossible to observe, the serial interval (the time from the primary case becoming symptomatic to the secondary case becoming symptomatic) is used as a proxy for the generation interval (the time from the primary case becoming infected to the secondary case becoming infected). Because of the memorable nature of rabies' infecting events (i.e., a bite), the generation interval for rabies can be constructed using contact-tracing data (Figure 1.1). Curiously, rabies' generation interval has been found to be significantly longer than its serial interval (32.1 days compared to 21.7 days, respectively; Li 2019), due to dogs with longer incubation periods producing a larger number of secondary cases.

Upon reaching the brain, the aforementioned symptoms emerge, at which point it is too late to save the patient; it is therefore important that patients receive post-exposure prophylaxis (PEP) as soon as possible after being bitten by an animal that is suspected to be rabid. PEP is almost 100% effective in preventing the onset of rabies as long as

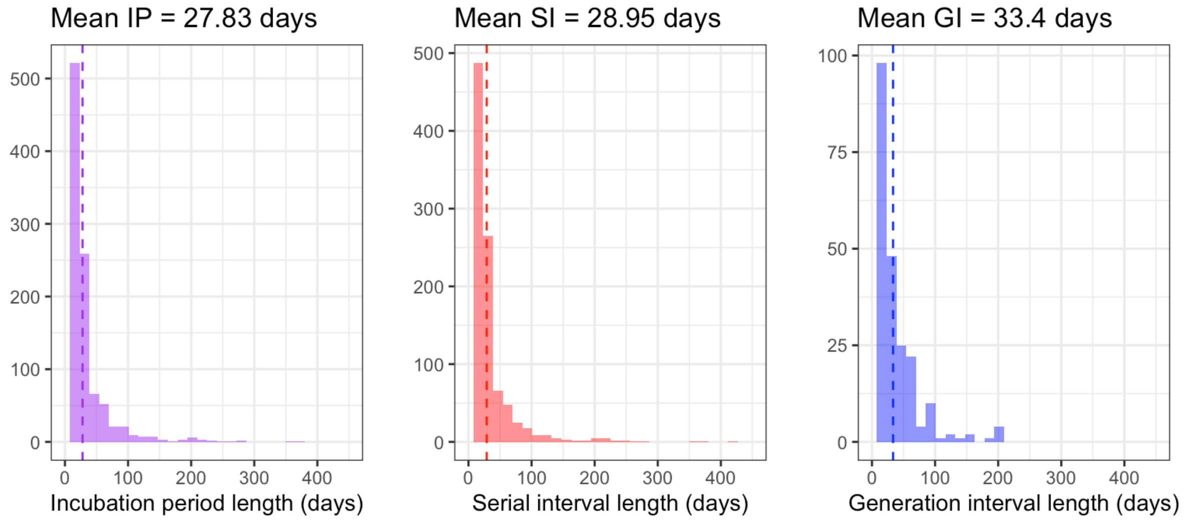

Figure 1.1: Histograms comparing the incubation period, serial interval and generation interval lengths of rabies cases in a Tanzanian dataset. While not as extreme as in Li et al.'s study, the mean generation interval is longer than the mean serial interval.

it is administered quickly after a bite. It is recommended that the wound is washed, a number of vaccinations are given either intradermally or intramuscularly depending on the protocol used, and in some cases rabies immunoglobulins can also be given (World Health Organization 2018). The infectious-symptomatic period lasts for a matter of days before the victim inevitably dies of the disease. The only documented survivors of symptomatic rabies across all reservoir species are a handful of patients who survived after receiving either PEP before becoming symptomatic or the Milwaukee protocol (Jackson 2013). Under experimental conditions vampire bats have also survived for over two years after the virus was detected in their saliva, but these animals never became symptomatic and their saliva only contained detectable titres at one time point (Aguilar-Setien et al. 2005).

In many areas in Africa and Asia, domestic dogs are the main reservoir of rabies. Outbreaks can be managed through mass dog vaccination, where a target of 70% of the dog population being vaccinated is estimated to be sufficient to prevent outbreaks (Cleaveland et al. 2018). The culling of dogs may also occur in response to an outbreak, although this appears to be ineffective or actively detrimental as a control strategy due to dog owners concealing or moving their potentially rabid animals in response to the cull order (Putra et al. 2013). The overall burden of dog rabies is unknown, although annual incidence is estimated to not exceed 1% of the dog population in affected areas, with estimated prevalence remaining below 0.15% in the Serengeti district of Tanzania (Mancy et al. 2022). The R_0 of canine rabies is estimated to be between 1 and 2 (Hampson et al. 2009; Townsend et al. 2013a), with more recent recalculations updating this interval to between 1 and 2.5 (Li et al. 2024); in raccoons, R_e is just above 1 (Biek et al. 2007).

Rabies' low prevalence combined with its low R_0 allows it to persist despite being fatal by avoiding susceptible depletion, but also makes it a prime candidate for elimination through vaccination campaigns. The World Health Organization, along with the Food and Agriculture Organization of the United Nations, the World Organisation for Animal Health, and the Global Alliance for Rabies Control, have proposed a goal of reducing global human rabies deaths to zero by the year 2030 (World Health Organization et al. 2018). This plan includes increasing access to post-exposure prophylaxis, scaling up and sustaining dog vaccination efforts, and improving education surrounding the prevention of dog bites and rabies infections, as well as improving data collection to better inform decision making.

1.2.1 Genomic epidemiology and contact tracing

Genomic epidemiology is the collecting and monitoring of whole genome sequence data over the course of an outbreak, often to track transmission dynamics and the emergence of new variants (Hill et al. 2023). Genomic epidemiology has become an increasingly important part of the response to disease outbreaks; perhaps the most well known example is the massive sequencing effort of the SARS-CoV-2 pandemic and subsequent tracking of new variants of concern (Robishaw et al. 2021). Prior to this, genomic epidemiology had also been used successfully during outbreaks of HIV, Ebola and Zika (Holmes et al. 1995; Gardy and Loman 2018), and can be implemented both retrospectively or while the outbreak is ongoing (Jackson et al. 2016).

The RABV genome is just under 12 kilobases in length, and made up of five genes; the nucleoprotein (N), phosphoprotein (P), matrix protein (M), glycoprotein (G) and the polymerase (or “large” protein; L) genes (Tordo et al. 1988). Until recently, most sequences generated from rabies samples were only partial genome sequences, most commonly either the N or G gene, but increasingly whole genomes are being generated and used in analysis (Jaswant et al. 2024). This is in part due to the development of more portable sequencing technologies, such as the Oxford Nanopore MinION, allowing RNA extraction and sequencing to be conducted close to where the sample was collected using a “lab-in-a-suitcase” setup (Brunker et al. 2020), and improvements in protocols reducing costs and time associated with sequencing (Bautista et al. 2023; Gigante et al. 2020). At the time of writing[1], 29,859 RABV sequences are available on the RABV-GLUE sequence database (Campbell et al. 2022), 3,425 of which are whole genome sequences.

The resulting RABV sequence data is usually used in two ways. The first is to discover the origins of an outbreak: when it started, how many introductions of the disease there have been to the area, and where the introductions came from. At its simplest, this analysis involves constructing phylogenetic trees from the sequences acquired from the outbreak along with sequences from outwith the outbreak and determining the location from which the closest related sequences were collected. For example, a fox rabies outbreak in Newfoundland was determined to likely have originated from a single introduction as the sequences were very similar to one another, but sequences from Greenland being interspersed with Canadian sequences in the phylogenetic tree suggests that rabid animals may have been crossing the sea ice between the two landmasses (Nadin-Davis et al. 2008a).

Using phylodynamic software that explicitly takes temporal and spatial data into account, such as BEAST (Drummond et al. 2012), can provide more in-depth investigations into the origin of the outbreak. For example, Brunker et al. (2015) used phylogeographic methods to determine that long distance dispersal events and co-circulation of lineages were common within Tanzania, and Lushasi et al. (2023) used sequence data to determine

[1] Data retrieved 16th September 2024

that an outbreak on Pemba Island resulted from two separate introductions from mainland Tanzania. Nadin-Davis et al. (2017) used whole-genome sequencing to determine that a raccoon rabies outbreak in Ontario, Canada likely resulted from a single incursion, rather than repeat incursions from New York state, USA, which previously could not be determined using partial genome data (Nadin-Davis et al. 2006). These more complex analyses require, however, that the sequence dataset has sufficient "temporal signal" (i.e., genetic divergence increases observably through time; Biek et al. 2015), and some rabies datasets lack this (Fusaro et al. 2013; Wang et al. 2019; Zhang et al. 2017; Faye et al. 2022; Caraballo et al. 2021).

The second way genomic data is commonly used is to determine relationships between cases by reconstructing transmission trees. These methods involve determining the probability that pairs of rabies cases are linked by comparing differences in the dates of symptom emergence and geographic locations to probability distributions of the generation interval and distance kernel. Transmission tree reconstruction methods that utilise only temporal and spatial data have been previously used in Tanzania (Lushasi et al. 2021), but if multiple co-circulating lineages are present it is possible to incorrectly link completely unrelated cases unless genomic data is taken into account. Incorporating genomic data can involve directly comparing the pairwise number of single nucleotide polymorphisms (SNPs) between cases and eliminating any links with a SNP distance above a cut-off value, as used by Cori et al. (2018), or by narrowing down the list of possibly related cases by splitting the sequences into separate clusters or lineages, such as the lineages outlined in the MADDOG nomenclature system (Campbell et al. 2022), and then using spatial- and temporal-data based transmission tree reconstruction methods within each cluster, as used by Yuson et al. (2024). These two methods essentially complete the same process but using a different order of operations.

1.2.2 Phylogenetics and host-associated clades

Genomic data has revealed that RABV can be split into eight major clades: the “Bats”, “RAC-SK”, “Cosmopolitan”, “Africa-2”, “Africa-3”, “Arctic”, “Asian” and “Indian Subcontinent” clades (Figure 1.2), some of which can be further split into a total of 33 minor clades (Troupin et al. 2016). As their names suggest, the “Bats” and “RAC-SK” major clades are mainly associated with bats and raccoons or skunks respectively, while the remaining major clades are associated with other carnivores in certain geographic areas. Palearctic bats are thought to be the original rabies reservoir (Badrane and Tordo 2001; Hayman et al. 2016), with the host shift event from bats to carnivores that resulted in the modern-day carnivore-associated clades being estimated to have occurred approximately 600 years ago (Troupin et al. 2016). The global spread of rabies may have been a direct result of the human-mediated transport of dogs during the height of European colonialism (Holtz et al. 2023), and the human movement of dogs over long distances continues to be a major source of rabies introductions to this day (Talbi et al. 2010).

While rabies can infect any mammal, sustained transmission has only been observed in a limited number of species (Holmes et al. 2002), with most minor rabies clades being associated with transmission within a specific host species (Troupin et al. 2016). For example, the Asian SEA2b and SEA5 clades circulate in Chinese ferret badger populations (Liu et al. 2010; Zhang et al. 2013), and the Africa-3 and Cosmopolitan AM2a clades circulate in mongooses in both southern Africa (Van Zyl et al. 2010) and the Caribbean (Nadin-Davis et al. 2008b) respectively, while separate clades also co-circulate in the dog populations of all of these regions. Despite extended periods of species-specific transmission, these clades are still able to infect other species, including humans, although in some cases transmissibility is reduced between species depending on the direction of transmission. For example, raccoon rabies strains can infect skunks more easily than skunk rabies strains can infect raccoons, in both natural and experimental infections (Hill et al. 1993; Wallace et al. 2014).

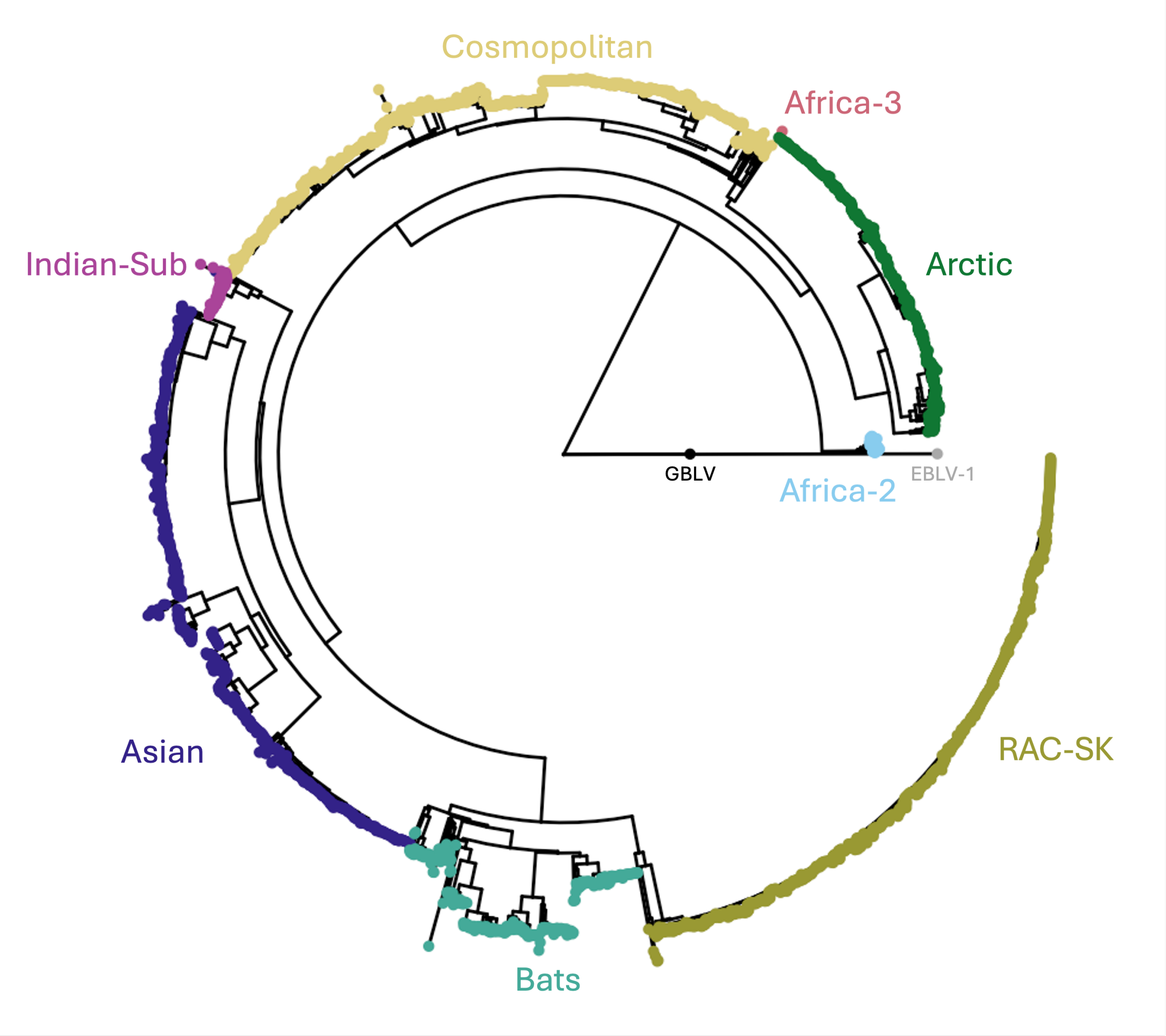

Figure 1.2: Phylogeny of the eight major rabies clades, European Bat Lyssavirus 1 (EBLV-1) and Gannoruwa Bat Lyssavirus (GBLV). The maximum-likelihood tree was constructed from complete N gene sequences with GTR+F+R5 model.

There has been debate as to whether this species-specific transmission is due to genetic adaptation, either pre- or post- host shift, or ecological circumstances (Marston et al. 2018; Mollentze et al. 2014); determining what factors contribute to rabies host-shift events could help to prevent them in the future. Positive selection in general is very rare in RABV, including on the glycoprotein, where non-synonymous diversity appears to be more highly constrained than in other RNA viruses (Holmes et al. 2002). Evidence of positive selection in relation to rabies host shifts is mixed, with some studies showing no evidence of adaptation to new host species (Kuzmin et al. 2012; Troupin et al. 2016) and others observing some positively selected sites but with low repetition across different host shift events (Streicker et al. 2012b). Troupin et al. (2016) also found, however, that an amino acid change from leucine to serine at nucleoprotein position 374 appears almost exclusively and ubiquitously in two separately emerging ferret badger-associated clades, which may suggest pre-adaptation is required to infect this new host. The reduced transmissibility of a skunk-associated strain of rabies to raccoons compared to skunks under experimental conditions also suggests that this species-specificity may not be purely due to the behaviour or geographic range of the host species (Hill et al. 1993). The focus by many of these studies on non-synonymous substitutions may miss other possible mechanisms of genetic adaptation to the host, such as changes in codon usage (Bahir et al. 2009), which may contribute to host specificity.

1.3 Thesis aims & organisation

This thesis aims to leverage the growing number of rabies genome sequences to explore how rabies evolves, and how we can use this invaluable data to assist in reaching the global goal of achieving zero human rabies deaths by 2030 (Minghui et al. 2018). This thesis consists of an introductory chapter, three data chapters in journal article format, and a discussion chapter. As each data chapter contains its own in-depth introduction and discussion, chapters 1 and 5 are kept brief.

The fact that viral replication and mutation are closely linked (Belshaw et al. 2008), taken together with RABV's variable lengths of time spent within the axons of nerve cells, implies that the difference in replication rate between the incubation and infectious periods could affect how RABV evolves. This, along with the difficulty oftentimes encountered while trying to conduct temporal phylogenetic analyses on rabies datasets, begs the question of whether rabies conforms to the most commonly used molecular clock models. In chapter 2 I use a simulation framework to explore how patterns of genetic divergence over time vary in rabies outbreaks where substitutions arise on a per-unit time (strict molecular clock) or a per-generation (controlling for variation in the incubation period) basis. I compare these synthetic patterns to those observed in genetic data from the Serengeti district of Tanzania, and estimate a per-generation substitution rate from these data. This chapter has been published in *PLoS Pathogens* as “Examining the molecular clock hypothesis for the contemporary evolution of the rabies virus” (DOI: 10.1371/journal.ppat.1012740, November 2024) with the following author list: Rowan Durrant, Christina Cobbold, Kirstyn Brunker, Kathryn Campbell, Jonathan Dushoff, Elaine Ferguson, Gurdeep Jaswant, Ahmed Lugelo, Kennedy Lushasi, Lwitiko Sikana, and Katie Hampson. Rowan Durrant (RD) wrote all R code except where otherwise stated, generated and analysed the data, and wrote the manuscript. Jonanthan Dushoff wrote the R code to generate lognormal Bayesian posteriors. Elaine Ferguson provided epidemiological simulation output. CC, Kirstyn Brunker, Kathryn Campbell, Jonathan Dushoff (JD), Elaine Ferguson (EF), and KH assisted in editing the manuscript for publication. During the review process the work in this chapter received feedback from two anonymous reviewers.

Estimates of the size of an outbreak, either as part of the initial outbreak response or retroactively, can inform the scale of the outbreak response needed or can be used to evaluate how successful contact tracing efforts were (or indeed surveillance performance more generally is) in detecting transmission. Commonly used methods of outbreak size estimation such as seroprevalence surveys are unsuitable for rabies outbreaks, and existing phylodynamic methods are computationally expensive and require that the dataset has sufficient temporal signal in order to conduct them. In chapter 3 I develop a simple mathematical method of estimating outbreak sizes that only requires a maximum-likelihood phylogenetic tree and an estimate of the per-generation substitution rate. I test this method using the simulation framework developed in chapter 2 and apply it to genetic data from a recent rabies outbreak in the Philippines. This chapter is in preparation for publication with the following provisional author list: Rowan Durrant, Jonathan Dushoff, Christina Cobbold, Daniel Haydon, Elaine Ferguson, Criselda Bautista, Kirstyn Brunker, Daria Manalo, Mary Elizabeth Miranda, Mirava Yuson and Katie Hampson. RD wrote all R code except where otherwise stated, generated and analysed the data and wrote the manuscript. EF provided some epidemiological simulation output. BEAST log files were provided by Criselda Bautista. JD assisted with early stages of deriving the equation and wrote the R code to generate lognormal Bayesian posteriors. JD and Daniel Haydon provided useful discussion and feedback. Mirava Yuson, Daria Manalo and Mary Elizabeth Miranda provided the Romblon genome sequences.

Most rabies clades circulate predominantly within a specific host species, but there is debate as to whether this is purely due to ecology or whether genetic adaptation to the host plays a role. In chapter 4 I leave the simulation framework to explore how the rabies virus evolved in practice and investigate whether there is evidence of genetic adaptation to the host hidden in synonymous substitutions. I explore differences in codon usage and CpG content in a range of bat- and carnivore- associated rabies clades and explore what forces could be driving these differences, including host immune mechanisms. This chapter is in preparation for publication with the following provisional author list: Rowan Durrant, Jonathan Dushoff, Christina Cobbold, and Katie Hampson. All conceptualisation, analysis and writing was conducted by RD, and RD wrote all R code except where otherwise stated. JD and Denise Marston provided useful discussion and feedback. Spyros Lytras shared code for the PhylogeneticEM analysis.

Finally, in chapter 5 I discuss how my findings contribute to our understanding of rabies evolution and elimination, the implications of these findings, and what further work is required to answer questions that arise from these.

Chapter 2

Examining the molecular clock hypothesis for the contemporary evolution of the rabies virus

2.1 Abstract

The molecular clock hypothesis assumes that mutations accumulate on a genome at a constant rate over time, but this assumption does not always hold true. While modelling approaches exist to accommodate deviations from a strict molecular clock, assumptions about rate variation may not fully represent the underlying evolutionary processes. There is considerable variability in rabies virus (RABV) incubation periods, ranging from days to over a year, during which viral replication may be reduced. This prompts the question of whether modelling RABV on a per-infection generation basis might be more appropriate. We investigate how variable incubation periods affect root-to-tip divergence under per-unit time and per-generation models of mutation. Additionally, we assess how well these models represent root-to-tip divergence in time-stamped RABV sequences. We find that at low substitution rates (<1 substitution per genome per generation) divergence patterns between these models are difficult to distinguish, while above this threshold differences become apparent across a range of sampling rates. Using a Tanzanian RABV

dataset, we calculate the mean per-generation substitution rate to be 0.17 substitutions per genome per generation. At RABV's substitution rate, the per-generation substitution model is unlikely to represent rabies evolution substantially differently than the molecular clock model when examining contemporary outbreaks; over enough generations for any divergence to accumulate, extreme incubation periods average out. However, measuring substitution rates as per-generation holds potential in applications such as inferring transmission trees and predicting lineage emergence.

2.2 Introduction

The molecular clock hypothesis assumes that the genomes of organisms accumulate neutral mutations at a constant rate over time, either across all lineages (the "strict molecular clock") or within each individual lineage but with some degree of variation between them (clock models with this assumption include the relaxed and multirate clock models; Gojobori et al. 1990; Drummond et al. 2006; Ho and Duchêne 2014).The ability to sample viral sequences through time, and the application of the molecular clock hypothesis to these sequences, has led to massive advances in using viral genetic data to investigate disease outbreaks (Drummond et al. 2003a). The substitution rate, measured in substitutions per site per unit time, can be used to estimate how long ago pathogens diverged (Pybus and Rambaut 2009), and the date of infection of individual infected hosts (Wróbel et al. 2006). Combining the analysis of epidemiological and genetic data has allowed further insights into the history of outbreaks (Grenfell et al. 2004), and the introduction of geographic data provides estimates as to rates of spatial spread and the frequency and source of introductions (Gire et al. 2014; Kamath et al. 2016). However, in order to conduct these phylogenetic analyses, genetic divergence must increase appreciably over time in the dataset under investigation (Drummond et al. 2003b). Whether or not the viral population is measurably evolving, and thus whether a dataset of sequences contains sufficient temporal signal for Bayesian phylogenetic analysis, depends mainly on the evolutionary rate, the sequence length and the length of time sequences are sampled over being sufficiently high. Various methods exist to assess temporal signal, the most commonly used being root-to-tip divergence plots (Korber et al. 2000; Buonagurio et al. 1986) implemented in tools such as TempEst (Rambaut et al. 2016), but these also include Bayesian evaluation of temporal signal (BETS; Duchêne et al. 2020a) and the date-randomisation test (Duchêne et al. 2015).

The rabies virus (RABV) is a negative-strand RNA virus, with a genome size of approximately 12 kilobases. While RNA viruses generally have high mutation rates due to a lack of proofreading by RNA polymerases, RABV has a substitution rate at the lower end of normal for single-stranded RNA viruses of between 1×10^{-4} and 5×10^{-4} substitutions per site per year (Holmes et al. 2002; Biek et al. 2015; Layan et al. 2021). This may be due to strong purifying selection (Holmes et al. 2002), or due to peculiarities of the rabies virus. For example, the RABV genome is longer than average for RNA viruses, and genome length and evolutionary rate are negatively correlated (Duchêne and Holmes 2018), although this relationship appears to be weaker in single-stranded RNA viruses (Sanjuán 2012). A more unusual feature of RABV is that infections can exhibit extended incubation periods within the host. The median generation interval (the time between one individual becoming infected and then infecting another) is estimated to be 17.3 days in domestic dogs (Mancy et al. 2022), with other studies estimating mean serial intervals of 26.3 days (Hayes et al. 2022) and 45.0 days (Kurosawa et al. 2017). Symptoms, infectivity, and death from rabies, however, can occasionally occur years after the initial infection event (Boland et al. 2014). The length of the incubation period is influenced by the route of exposure, with bites to the head and neck leading to more rapid disease progression than bites to lower extremities (Dimaano et al. 2011). RABV can remain in the muscle at the bite site for prolonged lengths of time before invading the host's motor neurons and progressing through the nervous system, with limited, if any, infection of other muscle fibres (Charlton et al. 1997). While some replication in the muscle cells has been observed (Yamaoka et al. 2013), RABV replication at the inoculation site is not necessary for neural invasion (Shankar et al. 1991). It is currently unknown precisely how the RABV replication rate in the host muscle cells and peripheral nervous system compares to the massive replication rate within the cells of the central nervous system and brain. However, work suggests that RABV replication in muscle cells may be reduced (Schnell et al. 2010), and RABV replication in cultured rat sensory neurons may be 10- to 100-fold lower than replication rates in rat and mouse CNS neurons (Lycke and Tsiang 1987). Rabies infections that involve long incubation periods may, therefore, not lead to more accumulated mutations than those with shorter incubation periods, as viral mutation is strongly influenced by the replication process (Belshaw et al. 2008).

Changes in mutation rates through time on an individual case level due to long incubation periods may affect how we analyse RABV sequence data and interpret these analyses on a population level. A relaxed molecular clock is usually required to carry out phylogenetic analyses on rabies datasets, and it is not uncommon for there to be difficulties in applying these analyses due to “insufficient temporal signal”; usually referring to either no or a negative relationship between genetic divergence and time, or this relationship having a lot of noise and a very low R^2 (Fusaro et al. 2013; Wang et al. 2019; Zhang et al. 2017; Faye et al. 2022; Caraballo et al. 2021). RABV shows variation in substitution rate between lineages (Troupin et al. 2016; Streicker et al. 2012a; Layan et al. 2021), which may be driven in part by differences in incubation periods. If the variable incubation period of rabies infections does cause deviation from the molecular clock model (exceeding the variation captured by relaxed or multirate clock models), this may negatively affect the accuracy of time-scaled phylogenetic trees and emergence date predictions. Conversely, if mutation does continue at a consistent rate during the incubation period, attention should be paid to extremely long incubators which could drive the emergence of new variants, as seen recently in chronic SARS-CoV-2 infections (Kemp et al. 2021; Choi et al. 2020).

We hypothesised that reduced replication (and thus mutation) during the incubation period could cause rabies evolution to be better represented by a per-generation model of mutation than by the strict molecular clock model. We aim to clarify the nature of contemporary RABV evolution using *in silico* methods, comparing the root-to-tip divergence of sequences generated from synthetic outbreaks under per-unit time or per-generation mutation models, and comparing these to RABV genomic data from Tanzania. We also aim to calculate a per-generation substitution rate for RABV for future use as a parameter in transmission tree reconstruction algorithms.

2.3 Methods

We investigate two contrasting mutational models for RABV – i.e., substitutions occurring on a per-generation vs. per-unit-time basis – using a simulation approach. We first generated synthetic RABV outbreaks using a branching process model (Mancy et al. 2022) and then simulated these two mutation processes over the resulting transmission trees. From the synthetic sequences generated, we examined root-to-tip divergence and calculated the variance explained ($\boldsymbol{R}^2$) from linear regressions, and compared these to the root-to-tip divergence of a set of RABV whole genome sequences from Tanzania. Finally, we developed a method to estimate the per-generation substitution rate for RABV and tested this on synthetic data before applying it to the Tanzanian RABV dataset.

2.3.1 Rabies outbreak simulation

We simulated RABV mutation on branching-process simulations of rabies outbreaks. Outbreaks were simulated 100 times over a spatially explicit representation of the Mara Region in northern Tanzania. In the Serengeti District of this region, where contact tracing data were available, the model was initialised with the three cases that occurred in the mean generation interval (g=27 days, based on contact tracing data) prior to 2017 (simulations were run over a dog population representing that in the Mara region between 2017 and 2024). In the rest of the Mara region, where there were no data to guide initialisation, we seeded with (0.01Dg)/365 cases, where D is the initial dog population in that area. If $\boldsymbol{R_e}$ is equal to 1 (endemic transmission), this results in roughly 1% of the population becoming rabid over a year; contact tracing data suggest that incidence typically does not exceed that level (Hampson et al. 2009). This led to a total of 273 initial cases in the region. Each case was assigned a number of offspring cases drawn from a negative binomial distribution (ibid.) with a mean R_0 of 1.05 and a dispersion parameter of 1.33. This R_0 value was chosen to result in a median number of cases each month that was roughly constant over time (over the 100 simulations), mimicking endemic disease, with the dispersion parameter giving a distribution in secondary case numbers that reflects observed patterns in rabies transmission, where most cases result in no onward transmission and an increasingly small proportion of cases result in an increasingly large number of secondary cases.. Movement of rabid dogs from their home locations to and between transmission locations followed a random walk with step lengths drawn from a Weibull distribution (shape = 0.41; scale = 0.13). We simulated occasional long-distance transport of dogs to a random location prior to their first transmission in 2% of cases (Mancy et al. 2022). At each of a rabid dog's transmission locations, another dog was randomly selected within the local 1km^2 grid cell. If this dog was susceptible (i.e., not vaccinated or already incubating infection from a prior transmission event), rabies was transmitted. While not explicitly explored in this study, a model with a spatial element was chosen as the underlying epidemiological model as these more accurately reflect rabies dynamics such as local susceptible depletion and long distance movement, which allow rabies to persist at low incidence in the population (ibid.). A generation interval was drawn for each new infection from a lognormal distribution (meanlog = 2.96; sdlog = 0.82, fit to contact tracing data from Serengeti District, Tanzania (ibid.); Figure 2.1), describing the time delay before it also became rabid and made its assigned transmissions. The step-length distribution was also fitted using this contact tracing data (ibid.). Branching process simulations were continued until 7 years had passed or rabies went extinct. Each synthetic case was assigned an individual ID, and for every case (except initial seed cases) we recorded the ID of the associated progenitor case. Dates of infection and transmission were recorded for each case.

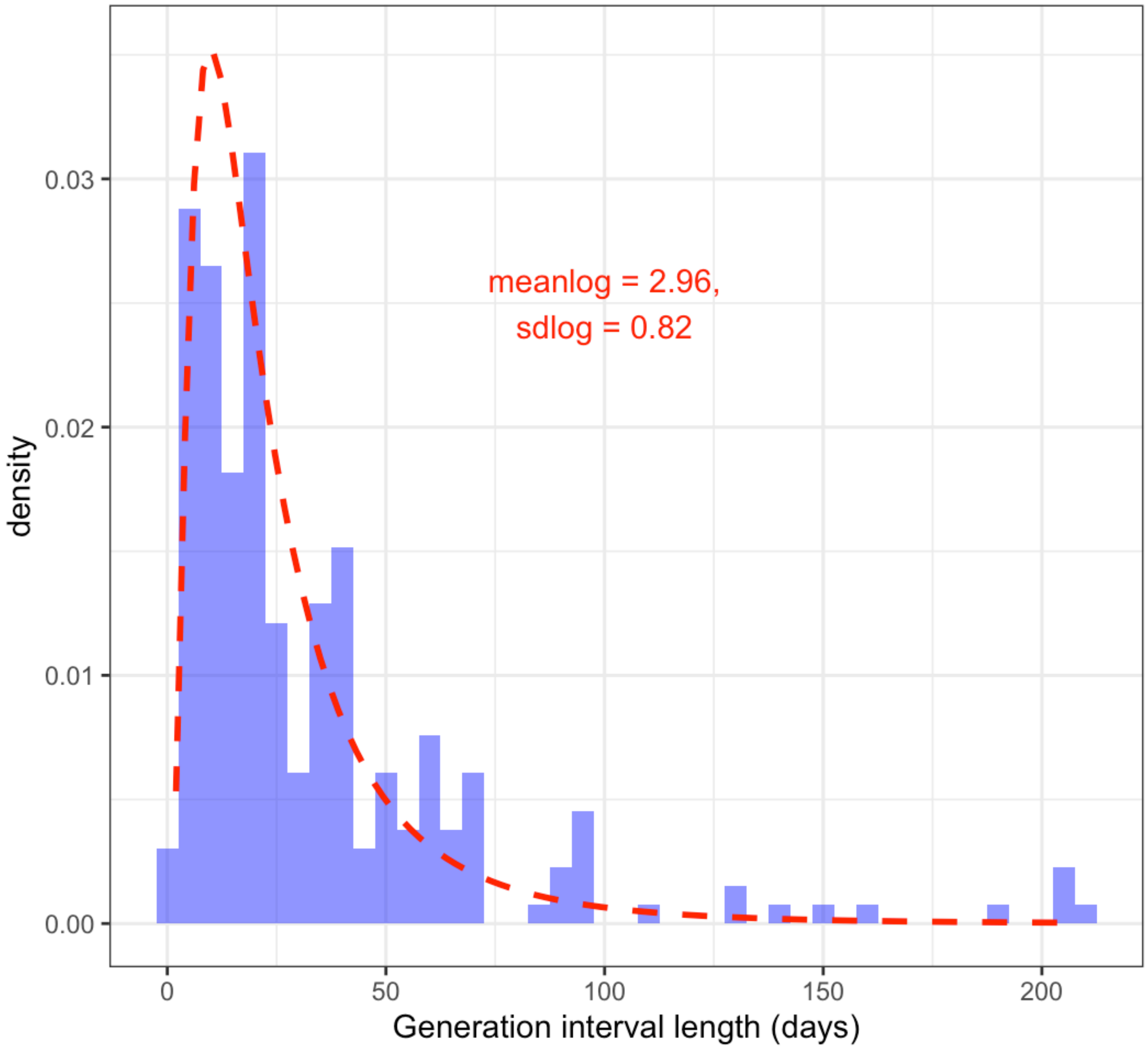

Figure 2.1: Histogram of generation intervals from the Tanzanian contact tracing dataset with the lognormal distribution used in simulations.

We isolated complete transmission trees descending from each of the 273 initial cases from within one randomly selected synthetic outbreak. Transmission trees that contained over 100 cases (9 out of 273 trees in total, ranging in size from 533 - 19,382 cases) were then used to generate synthetic sequence data. Across these trees, we see a mean generation interval of 26.6 days, and 2.5 and 97.5 percentiles of 3.90 and 94.11 days (Figure 2.2). For each of the 9 trees, the index case was assigned an initial 12kb genome sequence. Under the per-unit time mutation model, we determined the expected number of mutations by multiplying the substitution rate, the genome length and the length of the generation interval, for each case along the resulting transmission chain (because we assume mutations are neutral, the individual-level mutation rate is the same as the population-level substitution rate). The realised number of mutations was then drawn from a Poisson distribution, with this mean. We then randomly chose positions to change and new nucleotides to change them to. The resulting synthetic sequence data is referred to as the "time-based sequence data". The generation-based model of mutation works as above, with the exception that the expected number of substitutions in a generation is constant regardless of the generation interval length, and produces the synthetic "generation-based sequence data".

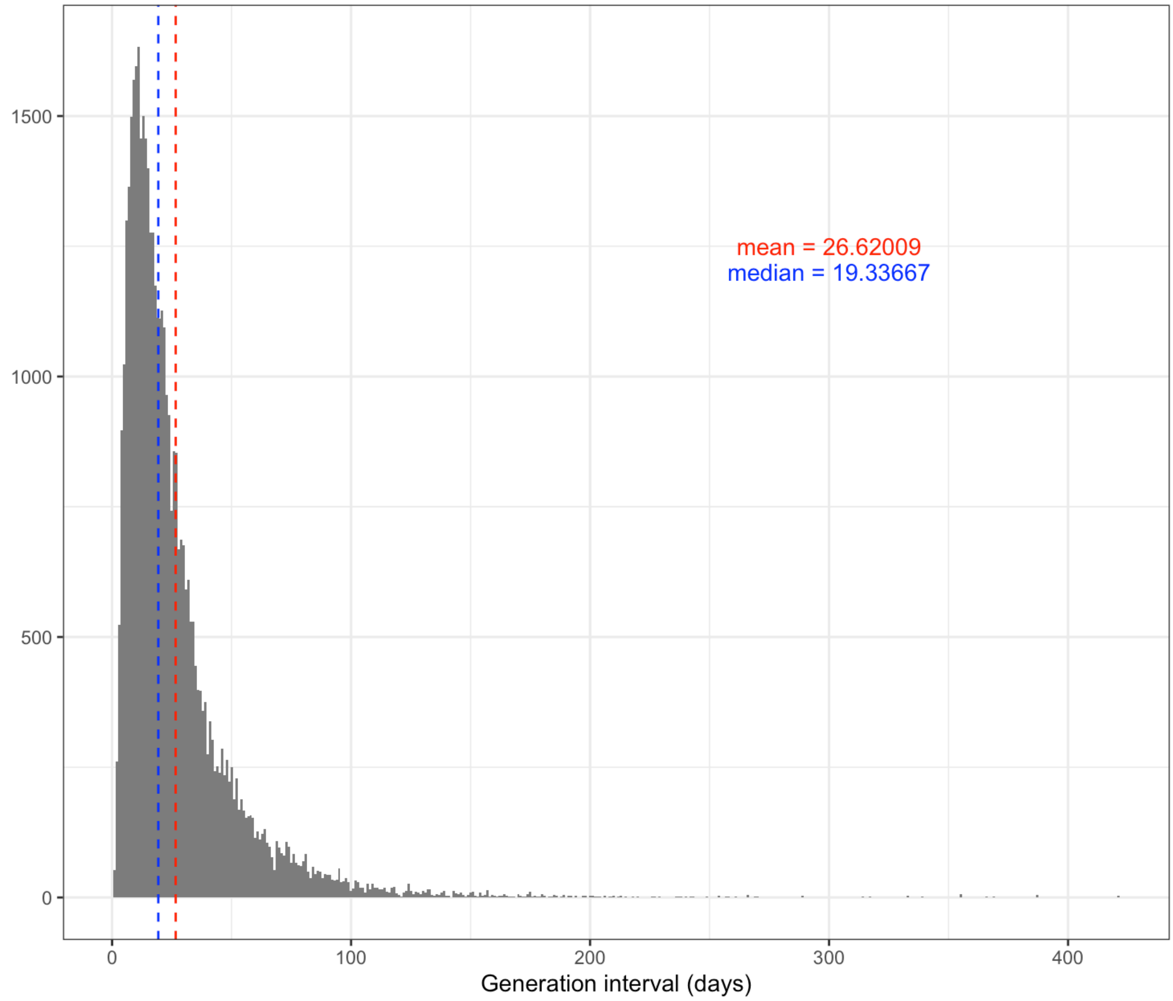

Figure 2.2: Histogram of generation intervals from the simulated outbreaks. Vertical dashed lines represent the median (blue) and mean (red) generation interval.

2.3.2 Divergence rate analysis

To investigate patterns of temporal divergence under the mutation models described above, we generated synthetic data with values of substitution rates ranging from 0.05 to 3 substitutions per genome per generation (or the per unit time substitution rate equivalent) and 4 population sampling regimes (from 1% of cases to 20%, informed by a previous study that estimated that routine surveillance for rabies rarely confirms more than 10% of circulating cases; Townsend et al. 2013b). We calculated the genetic divergence as the number of nucleotide differences from the index case to each sampled case. For each of the nine transmission trees, we then compared genetic divergence with time under each scenario (substitution rate and sampling regime combination), using linear regression through the origin. In order to compare our synthetic patterns of divergence over time to real rabies data, a root-to-tip divergence plot was also generated for a dataset of real RABV sequences (data from Lushasi et al. (2023); Figure 2.3A) using TempEst (v1.5.3; Rambaut et al. 2016), with the best-fit root located (Figure 2.3B). These rabies cases occurred between 2001 and 2017 and were primarily from the Serengeti district and Pemba Island, with the remaining sequences from elsewhere in Tanzania (Figure 2.3A inset). Sequence acquisition and tree building methods are detailed in Lushasi et al. (2023).

2.3.3 Calculating the per-generation substitution rate

We updated a method of calculating the per-generation substitution rate previously used in eukaryotes (Slatkin and Hudson 1991) by using Bayesian posterior estimates of the substitution rate and generation interval. We assessed this method's accuracy using the synthetic outbreak sequence data, before applying it to the aforementioned set of RABV whole genome sequences.

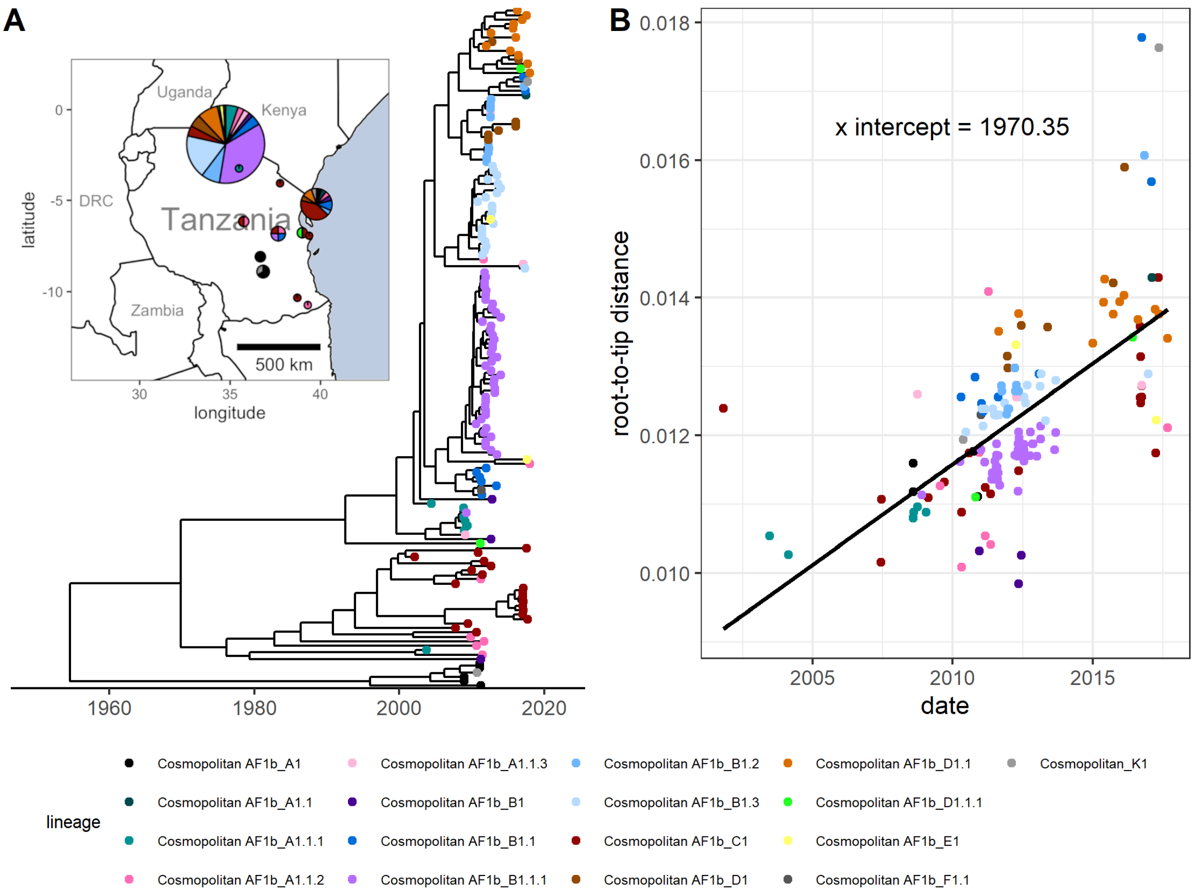

Figure 2.3: Phylogeny of real rabies virus whole genome sequences from Tanzania and root-to-tip divergence. (A) The time-scaled tree Lushasi et al. (2023) used to generate the root-to-tip divergence plot and to calculate the per-generation substitution rate. The inset map shows the approximate locations that the samples were collected from and the lineages present in each location. Map point size represents the number of sequences in this dataset from district centroid locations. Base map data is from Natural Earth (`naturalearthdata.com`), via the `maps` R package. (B) The corresponding root-to-tip divergence plot. Point colours represent RABV lineage.

To estimate the mean per-generation substitution rate, we analysed sequence data with BEAST, and multiplied the posterior rate estimate for each MCMC sample (excluding the burn-in period) by the generation interval lengths sampled from the posterior of a simple Bayesian analysis and then multiplied again by the genome length. The mean and 95% credible interval of the estimate of the per-generation substitution rate for the RABV dataset was calculated by taking the mean and the 2.5% and 97.5% percentiles of the resulting multiplied posteriors.

To evaluate the accuracy of this method in estimating the mean per-generation substitution rate, we also applied it to synthetic sequence data generated from outbreaks using the per-generation mutation model as described above, under different substitution rates (11 values ranging from 0.05 substitutions per generation to 3 substitutions per generation) and case sampling rates (1%, 5%, 10% or 20% of cases sampled) across the 9 transmission trees that contained at least 100 cases. Subsampled synthetic datasets containing more than 2000 sequences were not analysed as this number exceeds the total whole-genome RABV sequences currently available on the RABV-GLUE database (a repository of RABV sequences regularly updated from the NCBI Nucleotide database, combined with metadata from NCBI and the source publications, and assigned major and minor clades; Campbell et al. 2022), and is unrealistic in the context of examining individual rabies outbreaks. BEAST log files were generated from these sequences using BEAST-Gen version 1.0.2 and BEAST version 1.10.4 (Suchard et al. 2018). We chose to use a Jukes-Cantor (JC) substitution model with a strict clock, no site heterogeneity (due to our per-generation mutation model used in the simulations having equal probability of any site or base being chosen) and assumed a constant population size. We used a tracelog frequency of 1000 and a sufficiently long chain length determined for each dataset so that the effective sample size of each parameter exceeded 200 when analysed using Tracer (Rambaut et al. 2018) with a 10% burn-in period. We applied the substitution rate calculation method to these phylogenetic trees, and assessed the accuracy of the resulting mean per-generation substitution rates by comparing them to the parameter values used to generate the synthetic sequences, using the natural log of the ratio:

$$Deviation = \ln\left(\frac{\mu_e}{\mu_a}\right) \tag{2.1}$$

where μ_e is the mean estimated per-generation substitution rate and μ_a is the actual substitution rate, where a deviation of zero means perfect accuracy.

The same method was applied to the dataset of 153 RABV sequences sampled from across Tanzania (data from Lushasi et al. (2023); Figure 2.3A). The mean per-generation substitution rate was calculated, and distributions were fitted from the multiplied generation interval and substitution rate posteriors (generation interval posterior based on values from Mancy et al. (2022) for the Tanzanian dataset, extracted directly from the lognormal distribution used in simulations; substitution rate posteriors taken from the BEAST log file of the time-scaled tree from Lushasi et al. (2023)) and genome length as described above. We compared different distributions (Gamma and Lognormal) for estimating substitution rates and selected the best fitting distribution by AIC. We also calculated the probabilities of between 0 and 10 SNP differences occurring across 1, 5 or 10 infection generations. For this calculation we simulated mutations arising at a Poisson rate with lambda drawn from the fitted substitution rate distribution. The means and 95% confidence intervals were calculated from the 10,000 simulations.

2.3.4 Data and code availability

All code is available at `https://github.com/RowanDurrant/Rabies-Mutation`. Analyses were conducted using the R programming language (R Core Team 2020). The beta regression curve and prediction interval in Figure 2.4C was generated using the `betareg` R package (Cribari-Neto and Zeileis 2010). RABV lineages were assigned using MADDOG (Campbell et al. 2022).

2.4 Results

2.4.1 Root-to-tip divergence analysis

At higher per-generation substitution rates (1 substitution per genome per generation and above), distinct differences can be seen between root-to-tip divergence plots from the two models of mutation (Figure 2.4A). The synthetic data generated from the per-generation mutation model shows “stray” clusters or ridges of points both above and below the main funnel of points, illustrated in the example in Figure 2.4A. Divergence plots from synthetic data generated from the time-based model of mutation have less variance and do not exhibit this pattern. At lower substitution rates (below 1 substitution per generation), no such pattern is clearly distinguishable (Figure 2.4B). When the cases represented by the high-divergence points from the per-generation model in Figure 2.4A are visualised in a transmission tree, they are mainly confined to a single transmission chain (Figure 2.5).

Root-to-tip divergence plots derived from synthetic transmission trees using the time-based mutation model had, on average, higher R^2 values than those from synthetic transmission trees using the per-generation mutation model, although this is more difficult to distinguish below a substitution rate of 0.5 substitutions per genome per generation (Figure 2.4C). As the substitution rate increases, the R^2 values across both mutation models increase. The case sampling rate appears to have little effect on R^2 (Figure 2.6). The root-to-tip divergence plot of the Tanzanian RABV dataset more closely resembles those of lower substitution rate simulations, where it is difficult to determine any difference between the models of mutation (Figure 2.3). While most lineages surround the regression line, some (for example Cosmopolitan AF1b_B1) group below the line, but without forming a distinguishable “ridge”.

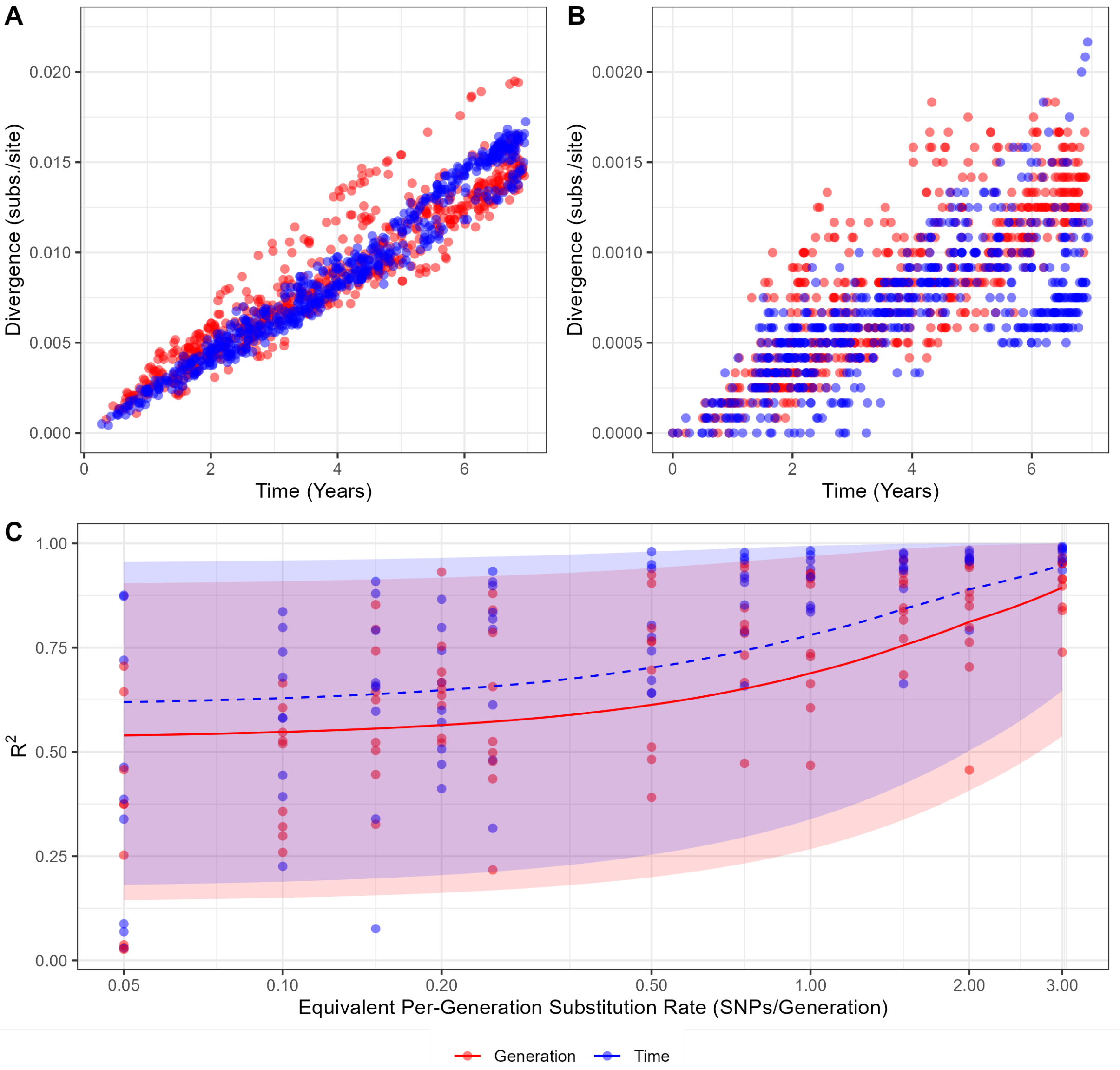

Figure 2.4: Temporal genetic divergence varies under two models of mutation. (A) Root-to-tip divergence plots for synthetic sequences produced using time-based and generation-based mutation models, equivalent to 2 substitutions per genome per generation and (B) equivalent to 0.2 substitutions per genome per generation. Note that the y-axis scales differ by an order of magnitude between A and B. These data are from running mutation models over the same single synthetic outbreak where 5% of cases were sequenced (i.e., 621 cases sequenced of 12,434 total). (C) The R^2 values obtained from regression through the origin of root-to-tip divergence of synthetic data from the time-based and generation-based models. Point colour indicates the mutation model used to generate the data. Lines represent beta regressions with logit links fit to data points, and shading represents the 95% prediction interval. The X axis is log scaled. 5% of cases were sequenced here; the proportion of cases sequenced had little effect on R^2 (Figure 2.6).

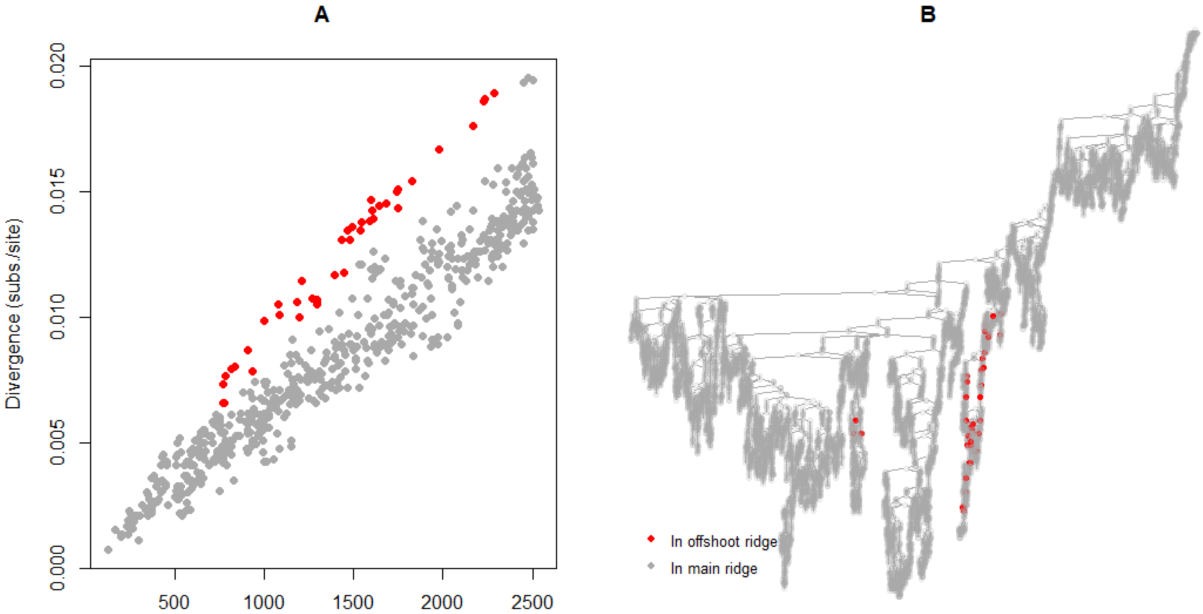

Figure 2.5: Points in the offshoot ridge predominantly occur in one transmission tree. (A) root-to-tip divergence plot (2 substitutions per genome per generation, 5% of cases sequenced) with offshoot ridge points highlighted in red. Offshoot ridge points are defined in this plot as having a divergence rate above 8×10^{-6} substitutions per day and occurring after day 750. (B) transmission tree of the simulated outbreak with offshoot ridge cases highlighted in red. Graph edge length is not proportional to time or divergence.

2.4.2 Substitution rate calculation

The accuracy of our method used to calculate per-generation substitution rate remains similar at all but the lowest values of substitution rate (Figure 2.7), with a tendency to underestimate the substitution rate (meaning that the estimated substitution rate is below the substitution rate parameter used to generate the synthetic data; mean natural log of the ratio of -0.18 and root-mean-square of 0.54, where values of 0 indicate perfect estimates). Accuracy appears to be more influenced by the number of sequences used in the BEAST analysis than by the case sampling rate itself; the mean natural log of the ratio falls to -0.36 when fewer than 50 sequences are used (root-square-mean of 0.95).

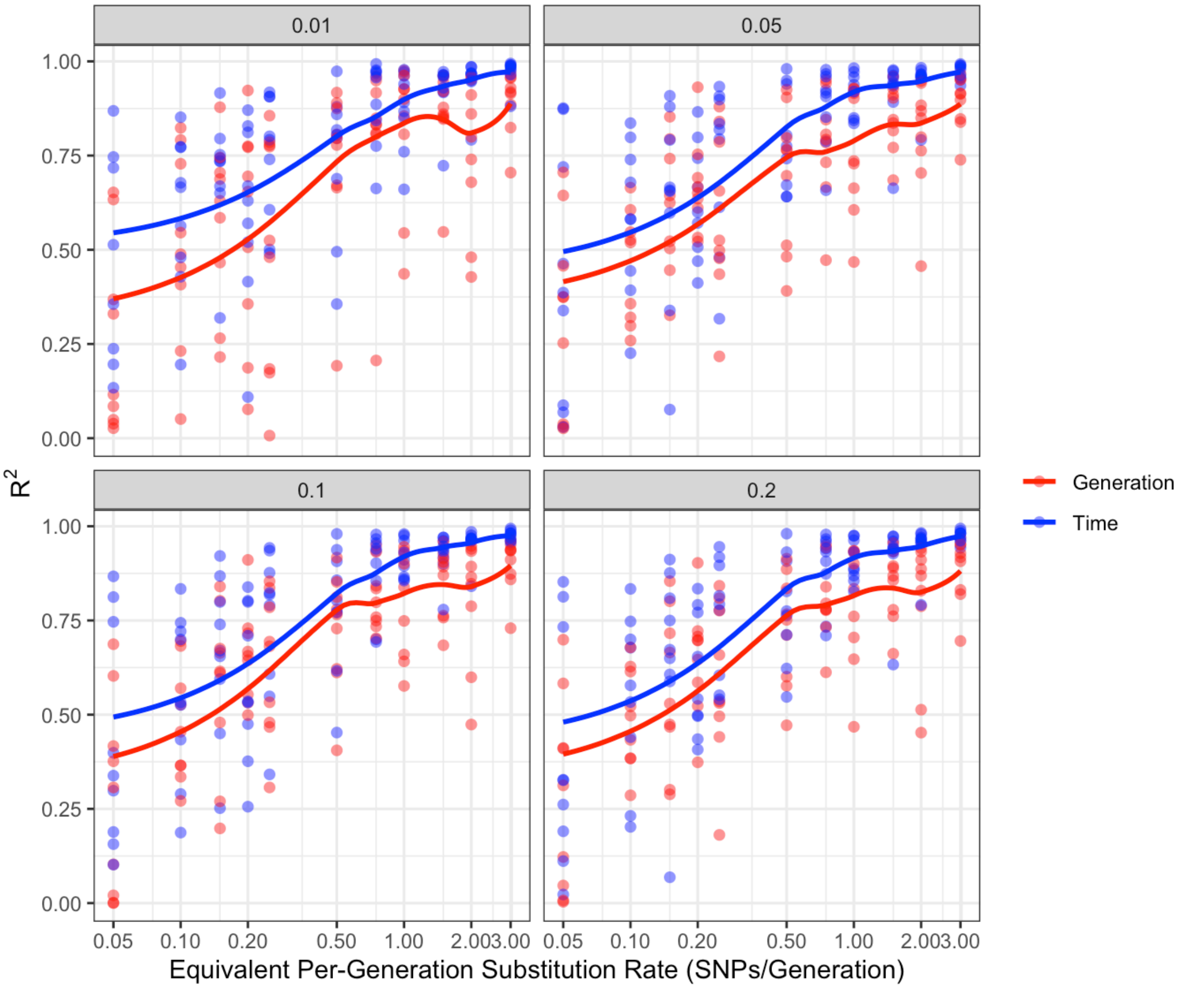

Figure 2.6: The proportion of cases sequenced does not impact the R^2 of root-to-tip divergence plots from synthetic data. Plot is faceted by the proportion of cases sequenced, point colour represents the mutation model.

The Tanzanian RABV dataset from which we estimated the per-generation substitution rate contains 153 sequences in total, and the accompanying time-scaled phylogenetic tree has a root-to-tip height of approximately 65 years, although the sequences spanned just 16 years as they were sampled from 2001 to 2017 (with 46.7% from years 2011-2012). These sequences were largely complete; 98% of sequences were >95% complete (>11,327 kb in length). The mean per generation substitution rate of RABV in this dataset was estimated to be 0.171 (95% credible interval: 0.127 - 0.219). The best fitting distribution by AIC to the output of the multiplied Bayesian posteriors was a Gamma distribution with the parameters shape = 51.69 and rate = 301.8 (Table 2.1).

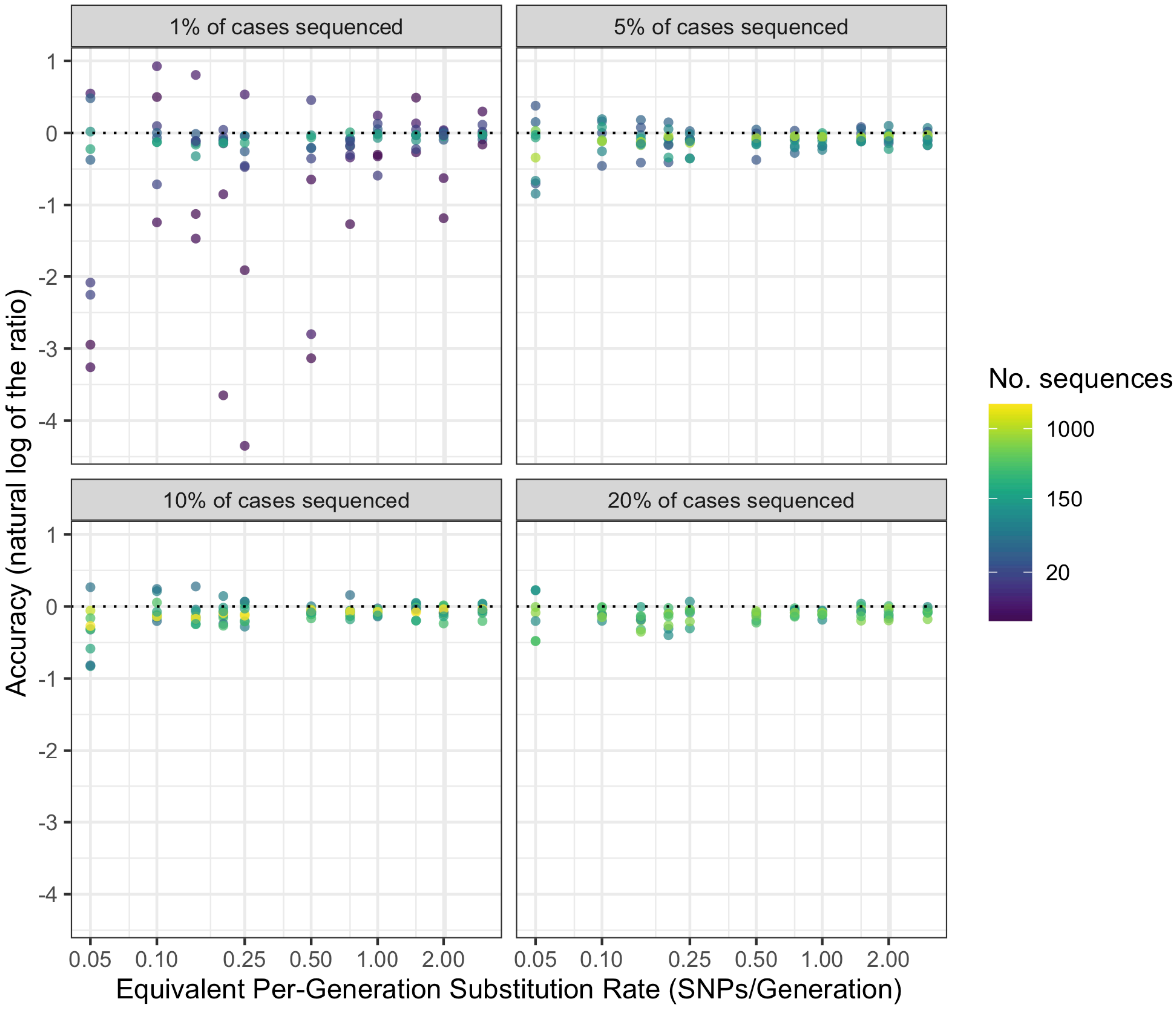

Figure 2.7: Accuracy of per-generation substitution rate predictions for different numbers of sequences, substitution rates and sampling rates. Facets indicate the proportion of cases sequenced. The dotted line represents perfect accuracy. The x-axis and colour scale are log transformed.

Distribution	Parameters	Δ AIC
Gamma	shape = 51.69, rate = 301.8	0
Lognormal	meanlog = -1.774, sdlog = 0.1402	1277.4

Table 2.1: Parameters and distributions used to estimate the per generation substitution rate. The distributions fitted to the multiplied Bayesian output are shown, along with their parameters and AIC.

Using the estimated per-generation substitution rates, we calculated the probability of different numbers of substitutions occurring over 1, 5 and 10 generations, drawing the per-generation substitution rate from the above distribution (Figure 2.8). Over many generations it is still quite likely for zero substitutions to occur; after 10 generations, the probability of zero substitutions having occurred is 0.19.

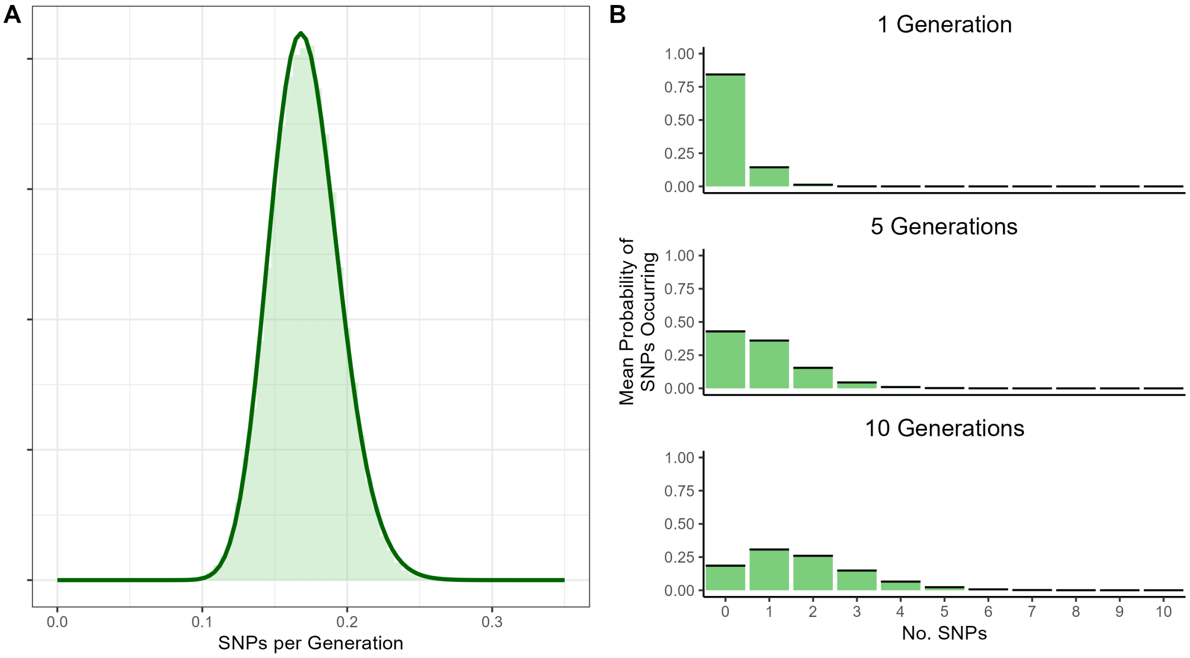

Figure 2.8: Probability distributions of the mean per-generation substitution rate and substitutions occurring over generations. (A) estimated probability distribution of the per-generation substitution rate from Tanzanian RABV sequences. (B) probability distribution of substitutions occurring over 1, 5 and 10 generations. The λ value for a Poisson rate of substitution occurrence is drawn from the substitutions per generation distribution fitted in Figure 2.8A. Black bars represent the 95% confidence intervals (which are very tight).

2.5 Discussion

Some RABV sequence datasets show little to no temporal signal despite the sequences being gathered over many years, which we hypothesised could be due in part to its variable incubation periods. We hypothesised that a per-generation model of mutation may be more representative of RABV evolution than a purely time-based model. We found that substantial differences in root-to-tip divergence patterns between synthetic outbreaks using generation-based and time-based models of mutation could be observed only at high underlying substitution rates. The substitution rate for the Tanzanian RABV sequences examined (~0.17 substitutions per genome per generation) was in the range where divergence patterns in the two models were extremely similar. We can thus assume that the two models will give extremely similar results on the relevant time scale. As we observed increasing divergence over time with reasonable R^2 values within this substitution rate range, it implies that variable incubation periods alone do not fully account for the lack of temporal signal. Therefore, other factors such as insufficiently long sampling windows for the substitution rate are likely to be responsible (Duchêne et al. 2020b). This is an important consideration for analysing RABV sequences from new outbreaks, or from endemic areas where sampling is opportunistic. As RABV has a substitution rate lower than many other RNA viruses, longer sampling windows are required to construct a dataset with sufficient temporal signal to conduct Bayesian phlyogenetic analyses.

The observation of little difference between root-to-tip divergence plots derived from the two mutation models at substitution rates below 1 substitution per genome per generation is likely due to averaging; multiple generations of infection are expected to have occurred per substitution that arises on the viral genome. Over the many generations needed before significant levels of viral genetic diversity are reached, the influence of any unusually long incubation periods will be damped by the opposite influence of unusually short incubation periods, eventually becoming indistinguishable from clock-like behaviour. On the other hand, at higher substitution rates ridges form on the root-to-tip divergence plots under the per-generation model of mutation but not under the per unit time model. Increasing both the substitution rate and the variability of the generation interval length would likely lead to the greatest differences between root-to-tip divergence plots between the two mutation models, due to the combined increase in the frequency of substitutions occurring during unusually long or short incubation periods.

While not affecting the overall substitution rate, these ridges reduce the overall R^2, and may be better analysed using a separate local clock (Featherstone et al. 2023). The cases in these ridges almost all descend from a common ancestor (Figure 2.5), suggesting that a single unusually long or short incubation period can affect which phylogenetic analyses we perform. Ridges caused by these incubation periods can be distinguished from ridges caused by rate variation between lineages as they will be parallel to the main cluster of points in the plot, whereas points belonging to lineages with a different substitution rate will have a different slope. Studies examining the number of substitutions occurring between successive sequenced cases, and whether this increases when the secondary case's incubation period is unusually long, could clarify the exact relationship between substitutions, generations, and time in the rabies virus. More detailed data will be required to investigate this further.

We calculated RABV's mean per-generation substitution rate to be approximately 0.17 substitutions per genome per generation. This estimate is lower than those of other RNA viruses, such as SARS (2 substitutions per genome per human passage; Vega et al. 2004), SARS-CoV-2 (0.52 substitutions per genome per 5.8-day generation interval; Braun et al. 2021) and the Ebola virus (0.875 substitutions per genome per 14-day generation interval; Kinganda-Lusamaki et al. 2020). RNA viruses that undergo periods of reduced replication or complete latency often show reduced per-unit time substitution rates, with one extreme example being HTLV-1/2 (Holmes 2003; Van Dooren et al. 2001). However, we would not expect these latent periods to affect the per-generation rate, as the longer generation interval would take this into account. The low per-generation substitution rate seen in rabies is therefore likely due to mutation being constrained by other factors, such as strong purifying selection (Holmes et al. 2002), and likely contributes to the observed lack of temporal signal. Previous studies suggest that for viruses in this substitution rate range, sampling windows of up to 30 years may be required to overcome the phylodynamic threshold (Duchêne et al. 2015); for comparison, SARS-CoV-2 was observed to have sufficient temporal signal to conduct these analyses within two months of the start of the pandemic (Duchêne et al. 2020b).

We can predict from the estimated per-generation substitution rate that identical sequences are likely to have more than 5 intermediate generations between them; the probability of fewer than five generations occurring before a mutation occurs is 0.49, as determined by repeated sampling of a Poisson distribution with a lambda of 0.17. While the low substitution rate means that comparing the number of SNPs between sequences alone may not be an effective method of determining infector-infectee relationships, it can be used in conjunction with temporal and location data to make more accurate predictions of transmission events by ruling out relationships between more distantly related transmission chains co-circulating in the same area, as in Cori et al. (2018). Notably, our Poisson distribution of the number of substitutions occurring in one generation is visually very similar to the genetic signature distribution reported in Cori et al.'s Figure S1, despite different methods and RABV datasets being used in their calculations. It is likely, however, that our estimate of the per-generation substitution rate is lower than the mean number of SNPs expected between sequences from a primary and secondary case, due to the time-based substitution rate being affected by purifying selection (Duchêne et al. 2014). Further analysis comparing the estimated per-generation substitution rate to realised SNP distances between primary-secondary case pairs could quantify this difference.

While the Jukes-Cantor model was the most appropriate to use on our synthetic data due to the simplicity of the mutation models, phylogenetic analyses on real RABV genome data usually use a more complex model, such as the GTR + G substitution model used to generate the Tanzanian tree shown in this study (Lushasi et al. 2023). This, along with the simplicity of our mutation model as well as sampling biases in the real dataset, may affect how comparable synthetic root-to-tip divergence plots are to the real data.

While the molecular clock has proven critical for gaining insights into the history and dynamics of disease outbreaks, the epidemiological characteristics of a virus should be considered when choosing how to measure viral evolution. In this study, we determine that the per-generation model is not likely to produce substantially different results from the strict molecular clock model when analysing contemporary RABV evolution. We also estimate the mean per-generation substitution rate of RABV for future use in transmission tree reconstruction and efforts to estimate outbreak sizes and lineage emergence rates.

Chapter 3

A simple method of estimating rabies outbreak sizes from phylogenies

3.1 Abstract

The ability to estimate the size of a disease outbreak is beneficial both for coordinating an outbreak response and for retrospectively evaluating the effectiveness of response measures and surveillance. Seroprevalence surveys and epidemiological modelling can be used for this purpose, but can be computationally expensive or unsuitable for certain diseases, and current phylogenetic methods do not explicitly estimate outbreak size. We develop a simple formula to estimate the proportion of cases in an outbreak that have been sequenced using only an estimate of the per-generation mutation rate and the mean branch length of a phylogenetic tree, which can then be used to calculate the outbreak size. We validate this method by applying it to simulated outbreaks, and then to an ongoing rabies outbreak in the Romblon province of the Philippines. We find that in 95% of simulated outbreaks the outbreak size is estimated to within 28.8% of its true value, with a mean deviation of 16.3% from the true value, and that this method can be applied to outbreaks with a range of substitution rates and values of R_0 appropriate for rabies.

We estimated that approximately 14.7% of cases from the Romblon province had been sequenced, suggesting better than average case detection and case numbers consistent with previous analyses. This method should make rabies outbreak size estimation more accessible, but further work is required to improve its applicability to other diseases.

3.2 Introduction

In the early stages of responding to a disease outbreak it is often difficult to grasp the scale of the problem, but having an accurate estimate of the number of infections is beneficial in helping to plan resource distribution, contact tracing, and vaccination efforts. Current methods for estimating outbreak sizes or the prevalence of a disease in a population often involve epidemiological modelling (Choi and Ki 2020; Fraser et al. 2009; Shutt et al. 2017; Zhang et al. 2020), randomised testing (Pouwels et al. 2021), or conducting serological surveys (Gérardin et al. 2008), but these are not always appropriate for every outbreak.

Generating genomic data has increasingly become an important part of the outbreak response process, with a recent example being the massive testing and sequencing efforts of suspected COVID-19 cases across the world during the pandemic (Ko et al. 2022; Meredith et al. 2020). As the capacity for viral genome sequencing increases, the resulting data will have the potential to become another useful tool for estimating outbreak sizes. Existing methods leveraging genomic data include Bayesian skygrid (Gill et al. 2013; Hill and Baele 2019) and skyline models (Drummond et al. 2005; Stadler et al. 2013), and GlnPipe (Smith et al. 2021). In actuality, however, these methods estimate the effective population size (N_e), which is an abstract value usually much lower than the real outbreak size and scaled by an unknown quantity (Waples 2022), and as such are mainly used to investigate relative growth rates over time (Faria et al. 2014). To our knowledge, no methods have been published that use genetic information to explicitly estimate outbreak size at the time of writing.

Rabies is an example of a disease where obtaining estimates of outbreak sizes would be useful, but so far largely unsuccessful. As dogs in the regions where rabies is most prevalent are often free-roaming, many rabies infections go unobserved unless they result in a human exposure; human cases of rabies are also largely under-reported (Nel 2013). Because of this, it is difficult to estimate how many rabies cases in total truly occur in any given outbreak, although the prevalence of rabies is assumed to not exceed 1% of the dog population (Mancy et al. 2022). As rabies is fatal once the victim becomes symptomatic and infectious, serological surveys cannot be used to estimate historical rabies prevalence. Testing for rabies in non-human animals involves acquiring a brain sample which requires the animal to be deceased, which is not practical for randomised testing. Furthermore, epidemiological modelling methods often fail to accurately replicate rabies' ability to persist at low prevalences, and require computational resources and specialised skill sets that may make them inaccessible to local practitioners.

Genomic surveillance has become a common part of rabies outbreak responses in recent years (Gibson et al. 2022; Lushasi et al. 2023; Zinsstag et al. 2017), with the development of portable sequencing technologies and more cost-effective protocols leading to an increase in capacity for in-country whole genome sequencing (Bautista et al. 2023; Brunker et al. 2020). While genomic data is becoming increasingly available, the rabies virus evolves relatively slowly, leading to some datasets having insufficient temporal signal to carry out advanced phylogenetic analyses such as skyline and skygrid methods on emerging outbreaks despite samples being collected over multiple years, as discussed in chapter 2. On top of this, these methods require specialist knowledge that it is unreasonable to assume local responders would possess, and as previously mentioned any estimates of N_e are likely to be much lower than the true outbreak size. Developing a simple method of outbreak size estimation which utilises genomic data could aid in making outbreak estimation more accessible to responders in the worst affected areas.

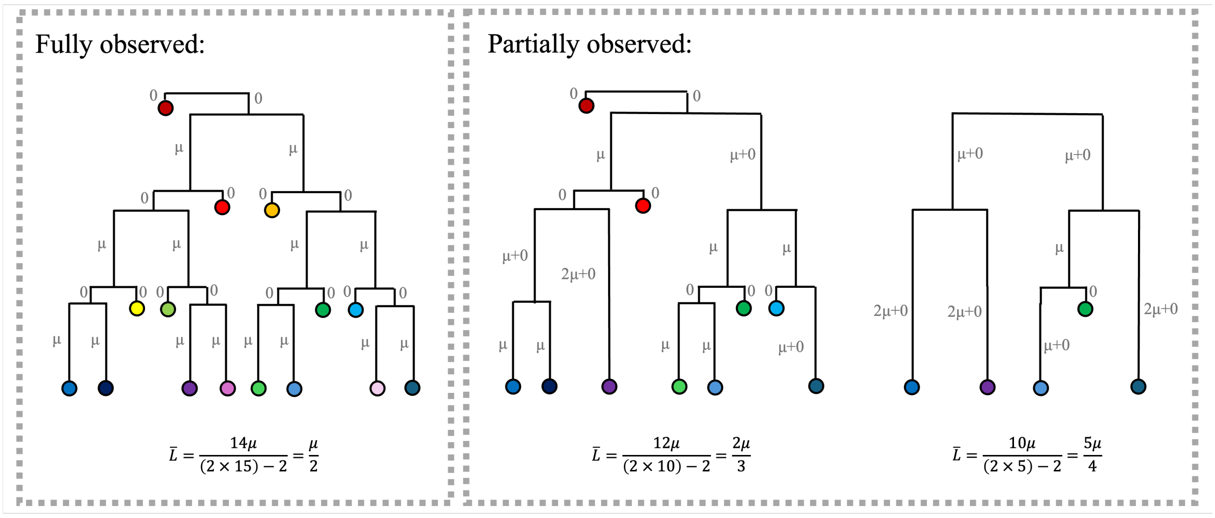
Fully observed:
Partially observed:
$\bar{L} = \frac{14\mu}{(2 \times 15) - 2} = \frac{\mu}{2}$
$\bar{L} = \frac{12\mu}{(2 \times 10) - 2} = \frac{2\mu}{3}$
$\bar{L} = \frac{10\mu}{(2 \times 5) - 2} = \frac{5\mu}{4}$
μ
0
μ+0
2μ+0

Figure 3.1: Diagram showing a phylogenetic tree of a hypothetical outbreak where every new case gains μ mutations, and trees with 50% and 75% of tips removed.

A feature of phylogenetic trees which may aid in this goal is that two branches are removed when a tip is removed from a perfectly observed tree (a tree where every case in the outbreak of interest has been sequenced), but the tree length (the sum of all branch lengths) will only be reduced by the length of one of these branches (and in some cases, this reduction will be 0). This means that as tips are removed from the fully observed tree, the mean branch length (tree length divided by the number of branches) will almost always increase. This is shown in the following phylogenetic tree of a hypothetical outbreak (Figure 3.1), where each new case gains μ mutations, and $\bar{L}$ is the mean branch length, which increases as 50% and 75% of cases are removed from the tree in this example. If there is a consistent relationship between the proportion of cases sequenced and the mean branch length across outbreaks and sequence datasets of different sizes, it could be exploited to estimate the proportion of cases in an outbreak that have been sequenced, and thus the outbreak size.

In this study we develop a simple method of estimating rabies outbreak sizes using only a phylogenetic tree of viral sequences acquired from the outbreak and an estimate of the per-generation substitution rate, assuming that the sequences have been collected in an unbiased manner across the course of the outbreak. We test this method within our simulation frameworks, and then apply it to data from a recent outbreak of rabies in the Romblon province of the Philippines.

3.3 Methods

3.3.1 Simulation framework

Rabies outbreaks were simulated in R (R Core Team 2020) using a branching process model (as in chapter 2; Durrant et al. 2024). Simulations were allowed to run for seven years (or sooner if rabies went extinct before this time) after which time they were stopped, allowing us to treat the simulated data as an ongoing outbreak. Synthetic genetic data were generated based on the resulting synthetic outbreaks, using a mutation model that has variable substitution rates per site. Synthetic outbreaks were generated with two values of R_0 (1.05 and 1.7) and genetic data was generated with three substitution rates (2×10^{-3}, 2×10^{-4}, or 2×10^{-6} substitutions per site per year). 273 outbreaks originating from a single introduction were generated for each value of R_0, but only outbreaks containing over 500 cases were kept for analyses to ensure that phylogenetic trees could be generated for datasets with the sparsest sampling strategy used. This resulted in 6 outbreaks with an R_0 of 1.05 and 40 outbreaks with an R_0 of 1.7, ranging in size from 1,048 to 15,287 cases.

The per-generation substitution rate (substitutions per site per generation) was calculated by multiplying the generation interval length (converted from days, as typically reported, into years) by the unit-time based substitution rate (substitutions per site per year). In order to produce synthetic sequences which more closely replicate real rabies sequences, the site of each mutation in the genome was selected based on site-specific substitution rates, and the substitution itself based on transversion rates, both estimated from rabies Cosmopolitan AF1b clade genetic data. This clade was chosen as it is one of the most well sequenced dog-associated rabies clades. A dataset of 127 rabies whole genome sequences from the Cosmopolitan AF1b clade was acquired from RABV-GLUE (Campbell et al. 2022), and site-specific substitution rates and transversion rates for these sequences were estimated in IQtree (Nguyen et al. 2015). Simulated mutations were neutral and had no impact on transmission or sequencing probability. Subsets of the resulting synthetic sequences were randomly sampled to emulate between 1% and 90% of cases being sequenced. Maximum likelihood trees were generated using IQtree, and the mean branch length was calculated in R from trees imported using the `ape` package (Paradis and Schliep 2019).

Two further synthetic outbreaks with "extreme" topologies were also generated: a "star" shaped outbreak, where the initial case has a single generation of 999 secondary cases; and a "line" outbreak, where each case has exactly one secondary case for 1000 generations. Both of these outbreaks have a substitution rate of 2×10^{-4} substitutions per site per year, with substitutions assigned with no site or transversion bias across a 11923 base genome. In order to investigate the effect of higher R_0s than are realistic for rabies on our method, synthetic outbreaks were also generated with R_0 values of 1.05, 1.5, 2, 3, 4, 6 and 8 over a sequencing probability of 10 to 90%. These outbreaks are generated using a very simple branching process model, with the number of secondary cases for each case being drawn from a Poisson distribution and the lambda parameter being the given value of R_0. Each case gained exactly 1 SNP at a random position on a 10,000 base genome. The branching process was run until the outbreak size after the latest full generation of cases had been assigned exceeded 2000 cases.

3.3.2 Accuracy testing

We tested the accuracy of the method developed in this study using synthetic outbreak and genetic data, where sequencing was conducted randomly across the entire course of the outbreak. Accuracy testing was undertaken on synthetic outbreaks with R_0 values of 1.05 or 1.7, which are both within the plausible range of R_0s estimated for rabies outbreaks (Li et al. 2024). We also tested how the method performed across a range of substitution rates, of either 2×10^{-3}, 2×10^{-4} or 2×10^{-6} substitutions per site per year, to represent plausible substitution rates across different viruses, of which 2×10^{-4} substitutions per site per year represents the substitution rate of rabies. We calculate the ratio of the estimated proportion of cases sequenced to the actual proportion of cases sequenced, where a value of 1 represents perfect estimation.

3.3.3 Application to the 2022 Romblon rabies outbreak

In September 2022 the first new rabies case since 2012 was detected on Tablas Island in the Romblon province of the Philippines, which had previously been declared rabies free. The outbreak response has included enhanced surveillance through integrated bite case management (Swedberg et al. 2023), resulting in 45 confirmed cases between September 2022 and September 2023; two of these cases were human deaths, with the remaining 43 being animal cases. From these confirmed cases 24 whole genome sequenced were produced using an Oxford Nanopore MinION-based protocol (Bautista et al. 2023). Further details of the outbreak and the ensuing response, including details of how the RABV sequences were generated, are available from Yuson et al. (2024).

One thousand maximum likelihood trees were generated using all available sequences from Romblon using IQTree, and for each a subtree consisting of 14 sequences identified as potentially resulting from a single introduction to the region ("cluster 1") was extracted. The remaining sequences were identified as likely resulting from separate introductions, and were too few in number to make accurate estimates from (1, 1, 3 and 5 sequences per cluster). We took the mean branch length from across these 1000 subtrees and calculated the per-generation substitution rate by multiplying the mean generation interval estimated from the reconstructed transmission tree (ibid.) by a Philippines-wide substitution rate estimate. This was generated from 299 whole genome sequences using a BEAST (Drummond et al. 2012) coalescent Bayesian skyline plot with a GTR F+I+G4 uncorrelated relaxed substitution model.

To construct a 95% credible interval we sampled 10,000 values from the posterior distribution for substitution rate (acquired from the BEAST log file), the posterior for the generation interval (by fitting, and then sampling from, a lognormal distribution based on the transmission tree estimates) and the mean branch lengths from the 1000 subtrees and calculated the proportion of cases sequenced for each combination of sampled values. We then extracted the 2.5% and 97.5% percentiles from the resulting 10,000 values.

3.4 Results

3.4.1 Curve fit

In synthetic outbreaks, the mean branch lengths of trees that contain sequences subsampled randomly across their entire duration were observed to increase as the proportion of cases sequenced decreased (Figure 3.2). This trend was observed to follow a curve of approximately the equation:

$$\bar{L} = \frac{\mu}{2\sqrt{\rho}} \tag{3.1}$$

where ρ is the proportion of cases sequenced, $\bar{L}$ is the mean branch length of the phylogenetic tree, and μ is the per-generation substitution rate. This relationship is very strong in synthetic outbreaks with substitution rates of both 2×10^{-3} and 2×10^{-4} substitutions per site per year; the association is less clear at 2×10^{-6} substitutions per site per year, however. This is likely due to stochasticity in the simulation process producing outbreaks with a realised substitution rate of less than 2×10^{-6} substitutions per site per year, or insufficient accumulation of genetic diversity over the simulated time frame resulting in poor trees.

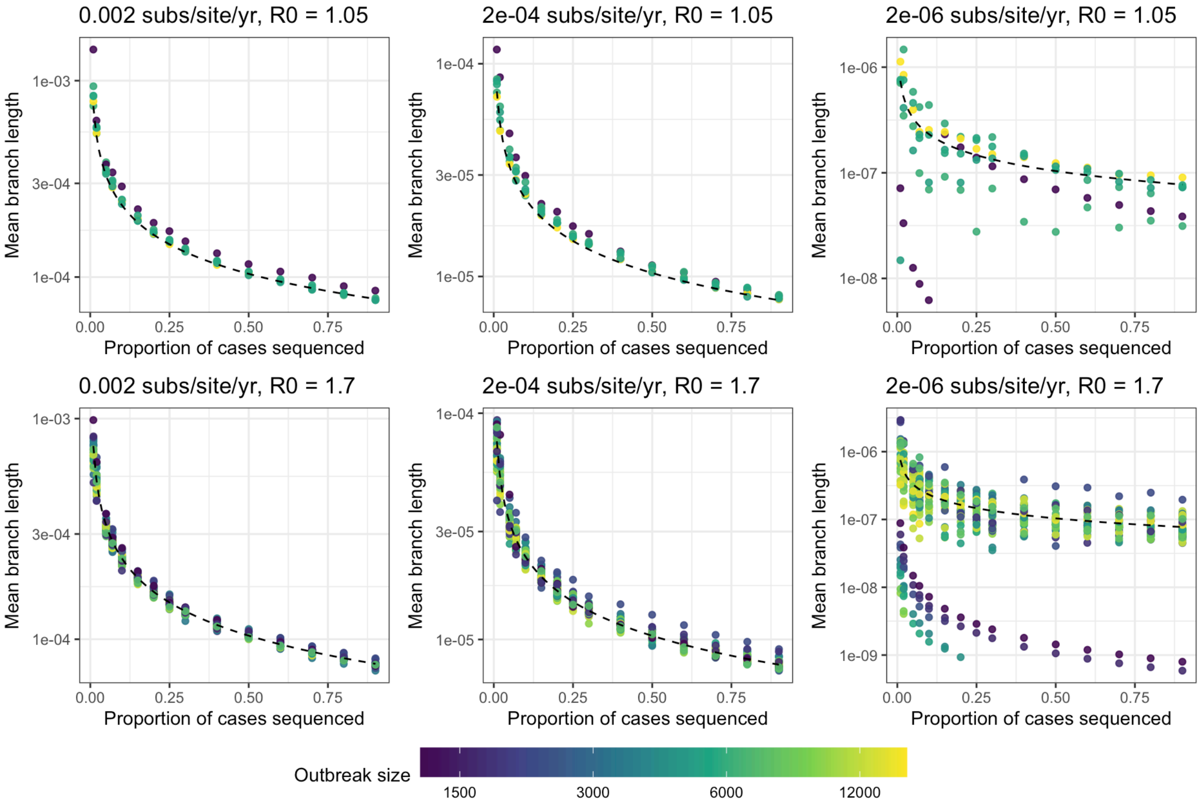

Figure 3.2: The mean branch length of a maximum likelihood tree generated from synthetic data plotted against the proportion of cases in the outbreak that were sequenced. The curve is of Equation 3.1. Point colour represents the total size of the outbreak; facet titles describe the parameters of the simulation. The Y axis and colour are $\log_{10}$ scaled.

3.4.2 Extreme outbreak shapes

The relationship between the mean branch length and the proportion of cases sequenced also ceases to fit well to Equation 3.1 when the phylogenetic tree has an “extreme” shape, such as those generated from the “line” outbreak where every case has exactly 1 secondary case for 1000 generations, and from the “star” outbreak where the one initial case has a single generation of 999 secondary cases (Figure 3.3A). This may be due to how well the tree length over different proportions of cases sequenced approximates a square-root relationship; while the synthetic outbreaks with rabies-appropriate R_0 values approximate this well, the "star" outbreak tree length has a linear relationship with the proportion of cases sequenced, and the tree length of the "line" outbreak is seemingly unaffected by the proportion of cases sequenced (Figure 3.3B).

This suggests that there is a suitable range of R_0 values where Equation 3.1 will apply sufficiently well to an outbreak; further simple simulations suggest that outbreaks with an R_0 between 1.05 and 2 approximate the square-root relationship well, whereas outbreaks with an R_0 above this begin to more closely resemble a linear relationship (Figure 3.4).

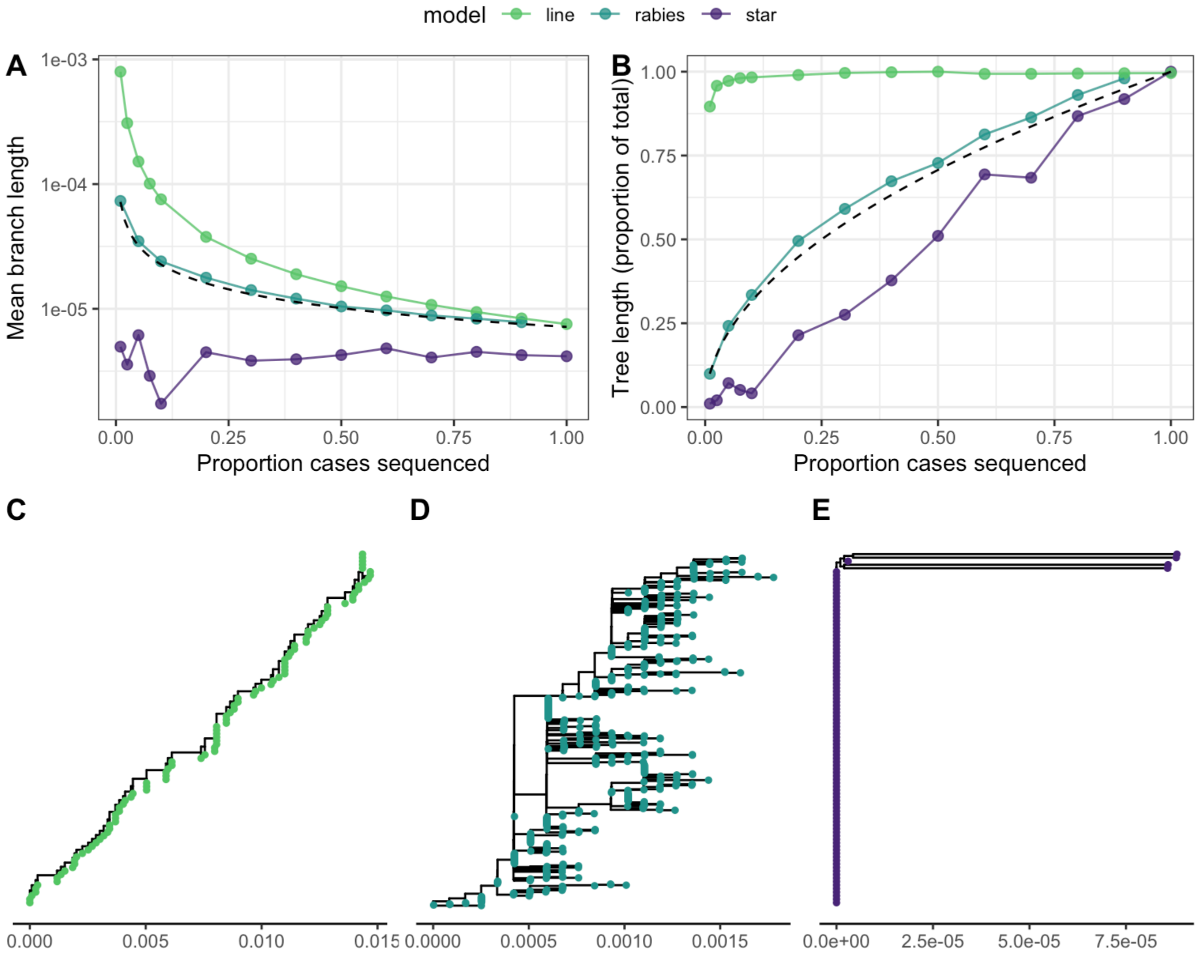

Figure 3.3: Comparison of the relationship between the mean branch length and the proportion of cases sequenced for different tree topologies. (A) Points represent the mean branch length of a phylogenetic tree generated with a certain proportion of cases sequenced. The dashed line is of Equation 3.1. (B). Points represent the tree length of a phylogenetic tree. The dashed line represents the square root of the proportion of cases sequenced; the tree length has been normalised to one in order to better compare the three outbreak simulations. (C-E) The phylogenetic trees shown are generated with 10% of cases sequenced from the "line" outbreak (C), a rabies outbreak with an R_0 of 1.05, (D) and the "star" outbreak (E). The x-axes are scaled in substitutions per site.

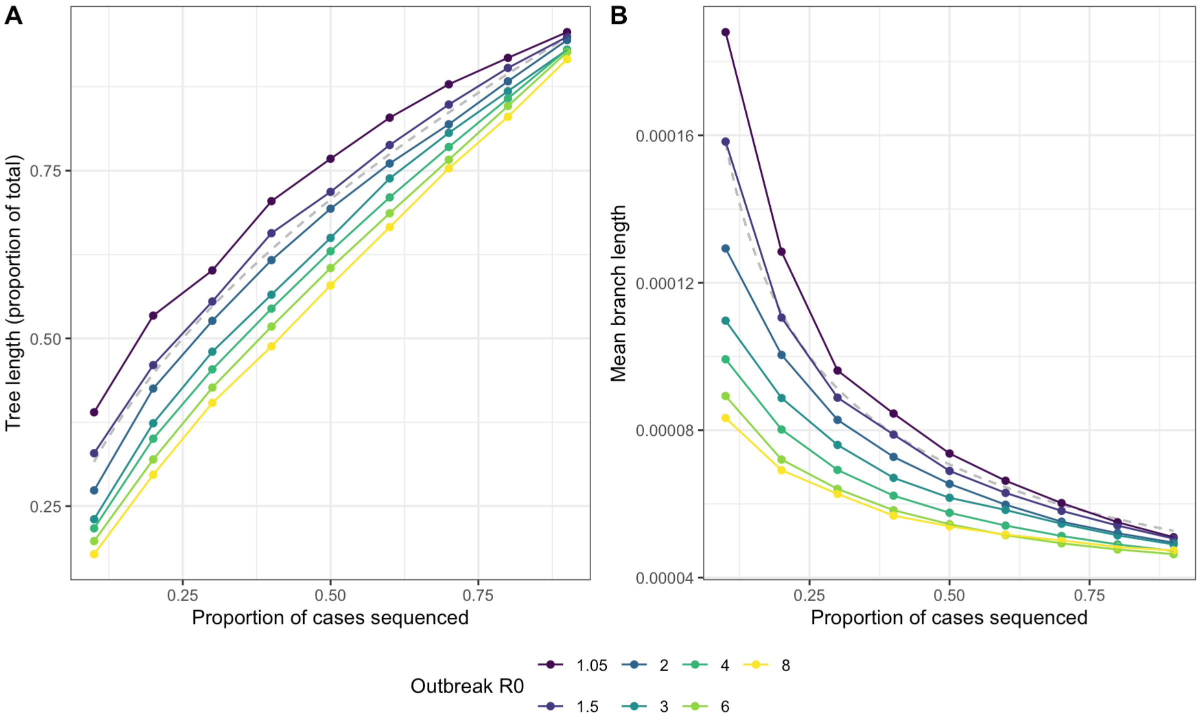

Figure 3.4: Outbreaks with R_0s of 1.5 and 2 best fit the square-root relationship with tree length. (A) The relationship between the proportion of cases sequenced and observed tree length in outbreaks of varying R_0 compared to a square root relationship (dashed grey line). (B) The relationship between the proportion of cases sequenced and the observed mean branch length compared to Equation 3.1 (dashed grey line).

3.4.3 Deriving Equation 3.1

As established in the last section, the tree length when the proportion of cases sequenced is ρ in an outbreak with an R_0 typical of rabies can be approximated as:

$$T_{\rho} = T_1 \times \sqrt{\rho} \tag{3.2}$$

Where T_{ρ} is the tree length when the proportion of cases sequenced is ρ, and T_1 is the tree length of the fully observed tree. This tree length can also be expressed as:

$$T_1 = \mu(N-1) \tag{3.3}$$

Where μ is the per-generation substitution rate and N is the total outbreak size ($N-1$ being the number of transmission events within the outbreak). The mean branch length at ρ (the tree length at ρ divided by the number of branches) can thus be expressed as:

$$\bar{L} = \frac{\mu(N-1)\sqrt{\rho}}{2\rho N - 2} \tag{3.4}$$

as the number of branches in a bifurcating phylogenetic tree is 2 times that number of tips minus 2. If we assume that N and ρN are sufficiently large that $N \approx N-1$ and $\rho N \approx \rho N - 1$, so that $\frac{N-1}{\rho N-1} \approx \frac{1}{\rho}$, we can then simplify Equation 3.4 to Equation 3.1.

As ρN is the number of sequences in our phylogenetic tree, this presents us with the possibility of calculating a minimum number of sequences required to sufficiently satisfy this assumption. As we simplify $\frac{N-1}{\rho N-1}$ to $\frac{1}{\rho}$, the ratio between these two values should be as close to 1 as possible to minimise the error introduced by this assumption. By calculating the $\frac{1}{\rho}$ to $\frac{N-1}{\rho N-1}$ ratio for differing values of ρ and N, we find that at approximately 20 sequences are required to keep the ratio above 0.95, or approximately 10 sequences for a ratio above 0.9 (Figure 3.5).

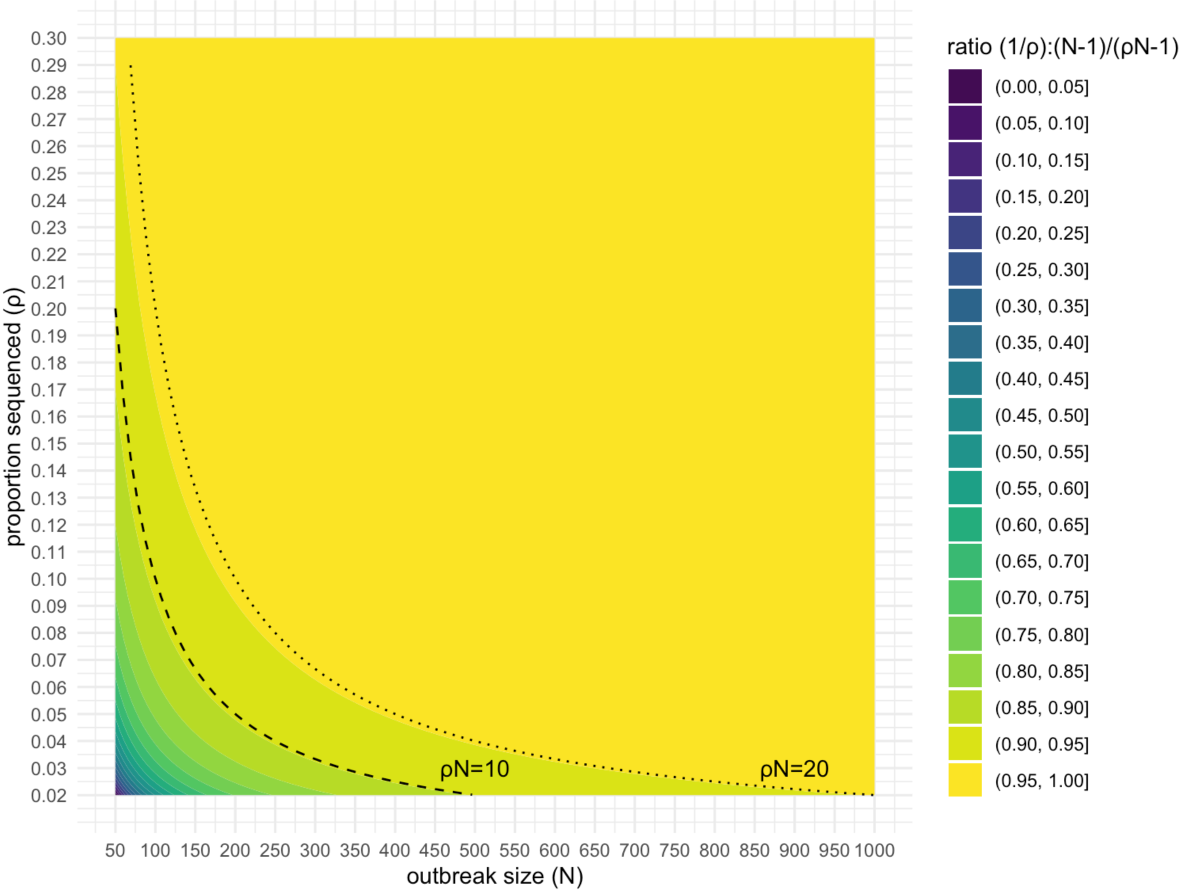

Figure 3.5: a dataset of more than 20 sequences is required to keep the introduced error from the assumption that $\frac{N-1}{\rho N-1} \approx \frac{1}{\rho}$ below 5%. Contour colours represent the $\frac{1}{\rho}$ to $\frac{N-1}{\rho N-1}$ ratio. The dotted curve represents 20 sequences; the dashed curve represents 10 sequences.

3.4.4 Accuracy testing

Equation 3.1 can be rearranged as:

$$\rho = (\frac{\mu}{2\bar{L}})^2 \tag{3.5}$$

to give an estimate of the proportion of cases sequenced. The estimated total number of cases in the outbreak can then be calculated as the total number of sequences divided by the estimated proportion of cases sequenced. Applying Equation 3.5 to synthetic outbreaks results in estimated outbreak sizes within 28.8% of the true proportion sequenced in 95% of outbreaks (Figure 3.6), and a mean deviation of 16.3% from the true value. Accuracy decreases as the estimated proportion of cases sequenced falls below 0.1. The method tended to under-estimate the proportion of cases sequenced in simulated outbreaks with an R_0 of 1.05, whereas estimates appeared equally above and below the true value at an R_0 of 1.7. Based on Figure 3.4, we would expect this method to over-estimate the proportion of cases sequenced more severely as the R_0 of the outbreak increases above 2.

3.4.5 Application to the 2022 Romblon rabies outbreak

We estimated that 14.9% (95%CI: 11.6%, 19.7%) of cases from the Romblon outbreak were sequenced, suggesting that during this outbreak period 163 cases (95%CI: 122, 206) occurred in total, 95 (95%CI: 71, 120) of which belonged to cluster 1 (Figure 3.7). This is consistent with the number of undetected cases estimated previously (Yuson et al. 2024). Specifically, the maximum-likelihood trees constructed using the 14 cluster 1 sequences had a mean branch length of 2.90×10^{-5} substitutions per site. Using the Philippines-wide substitution rate estimate of 3.07×10^{-4} substitutions per site per year and an estimated mean generation interval of 26.6 days, we calculated a per generation substitution rate of 2.24×10^{-5} substitutions per site per generation. Dividing the per-generation substitution rate by twice the mean branch length and squaring the result gives an estimated proportion of cases sequenced of 0.147, and dividing the number of sequences (14 from cluster

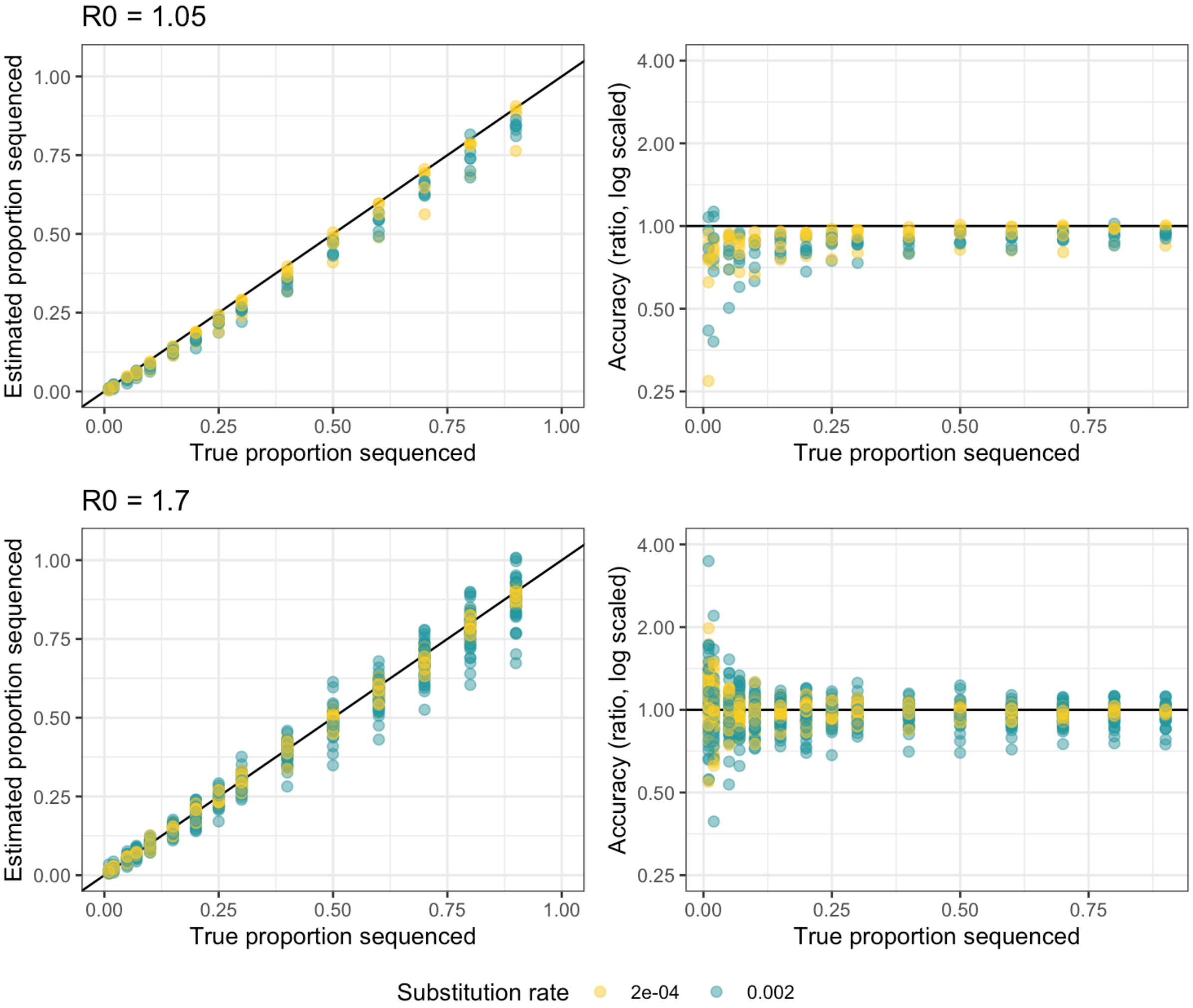

Figure 3.6: The accuracy of our method when applied to simulated rabies outbreaks with two values of R_0 and substitution rate.

1) by this value results in 95 total cases being estimated in cluster 1. Assuming that there was no bias towards a certain cluster being more likely to be sequenced than any other, adding the 10 sequences from the other clusters to the total number of sequences (24 in total) gives an estimate of 163 cases across the introductions to Romblon.

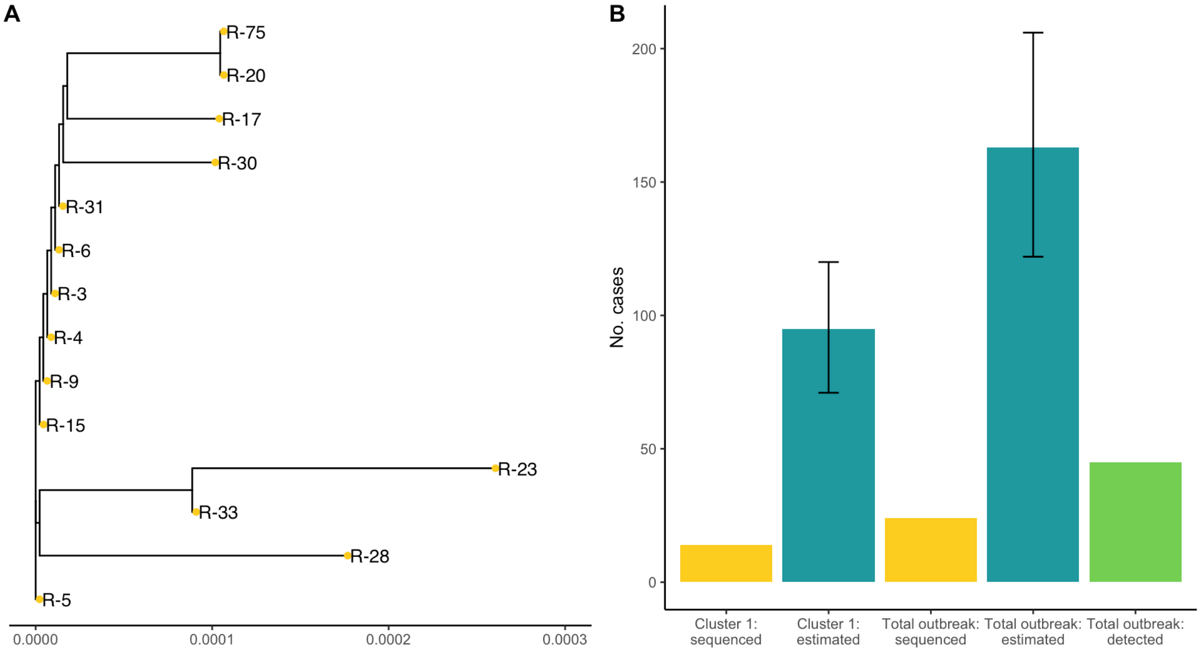

Figure 3.7: Estimated case numbers in Romblon cluster 1 and across the entire outbreak suggest extensive undetected transmission of rabies. A: Example of a cluster 1 subtree used to estimate the outbreak size. The x-axis is scaled in substitutions per site. B: Comparison of the numbers of sequenced, detected and estimated cases within cluster 1 and across the entire outbreak. Error bars represent the 95% credible interval.

3.5 Discussion

We developed a simple method of estimating outbreak sizes which is simple, computationally inexpensive, reasonably accurate, and only requires information routinely gathered during an outbreak response that incorporates genomic surveillance. This presents an improvement in usability over previous methods. Using this method, we estimated that cluster 1 of the ongoing Romblon outbreak contains 95 cases, with 163 being estimated for the total outbreak. When compared with the 45 confirmed cases, this suggests that there has been extensive undetected transmission in Romblon, which reflects the findings of the main outbreak report (Yuson et al. 2024).

This method relies on the observed tree length having an approximately square-root relationship with the proportion of cases sequenced. Interestingly, such a relationship has been previously noted in a study downsampling human genome data (Karczewski et al. 2020), but other than this little attention has been paid to this relationship. In simple branching process simulations (i.e., the same as described above but without a spatial or temporal element, and where every case is assumed to generate one SNP), the observed tree length points pass from above to below the square root curve as R_0 increases from below to above approximately 1.5. This suggests that it is unlikely that this is truly a square-root relationship, but an exponential relationship that takes R_0 into account to some degree, and in its current form this method will work best for outbreaks of diseases with an R_0 between 1 and 2. Further work is required to determine what the true relationship between tree length, R_0 and proportion of cases sequenced is to improve estimation accuracy and widen this method's applicability to diseases with higher values of R_0.

As R_0 is kept stable during these simulations, future work should also consider the effect of R_t changing through time, for example as a result of increased vaccination, on the accuracy of this method, and whether it can be accounted for. The effect of overdispersion of the secondary case distribution while maintaining the same R_0 should also be investigated. As in the simplified system shown in Figure 3.1, tree length only decreases when a tip with no descendants is removed, and as such the relationship between tree length and the proportion of cases sequenced may rely on the proportion of cases in the tree with no descendants. In this case, the overdispersion of secondary cases is likely to have some effect on how well the square-root assumption of the relationship between tree length and proportion of cases sequenced applies to outbreaks a suitable R_0 range for this assumption. This method also relies on the assumption that sequencing effort is even over time and unbiased; in reality, this is unlikely to occur for rabies outbreaks in settings such as Romblon, as sampling is usually only carried out when a human bite case is reported; in contrast, in the Serengeti district the opportunistic sampling of animals found dead may represent a less biased approach. Increased sequencing effort in later stages of the outbreak, as is more likely to occur during rabies outbreak responses, may result in an overestimation in the proportion of cases sequenced, as some lineages may go extinct before they can be sampled from.

An important aspect to consider when applying this method is the possible error introduced through estimating the per-generation substitution rate, especially when using a wider dataset's substitution rate to calculate this value, as it is necessary to do with new rabies outbreaks where there is insufficient temporal signal to conduct Bayesian phylogenetic analyses (Durrant et al. 2024). As purifying selection has a strong effect on RNA virus genomes (Hughes and Hughes 2007), and the substitution rate estimate decreases over a wider temporal sampling window because of this (Duchêne et al. 2014), the estimated per-generation substitution rate using the wider dataset may be lower than the per-generation substitution rate of the dataset being investigated. Further analysis is required to determine whether this has a significant impact on outbreak size estimates using this method, and how more accurate estimates can be acquired.

The error introduced from the assumption made that N and ρN are sufficiently high during the derivation is also important to consider. We would recommend using a minimum of 10 sequences to conduct this analysis to conserve funds and resources, but using 20 sequences or more would be optimal. The effect of this assumption is likely why we see a decline in accuracy when analysing outbreaks with a proportion of cases sequenced below 0.1. We also observe that in synthetic outbreaks with an R_0 of 1.7, some outbreaks with a substitution rate of 2×10^{-3} substitutions per site per year appear to have a lower accuracy than those with a substitution rate of 2×10^{-4} substitutions per site per year. This may be due to the combination of a high substitution rate, a substitution model that incorporates site-specific rates and large outbreak sizes causing site saturation to be achieved over the relatively long time scale of the simulation.

We describe a novel, simple method of estimating rabies outbreak sizes from phylogenetic trees, which we hope can assist in expediting the roll-out of rabies control measures at an appropriate scale in the case of an outbreak, and be used more generally in assessing surveillance performance. This will be essential in monitoring our progress towards the global goal of achieving zero human rabies deaths by 2030.

Chapter 4

Differences in codon usage and CpG content between host-species-specific rabies clades

4.1 Abstract

Viruses favour the usage of certain codons due to their nucleotide content, translational efficiency, and pressure from the host immune system. Rabies is a negative strand RNA virus which can infect a broad range of mammalian hosts, with some clades circulating predominantly in specific host species. While previous studies have investigated codon usage in the rabies virus, sequences have only been split into bat- or carnivore-related groups, or into the major clades. Here we investigate whether codon usage varies between minor rabies clades that circulate in specific host species and what drives these differences. We acquired a dataset of publicly available nucleoprotein gene sequences and investigated how biased codon usage was between clades, how these biases differ, and calculated the CpG content of each clade. We found that codon usage varies most between bat- and carnivore-associated RABV groups, and varies more subtly between host-species-specific minor clades within these groups. Codons containing CpG dinucleotides were found to be among the most influential in a principal component analysis, and we found that

CpG content was considerably higher in carnivore-associated rabies clades than in bat-associated clades. We also found that bat-associated rabies clades appear to be under higher selection pressure from the host's zinc-finger antiviral protein than other clades, warranting further investigation of the mechanism underpinning this change.

4.2 Introduction

Despite encoding the same amino acids, synonymous codons are often not used equally. This phenomenon, called codon usage bias, is driven mainly by two factors (Hershberg and Petrov 2008): the first being mutation pressure, where codon usage is biased due to the non-randomness of mutation patterns, such as biases in guanine and cytosine content (GC content) in the genetic code. The second is natural selection, where the codon used affects the fitness of the organism; for example, using codons corresponding to more abundant tRNAs (either belonging to the organism itself, or to the host in the case of viruses) to increase translational efficiency. This is more common in eukaryotes and prokaryotes, but codon bias driven by natural selection has also been observed in some viruses, such as henipaviruses (Kumar et al. 2018). In general, however, codon bias appears to be shaped more by mutational pressure in both RNA and DNA viruses (Jenkins and Holmes 2003; Shackelton et al. 2006). Positive-strand RNA viruses with a narrow host range have stronger codon usage biases than those with a broad host range, and have a more similar codon usage to their host species (Tian et al. 2018), likely to maximise translational efficiency, whereas negative-strand RNA viruses do not appear to mimic their host species' codon usage (Rima 2015), despite both being reliant on the host's translation machinery. This may be due to the transcription of the negative-strand virus' RNA by its own polymerase being dependent on the RNA's stability, which can be influenced by codon usage biases (Gumpper et al. 2019). Despite this, using machine learning methods, codon usage together with nucleotide and dinucleotide content have previously been found to accurately predict the hosts of coronaviruses (Brierley and Fowler 2021) and other RNA viruses (both positive- and negative-strand; Babayan et al. 2018).

The rabies virus (RABV) is a negative strand RNA virus that can infect a broad range of mammal species. The RABV genome consists of five genes, which encode the nucleoprotein (N), phosphoprotein (P), matrix protein (M), glycoprotein (G) and RNA polymerase (L; "large" protein). RABV can be broadly split into bat- and carnivore-associated clades, with the main bat to carnivore host shift being predicted to have occurred approximately 600 years ago (Troupin et al. 2016). Bat to carnivore shifts have also occurred in more recent history, with one example being the shift from bats to skunks and raccoons which is estimated to have occurred 250 years ago (Ding et al. 2017), and spillover events continue to occur in the present day (Kuzmin et al. 2012). Rabies clades are known to sometimes circulate among specific host species; for example, sustained transmission cycles have been observed in Chinese ferret badger populations (Zhang et al. 2013) associated with the minor clade Asian SEA2b, as opposed to the clade circulating sympatrically in dogs, Asian SEA2a. There has been debate as to whether these host-species-specific transmission cycles are due to adaptation or ecological circumstances (Marston et al. 2018; Mollentze et al. 2014), with evidence of positive selection over host shifts being mixed (Kuzmin et al. 2012; Troupin et al. 2016; Streicker et al. 2012b). These studies have focused on non-synonymous substitutions, and not the possible effect of synonymous substitutions (through changes in codon usage bias or dinucleotide bias) on host adaptation.

Previous codon usage studies have been conducted with the rabies virus as a focus. RABV has the lowest codon usage bias among lyssaviruses (Zhang et al. 2018), perhaps due to its broad host range compared to other members of the family (Marston et al. 2018). While natural selection is thought to contribute the most to codon bias on the RABV N gene (He et al. 2017) and over the whole genome (Li et al. 2023), mutation pressure dominates on the G gene (Morla et al. 2016). Viral replication is inhibited when the RABV matrix protein is reconstructed with suboptimal codon usage (Luo et al. 2020a), indicating that codon biases are an important factor in rabies evolution. These previous studies either subdivide the sequences into dog-, bat-, or human-derived sequences based on the species of the host animal the sample was acquired from, or into the major rabies clades (Bat, Africa-2, Arctic, Cosmopolitan, Asian and Indian subcontinent major clades in He et al. (2017); these plus Africa-3 and South-East Asian in Li et al. (2023)). Differences in codon usage bias between host-specific minor clades have not been investigated.

In this study we aimed to quantify differences in codon usage biases between host-species-specific RABV minor clades, and investigate what factors underpin these differences. We hypothesised that clades that diverged more deeply in the virus' evolutionary history (i.e., from bats to carnivores) would show clear codon usage biases, while more recent host shifts (i.e., Chinese ferret badgers) would show less obvious differences to their most closely related clades due to genetic drift rather than to adaptation to their host, as host mimicry does not appear to drive codon usage biases in negative-strand RNA viruses.

4.3 Methods

4.3.1 Data acquisition

We downloaded complete RABV N gene sequences from the RABV-GLUE sequence database (Campbell et al. 2022). We searched for sequences with 100% N gene coverage from host-specific minor clades, including those mentioned in Troupin et al. (2016), with the addition of bat-associated clades. The full list of clades and the search terms used to filter sequences on RABV-GLUE are shown in Table 4.1.

RABV-GLUE Clade	Host	RABV-GLUE host search term

Asian SEA2a	Domestic dog	“Canis familiaris”, “Canis lupus familiaris”
Asian SEA2b	Chinese ferret badger	“Melogale moschata”, “Chinese ferret badger”
Arctic A	Arctic fox	“Vulpes lagopus”
Cosmopolitan AM2a	Mongoose (Caribbean)	“Mongoose”, “Herpestes”
Cosmopolitan AF1b	Domestic dog	“Canis familiaris”, “Canis lupus familiaris”
RAC-SK	Raccoon	“Procyon lotor”
RAC-SK	Skunk	“Mephitis”, “skunk”
Bat DR	Vampire bat	“Desmodus rotundus”
Bat TB1	Mexican free-tailed bat	“Tadarida brasiliensis”
Bat EF-E2	Big brown bat	“Eptesicus fuscus”
Bat LC	Hoary bat	“Lasiurus cinereus”

Table 4.1: RABV-GLUE clades and search terms used when downloading RABV sequences.

Raccoon sequences and raccoon-related skunk sequences were removed from the study after initial phylogenetic analysis suggested substantial transmission between raccoons and skunks, with the exception of sequences noted as belonging to the southern central skunk variant (SCSK) within the RAC-SK major clade, which did not contain any raccoon-derived sequences and were therefore retained. Five additional sequences from the Arctic A minor clade were added to the dataset which had been collected from *Vulpes lagopus* according to the original publications, but were mislabelled as originating from *Vulpes vulpes* on RABV-GLUE. Sequences containing ambiguous bases (i.e., N, R or Y) were removed from the dataset, leaving 430 complete N gene sequences in total. The GenBank accession numbers for these sequences are available in the supplementary materials (Supplementary Table 6).

4.3.2 Phylogenetic trees

Phylogenetic trees were constructed using IQtree (Nguyen et al. 2015) with a Gannoruwa bat lyssavirus sequence (accession no. NC_031988) used as an outgroup. The best fitting model of the tree shown was GTR+F+I+G4. Visualisation was carried out in R version 4.3.2 (R Core Team 2020) using the `ggtree` package (Yu et al. 2017).

4.3.3 Codon usage

In order to compare how biased the codon usage of each rabies clade is, the effective number of codons (a measure of overall codon usage bias accounting for amino acid composition, with values ranging from 20 to 61, where a value of 20 represents strict bias towards one codon per amino acid, and 61 represents no bias in codon usage; ENC) was calculated using the equations shown in Wright (1990). Raw codon usage and relative synonymous codon usage (RSCU) values were calculated for each species group using the CAIcal web server (Puigbò et al. 2008). The raw codon usage was analysed using principal component analysis (PCA) using the `stats` R package (R Core Team 2020). Nucleotide content, including GC content at each position, was also calculated using CAICal. GC3 content (guanine and cytosine content at the third codon position) was compared to both ENC values and GC12 content (guanine and cytosine content at the first and second codon positions) to assess to what extent codon usage biases were influenced by GC content. Codon adaptation index (CAI) values and expected CAI (eCAI; the upper limit of CAI of 500 sequences with identical amino acid usage and GC content but randomised codon usage, not exceeded by 95% of the sequences with a 95% confidence limit) values were calculated for each RABV clade against four host species using CAIcal to measure the similarity in codon usage between each rabies clade and the corresponding host species.

The reference codon usage datasets were obtained for *Canis familiaris*, *Vulpes lagopus*, *Eptesicus fuscus* and *Desmodus rotundus* from CoCoPUTs (Alexaki et al. 2019); other host species had no available codon usage data. The CAI values were divided by the eCAI values to give a normalised CAI (nCAI).

4.3.4 CpG content

The ratio of observed to expected CpG content for each sequence was calculated as in Gardiner-Garden and Frommer (1987). In order to determine when in rabies' evolutionary history changes in CpG content may have occurred, ancestral sequence reconstruction was carried out using IQtree and the ratio of observed to expected CpG content for these predicted ancestral sequences was also calculated. Corrected relative synonymous dinucleotide usage (RSDUc) values, a measure of dinucleotide usage that accounts for nucleotide and amino acid usage biases, were calculated using the `DinuQ` python package for CpG content (Lytras and Hughes 2020). These were then analysed using the R package `PhylogeneticEM` (Bastide et al. 2018), which predicts when environmental (i.e., host species or tissue in the case of viruses) shifts may have occurred in the phylogeny.

4.4 Results

The RABV phylogeny is broadly split into carnivore-associated clades and bat-associated clades, with the south-central skunk clade grouped more closely with the bat-related clades (Figure 4.1). All minor clades were monophyletic with the exception of the Asian SEA2a domestic dog minor clade, which contains the Asian SEA2b Chinese ferret badger (CFB) clade.

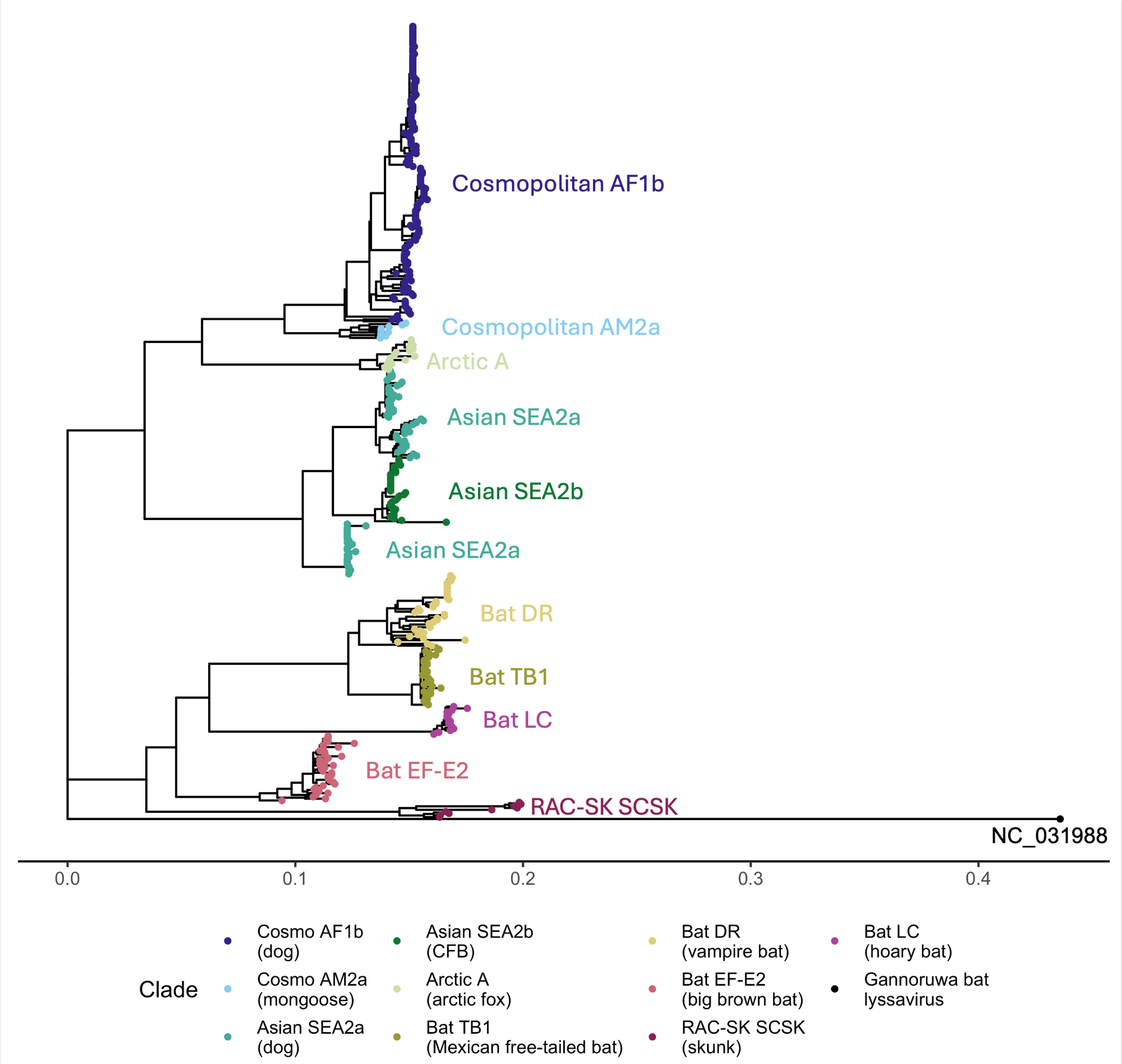

Figure 4.1: A maximum-likelihood phylogenetic tree of all the sequences used in analyses. A Gannoruwa bat lyssavirus sequence was used as an outgroup (NC_031988). Tips are coloured and annotated by host-specific minor clade. The x-axis is scaled in substitutions per site.

4.4.1 Codon usage

The effective number of codons (ENC) used by the rabies virus had a mean value of 54.0 (where a value of 20 denotes the strictest possible bias of one codon per amino acid, and 61 denotes no bias) with a standard deviation of 2.30, and varied between host-species-specific minor clades (Figure 4.2). The dog-associated AF1b rabies minor clade was the most biased in its codon usage (mean ENC of 51.8), with the big brown bat and arctic fox-associated clades also being more biased than other clades. The Chinese ferret badger-associated clade was the least biased (mean ENC of 57.6).

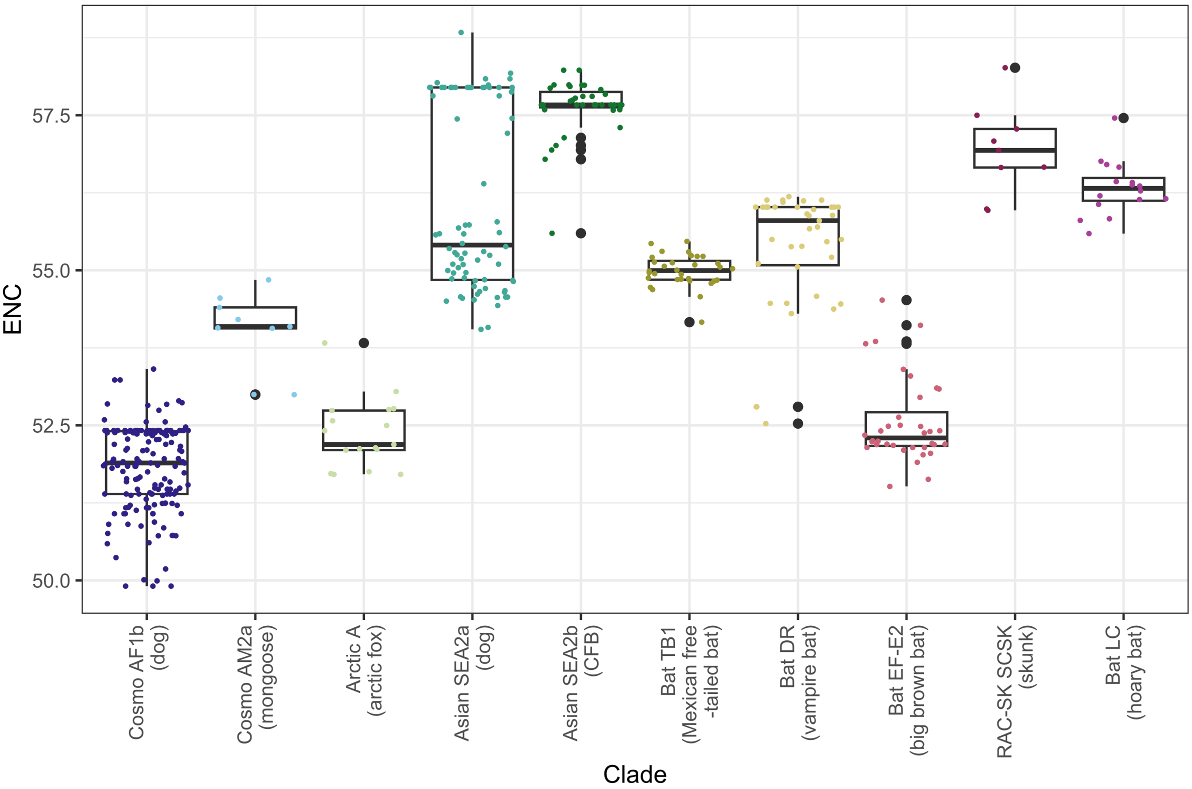

Figure 4.2: The effective number of codons (ENC) across different host-species-specific rabies clades. Points are coloured by species-specific minor clade. An ENC of 61 denotes no bias in codon usage, whereas an ENC of 20 denotes strict bias towards one codon per amino acid.

Certain amino acids appear to have stronger codon biases than others (Figure 4.3; Supplementary Table 7); arginine codon AGA has an RSCU value above 2 in every host species investigated (RSCU values above 1.6 suggest significant over-representation). Preferred codons varied between host-specific rabies clades for almost all amino acids; only arginine had a common preferred codon across all of the host-specific rabies groups investigated. Codons containing a CpG dinucleotide were almost universally underrepresented.

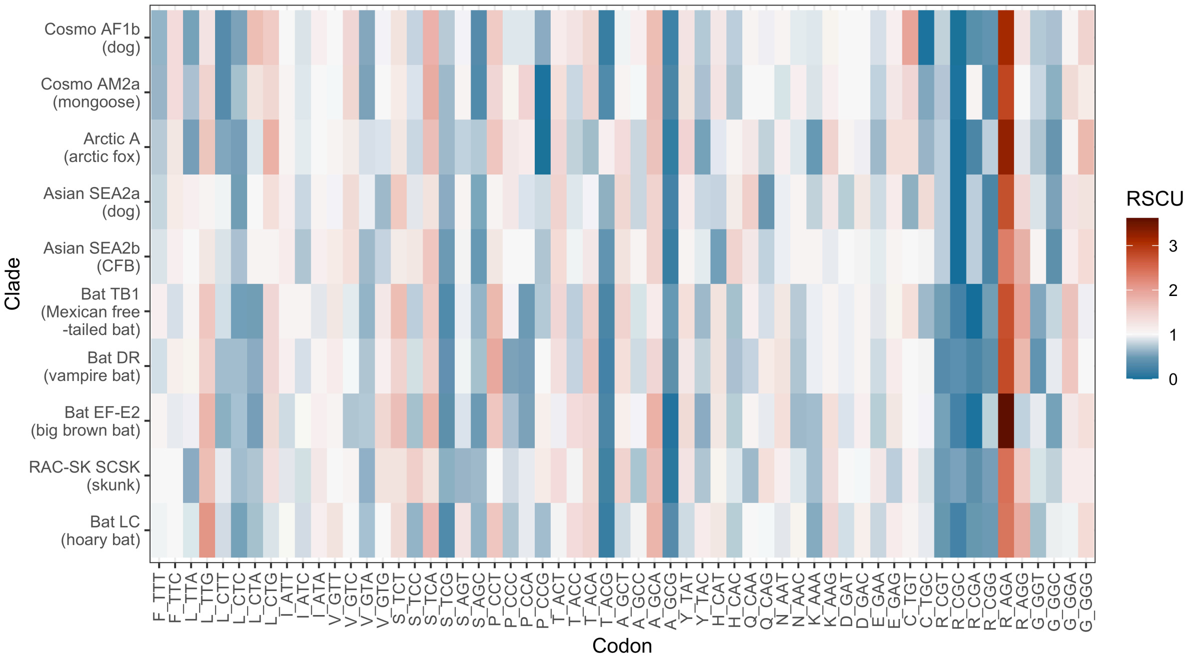

Figure 4.3: Strength of codon preference varies by amino acid. Cell colour represents RSCU value, where white indicates no codon preference, red indicates preference towards the codon, and blue preference against.

The ENC-GC3 plot (Figure 4.4A) is used to investigate whether the effective number of codons is a result of GC3 content (i.e., GC content in the third codon position), where the curve represents the expected ENC if codon bias was due purely to GC3 content. As almost all of these points sit below the expected curve and do not follow it, we can assume that codon bias is not due purely to GC3 bias, and may be due to other factors such as natural selection.

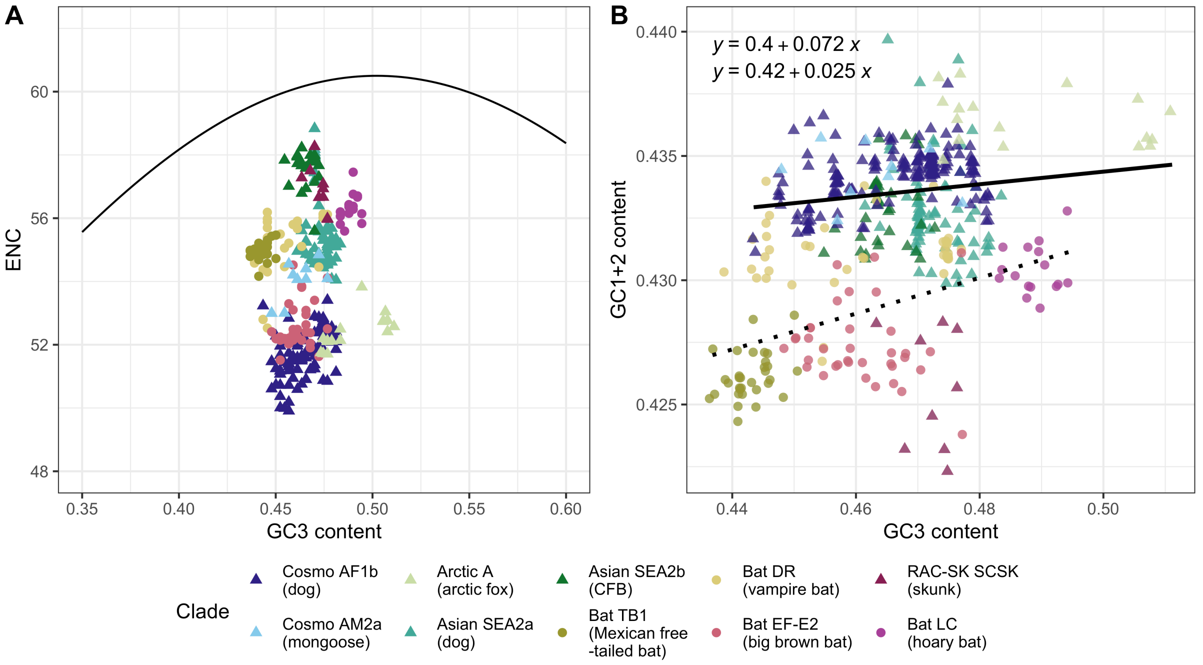

Figure 4.4: Codon usage is minimally influenced by mutation pressure. (A) ENC in relation to GC content at the third position. (B) GC12 content plotted against GC3 content for all sequences. Lines and equations represent the linear models fitted to data from each major host group, i.e. bats (dotted line, upper equation) and carnivores (solid line, lower equation). Points have been jittered to prevent overplotting. Point colour represents the rabies clade. Point shape represents whether the host species is a bat or a carnivore.

Plotting GC12 content against GC3 content can be used to determine how much GC content, and therefore mutation pressure, influences codon usage bias, where a linear regression slope of 1 suggests GC content alone is responsible for the observed codon usage biases, and a slope of 0 denotes no effect of GC content. The slope of the regression line fitted to all data points is 0.1009, suggesting that codon usage is not influenced by GC content to a great extent, matching the findings of the ENC-GC3 analysis. When split into bat- or carnivore-associated groups (Figure 4.4B), the slope of the regression line fitted to the bat-related points is 0.072 whereas the carnivore-related slope is 0.025. Taken together with the ENC-GC3 analysis, this suggests that codon usage biases in bat-related RABV are slightly more influenced by mutation pressure than in carnivore-related RABV, while still mostly being influenced by other factors.

We then conducted principal component analysis (PCA) on the raw codon usage values to further explore any patterns in codon usage between clades and investigate which codons have the most influence over these differences. The first and second principal axes accounted for 24.2% and 21.8% of the total variation in codon usage respectively, accounting for a total of 46.0% of the variation in codon usage (Figure 4.5). Some host-specific rabies clades appear to group by major clades, i.e., Asian and Bat/RAC-SK, with bat-related and carnivore-related clades broadly splitting along PC2 and Cosmopolitan and Asian clades splitting along PC1. Some clades, however, are more tightly clustered than others, and there is considerable overlap between the Cosmopolitan AM2a, RAC-SK SCSC and Bat LC clades, suggesting similar codon usage patterns. The codons exerting the most influence over PC1 (i.e., the Asian vs Cosmopolitan divide) were leucine CTT (loading of -0.240), histidine CAC (-0.206), and cysteine TGT and TGC (0.228 and -0.229 respectively). PC2 (i.e., the bat- vs carnivore- associated RABV divide) was most influenced by leucine CTA (-0.226), valine GTC (-0.247) and arginine CGT and CGC (-0.246 and 0.236 respectively).

To investigate whether RABV codon usage patterns resembled those of the clades' host species, we calculated the normalised codon adaptation index for each clade with *Canis familiaris*, *Vulpes lagopus*, *Desmodus rotundus* and *Eptesicus fuscus* codon usage values as references (Figure 4.6). The bat-associated rabies clades, along with the skunk clade, had higher normalised CAI values than the carnivore-associated clades, but only the Bat EF-E2, DR and TB1 clades had a mean nCAI above 1 for any host species, which is associated with significant similarity to the host's codon usage patterns. There was no evidence of RABV codon usage being more similar to that of bats than to carnivores or *vice versa.*

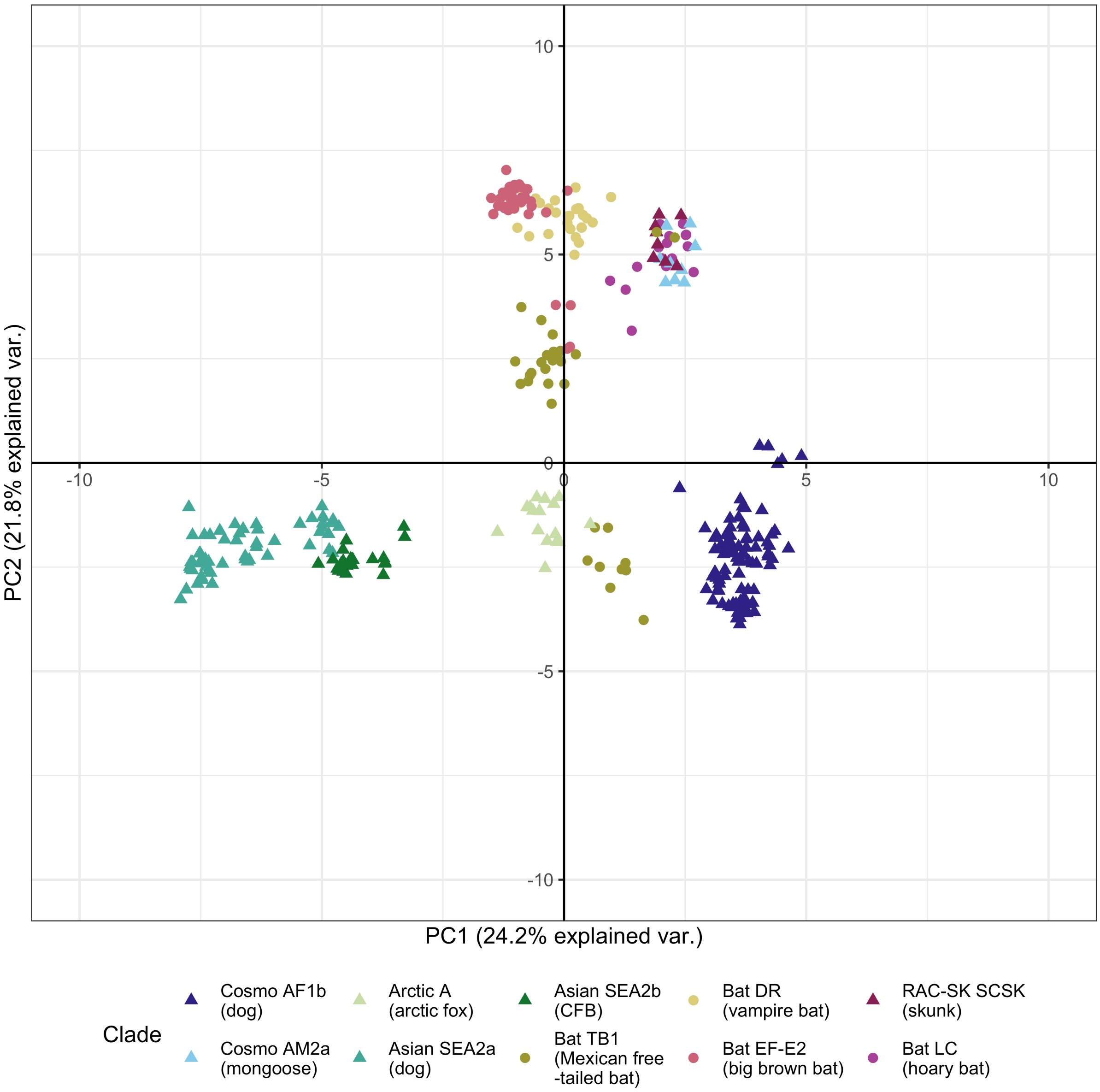

Figure 4.5: Principal component analysis of raw codon usage values. Point colour represents the rabies clade. Point shape represents whether the host species is a bat or a carnivore.

4.4.2 CpG content

As codons containing CpG dinucleotides were found to be among the most influential in the PCA analysis, and are known to be selected against by the host immune system in some viruses (Meagher et al. 2019), we investigated whether there were differences in CpG content between the RABV clades. CpG underrepresentation (i.e., a ratio of observed to expected CpG content less than 1) was seen across all host-species-specific minor clades

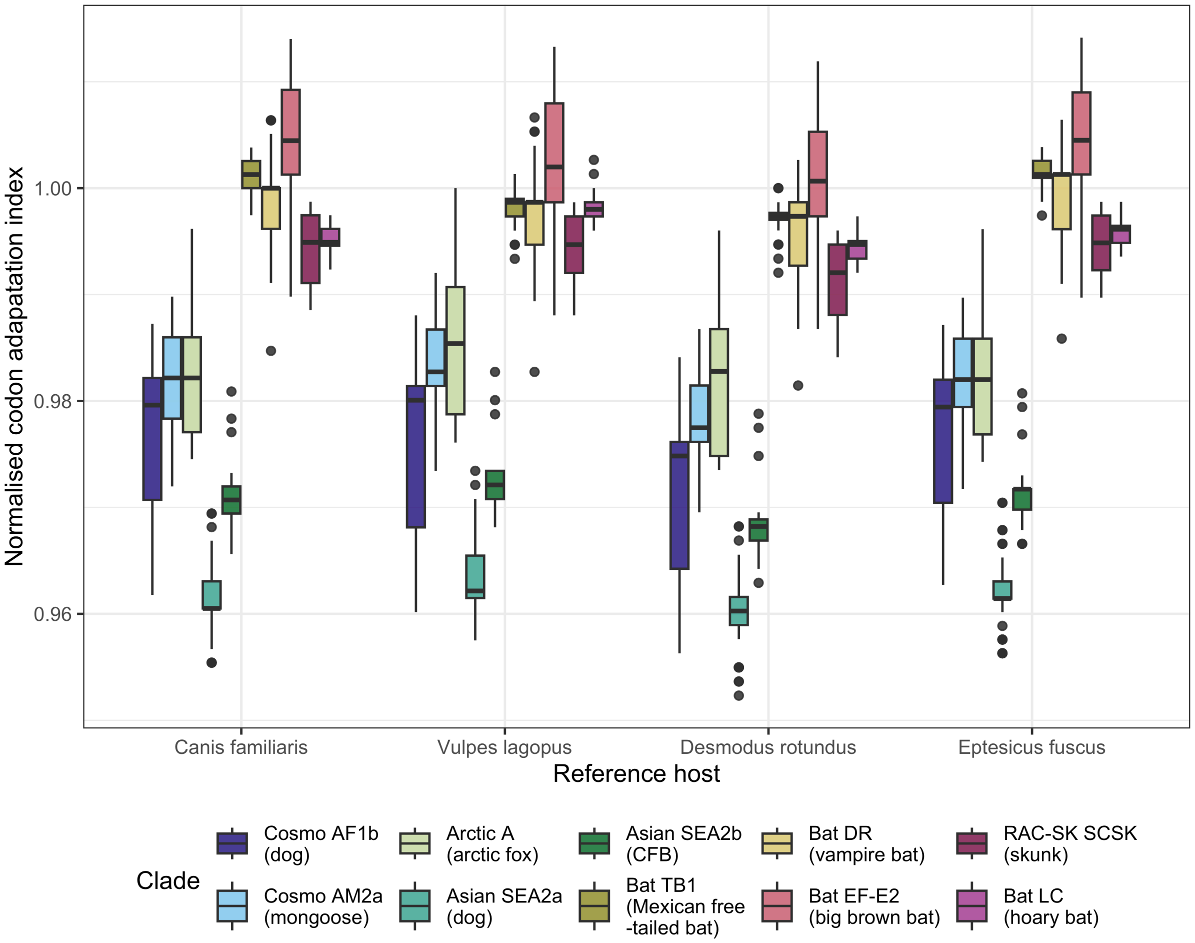

Figure 4.6: Codon adaptation index (CAI) of each host-specific rabies clade. *Canis familiaris*, *Vulpes lagopus*, *Desmodus rotundus* and *Eptesicus fuscus* codon usage values are used as reference sets; bar colour represents RABV minor clade.

(Figure 4.7), with bat-associated rabies clades having even lower CpG content than in carnivore-associated clades. The Bat TB1 and Bat LC clades in particular displayed high levels of CpG suppression, and this difference appears to be driven by the number of CpG dinucleotides as opposed to differences in the underlying GC content.

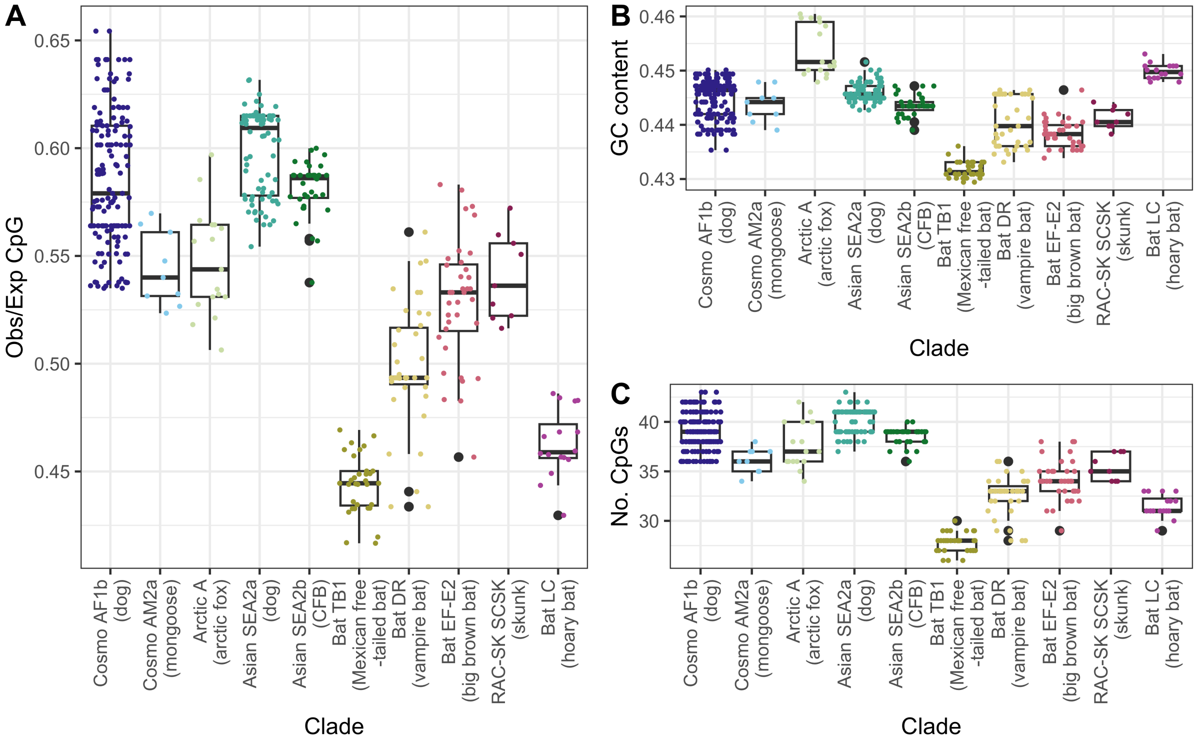

Figure 4.7: Ratio of observed to expected CpG content and GC content across host-specific rabies groups. (A) Ratio of observed to expected CpG content, where a value of 1 indicates no bias, >1 indicates overrepresentation of CpG, and <1 indicates underrepresentation. (B) GC content by host species. (C) Number of CpG dinucleotides per host species..

In order to investigate whether the low CpG content in bat-associated clades developed before or after the bat to carnivore host shift, we used ancestral sequence reconstruction and calculated the CpG content of these sequences. This predicted that low CpG content was the ancestral state of RABV (Figure 4.8), with the common ancestor of all the clades having an observed to expected CpG content ratio of 0.428, and higher CpG content emerging in carnivore-associated clades after the common ancestor of the carnivore-associated clades.

More thorough analysis using `PhylogeneticEM` with RSDUc values of each sequence predicts 16 ancestral environmental shifts (i.e., changes in the type of host or tissue inhabited) in our dataset that caused changes in CpG content across the 3 codon dinucleotide positions (Figure 4.9). In many cases, these shifts line up with a host shift event (i.e., the descendants of the environmental shift encompass an entire host-specific rabies clade); others appear to only have one descendant.

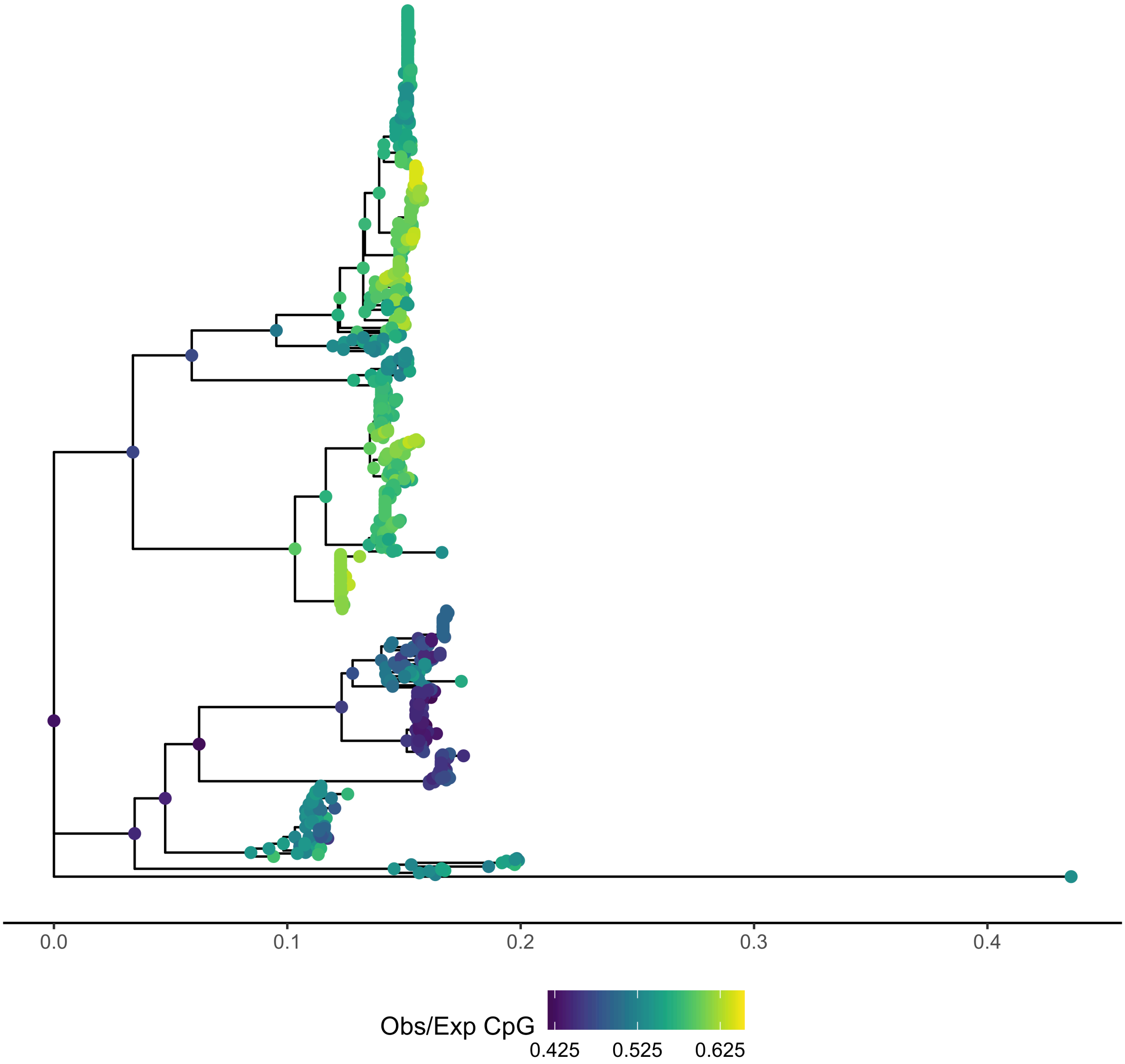

Figure 4.8: Low CpG content appears to be the ancestral state of RABV. Phylogenetic tree of all the N gene sequences with ancestral node sequences reconstructed and CpG content calculated. Point colour represents the observed CpG content as a proportion of the expected CpG content given guanine and cytosine content.

To determine whether this difference was caused by the hosts' zinc-finger antiviral protein (ZAP), we compared the frequency of ZAP's optimal binding motif in mice C(n7)G(n)CG (Luo et al. 2020b) to a similar but suboptimal motif C(n7)C(n)CG, as previously applied to SARS-CoV-2 (Afrasiabi et al. 2022). We found that three bat clades (Bat TB1, Bat DR and Bat EF-E2) showed very low levels of the optimal motif compared to the suboptimal motif, suggesting high levels of selection against this motif by ZAP (Figure 4.10). Surpris-

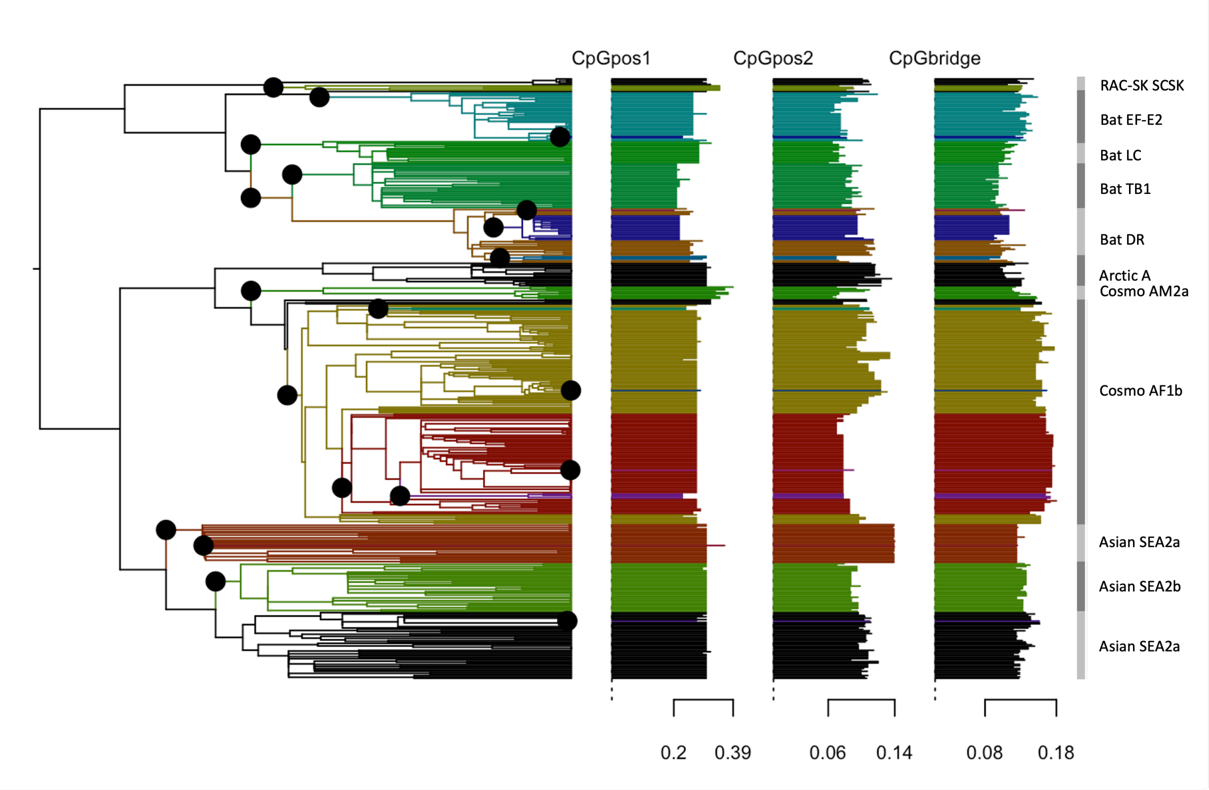

Figure 4.9: Multiple historic shifts in environment are predicted based on changes in CpG content at the three codon positions. CpGpos1 denotes the probability of a dinucleotide containing the first two nucleotides of a codon being CG; CpGpos2 notes the probability of a dinucleotide containing the final two nucleotides of a codon being CG; CpGbridge denotes the probability of a dinucleotide containing the final nucleotide of one codon and the first nucleotide of the next being CG. Black points on the tree are the predicted location of an environment shift. Coloured groups on the tree and bar plots represent the ancestors affected by each shift (colour is not indicative of CpG shift direction).

ingly, the sequences from the East African domestic dog-associated clade Cosmopolitan AF1b contained much higher levels of the optimal binding motif than the suboptimal motif, suggesting reduced selection pressure against the optimal motif and increased selection pressure against the suboptimal motif.

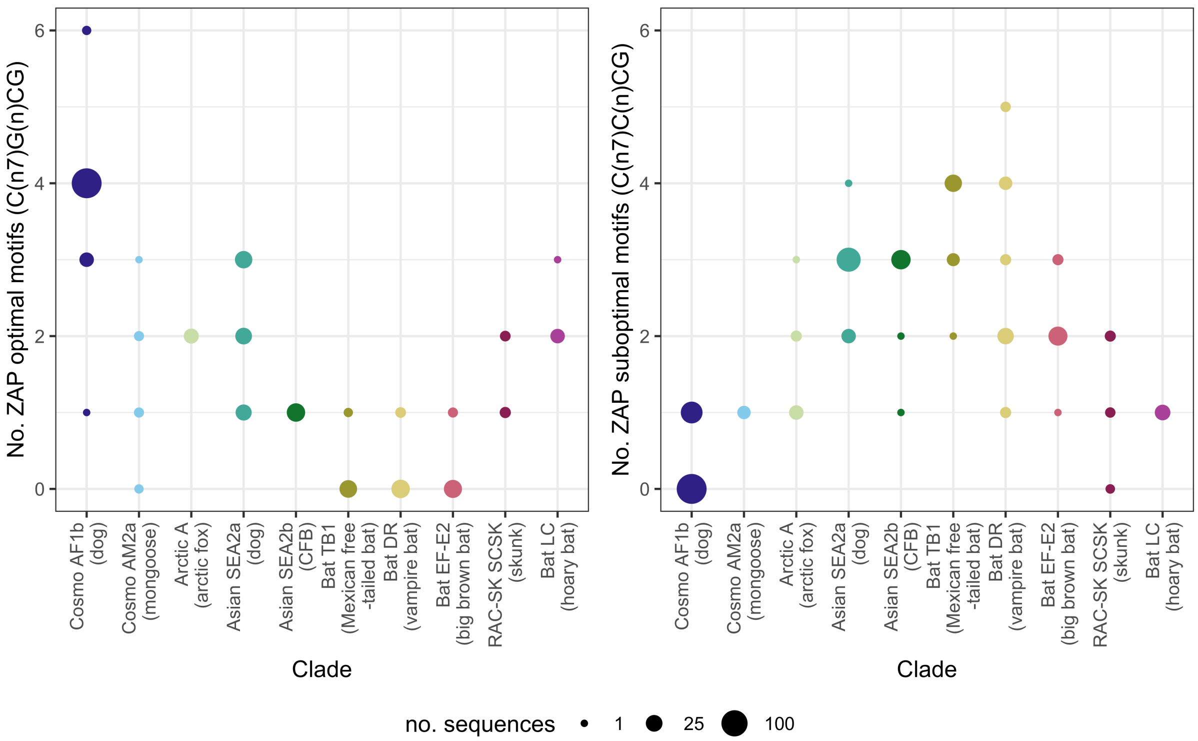

Figure 4.10: Frequency of optimal and suboptimal ZAP binding motifs suggest strong selection pressure from ZAP in bat-related clades. Point colour represents rabies clade. (A) Frequency of mouse-ZAP's optimal binding motif C(n7)G(n)CG. (B) Frequency of a suboptimal motif C(n7)C(n)CG.

4.5 Discussion

We investigated whether there were differences in codon usage biases between host-specific rabies clades, and whether these differences were suggestive of host adaptation. We found that codon usage differed mainly between the bat-associated and carnivore-associated clades as opposed to between species-specific clades that were more closely related, and that some of the codons determined to have the most influence over the differences between these clades contained CpG dinucleotides. Multiple significant shifts in CpG content were found to have occurred in RABV's evolutionary past, with carnivore-associated clades having higher CpG content than bat-related clades, which may be due to the easing of selection against CpG dinucleotides by ZAP.

CpG content in RNA viruses is almost universally lower than would be expected based on GC content due to selective pressure from the host's zinc-finger antiviral protein (ZAP) which specifically targets CpG dinucleotides (Takata et al. 2017). The bat-associated clades we investigated have CpG content considered average compared to other mammalian negative-strand RNA viruses, whereas the carnivore-associated clades' CpG content is comparatively high, while remaining within the known range for single-strand RNA viruses (Simmonds et al. 2013). Carnivores themselves don't have a higher CpG content than bats (Shaw et al. 2021), so this difference is likely not due to host mimicry. This finding is also opposite to similar analysis conducted in coronaviruses, where coronaviruses with bat hosts had a *higher* CpG content than coronaviruses in other mammals (Nchioua et al. 2020). We also found that bat-related rabies clades had very few C(n7)G(n)CG motifs compared to a negative control motif and to carnivore-related clades. While the reason behind the observed difference in CpG content and ZAP-optimal motifs between bat-associated rabies clades and carnivore-associated rabies clades is currently unknown, one possible explanation is that carnivore-associated rabies is less exposed to ZAP and therefore the selection pressure against CpG and the optimal binding motif is eased, leading to increased freedom to use codons or codon pairs that contain CpG dinucleotides may confer some advantage in infecting carnivore hosts.

While there are differences in codon usage biases between host-specific rabies clades, and RABV's codon usage biases are attributed mostly to natural selection which is often associated with translational efficiency, we believe it is unlikely that this is due to adaptation to the host's tRNA availability, and rather due to this easing of selective pressure against CpG dinucleotides in carnivore-related clades and genetic drift. One reason for this is that the Cosmopolitan AF1b and Asian SEA2a clades are the furthest clades apart along PC1 of our principal component analysis, despite both primarily infecting domestic dogs. The RAC-SK south-central skunk clade is also still grouped within the bat clades, suggesting that RABV's codon usage strategy did not need to change to effectively infect the new carnivore host.

The results of the codon adaptation index analyses suggest that bat-associated RABV codon usage is “better adapted” to both carnivore and bat hosts than carnivore-associated RABV, but that RABV overall is not better adapted to bats or carnivores. While initially it might seem interesting that rabies would become less well adapted over the bat to carnivore host shift, codon adaptation index analyses are likely not very informative when it comes to RNA viruses, as their codon usage is heavily influenced by immune evasion and transcription efficiency (Mordstein et al. 2021). This topic was further discussed on the Virological forum early in the COVID-19 pandemic, when studies were published claiming SARS-CoV-2 originated from various host species based on high CAI values (Andersen et al. 2020; Robertson et al. 2020). We included CAI analysis as a common and expected part of a codon usage bias study, especially when investigating the possibility of a pathogen’s adaptation to the host, but do not believe the associated results to be informative. It may be that the higher CAI values observed in the bat rabies clades are simply reflective of their lower CpG contents, as vertebrates also have low levels of CpG content due to the deamination of methylated CpG sites (Sved and Bird 1990).

As this study only investigates the codon usage and CpG content of the RABV nucleoprotein gene, further analysis should be conducted to confirm whether the patterns observed are present across the genome or vary between genes. Further study is also required to elucidate what mechanisms have caused the easing of the selection pressure against CpG during the bat to carnivore host shift, such as mechanistic differences between bat- and carnivore- derived ZAP and its associated proteins or differences in the probability of the exposure of the virus to ZAP. Other lyssaviruses should also be investigated to confirm whether this pattern is consistent across the family.

We found that differences in codon usage between host-species-specific rabies clades were unlikely to be due to adaptation to the host, but rather genetic drift over the scale of decades to centuries of evolution. Combined with a lack of evidence for positive selection over host shifts in previous studies, this suggests that host-specific transmission cycles may be due to ecological circumstances of the host rather than adaptive evolution at the minor clade level. Differences in CpG content between bat- and carnivore-associated clades may suggest underlying differences in host immune mechanisms between the host groups. This may have implications for the efficacy of CpG-enriched live attenuated vaccines between rabies clades should they be developed for the rabies virus, as they have been for Influenza A (Sharp et al. 2023).

Chapter 5

Discussion

Rabies sequence data are becoming increasingly available due to the growing accessibility of sequencing technologies (Brunker et al. 2020; Bautista et al. 2023; Gigante et al. 2020) and the sharing of sequences on platforms such as NCBI's GenBank (Sayers et al. 2020) and RABV-GLUE (Campbell et al. 2022). In this thesis I aimed to explore what this data can tell us about rabies' emergence and evolution, and how it can be used to aid in the global goal of achieving zero human rabies deaths by 2030 (Minghui et al. 2018). In order to achieve this, I developed a simulation framework that combines epidemiological and mutational models to produce large datasets of synthetic rabies genomes linked to outbreaks with known transmission dynamics. I then used these data to investigate whether variable incubation periods could be behind the lack of sufficient temporal signal required for phylodynamic analyses of rabies outbreaks to be conducted on some datasets, and to develop a computationally inexpensive method of estimating outbreak sizes from viral genetic data. I also used publicly available rabies sequence data to investigate whether the host-species-specific rabies transmission cycles sometimes observed have been facilitated by RABV's codon usage becoming adapted to specific hosts.

5.1 Main findings

It is often not possible to carry out Bayesian phylogenetic analysis of rabies datasets using software such as BEAST due to insufficient temporal signal; many phylogenetic studies of rabies do not carry out these analyses as a result (Jaswant et al. 2024). In chapter 2 I investigated whether this may be due to rabies' variable incubation periods leading to the molecular clock hypothesis being violated. I found that while variable incubation periods can disrupt root-to-tip diversity over time for viruses with a high substitution rate, there appears to be no observable effect on viruses with rabies' substitution rate. While the rabies virus is often described as fast evolving due to it being an RNA virus, in actuality it has a relatively low substitution rate compared to other RNA viruses (Biek et al. 2015). This can lead to many datasets being too temporally "shallow" for phylogenetic analysis to be conducted successfully due to a lack of measurable evolution over the timescale that RABV sequences are being collected; at rabies' substitution rate, sufficient temporal signal for phylodynamic analyses may only emerge over decades' worth of genomic data (Duchêne et al. 2015). The lack of this depth of data is likely the cause of the difficulties in conducting phylogenetic analyses, rather than the variability of the infectious period. This depth of data required to conduct phylogenetic analyses makes it difficult to accurately determine the substitution rate and date of emergence of new outbreaks.

In chapter 2 I argued that thinking about substitutions on a per-generation scale to mirror how we measure the generation interval for temporal data and the distance kernel for spatial data may be more useful than measuring substitutions per unit time in certain contexts. This way of thinking is more commonly used in studies of eukaryotes, but became instrumental to my work in chapter 3. In this chapter I found that the relationship between the mean branch length and the proportion of cases sequenced followed a formula within which the per-generation substitution rate was a component. This formula can be rearranged to estimate the proportion of cases sequenced and thus the total outbreak size, if the assumption that the tree length follows an approximately square-root relationship with the proportion of cases sequenced is met. I applied this method to a phylogenetic tree of rabies virus whole genomes from a recent outbreak in the Romblon province of the Philippines. I found that while there was likely to have been extensive undetected transmission in the province, the surveillance of this outbreak was much more successful in identifying cases than is usual for rabies outbreaks, with an estimated 26% of potential cases being identified compared to less than 10% in areas with endemic rabies circulation (Townsend et al. 2013b).

Despite having the ability to infect any mammal species, most rabies clades appear to primarily circulate in specific host species, even when other known rabies host species are present in the area; for example, the Asian SEA2a and SEA2b clades circulate sympatrically in domestic dog and Chinese ferret badger populations, respectively (Liu et al. 2010). There has been mixed evidence to support genetic adaptation of these clades to their preferred host, however, with positive selection events associated with host shifts being rare (Marston et al. 2018; Mollentze et al. 2014). While differences in codon usage biases between minor rabies virus clades were observed, it does not seem plausible that codon usage “adapts” to different host species. Very similar codon usage was observed between the Cosmopolitan AM2a, Bat LC and SCSK clades, which primarily infect mongooses, hoary bats and skunks respectively, all of which belong to different taxonomic families. Conversely, codon usage was very different between the Cosmopolitan AF1b and Asian SEA2a clades, which both primarily infect domestic dogs in East Africa and China respectively. It is therefore likely that these similarities and differences are due to coincidental genetic drift rather than adaptation. CpG content varied between carnivore- and bat-associated clades, with carnivore-associated clades having a higher CpG content than normal for RNA viruses. The host-specific minor clades differed in the frequency of zinc-finger antiviral protein (ZAP) optimal and suboptimal binding motifs, suggesting that some clades are less affected by ZAP, and the increase in CpG content following the bat-to-carnivore host shift may be due to a relaxing of selection pressure against CpG dinucleotides from the host immune system.

5.2 Implications

The findings of this thesis have led to some recommendations as to how we should handle rabies virus genomic data to improve surveillance and outbreak responses. While variable incubation period lengths did not appear to have an effect on temporal signal in RABV sequence datasets, this could complicate phylogenetic analysis in viruses with a higher substitution rate. I would recommend using resources such as Clockor2 (Featherstone et al. 2023) to assess the need for local clock models in datasets where variable incubation periods are paired with high substitution rates.

Due to the temporal depth of sequence data required to successfully conduct phylogenetic analysis on rabies virus data, using a wider dataset than the outbreak being investigated is commonly required for acquiring estimates of the time of emergence or the substitution rate. This is especially true in newly emerging outbreaks where analyses such as transmission tree reconstruction and outbreak size estimation are most useful. The time-dependent rate phenomenon, where substitution rate estimates from temporally deeper datasets are estimated to have a lower substitution rate perhaps due to purifying selection or site saturation (Duchêne et al. 2014), however, may become an issue in this case, where the substitution rate estimate of the wider dataset used may be different to the true substitution rate of the investigated outbreak. It is currently unknown to what extent purifying selection over long sampling windows affects phylogenetic analysis of rabies virus data, but this should be further investigated in the future and taken into account where necessary.

The equation presented in chapter 3 provides a simple, non-computationally expensive way to estimate the proportion of cases in an outbreak that have been sequenced, and in turn the total size of the outbreak. This enables outbreak size estimation to be built into the genomic surveillance pipeline without the use of any software or specialist skills that are not already routinely used during genomic surveillance efforts. This will be helpful in evaluating the effectiveness of responses and interventions, and will aid in tracking the decline in numbers of rabies cases as elimination is approached.

The finding that there were potentially 117 undetected rabies cases in Romblon is unfortunately unsurprising given that no routine surveillance was in place during the 10 year period where the province was declared rabies free; continued mass dog vaccination after this point, however, could have prevented the sustained onward transmission after the five potential introduction events. The surveillance of the Romblon outbreak had more success in detecting rabies cases than responses in other areas; surveillance achieved an estimated 26.3% detection rate (45 cases confirmed out of an estimated 163 total cases), whereas a detection rate of below 10% is to be expected in many rabies-endemic areas (Townsend et al. 2013b), with an estimated 12% of cases being detected across Tanzania (Hampson et al. 2016). The detection rate in Romblon falls short of the estimated detection rate of 83 - 95% of cases within the Serengeti district (Mancy et al. 2022), but this is likely due to the Serengeti district being the subject of an extensive, long-term contact tracing research effort, and this level of detection is unlikely to be required for effective rabies control.

Our finding that the differences in codon usage between host-species-specific rabies virus clades are likely due to genetic drift suggests that codon usage does not require adaptation in order to effectively infect new host species. Differences in CpG content and ZAP-optimal binding motifs, however, suggest that the evolution of bat-associated rabies clades has been affected by a host environment which is under a strong selection pressure by zinc-finger antiviral proteins, whereas this selection pressure may not be acting as strongly on carnivore-associated clades. This may be because carnivore-ZAP has a different mode of function to bat-ZAP, such as binding to a different optimal motif, or that carnivore rabies has traded being more susceptible to ZAP for some unknown evolutionary advantage. Unfortunately little is known about ZAP, especially in bats, and there have been no studies investigating how ZAP is activated by or interacts with RABV. This potential difference in ZAP action should be considered if a CpG-enriched live attenuated vaccine is ever developed for rabies, as it has been for Influenza A (Sharp et al. 2023), or when using a CpG oligodeoxynucleotide as a rabies vaccine adjuvant (Yu et al. 2018), as its efficacy may vary between bats and carnivores.

5.3 Future work

The findings in this thesis have raised questions that could be addressed through further study. It is currently unknown to what extent purifying selection and site saturation over long sampling windows affects phylogenetic analysis of rabies data, so the effect of the time-dependency of substitution rates over datasets of different temporal depths on rabies genomic analyses should be quantified and minimised. Studies of other organisms have compared rates of synonymous and non-synonymous mutations (Soares et al. 2009) to try and quantify and correct for this problem, but as I found in chapter 4 synonymous substitutions can also be deleterious if they result in the formation of a CpG dinucleotide and may also be subject to purifying selection in some RABV clades. Another method of quantifying this effect is comparing the substitution rates of terminal and internal branches (Ghafari et al. 2022). As the relationship between the width of the sampling window and the observed substitution rate has been observed to follow a predictable relationship and can therefore be corrected for using a power law rate decay model over very long time scales (Aiewsakun and Katzourakis 2015, 2016; Ghafari et al. 2021), it might be possible to correct for this effect over shorter, more rabies-relevant time scales in a similar way.

The exact relationship between the R_0, tree length and proportion of cases sequenced in an outbreak is still unknown at the time of writing. While assuming a square root relationship between tree length and the proportion of cases sequenced works well for outbreaks with rabies' usual R_0 range of between 1 and 2 (Li et al. 2024), the poor fit of the square root curve for outbreaks with values above this limit the method's applicability to pathogens with high R_0s such as measles or chicken pox (estimated R_0s of 12-18 and 8-9 respectively; Anderson and May 1982). Determining the true relationship will make this method applicable to a wider range of pathogens and is a priority for future research. The relationship between the tree length and proportion of cases sequenced may not be dependent on R_0 itself, but some related measure such as the proportion of cases with no descendants, as the removal of these tips from the tree is the main contributor to tree length reduction.

CpG dinucleotides and ZAP-optimal motifs vary in frequency across minor rabies clades, but it is unknown whether this is a response to differences in the form and function of ZAP itself between the host species, or a trade-off with an advantageous change that makes some clades more susceptible to ZAP. Further analysis to determine this could involve a laboratory study comparing the optimal binding motifs of ZAP in dogs (or another carnivore rabies host) and bats, or experimentally infecting bats and dogs and comparing how ZAP and its associated proteins are expressed over the course of the infection. This may help uncover details of how bat's immune systems differ from other mammals, which is still largely unknown.

5.4 Conclusion

Here I have shown that genomic data can be a powerful tool in aiding in rabies elimination, and that it has wider uses outwith advanced phylodynamic analyses as outlined in the introduction. Genomic data can be used to monitor efforts to eliminate canine rabies, and can tell us more about how rabies may behave differently in other host species when the focus of elimination efforts shifts from domestic dogs to wildlife reservoirs. Nahata et al. (2021) recommend that phylogeographic analyses of rabies sequence data become more widely used to investigate the geographic origins and spread of rabies outbreaks. These analyses, like the ones carried out in this thesis, will only help in the fight to eliminate rabies if they lead to the implementation of control measures in real time, or inform how we respond to future outbreaks. Without this, the money used on sequencing and analysing samples may be better spent on dog vaccination or improving PEP accessibility.

Appendices

A Chapter 4 Supplementary Tables

Clade	Accession numbers
Asian SEA2a	DQ666289, DQ666290, DQ866105, DQ866106, DQ866108, DQ866109, DQ866110, DQ866112, DQ866116, DQ866119, EU086182, EU086183, EU159386, EU828653, FJ561726, FJ561727, FJ561728, FJ866829, FJ866830, FJ866831, GQ472468, GQ472470, GQ472473, GQ472477, GU358653, HM486355, HM486356, HM486357, HM486358, HM486359, HM486361, HM486362, HM486364, HM486365, HM486366, HM486367, HM486368, HM486369, HM486370, HM486371, HM486372, HM486373, HM486374, HM486375, HM486379, HM486380, HM486381, JF819605, JF819615, JF819622, JN974823, JN974824, JN974828, JN974829, JN974838, JN974842, JN974845, JN974854, JN974870, JN974871, JN974872, JN974873, JQ970480, KT221107, KT894558, KT894572, KT894573, KT894577, KT894578, KX148264, KY451767, MG201879, MG201880, MG201884, MG201919, MG201921
Asian SEA2b	FJ598135, FJ712195, FJ712196, GU647092, KF726852, KP319184, KP319185, KP319186, KP319187, KP319188, KP319189, KP319190, KP319191, KP319192, KP319193, KP319194, KP319195, KP319196, KP319197, KP319198, KP319199, KP319200, KP319201, KP319202, KP319203, KP319204, KP319205, KP319206, KP319207, KP319208, KP319209, KP319210, KP319211, KP319212, KP319213

Cosmopolitan AF1b	AB284509, AB284510, AB284511, AB284512, AB284513, AB284514, AB285215, EU853582, EU853584, EU853585, EU853586, EU853587, EU853588, EU853589, EU853590, HM179504, HM179505, HM179506, KF155002, KR534217, KR534218, KR534219, KR534220, KR534228, KR534229, KR534230, KR534231, KR534233, KR534238, KR534244, KR534247, KR534249, KR534250, KR534251, KR534252, KR534253, KR906734, KR906735, KR906736, KR906739, KR906741, KR906742, KR906743, KR906744, KR906745, KR906746, KR906747, KR906748, KR906749, KR906750, KR906751, KR906752, KR906754, KR906755, KR906756, KR906757, KR906759, KR906760, KR906762, KR906763, KR906765, KR906766, KR906767, KR906768, KR906770, KR906771, KR906772, KR906773, KR906774, KR906776, KR906780, KR906782, KR906784, KR906789, KR906790, KR906792, KT336432, KT336433, KT336434, KT336435, KT336436, KT336437, KX148203, KX148204, KX148205, KX148206, KX148208, KY210222, KY210223, KY210224, KY210226, KY210227, KY210229, KY210231, KY210232, KY210234, KY210235, KY210238, KY210239, KY210243, KY210244, KY210245, KY210247, KY210248, KY210249, KY210250, KY210251, KY210252, KY210253, KY210254, KY210255, KY210256, KY210257, KY210258, KY210260, KY210263, KY210264, KY210265, KY210266, KY210267, KY210269, KY210270, KY210271, KY210272, KY210273, KY210278, KY210279, KY210280, KY210281, KY210282, KY210286, KY210287, KY210288, KY210289, KY210290, KY210292, KY210293, KY210294, KY210295, KY210296, KY210299, KY210300, KY210303, KY210304, KY210307, KY210311, KY563717, LC029890, MN726804, MN726826, MN726827, MT454631, MT454634, MT454635, MT454636, MT454637, MT454639, MT454641, MT454643, MT454644, MT454645
Cosmopolitan AM2a	AY854505, AY854525, AY854531, AY854544, AY854546, AY854558, AY854567, AY854573, AY854576
RAC-SK SCSC	AF483524, EU345002, EU345003, EU345004, JQ513553, JQ685938, JQ685968, MW055084, MW055085, MW055086, MW055087

Bat DR	AB519642, AF070449, AF351847, AF351852, AY854587, AY854592, AY854594, AY877433, AY877434, AY877435, GU592648, KF656696, KF656697, KF864234, KF864322, KF864397, KM594040, KM594041, KM594042, KP202393, KT023101, KU523255, MN968374, MN968375, MN968376, MN968377, MN968378, MN968379, MN968380, MN968381, MN968382, MN968383, MN968384, MN968385, MN968386, MN968387, MT891038, MW249020, MW249021
Bat EF-E2	AF351828, AF351831, AF351832, AF351833, AF351855, AF351861, AF351862, AY039227, AY039228, AY039229, GU644652, GU644654, GU644655, GU644656, GU644657, GU644658, GU644659, GU644660, GU644661, GU644662, GU644663, GU644664, GU644665, GU644666, GU644667, GU644668, GU644669, GU644670, GU644671, GU644676, GU644677, GU644684, GU644689, GU644690, GU644695, JQ685925
Bat LC	AF351845, AF351846, AF351858, AF394883, AF394884, GU644712, GU644713, GU644714, GU644715, GU644716, GU644717, GU644718, GU644719, GU644720, GU644721, JQ685947
Bat TB1	AF351849, AF394876, GU644760, GU644761, GU644762, GU644763, GU644764, GU644765, GU644766, GU644767, GU644768, GU644769, GU644770, GU644771, GU644772, GU644773, GU644774, GU644775, GU644776, GU644777, GU644778, GU644779, GU644780, GU644781, GU644782, GU644783, GU644784, GU644785, GU644786, GU644787, GU644788, JQ685905

Table 6: Accession numbers for the sequences used in our analysis by host species.

Amino acid	Codon	Asian SEA2b	Cosmo AF1b	Asian SEA2a	Arctic A	Cosmo AM2a	RAC-SK SCSC	Bat LC	Bat EF-E2	Bat TB1	Bat DR
Phe	TTT	0.885	0.636	0.819	0.772	0.642	**1.01**	0.972	**1.07**	**1.12**	0.867
	TTC	**1.12**	**1.36**	**1.18**	**1.23**	**1.36**	0.990	**1.03**	0.932	0.877	**1.13**
Leu	TTA	1.02	0.525	1.12	0.486	0.723	0.615	0.903	1.04	0.965	1.08
	TTG	**1.24**	1.26	1.07	1.66	**1.50**	**1.77**	**2.11**	**1.80**	**1.63**	**1.54**
	CTT	0.886	0.352	0.968	0.571	0.332	0.938	0.859	0.607	0.868	0.682
	CTC	0.719	0.540	0.513	0.528	0.704	0.693	0.528	0.709	0.519	0.678
	CTA	1.06	**1.72**	1.00	0.929	1.29	0.724	0.704	0.522	1.43	0.594
	CTG	1.07	1.60	**1.33**	**1.83**	1.45	1.26	0.891	1.40	1.43	1.42
Ile	ATT	**1.15**	1.01	**1.12**	**1.06**	**1.07**	0.948	1.03	0.853	1.04	**1.09**
	ATC	0.740	0.889	0.837	0.935	0.915	0.842	0.858	1.03	1.04	0.871
	ATA	1.11	**1.10**	1.04	1.00	1.02	**1.21**	**1.11**	**1.11**	0.921	1.04
Val	GTT	1.17	1.01	0.983	1.07	0.989	0.983	**1.28**	1.07	**1.22**	**1.16**
	GTC	**1.34**	**1.43**	**1.39**	**1.17**	**1.46**	1.09	1.00	0.737	0.932	0.987
	GTA	0.669	0.574	0.958	0.862	0.548	0.637	0.711	0.752	0.634	0.732
	GTG	0.812	0.984	0.667	0.899	1.01	**1.29**	1.00	**1.44**	1.21	1.12
Ser	TCT	1.19	1.27	**1.61**	1.24	1.18	1.33	1.41	1.70	**1.73**	1.52
	TCC	1.25	0.809	0.922	0.827	0.895	**1.47**	0.640	0.701	0.783	0.863

	TCA	**1.61**	**1.83**	1.28	**1.72**	**1.89**	1.22	**1.74**	**1.82**	1.58	1.53
	TCG	0.487	0.763	0.599	0.656	0.667	0.646	0.305	0.388	0.320	0.418
	AGT	0.980	0.982	1.04	0.826	1.05	0.677	1.27	0.894	0.955	0.972
	AGC	0.483	0.348	0.554	0.727	0.324	0.662	0.640	0.497	0.631	0.701
Pro	CCT	**1.24**	**1.57**	1.04	**1.63**	1.42	1.02	**1.63**	**1.62**	**1.77**	**1.94**
	CCC	1.00	0.916	**1.21**	1.22	1.08	0.905	0.735	0.720	0.984	0.541
	CCA	1.02	0.909	0.901	1.16	**1.47**	0.971	0.941	0.537	0.492	0.521
	CCG	0.736	0.601	0.850	0	0.0278	**1.11**	0.691	1.13	0.758	0.995
Thr	ACT	**1.37**	1.15	**1.49**	**1.49**	1.35	1.33	0.978	0.936	1.42	1.31
	ACC	0.971	1.25	0.907	0.756	0.837	1.00	1.37	1.36	0.777	0.816
	ACA	1.22	**1.45**	0.987	0.678	**1.49**	**1.45**	**1.51**	**1.41**	**1.55**	**1.63**
	ACG	0.438	0.160	0.615	1.08	0.320	0.210	0.139	0.268	0.258	0.251
Ala	GCT	1.28	0.968	1.44	1.36	1.06	1.45	0.852	1.13	**1.48**	1.35
	GCC	1.08	1.09	0.833	0.851	0.990	0.730	0.982	0.986	0.883	0.846
	GCA	**1.50**	**1.72**	**1.49**	**1.66**	**1.69**	**1.74**	**1.80**	**1.84**	1.15	**1.42**
	GCG	0.143	0.216	0.239	0.132	0.257	0.0839	0.363	0.0443	0.480	0.392
Tyr	TAT	**1.07**	**1.25**	**1.16**	**1.48**	**1.28**	**1.25**	0.851	**1.41**	**1.33**	**1.20**
	TAC	0.930	0.748	0.836	0.516	0.720	0.748	**1.15**	0.592	0.670	0.796
His	CAT	0.519	**1.22**	0.810	0.923	**1.27**	**1.08**	**1.23**	**1.23**	**1.30**	**1.28**

	CAC	**1.48**	0.780	**1.19**	**1.08**	0.735	0.918	0.769	0.771	0.700	0.722
Gln	CAA	**1.19**	**1.04**	**1.53**	**1.20**	1.00	0.728	**1.01**	**1.06**	**1.12**	0.810
	CAG	0.806	0.963	0.469	0.800	1.00	**1.27**	0.988	0.941	0.879	**1.19**
Asn	AAT	0.955	**1.08**	0.953	**1.14**	0.883	**1.09**	0.918	**1.33**	**1.24**	**1.27**
	AAC	**1.05**	0.924	**1.05**	0.826	**1.12**	0.914	**1.08**	0.671	0.764	0.726
Lys	AAA	**1.06**	0.914	0.793	0.479	0.885	0.703	0.589	0.678	0.960	0.955
	AAG	0.943	**1.09**	**1.21**	**1.52**	**1.11**	**1.30**	**1.41**	**1.32**	**1.04**	**1.05**
Asp	GAT	**1.02**	0.997	0.772	**1.04**	0.992	0.982	0.867	0.900	0.956	0.956
	GAC	0.977	**1.00**	**1.23**	0.964	**1.01**	**1.02**	**1.13**	**1.10**	**1.04**	**1.04**
Glu	GAA	0.932	0.880	0.842	0.683	0.805	**1.20**	0.833	0.768	0.920	0.841
	GAG	**1.07**	**1.12**	**1.16**	**1.32**	**1.20**	0.805	**1.17**	**1.23**	**1.08**	**1.16**
Cys	TGT	**1.02**	**1.97**	0.610	**1.36**	**1.43**	**1.13**	**1.06**	**1.01**	**1.33**	**1.02**
	TGC	0.981	0.0284	**1.39**	0.639	0.571	0.866	0.938	0.985	0.667	0.978
Arg	CGT	0.781	0.804	0.781	0.786	0.746	0.495	0.480	0.531	0.504	0.360
	CGC	0	0.0153	0	0	0.0303	0.522	0.240	0.225	0.252	0.403
	CGA	0.781	0.368	0.777	0.262	1.06	0.378	0.480	0.0290	0	0.276
	CGG	0.260	0.443	0.265	0.786	0.317	0.520	0.495	0.792	0.512	0.399
	AGA	**2.32**	**3.11**	**2.76**	**3.27**	**2.84**	**2.48**	**2.42**	**3.62**	**2.76**	**2.79**
	AGG	1.86	1.26	1.42	0.893	1.00	1.61	1.89	0.807	1.98	1.77

Gly	GGT	1.07	0.756	0.886	0.763	0.838	0.894	0.699	0.940	0.556	0.432
	GGC	0.401	0.716	0.548	0.452	0.603	0.773	0.949	0.529	0.556	0.934
	GGA	1.15	1.04	1.27	1.02	**1.38**	**1.20**	0.975	1.20	**1.66**	**1.62**
	GGG	**1.38**	**1.49**	**1.30**	**1.76**	1.18	1.13	**1.38**	**1.33**	0.966	1.01

Table 7: Mean RSCU values across different clades. Bold indicates the most preferred codon for each amino acid and rabies clade.

Bibliography

Afrasiabi, Ali et al. (Feb. 2022). ‘The low abundance of CpG in the SARS-CoV-2 genome is not an evolutionarily signature of ZAP’. en. In: *Scientific Reports* 12.1, p. 2420. ISSN: 2045-2322. DOI: 10.1038/s41598-022-06046-5. URL: https://www.nature.com/articles/s41598-022-06046-5 (visited on 08/04/2024).

Aguilar-Setien, A. et al. (June 2005). ‘Salivary excretion of rabies virus by healthy vampire bats.’ en. In: *Epidemiology and Infection* 133.3, p. 517. DOI: 10.1017/s0950268805003705. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2870282/ (visited on 08/04/2024).

Aiewsakun, Pakorn and Aris Katzourakis (June 2015). ‘Time dependency of foamy virus evolutionary rate estimates’. In: *BMC Evolutionary Biology* 15.1, p. 119. ISSN: 1471-2148. DOI: 10.1186/s12862-015-0408-z. URL: https://doi.org/10.1186/s12862-015-0408-z (visited on 02/10/2024).

– (July 2016). ‘Time-Dependent Rate Phenomenon in Viruses’. In: *Journal of Virology* 90.16, pp. 7184–7195. DOI: 10.1128/jvi.00593-16. URL: https://journals.asm.org/doi/full/10.1128/jvi.00593-16 (visited on 05/06/2024).

Alexaki, Aikaterini et al. (June 2019). ‘Codon and Codon-Pair Usage Tables (CoCoPUTs): Facilitating Genetic Variation Analyses and Recombinant Gene Design’. In: *Journal of Molecular Biology*. Computation Resources for Molecular Biology 431.13, pp. 2434–2441. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2019.04.021. URL: https://www.sciencedirect.com/science/article/pii/S0022283619302281 (visited on 11/01/2024).

Andersen, Kristian et al. (Jan. 2020). *nCoV-2019 codon usage and reservoir (not snakes v2) - SARS-CoV-2 coronavirus / nCoV-2019 Evolutionary History*. en. URL: https://virological.org/t/ncov-2019-codon-usage-and-reservoir-not-snakes-v2/339 (visited on 12/01/2024).

Anderson, Roy M. and Robert M. May (Feb. 1982). 'Directly Transmitted Infections Diseases: Control by Vaccination'. In: *Science* 215.4536, pp. 1053–1060. DOI: 10.1126/science.7063839. URL: https://www.science.org/doi/abs/10.1126/science.7063839 (visited on 30/08/2024).

Babayan, Simon A., Richard J. Orton and Daniel G. Streicker (Nov. 2018). 'Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes'. In: *Science* 362.6414, pp. 577–580. DOI: 10.1126/science.aap9072. URL: https://www.science.org/doi/full/10.1126/science.aap9072 (visited on 07/12/2023).

Badrane, Hassan and Noël Tordo (Sept. 2001). 'Host Switching in Lyssavirus History from the Chiroptera to the Carnivora Orders'. In: *Journal of Virology* 75.17, pp. 8096–8104. DOI: 10.1128/jvi.75.17.8096-8104.2001. URL: https://journals.asm.org/doi/full/10.1128/jvi.75.17.8096-8104.2001 (visited on 31/01/2024).

Bahir, Iris et al. (Jan. 2009). 'Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences'. In: *Molecular Systems Biology* 5.1, p. 311. ISSN: 1744-4292. DOI: 10.1038/msb.2009.71. URL: https://www.embopress.org/doi/full/10.1038/msb.2009.71 (visited on 11/08/2023).

Baker, Steven F, Aitor Nogales and Luis Martínez-Sobrido (June 2015). 'Downregulating Viral Gene Expression: Codon Usage Bias Manipulation for the Generation of Novel Influenza A Virus Vaccines'. In: ISSN: 1746-0794.

Bastide, Paul et al. (July 2018). 'Inference of Adaptive Shifts for Multivariate Correlated Traits'. In: *Systematic Biology* 67.4, pp. 662–680. ISSN: 1063-5157. DOI: 10.1093/sysbio/syy005. URL: https://doi.org/10.1093/sysbio/syy005 (visited on 23/11/2023).

Bautista, Criselda et al. (Aug. 2023). 'Whole Genome Sequencing for Rapid Characterization of Rabies Virus Using Nanopore Technology'. en. In: *JoVE (Journal of Visualized Experiments)* 198, e65414. ISSN: 1940-087X. DOI: 10.3791/65414. URL: https://www.jove.com/v/65414/author-spotlight-cost-effective-genomic-workflow-for-advancing-rabies (visited on 23/11/2023).

Belshaw, Robert et al. (Apr. 2008). 'Pacing a small cage: mutation and RNA viruses'. en. In: *Trends in Ecology & Evolution* 23.4, pp. 188–193. ISSN: 0169-5347. DOI: 10.1016/j.tree.2007.11.010. URL: https://www.sciencedirect.com/science/article/pii/S0169534708000554 (visited on 13/10/2022).

Biek, Roman et al. (May 2007). 'A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus'. en. In: *Proceedings of the National Academy of Sciences* 104.19, pp. 7993–7998. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0700741104. URL: https://www.pnas.org/content/104/19/7993 (visited on 06/04/2021).

Biek, Roman et al. (June 2015). 'Measurably evolving pathogens in the genomic era'. English. In: *Trends in Ecology & Evolution* 30.6, pp. 306–313. ISSN: 0169-5347. DOI: 10.1016/j.tree.2015.03.009. URL: https://www.cell.com/trends/ecology-evolution/abstract/S0169-5347(15)00068-3 (visited on 15/11/2021).

Boland, Torrey A. et al. (2014). 'Phylogenetic and epidemiologic evidence of multiyear incubation in human rabies'. en. In: *Annals of Neurology* 75.1, pp. 155–160. ISSN: 1531-8249. DOI: 10.1002/ana.24016. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.24016 (visited on 31/10/2022).

Braun, Katarina et al. (Apr. 2021). *Limited within-host diversity and tight transmission bottlenecks limit SARS-CoV-2 evolution in acutely infected individuals*. en. DOI: 10.1101/2021.04.30.440988. URL: https://www.biorxiv.org/content/10.1101/2021.04.30.440988v1 (visited on 09/02/2023).

Brierley, Liam and Anna Fowler (Apr. 2021). 'Predicting the animal hosts of coronaviruses from compositional biases of spike protein and whole genome sequences through machine learning'. en. In: *PLOS Pathogens* 17.4, e1009149. ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1009149. URL: https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1009149 (visited on 02/08/2023).

Broadbent, Andrew J. et al. (Jan. 2016). 'Evaluation of the attenuation, immunogenicity, and efficacy of a live virus vaccine generated by codon-pair bias de-optimization of the 2009 pandemic H1N1 influenza virus, in ferrets'. In: *Vaccine* 34.4, pp. 563–570. ISSN: 0264-410X. DOI: `10.1016/j.vaccine.2015.11.054`. URL: `https://www.sciencedirect.com/science/article/pii/S0264410X15017041` (visited on 16/01/2025).

Brunker, Kirstyn et al. (Mar. 2015). 'Elucidating the phylodynamics of endemic rabies virus in eastern Africa using whole-genome sequencing'. In: *Virus Evolution* 1.vev011. ISSN: 2057-1577. DOI: `10.1093/ve/vev011`. URL: `https://doi.org/10.1093/ve/vev011` (visited on 20/04/2021).

Brunker, Kirstyn et al. (May 2020). 'Rapid in-country sequencing of whole virus genomes to inform rabies elimination programmes'. en. In: *Wellcome Open Research* 5, p. 3. ISSN: 2398-502X. DOI: `10.12688/wellcomeopenres.15518.2`. URL: `https://wellcomeopenresearch.org/articles/5-3/v2` (visited on 24/02/2021).

Buonagurio, Deborah A. et al. (May 1986). 'Evolution of Human Influenza A Viruses Over 50 Years: Rapid, Uniform Rate of Change in NS Gene'. In: *Science* 232.4753, pp. 980–982. DOI: `10.1126/science.2939560`. URL: `https://www.science.org/doi/10.1126/science.2939560` (visited on 14/03/2024).

Campbell, Kathryn et al. (May 2022). 'Making genomic surveillance deliver: A lineage classification and nomenclature system to inform rabies elimination'. en. In: *PLOS Pathogens* 18.5, e1010023. ISSN: 1553-7374. DOI: `10.1371/journal.ppat.1010023`. URL: `https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1010023` (visited on 06/05/2022).

Caraballo, Diego A. et al. (Dec. 2021). 'A Novel Terrestrial Rabies Virus Lineage Occurring in South America: Origin, Diversification, and Evidence of Contact between Wild and Domestic Cycles'. en. In: *Viruses* 13.12, p. 2484. ISSN: 1999-4915. DOI: `10.3390/v13122484`. URL: `https://www.mdpi.com/1999-4915/13/12/2484` (visited on 04/07/2023).

Charlton, K. M. and G. A. Casey (July 1979). 'Experimental rabies in skunks: immunofluorescence light and electron microscopic studies'. eng. In: *Laboratory investigation; a journal of technical methods and pathology* 41.1, pp. 36–44. ISSN: 1530-0307.

Charlton, K. M. et al. (June 1997). 'The long incubation period in rabies: delayed progression of infection in muscle at the site of exposure'. en. In: *Acta Neuropathologica* 94.1, pp. 73–77. ISSN: 1432-0533. DOI: 10.1007/s004010050674. URL: https://doi.org/10.1007/s004010050674 (visited on 17/11/2021).

Cheng, Benson Yee Hin et al. (Mar. 2015). 'Development of Live-Attenuated Arenavirus Vaccines Based on Codon Deoptimization'. In: *Journal of Virology* 89.7, pp. 3523–3533. DOI: 10.1128/jvi.03401-14. URL: https://journals.asm.org/doi/full/10.1128/jvi.03401-14 (visited on 16/01/2025).

Choi, Bina et al. (Dec. 2020). 'Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host'. In: *New England Journal of Medicine* 383.23, pp. 2291–2293. ISSN: 0028-4793. DOI: 10.1056/NEJMc2031364. URL: https://www.nejm.org/doi/10.1056/NEJMc2031364 (visited on 20/04/2023).

Choi, Sunhwa and Moran Ki (Mar. 2020). 'Estimating the reproductive number and the outbreak size of COVID-19 in Korea'. In: *Epidemiology and Health* 42, e2020011. ISSN: 2092-7193. DOI: 10.4178/epih.e2020011. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7285447/ (visited on 12/12/2023).

Cleaveland, S. et al. (Aug. 2018). 'Proof of concept of mass dog vaccination for the control and elimination of canine rabies'. In: *Revue scientifique et technique (International Office of Epizootics)* 37.2, pp. 559–568. ISSN: 0253-1933. DOI: 10.20506/rst.37.2.2824. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7612386/ (visited on 11/09/2024).

Cori, Anne et al. (Dec. 2018). 'A graph-based evidence synthesis approach to detecting outbreak clusters: An application to dog rabies'. en. In: *PLOS Computational Biology* 14.12, e1006554. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006554. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006554 (visited on 19/04/2021).

Cribari-Neto, Francisco and Achim Zeileis (Apr. 2010). 'Beta Regression in R'. en. In: *Journal of Statistical Software* 34.1, pp. 1–24. ISSN: 1548-7660. DOI: 10.18637/jss.v034.i02. URL: https://www.jstatsoft.org/index.php/jss/article/view/v034i02 (visited on 25/09/2020).

Dimaano, Efren M. et al. (July 2011). 'Clinical and epidemiological features of human rabies cases in the Philippines: a review from 1987 to 2006'. en. In: *International Journal of Infectious Diseases* 15.7, e495–e499. ISSN: 1201-9712. DOI: 10.1016/j.ijid.2011.03.023. URL: https://www.sciencedirect.com/science/article/pii/S1201971211000889 (visited on 13/04/2023).

Ding, Nai-Zheng et al. (Mar. 2017). 'A permanent host shift of rabies virus from Chiroptera to Carnivora associated with recombination'. en. In: *Scientific Reports* 7.1, p. 289. ISSN: 2045-2322. DOI: 10.1038/s41598-017-00395-2. URL: https://www.nature.com/articles/s41598-017-00395-2 (visited on 31/01/2024).

Drummond, A. J. et al. (May 2005). 'Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences'. In: *Molecular Biology and Evolution* 22.5, pp. 1185–1192. ISSN: 0737-4038. DOI: 10.1093/molbev/msi103. URL: https://doi.org/10.1093/molbev/msi103 (visited on 02/05/2024).

Drummond, Alexei, Pybus Oliver G. and Andrew Rambaut (Jan. 2003a). 'Inference of Viral Evolutionary Rates from Molecular Sequences'. en. In: *Advances in Parasitology* 54, pp. 331–358. URL: https://www.sciencedirect.com/science/article/pii/S0065308X03540088 (visited on 07/02/2023).

Drummond, Alexei J. and Andrew Rambaut (Nov. 2007). 'BEAST: Bayesian evolutionary analysis by sampling trees'. en. In: *BMC Evolutionary Biology* 7.1, p. 214. ISSN: 1471-2148. DOI: 10.1186/1471-2148-7-214. URL: https://doi.org/10.1186/1471-2148-7-214 (visited on 16/01/2025).

Drummond, Alexei J. et al. (Sept. 2003b). 'Measurably evolving populations'. en. In: *Trends in Ecology & Evolution* 18.9, pp. 481–488. ISSN: 0169-5347. DOI: 10.1016/S0169-5347(03)00216-7. URL: https://www.sciencedirect.com/science/article/pii/S0169534703002167 (visited on 27/06/2022).

Drummond, Alexei J. et al. (Mar. 2006). 'Relaxed Phylogenetics and Dating with Confidence'. en. In: *PLOS Biology* 4.5, e88. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0040088. URL: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0040088 (visited on 22/08/2023).

Drummond, Alexei J. et al. (Aug. 2012). 'Bayesian Phylogenetics with BEAUti and the BEAST 1.7'. In: *Molecular Biology and Evolution* 29.8, pp. 1969–1973. ISSN: 0737-4038. DOI: 10.1093/molbev/mss075. URL: https://doi.org/10.1093/molbev/mss075 (visited on 10/06/2021).

Duchêne, Sebastian et al. (Nov. 2020a). 'Bayesian Evaluation of Temporal Signal in Measurably Evolving Populations'. In: *Molecular Biology and Evolution* 37.11, pp. 3363–3379. ISSN: 0737-4038. DOI: 10.1093/molbev/msaa163. URL: https://doi.org/10.1093/molbev/msaa163 (visited on 14/03/2024).

Duchêne, Sebastian et al. (July 2020b). 'Temporal signal and the phylodynamic threshold of SARS-CoV-2'. In: *Virus Evolution* 6.2, veaa061. ISSN: 2057-1577. DOI: 10.1093/ve/veaa061. URL: https://doi.org/10.1093/ve/veaa061 (visited on 01/03/2024).

Duchêne, Sebastián and Edward C Holmes (Feb. 2018). 'Estimating evolutionary rates in giant viruses using ancient genomes'. In: *Virus Evolution* 4.1, vey006. ISSN: 2057-1577. DOI: 10.1093/ve/vey006. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5829572/ (visited on 20/02/2024).

Duchêne, Sebastián, Edward C. Holmes and Simon Y. W. Ho (July 2014). 'Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates'. In: *Proceedings of the Royal Society B: Biological Sciences* 281.1786, p. 20140732. DOI: 10.1098/rspb.2014.0732. URL: https://royalsocietypublishing.org/doi/10.1098/rspb.2014.0732 (visited on 25/04/2022).

Duchêne, Sebastián et al. (July 2015). 'The Performance of the Date-Randomization Test in Phylogenetic Analyses of Time-Structured Virus Data'. In: *Molecular Biology and Evolution* 32.7, pp. 1895–1906. ISSN: 0737-4038. DOI: 10.1093/molbev/msv056. URL: https://doi.org/10.1093/molbev/msv056 (visited on 14/03/2024).

Durrant, Rowan et al. (Nov. 2024). 'Examining the molecular clock hypothesis for the contemporary evolution of the rabies virus'. In: *PLOS Pathogens* 20.11, e1012740. ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1012740. URL: https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1012740 (visited on 19/12/2024).

Faria, Nuno R. et al. (Oct. 2014). 'The early spread and epidemic ignition of HIV-1 in human populations'. In: *Science (New York, N.Y.)* 346.6205, pp. 56–61. ISSN: 0036-8075. DOI: 10.1126/science.1256739. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4254776/ (visited on 27/05/2024).

Farris, James S. (Mar. 1970). 'Methods for Computing Wagner Trees'. In: *Systematic Biology* 19.1, pp. 83–92. ISSN: 1063-5157. DOI: 10.1093/sysbio/19.1.83. URL: https://doi.org/10.1093/sysbio/19.1.83 (visited on 10/01/2025).

Faye, Martin et al. (2022). 'Rabies surveillance in Senegal 2001 to 2015 uncovers first infection of a honey-badger'. en. In: *Transboundary and Emerging Diseases* 69.5, e1350–e1364. ISSN: 1865-1682. DOI: 10.1111/tbed.14465. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/tbed.14465 (visited on 04/07/2023).

Featherstone, Leo A. et al. (July 2023). *Clockor2: Inferring global and local strict molecular clocks using root-to-tip regression.* en. DOI: 10.1101/2023.07.13.548947. URL: https://www.biorxiv.org/content/10.1101/2023.07.13.548947v1 (visited on 17/08/2023).

Fekadu, Makonnen, John H. Shaddock and George M. Baer (May 1982). 'Excretion of Rabies Virus in the Saliva of Dogs'. In: *The Journal of Infectious Diseases* 145.5, pp. 715–719. ISSN: 0022-1899. DOI: 10.1093/infdis/145.2.715. URL: https://doi.org/10.1093/infdis/145.2.715 (visited on 08/01/2025).

Felsenstein, Joseph (Nov. 1981). 'Evolutionary trees from DNA sequences: A maximum likelihood approach'. en. In: *Journal of Molecular Evolution* 17.6, pp. 368–376. ISSN: 1432-1432. DOI: 10.1007/BF01734359. URL: https://doi.org/10.1007/BF01734359 (visited on 10/01/2025).

Fraser, Christophe et al. (June 2009). 'Pandemic Potential of a Strain of Influenza A (H1N1): Early Findings'. In: *Science* 324.5934, pp. 1557–1561. DOI: 10.1126/science.1176062. URL: https://www.science.org/doi/full/10.1126/science.1176062 (visited on 10/06/2024).

Fusaro, Alice et al. (July 2013). 'The introduction of fox rabies into Italy (2008–2011) was due to two viral genetic groups with distinct phylogeographic patterns'. en. In: *Infection, Genetics and Evolution* 17, pp. 202–209. ISSN: 1567-1348. DOI: 10.1016/j.meegid.2013.03.051. URL: https://www.sciencedirect.com/science/article/pii/S1567134813001512 (visited on 04/07/2023).

Gardiner-Garden, M. and M. Frommer (July 1987). 'CpG Islands in vertebrate genomes'. en. In: *Journal of Molecular Biology* 196.2, pp. 261–282. ISSN: 0022-2836. DOI: 10.1016/0022-2836(87)90689-9. URL: https://www.sciencedirect.com/science/article/pii/0022283687906899 (visited on 11/08/2023).

Gardy, Jennifer L. and Nicholas J. Loman (Jan. 2018). 'Towards a genomics-informed, real-time, global pathogen surveillance system'. en. In: *Nature Reviews Genetics* 19.1, pp. 9–20. ISSN: 1471-0064. DOI: 10.1038/nrg.2017.88. URL: https://www.nature.com/articles/nrg.2017.88 (visited on 20/08/2024).

Ghafari, Mahan et al. (Nov. 2021). 'A mechanistic evolutionary model explains the time-dependent pattern of substitution rates in viruses'. In: *Current Biology* 31.21, 4689–4696.e5. ISSN: 0960-9822. DOI: 10.1016/j.cub.2021.08.020. URL: https://www.sciencedirect.com/science/article/pii/S0960982221011246 (visited on 15/03/2024).

Ghafari, Mahan et al. (Jan. 2022). 'Purifying Selection Determines the Short-Term Time Dependency of Evolutionary Rates in SARS-CoV-2 and pH1N1 Influenza'. In: *Molecular Biology and Evolution* 39.2, msac009. ISSN: 0737-4038. DOI: 10.1093/molbev/msac009. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8826518/ (visited on 19/08/2024).

Ghosh, J. B. et al. (2009). 'Acute flaccid paralysis due to rabies'. In: *Journal of Pediatric Neurosciences* 4.1, pp. 33–35. ISSN: 1817-1745. DOI: 10.4103/1817-1745.49106. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3162835/ (visited on 04/09/2024).

Gibson, A. D. et al. (May 2022). 'Elimination of human rabies in Goa, India through an integrated One Health approach'. en. In: *Nature Communications* 13.1, p. 2788. ISSN: 2041-1723. DOI: 10.1038/s41467-022-30371-y. URL: https://www.nature.com/articles/s41467-022-30371-y (visited on 12/07/2024).

Gigante, Crystal M. et al. (Nov. 2020). ‘Portable Rabies Virus Sequencing in Canine Rabies Endemic Countries Using the Oxford Nanopore MinION’. en. In: *Viruses* 12.11, p. 1255. ISSN: 1999-4915. DOI: 10.3390/v12111255. URL: https://www.mdpi.com/1999-4915/12/11/1255 (visited on 01/08/2024).

Gill, Mandev S. et al. (Mar. 2013). ‘Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci’. In: *Molecular Biology and Evolution* 30.3, pp. 713–724. ISSN: 0737-4038. DOI: 10.1093/molbev/mss265. URL: https://doi.org/10.1093/molbev/mss265 (visited on 27/05/2024).

Gire, Stephen K. et al. (Sept. 2014). ‘Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak’. In: *Science* 345.6202, pp. 1369–1372. DOI: 10.1126/science.1259657. URL: https://www.science.org/doi/full/10.1126/science.1259657 (visited on 27/06/2022).

Gojobori, T, E N Moriyama and M Kimura (Dec. 1990). ‘Molecular clock of viral evolution, and the neutral theory.’ In: *Proceedings of the National Academy of Sciences* 87.24, pp. 10015–10018. DOI: 10.1073/pnas.87.24.10015. URL: https://www.pnas.org/doi/abs/10.1073/pnas.87.24.10015 (visited on 07/02/2023).

Goonawardane, Niluka, Dung Nguyen and Peter Simmonds (Jan. 2021). ‘Association of Zinc Finger Antiviral Protein Binding to Viral Genomic RNA with Attenuation of Replication of Echovirus 7’. In: *mSphere* 6.1, 10.1128/msphere.01138–20. DOI: 10.1128/msphere.01138-20. URL: https://journals.asm.org/doi/10.1128/msphere.01138-20 (visited on 14/01/2025).

Gorman, O T et al. (Oct. 1990). ‘Evolution of influenza A virus PB2 genes: implications for evolution of the ribonucleoprotein complex and origin of human influenza A virus’. In: *Journal of Virology* 64.10, pp. 4893–4902. DOI: 10.1128/jvi.64.10.4893-4902.1990. URL: https://journals.asm.org/doi/abs/10.1128/jvi.64.10.4893-4902.1990 (visited on 09/01/2025).

Greenbaum, Benjamin D., Raul Rabadan and Arnold J. Levine (June 2009). ‘Patterns of Oligonucleotide Sequences in Viral and Host Cell RNA Identify Mediators of the Host Innate Immune System’. en. In: *PLOS ONE* 4.6, e5969. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0005969. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0005969 (visited on 06/01/2025).

Grenfell, Bryan T. et al. (Jan. 2004). ‘Unifying the Epidemiological and Evolutionary Dynamics of Pathogens’. en. In: *Science* 303.5656, pp. 327–332. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1090727. URL: https://science.sciencemag.org/content/303/5656/327 (visited on 19/04/2021).

Gumpper, Ryan H., Weike Li and Ming Luo (Feb. 2019). ‘Constraints of Viral RNA Synthesis on Codon Usage of Negative-Strand RNA Virus’. In: *Journal of Virology* 93.5, e01775–18. ISSN: 0022-538X. DOI: 10.1128/JVI.01775-18. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6384081/ (visited on 22/01/2025).

Gérardin, Patrick et al. (July 2008). ‘Estimating Chikungunya prevalence in La Réunion Island outbreak by serosurveys: Two methods for two critical times of the epidemic’. en. In: *BMC Infectious Diseases* 8.1, p. 99. ISSN: 1471-2334. DOI: 10.1186/1471-2334-8-99. URL: https://doi.org/10.1186/1471-2334-8-99 (visited on 10/06/2024).

Haas, Jürgen, Eun-Chung Park and Brian Seed (Mar. 1996). ‘Codon usage limitation in the expression of HIV-1 envelope glycoprotein’. English. In: *Current Biology* 6.3, pp. 315–324. ISSN: 0960-9822. DOI: 10.1016/S0960-9822(02)00482-7. URL: https://www.cell.com/current-biology/abstract/S0960-9822(02)00482-7 (visited on 19/12/2024).

Hampson, Katie et al. (Mar. 2009). ‘Transmission Dynamics and Prospects for the Elimination of Canine Rabies’. en. In: *PLOS Biology* 7.3, e1000053. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1000053. URL: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1000053 (visited on 24/02/2021).

Hampson, Katie et al. (Apr. 2015). ‘Estimating the Global Burden of Endemic Canine Rabies’. en. In: *PLOS Neglected Tropical Diseases* 9.4, e0003709. ISSN: 1935-2735. DOI: 10.1371/journal.pntd.0003709. URL: https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0003709 (visited on 16/04/2021).

Hampson, Katie et al. (Dec. 2016). *Surveillance to Establish Elimination of Transmission and Freedom from Dog-mediated Rabies*. en. DOI: 10.1101/096883. URL: https://www.biorxiv.org/content/10.1101/096883v1 (visited on 11/09/2024).

Hanlon, Cathleen A., Michael Niezgoda and Charles E. Rupprecht (Jan. 2003). '5 - Animal Rabies'. In: *Rabies*. Ed. by Alan C. Jackson and William H. Wunner. San Diego: Academic Press, pp. 163–218. ISBN: 978-0-12-379077-4. URL: https://www.sciencedirect.com/science/article/pii/B9780123790774500079 (visited on 18/09/2024).

Hasegawa, Masami, Hirohisa Kishino and Taka-aki Yano (Oct. 1985). 'Dating of the human-ape splitting by a molecular clock of mitochondrial DNA'. en. In: *Journal of Molecular Evolution* 22.2, pp. 160–174. ISSN: 1432-1432. DOI: 10.1007/BF02101694. URL: https://doi.org/10.1007/BF02101694 (visited on 10/01/2025).

Hayes, Sarah et al. (Dec. 2022). 'Understanding the incidence and timing of rabies cases in domestic animals and wildlife in south-east Tanzania in the presence of widespread domestic dog vaccination campaigns'. In: *Veterinary Research* 53.1, p. 106. ISSN: 1297-9716. DOI: 10.1186/s13567-022-01121-1. URL: https://doi.org/10.1186/s13567-022-01121-1 (visited on 15/08/2023).

Hayman, David T. S. et al. (Dec. 2016). 'The Global Phylogeography of Lyssaviruses - Challenging the 'Out of Africa' Hypothesis'. en. In: *PLOS Neglected Tropical Diseases* 10.12, e0005266. ISSN: 1935-2735. DOI: 10.1371/journal.pntd.0005266. URL: https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0005266 (visited on 13/09/2024).

He, Wanting et al. (Oct. 2017). 'Codon usage bias in the N gene of rabies virus'. en. In: *Infection, Genetics and Evolution* 54, pp. 458–465. ISSN: 1567-1348. DOI: 10.1016/j.meegid.2017.08.012. URL: https://www.sciencedirect.com/science/article/pii/S1567134817302691 (visited on 02/08/2023).

Hershberg, Ruth and Dmitri A. Petrov (2008). 'Selection on Codon Bias'. In: *Annual Review of Genetics* 42.1, pp. 287–299. DOI: 10.1146/annurev.genet.42.110807.091442. URL: https://doi.org/10.1146/annurev.genet.42.110807.091442 (visited on 07/12/2023).

Hill Jr., Richard E. et al. (July 1993). 'Further Studies on the Susceptibility of Raccoons (Procyon lotor) to a Rabies Virus of Skunk Origin and Comparative Susceptibility of Striped Skunks (Mephitis mephitis)'. In: *Journal of Wildlife Diseases* 29.3, pp. 475–477. ISSN: 0090-3558. DOI: 10.7589/0090-3558-29.3.475. URL: https://doi.org/10.7589/0090-3558-29.3.475 (visited on 24/08/2023).

Hill, Verity and Guy Baele (Nov. 2019). 'Bayesian Estimation of Past Population Dynamics in BEAST 1.10 Using the Skygrid Coalescent Model'. In: *Molecular Biology and Evolution* 36.11, pp. 2620–2628. ISSN: 0737-4038. DOI: 10.1093/molbev/msz172. URL: https://doi.org/10.1093/molbev/msz172 (visited on 27/05/2024).

Hill, Verity et al. (June 2023). 'Toward a global virus genomic surveillance network'. English. In: *Cell Host & Microbe* 31.6, pp. 861–873. ISSN: 1931-3128. DOI: 10.1016/j.chom.2023.03.003. URL: https://www.cell.com/cell-host-microbe/abstract/S1931-3128(23)00107-5 (visited on 31/07/2024).

Ho, Simon Y. W. and Sebastián Duchêne (2014). 'Molecular-clock methods for estimating evolutionary rates and timescales'. en. In: *Molecular Ecology* 23.24, pp. 5947–5965. ISSN: 1365-294X. DOI: 10.1111/mec.12953. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.12953 (visited on 14/03/2024).

Holmes, Edward C. (Apr. 2003). 'Molecular Clocks and the Puzzle of RNA Virus Origins'. In: *Journal of Virology* 77.7, pp. 3893–3897. DOI: 10.1128/JVI.77.7.3893-3897.2003. URL: https://journals.asm.org/doi/full/10.1128/JVI.77.7.3893-3897.2003 (visited on 31/01/2023).

Holmes, Edward C. et al. (Jan. 1995). 'The Molecular Epidemiology Of Human Immunodeficiency Virus Type 1 In Edinburgh'. In: *The Journal of Infectious Diseases* 171.1, pp. 45–53. ISSN: 0022-1899. DOI: 10.1093/infdis/171.1.45. URL: https://doi.org/10.1093/infdis/171.1.45 (visited on 04/09/2024).

Holmes, Edward C. et al. (Jan. 2002). 'Genetic Constraints and the Adaptive Evolution of Rabies Virus in Nature'. en. In: *Virology* 292.2, pp. 247–257. ISSN: 0042-6822. DOI: 10.1006/viro.2001.1271. URL: https://www.sciencedirect.com/science/article/pii/S0042682201912711 (visited on 17/11/2021).

Holtz, Andrew et al. (July 2023). 'Integrating full and partial genome sequences to decipher the global spread of canine rabies virus'. en. In: *Nature Communications* 14.1, p. 4247. ISSN: 2041-1723. DOI: 10.1038/s41467-023-39847-x. URL: https://www.nature.com/articles/s41467-023-39847-x (visited on 02/08/2023).

Huelsenbeck, John P. and Fredrik Ronquist (Aug. 2001). 'MRBAYES: Bayesian inference of phylogenetic trees'. In: *Bioinformatics* 17.8, pp. 754–755. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/17.8.754. URL: https://doi.org/10.1093/bioinformatics/17.8.754 (visited on 15/01/2025).

Hughes, Austin L. and Mary Ann K. Hughes (Dec. 2007). 'More Effective Purifying Selection on RNA Viruses than in DNA Viruses'. In: *Gene* 404.1-2, pp. 117–125. ISSN: 0378-1119. DOI: 10.1016/j.gene.2007.09.013. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2756238/ (visited on 30/05/2024).

Jackson, Alan C (Feb. 2011). 'Update on rabies'. In: *Research and Reports in Tropical Medicine* 2, pp. 31–43. ISSN: null. DOI: 10.2147/RRTM.S16013. URL: https://www.tandfonline.com/doi/abs/10.2147/RRTM.S16013 (visited on 13/12/2023).

Jackson, Alan C. (July 2013). 'Current and future approaches to the therapy of human rabies'. In: *Antiviral Research* 99.1, pp. 61–67. ISSN: 0166-3542. DOI: 10.1016/j.antiviral.2013.01.003. URL: https://www.sciencedirect.com/science/article/pii/S0166354213000181 (visited on 08/01/2025).

Jackson, Brendan R. et al. (Aug. 2016). 'Implementation of Nationwide Real-time Whole-genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation'. In: *Clinical Infectious Diseases* 63.3, pp. 380–386. ISSN: 1058-4838. DOI: 10.1093/cid/ciw242. URL: https://doi.org/10.1093/cid/ciw242 (visited on 04/09/2024).

Jaswant, Gurdeep et al. (May 2024). 'Viral sequencing to inform the global elimination of dog-mediated rabies - a systematic review'. en. In: *One Health & Implementation Research* 4.2, pp. 15–37. ISSN: ISSN 2769-6413 (Online). DOI: 10.20517/ohir.2023.61. URL: https://www.oaepublish.com/articles/ohir.2023.61 (visited on 10/09/2024).

Jenkins, Gareth M and Edward C Holmes (Mar. 2003). 'The extent of codon usage bias in human RNA viruses and its evolutionary origin'. en. In: *Virus Research* 92.1, pp. 1–7. ISSN: 0168-1702. DOI: 10.1016/S0168-1702(02)00309-X. URL: https://www.sciencedirect.com/science/article/pii/S016817020200309X (visited on 02/08/2023).

Jenkins, Gareth M. et al. (Feb. 2002). 'Rates of Molecular Evolution in RNA Viruses: A Quantitative Phylogenetic Analysis'. en. In: *Journal of Molecular Evolution* 54.2, pp. 156–165. ISSN: 1432-1432. DOI: 10.1007/s00239-001-0064-3. URL: https://doi.org/10.1007/s00239-001-0064-3 (visited on 10/01/2025).

Johnson, Nicholas, Anthony Fooks and Kenneth McColl (Dec. 2008). 'Reexamination of Human Rabies Case with Long Incubation, Australia'. en. In: *Emerging Infectious Diseases* 14.12, p. 1950. DOI: 10.3201/eid1412.080944. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2634648/ (visited on 13/12/2023).

Jukes, Thomas H. and Charles R. Cantor (1969). 'Evolution of Protein Molecules'. In: *Mammalian Protein Metabolism.* Vol. 3, pp. 21–132.

Kamath, Pauline L. et al. (May 2016). 'Genomics reveals historic and contemporary transmission dynamics of a bacterial disease among wildlife and livestock'. en. In: *Nature Communications* 7.1, p. 11448. ISSN: 2041-1723. DOI: 10.1038/ncomms11448. URL: https://www.nature.com/articles/ncomms11448 (visited on 27/06/2022).

Karczewski, Konrad J. et al. (May 2020). 'The mutational constraint spectrum quantified from variation in 141,456 humans'. en. In: *Nature* 581.7809, pp. 434–443. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2308-7. URL: https://www.nature.com/articles/s41586-020-2308-7 (visited on 13/03/2024).

Karlin, S, W Doerfler and L R Cardon (May 1994). 'Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses?' In: *Journal of Virology* 68.5, pp. 2889–2897. DOI: 10.1128/jvi.68.5.2889-2897.1994. URL: https://journals.asm.org/doi/10.1128/jvi.68.5.2889-2897.1994 (visited on 11/08/2023).

Kemp, Steven A. et al. (Apr. 2021). 'SARS-CoV-2 evolution during treatment of chronic infection'. en. In: *Nature* 592.7853, pp. 277–282. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03291-y. URL: https://www.nature.com/articles/s41586-021-03291-y (visited on 20/04/2023).

Kimura, M. (Feb. 1968). 'Evolutionary rate at the molecular level'. eng. In: *Nature* 217.5129, pp. 624–626. ISSN: 0028-0836. DOI: 10.1038/217624a0.

Kimura, Motoo (May 1977). 'Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution'. en. In: *Nature* 267.5608, pp. 275–276. ISSN: 1476-4687. DOI: 10.1038/267275a0. URL: https://www.nature.com/articles/267275a0 (visited on 20/12/2024).

– (1979). 'The Neutral Theory of Molecular Evolution'. In: *Scientific American* 241.5, pp. 98–129. ISSN: 0036-8733. URL: https://www.jstor.org/stable/24965339 (visited on 20/12/2024).

Kinganda-Lusamaki, Eddy et al. (June 2020). *Operationalizing genomic epidemiology during the Nord-Kivu Ebola outbreak, Democratic Republic of the Congo*. en. DOI: 10.1101/2020.06.08.20125567. URL: https://www.medrxiv.org/content/10.1101/2020.06.08.20125567v1 (visited on 09/02/2023).

Ko, Ko et al. (Feb. 2022). 'Mass Screening of SARS-CoV-2 Variants using Sanger Sequencing Strategy in Hiroshima, Japan'. en. In: *Scientific Reports* 12.1, p. 2419. ISSN: 2045-2322. DOI: 10.1038/s41598-022-04952-2. URL: https://www.nature.com/articles/s41598-022-04952-2 (visited on 08/10/2024).

Konopka-Anstadt, Jennifer L. et al. (Mar. 2020). 'Development of a new oral poliovirus vaccine for the eradication end game using codon deoptimization'. en. In: *npj Vaccines* 5.1, pp. 1–9. ISSN: 2059-0105. DOI: 10.1038/s41541-020-0176-7. URL: https://www.nature.com/articles/s41541-020-0176-7 (visited on 16/01/2025).

Korber, B. et al. (June 2000). 'Timing the Ancestor of the HIV-1 Pandemic Strains'. In: *Science* 288.5472, pp. 1789–1796. DOI: 10.1126/science.288.5472.1789. URL: https://www.science.org/doi/full/10.1126/science.288.5472.1789 (visited on 28/02/2024).

Kumar, Naveen et al. (Nov. 2018). 'Evolution of Codon Usage Bias in Henipaviruses Is Governed by Natural Selection and Is Host-Specific'. en. In: *Viruses* 10.11, p. 604. ISSN: 1999-4915. DOI: `10.3390/v10110604`. URL: `https://www.mdpi.com/1999-4915/10/11/604` (visited on 07/12/2023).

Kurosawa, Aiko et al. (Mar. 2017). 'The rise and fall of rabies in Japan: A quantitative history of rabies epidemics in Osaka Prefecture, 1914–1933'. en. In: *PLOS Neglected Tropical Diseases* 11.3, e0005435. ISSN: 1935-2735. DOI: `10.1371/journal.pntd.0005435`. URL: `https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0005435` (visited on 15/08/2023).

Kuzmin, Ivan V. et al. (June 2012). 'Molecular Inferences Suggest Multiple Host Shifts of Rabies Viruses from Bats to Mesocarnivores in Arizona during 2001–2009'. en. In: *PLOS Pathogens* 8.6, e1002786. ISSN: 1553-7374. DOI: `10.1371/journal.ppat.1002786`. URL: `https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1002786` (visited on 31/01/2024).

Lakhanpal, Urmila and R C Sharma (Dec. 1985). 'An Epidemiological Study of 177 Cases of Human Rabies'. In: *International Journal of Epidemiology* 14.4, pp. 614–617. ISSN: 0300-5771. DOI: `10.1093/ije/14.4.614`. URL: `https://doi.org/10.1093/ije/14.4.614` (visited on 13/12/2023).

Layan, Maylis et al. (May 2021). 'Mathematical modelling and phylodynamics for the study of dog rabies dynamics and control: A scoping review'. en. In: *PLOS Neglected Tropical Diseases* 15.5, e0009449. ISSN: 1935-2735. DOI: `10.1371/journal.pntd.0009449`. URL: `https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0009449` (visited on 27/01/2023).

Lentz, Thomas L. et al. (Jan. 1982). 'Is the Acetylcholine Receptor a Rabies Virus Receptor?' In: *Science* 215.4529, pp. 182–184. DOI: `10.1126/science.7053569`. URL: `https://www.science.org/doi/abs/10.1126/science.7053569` (visited on 13/12/2023).

Li, Gen et al. (Oct. 2023). 'Analyzing the Evolution and Host Adaptation of the Rabies Virus from the Perspective of Codon Usage Bias'. en. In: *Transboundary and Emerging Diseases* 2023, e4667253. ISSN: 1865-1674. DOI: `10.1155/2023/4667253`. URL: `https://www.hindawi.com/journals/tbed/2023/4667253/` (visited on 04/12/2023).

Li, Michael (July 2019). 'Methods for modelling the spread of infectious disease'. en. PhD thesis. McMaster University.

Li, Michael et al. (Apr. 2024). *Reassessing global historical R0 estimates of canine rabies*. en. DOI: 10.1101/2024.04.11.589097. URL: https://www.biorxiv.org/content/10.1101/2024.04.11.589097v1 (visited on 23/05/2024).

Li, W H, M Tanimura and P M Sharp (July 1988). 'Rates and dates of divergence between AIDS virus nucleotide sequences.' In: *Molecular Biology and Evolution* 5.4, pp. 313–330. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a040503. URL: https://doi.org/10.1093/oxfordjournals.molbev.a040503 (visited on 09/01/2025).

Liu, Ye et al. (Aug. 2010). 'Ferret badger rabies origin and its revisited importance as potential source of rabies transmission in Southeast China'. In: *BMC Infectious Diseases* 10.1, p. 234. ISSN: 1471-2334. DOI: 10.1186/1471-2334-10-234. URL: https://doi.org/10.1186/1471-2334-10-234 (visited on 17/08/2023).

Lorenzo, María M. et al. (May 2022). 'Vaccinia Virus Attenuation by Codon Deoptimization of the A24R Gene for Vaccine Development'. In: *Microbiology Spectrum* 10.3, e00272–22. DOI: 10.1128/spectrum.00272-22. URL: https://journals.asm.org/doi/full/10.1128/spectrum.00272-22 (visited on 16/01/2025).

Luo, Jun et al. (Jan. 2020a). 'The Deoptimization of Rabies Virus Matrix Protein Impacts Viral Transcription and Replication'. en. In: *Viruses* 12.1, p. 4. ISSN: 1999-4915. DOI: 10.3390/v12010004. URL: https://www.mdpi.com/1999-4915/12/1/4 (visited on 02/08/2023).

Luo, Xiu et al. (Jan. 2020b). 'Molecular Mechanism of RNA Recognition by Zinc-Finger Antiviral Protein'. In: *Cell Reports* 30.1, 46–52.e4. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2019.11.116. URL: https://www.sciencedirect.com/science/article/pii/S2211124719316390 (visited on 08/04/2024).

Lushasi, Kennedy et al. (2021). 'Reservoir dynamics of rabies in south-east Tanzania and the roles of cross-species transmission and domestic dog vaccination'. en. In: *Journal of Applied Ecology* 58.11, pp. 2673–2685. ISSN: 1365-2664. DOI: 10.1111/1365-2664.13983. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1365-2664.13983 (visited on 23/06/2023).

Lushasi, Kennedy et al. (May 2023). 'Integrating contact tracing and whole-genome sequencing to track the elimination of dog-mediated rabies: an observational and genomic study'. In: *eLife* 12. Ed. by Jennifer Flegg, e85262. ISSN: 2050-084X. DOI: 10.7554/eLife.85262. URL: https://doi.org/10.7554/eLife.85262 (visited on 31/05/2023).

Lycke, E and H Tsiang (Sept. 1987). 'Rabies virus infection of cultured rat sensory neurons'. In: *Journal of Virology* 61.9, pp. 2733–2741. DOI: 10.1128/jvi.61.9.2733-2741.1987. URL: https://journals.asm.org/doi/abs/10.1128/jvi.61.9.2733-2741.1987 (visited on 10/08/2023).

Lytras, Spyros and Joseph Hughes (Apr. 2020). 'Synonymous Dinucleotide Usage: A Codon-Aware Metric for Quantifying Dinucleotide Representation in Viruses'. en. In: *Viruses* 12.4, p. 462. ISSN: 1999-4915. DOI: 10.3390/v12040462. URL: https://www.mdpi.com/1999-4915/12/4/462 (visited on 20/11/2023).

Mancy, Rebecca et al. (Apr. 2022). 'Rabies shows how scale of transmission can enable acute infections to persist at low prevalence'. In: *Science* 376.6592, pp. 512–516. DOI: 10.1126/science.abn0713. URL: https://www.science.org/doi/full/10.1126/science.abn0713 (visited on 31/01/2023).

Marston, Denise A et al. (Feb. 2018). 'The lyssavirus host-specificity conundrum — rabies virus — the exception not the rule'. In: *Current Opinion in Virology*. Emerging viruses: intraspecies transmission • Viral Immunology 28, pp. 68–73. ISSN: 1879-6257. DOI: 10.1016/j.coviro.2017.11.007. URL: https://www.sciencedirect.com/science/article/pii/S1879625717301013 (visited on 12/07/2024).

Meagher, Jennifer L. et al. (Nov. 2019). 'Structure of the zinc-finger antiviral protein in complex with RNA reveals a mechanism for selective targeting of CG-rich viral sequences'. In: *Proceedings of the National Academy of Sciences* 116.48, pp. 24303–24309. DOI: 10.1073/pnas.1913232116. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1913232116 (visited on 14/08/2023).

Meredith, Luke W et al. (Nov. 2020). 'Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study'. In: *The Lancet Infectious Diseases* 20.11, pp. 1263–1271. ISSN: 1473-3099. DOI: 10.1016/S1473-3099(20)30562-4. URL: https://www.sciencedirect.com/science/article/pii/S1473309920305624 (visited on 08/10/2024).

Minghui, Ren et al. (Aug. 2018). 'New global strategic plan to eliminate dog-mediated rabies by 2030'. In: *The Lancet Global Health* 6.8, e828–e829. ISSN: 2214-109X. DOI: 10.1016/S2214-109X(18)30302-4. URL: https://www.sciencedirect.com/science/article/pii/S2214109X18303024 (visited on 26/09/2024).

Mollentze, Nardus, Roman Biek and Daniel G Streicker (Oct. 2014). 'The role of viral evolution in rabies host shifts and emergence'. In: *Current Opinion in Virology*. Antivirals and resistance / Virus evolution 8, pp. 68–72. ISSN: 1879-6257. DOI: 10.1016/j.coviro.2014.07.004. URL: https://www.sciencedirect.com/science/article/pii/S1879625714001485 (visited on 12/07/2024).

Mordstein, Christine et al. (Sept. 2021). 'Transcription, mRNA Export, and Immune Evasion Shape the Codon Usage of Viruses'. In: *Genome Biology and Evolution* 13.9, evab106. ISSN: 1759-6653. DOI: 10.1093/gbe/evab106. URL: https://doi.org/10.1093/gbe/evab106 (visited on 11/08/2023).

Morla, Sudhir, Aditi Makhija and Sachin Kumar (June 2016). 'Synonymous codon usage pattern in glycoprotein gene of rabies virus'. en. In: *Gene* 584.1, pp. 1–6. ISSN: 0378-1119. DOI: 10.1016/j.gene.2016.02.047. URL: https://www.sciencedirect.com/science/article/pii/S0378111916301433 (visited on 02/08/2023).

Nadin-Davis, S. A., F. Muldoon and A. I. Wandeler (June 2006). 'A molecular epidemiological analysis of the incursion of the raccoon strain of rabies virus into Canada'. en. In: *Epidemiology & Infection* 134.3, pp. 534–547. ISSN: 1469-4409, 0950-2688. DOI: 10.1017/S0950268805005108. URL: https://www.cambridge.org/core/journals/epidemiology-and-infection/article/molecular-epidemiological-analysis-of-the-incursion-of-the-raccoon-strain-of-rabies-virus-into-canada/94E0465BF26C3A1912DEB07CF27E9624 (visited on 01/08/2024).

Nadin-Davis, Susan et al. (Jan. 2008a). ‘Origins of the Rabies Viruses associated with an Outbreak in Newfoundland during 2002–2003’. In: *Journal of Wildlife Diseases* 44.1, pp. 86–98. ISSN: 0090-3558. DOI: 10.7589/0090-3558-44.1.86. URL: https://doi.org/10.7589/0090-3558-44.1.86 (visited on 04/09/2024).

Nadin-Davis, Susan A. et al. (Jan. 2008b). ‘A molecular epidemiological study of rabies in Puerto Rico’. In: *Virus Research* 131.1, pp. 8–15. ISSN: 0168-1702. DOI: 10.1016/j.virusres.2007.08.002. URL: https://www.sciencedirect.com/science/article/pii/S0168170207002997 (visited on 05/08/2024).

Nadin-Davis, Susan A. et al. (Mar. 2017). ‘Application of high-throughput sequencing to whole rabies viral genome characterisation and its use for phylogenetic re-evaluation of a raccoon strain incursion into the province of Ontario’. In: *Virus Research* 232, pp. 123–133. ISSN: 0168-1702. DOI: 10.1016/j.virusres.2017.02.007. URL: https://www.sciencedirect.com/science/article/pii/S0168170216307547 (visited on 01/08/2024).

Nahata, Kanika D. et al. (Aug. 2021). ‘On the Use of Phylogeographic Inference to Infer the Dispersal History of Rabies Virus: A Review Study’. en. In: *Viruses* 13.8, p. 1628. ISSN: 1999-4915. DOI: 10.3390/v13081628. URL: https://www.mdpi.com/1999-4915/13/8/1628 (visited on 03/01/2025).

Nchioua, Rayhane et al. (Oct. 2020). ‘SARS-CoV-2 Is Restricted by Zinc Finger Antiviral Protein despite Preadaptation to the Low-CpG Environment in Humans’. In: *mBio* 11.5, 10.1128/mbio.01930–20. DOI: 10.1128/mbio.01930-20. URL: https://journals.asm.org/doi/full/10.1128/mbio.01930-20 (visited on 27/03/2024).

Nel, Louis H. (Apr. 2013). ‘Discrepancies in Data Reporting for Rabies, Africa - Volume 19, Number 4—April 2013 - Emerging Infectious Diseases journal - CDC’. en-us. In: 9.4. DOI: 10.3201/eid1904.120185. URL: https://wwwnc.cdc.gov/eid/article/19/4/12-0185_article (visited on 06/06/2024).

Nguyen, Lam-Tung et al. (Jan. 2015). ‘IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies’. In: *Molecular Biology and Evolution* 32.1, pp. 268–274. ISSN: 0737-4038. DOI: 10.1093/molbev/msu300. URL: https://doi.org/10.1093/molbev/msu300 (visited on 08/09/2023).

Paradis, Emmanuel and Klaus Schliep (Feb. 2019). 'ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R'. In: *Bioinformatics* 35.3, pp. 526–528. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty633. URL: https://doi.org/10.1093/bioinformatics/bty633 (visited on 01/11/2021).

Pouwels, Koen B. et al. (Jan. 2021). 'Community prevalence of SARS-CoV-2 in England from April to November, 2020: results from the ONS Coronavirus Infection Survey'. English. In: *The Lancet Public Health* 6.1, e30–e38. ISSN: 2468-2667. DOI: 10.1016/S2468-2667(20)30282-6. URL: https://www.thelancet.com/journals/lanpub/article/PIIS2468-2667(20)30282-6/fulltext (visited on 10/09/2024).

Puigbò, Pere, Ignacio G. Bravo and Santiago Garcia-Vallve (Sept. 2008). 'CAIcal: A combined set of tools to assess codon usage adaptation'. In: *Biology Direct* 3.1, p. 38. ISSN: 1745-6150. DOI: 10.1186/1745-6150-3-38. URL: https://doi.org/10.1186/1745-6150-3-38 (visited on 03/08/2023).

Putra, Anak Agung Gde et al. (Apr. 2013). 'Response to a Rabies Epidemic, Bali, Indonesia, 2008–2011'. In: *Emerging Infectious Diseases* 19.4, pp. 648–651. ISSN: 1080-6040. DOI: 10.3201/eid1904.120380. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3647408/ (visited on 11/09/2024).

Pybus, Oliver G. and Andrew Rambaut (Aug. 2009). 'Evolutionary analysis of the dynamics of viral infectious disease'. en. In: *Nature Reviews Genetics* 10.8, pp. 540–550. ISSN: 1471-0064. DOI: 10.1038/nrg2583. URL: https://www.nature.com/articles/nrg2583 (visited on 24/06/2022).

R Core Team (2020). *R: A Language and Environment for Statistical Computing.* Vienna, Austria. URL: https://www.R-project.org/.

Rambaut, Andrew et al. (Jan. 2016). 'Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen)'. eng. In: *Virus Evolution* 2.1, vew007. ISSN: 2057-1577. DOI: 10.1093/ve/vew007.

Rambaut, Andrew et al. (Sept. 2018). 'Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7'. In: *Systematic Biology* 67.5, pp. 901–904. ISSN: 1063-5157. DOI: 10.1093/sysbio/syy032. URL: https://doi.org/10.1093/sysbio/syy032 (visited on 21/06/2023).

Rima, Bert K. (2015). 'Nucleotide sequence conservation in paramyxoviruses; the concept of codon constellation'. In: *Journal of General Virology* 96.5, pp. 939–955. ISSN: 1465-2099. DOI: 10.1099/vir.0.070789-0. URL: https://www.microbiologyresearch.org/content/journal/jgv/10.1099/vir.0.070789-0 (visited on 06/01/2025).

Robertson, David et al. (Jan. 2020). *nCoV's relationship to bat coronaviruses & recombination signals (no snakes) - no evidence the 2019-nCoV lineage is recombinant - SARS-CoV-2 coronavirus / nCoV-2019 Evolutionary History*. en. URL: https://virological.org/t/ncovs-relationship-to-bat-coronaviruses-recombination-signals-no-snakes-no-evidence-the-2019-ncov-lineage-is-recombinant/331 (visited on 18/01/2024).

Robishaw, Janet D. et al. (Sept. 2021). 'Genomic surveillance to combat COVID-19: challenges and opportunities'. English. In: *The Lancet Microbe* 2.9, e481–e484. ISSN: 2666-5247. DOI: 10.1016/S2666-5247(21)00121-X. URL: https://www.thelancet.com/journals/lanmic/article/PIIS2666-5247(21)00121-X/fulltext (visited on 20/08/2024).

Saitou, N and M Nei (Jan. 1986). 'Polymorphism and evolution of influenza A virus genes.' In: *Molecular Biology and Evolution* 3.1, pp. 57–74. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a040381. URL: https://doi.org/10.1093/oxfordjournals.molbev.a040381 (visited on 09/01/2025).

– (July 1987). 'The neighbor-joining method: a new method for reconstructing phylogenetic trees.' In: *Molecular Biology and Evolution* 4.4, pp. 406–425. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a040454. URL: https://doi.org/10.1093/oxfordjournals.molbev.a040454 (visited on 10/01/2025).

Sanderson, MJ (Dec. 1997). 'A Nonparametric Approach to Estimating Divergence Times in the Absence of Rate Constancy'. In: *Molecular Biology and Evolution* 14.12, p. 1218. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a025731. URL: https://doi.org/10.1093/oxfordjournals.molbev.a025731 (visited on 07/01/2025).

Sanjuán, Rafael (May 2012). 'From Molecular Genetics to Phylodynamics: Evolutionary Relevance of Mutation Rates Across Viruses'. In: *PLoS Pathogens* 8.5, e1002685. ISSN: 1553-7366. DOI: 10.1371/journal.ppat.1002685. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3342999/ (visited on 20/02/2024).

Sarich, Vincent M. and Allan C. Wilson (Dec. 1967). 'Immunological Time Scale for Hominid Evolution'. In: *Science* 158.3805, pp. 1200–1203. DOI: 10.1126/science.158.3805.1200. URL: https://www.science.org/doi/10.1126/science.158.3805.1200 (visited on 09/01/2025).

Sayers, Eric W et al. (Jan. 2020). 'GenBank'. In: *Nucleic Acids Research* 48.D1, pp. D84–D86. ISSN: 0305-1048. DOI: 10.1093/nar/gkz956. URL: https://doi.org/10.1093/nar/gkz956 (visited on 15/08/2024).

Schnell, Matthias J. et al. (Jan. 2010). 'The cell biology of rabies virus: using stealth to reach the brain'. en. In: *Nature Reviews Microbiology* 8.1, pp. 51–61. ISSN: 1740-1534. DOI: 10.1038/nrmicro2260. URL: https://www.nature.com/articles/nrmicro2260 (visited on 15/02/2023).

Shackelton, Laura A., Colin R. Parrish and Edward C. Holmes (May 2006). 'Evolutionary Basis of Codon Usage and Nucleotide Composition Bias in Vertebrate DNA Viruses'. en. In: *Journal of Molecular Evolution* 62.5, pp. 551–563. ISSN: 1432-1432. DOI: 10.1007/s00239-005-0221-1. URL: https://doi.org/10.1007/s00239-005-0221-1 (visited on 11/08/2023).

Shankar, V, B Dietzschold and H Koprowski (May 1991). 'Direct entry of rabies virus into the central nervous system without prior local replication'. In: *Journal of Virology* 65.5, pp. 2736–2738. DOI: 10.1128/jvi.65.5.2736-2738.1991. URL: https://journals.asm.org/doi/abs/10.1128/jvi.65.5.2736-2738.1991 (visited on 07/02/2023).

Sharp, Colin P. et al. (May 2023). 'CpG dinucleotide enrichment in the influenza A virus genome as a live attenuated vaccine development strategy'. In: *PLOS Pathogens* 19.5, e1011357. ISSN: 1553-7366. DOI: 10.1371/journal.ppat.1011357. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10191365/ (visited on 23/08/2024).

Shaw, Andrew E. et al. (Sept. 2021). 'The antiviral state has shaped the CpG composition of the vertebrate interferome to avoid self-targeting'. en. In: *PLOS Biology* 19.9, e3001352. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.3001352. URL: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3001352 (visited on 23/11/2023).

Shin, Young C. et al. (Nov. 2015). ‘Importance of codon usage for the temporal regulation of viral gene expression’. In: *Proceedings of the National Academy of Sciences* 112.45, pp. 14030–14035. DOI: 10.1073/pnas.1515387112. URL: https://www.pnas.org/doi/full/10.1073/pnas.1515387112 (visited on 16/01/2025).

Shope, R. E. (1982). ‘Rabies-related viruses.’ In: *The Yale Journal of Biology and Medicine* 55.3-4, pp. 271–275. ISSN: 0044-0086. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2596466/ (visited on 13/12/2023).

Shutt, Deborah P. et al. (Dec. 2017). ‘Estimating the reproductive number, total outbreak size, and reporting rates for Zika epidemics in South and Central America’. en. In: *Epidemics* 21, pp. 63–79. ISSN: 1755-4365. DOI: 10.1016/j.epidem.2017.06.005. URL: https://www.sciencedirect.com/science/article/pii/S1755436517300257 (visited on 30/05/2023).

Simmonds, Peter et al. (Sept. 2013). ‘Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla –selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses’. en. In: *BMC Genomics* 14.1, p. 610. ISSN: 1471-2164. DOI: 10.1186/1471-2164-14-610. URL: https://doi.org/10.1186/1471-2164-14-610 (visited on 02/02/2024).

Slatkin, M and R R Hudson (Oct. 1991). ‘Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations.’ In: *Genetics* 129.2, pp. 555–562. ISSN: 1943-2631. DOI: 10.1093/genetics/129.2.555. URL: https://doi.org/10.1093/genetics/129.2.555 (visited on 22/08/2024).

Smith, Maureen Rebecca et al. (Oct. 2021). ‘Rapid incidence estimation from SARS-CoV-2 genomes reveals decreased case detection in Europe during summer 2020’. en. In: *Nature Communications* 12.1, p. 6009. ISSN: 2041-1723. DOI: 10.1038/s41467-021-26267-y. URL: https://www.nature.com/articles/s41467-021-26267-y (visited on 15/01/2025).

Soares, Pedro et al. (June 2009). ‘Correcting for Purifying Selection: An Improved Human Mitochondrial Molecular Clock’. English. In: *The American Journal of Human Genetics* 84.6, pp. 740–759. ISSN: 0002-9297, 1537-6605. DOI: 10.1016/j.ajhg.2009.05.001. URL: https://www.cell.com/ajhg/abstract/S0002-9297(09)00163-3 (visited on 19/08/2024).

Stadler, Tanja et al. (Jan. 2013). 'Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV)'. In: *Proceedings of the National Academy of Sciences of the United States of America* 110.1, pp. 228–233. ISSN: 0027-8424. DOI: 10.1073/pnas.1207965110. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3538216/ (visited on 02/05/2024).

Streicker, Daniel G. et al. (May 2012a). 'Rates of Viral Evolution Are Linked to Host Geography in Bat Rabies'. en. In: *PLOS Pathogens* 8.5, e1002720. ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1002720. URL: https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1002720 (visited on 23/01/2023).

Streicker, Daniel G. et al. (Nov. 2012b). 'Variable evolutionary routes to host establishment across repeated rabies virus host shifts among bats'. In: *Proceedings of the National Academy of Sciences* 109.48, pp. 19715–19720. DOI: 10.1073/pnas.1203456109. URL: https://www.pnas.org/doi/full/10.1073/pnas.1203456109 (visited on 07/08/2024).

Suchard, Marc A et al. (Jan. 2018). 'Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10'. In: *Virus Evolution* 4.1, vey016. ISSN: 2057-1577. DOI: 10.1093/ve/vey016. URL: https://doi.org/10.1093/ve/vey016 (visited on 30/05/2022).

Surve, Rohini M., Hima S. Pendharkar and Sonia Bansal (June 2021). 'Paralytic rabies mimicking Guillain-Barré syndrome: the dilemma still prevails'. en. In: *Journal of Neurocritical Care* 14.1, pp. 52–56. ISSN: 2508-1349. DOI: 10.18700/jnc.210005. URL: http://e-jnc.org/journal/view.php?doi=10.18700/jnc.210005 (visited on 11/09/2024).

Sved, J and A Bird (June 1990). 'The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model.' In: *Proceedings of the National Academy of Sciences of the United States of America* 87.12, pp. 4692–4696. ISSN: 0027-8424. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC54183/ (visited on 20/09/2024).

Swedberg, Catherine et al. (Aug. 2023). 'Using Integrated Bite Case Management to estimate the burden of rabies and evaluate surveillance in Oriental Mindoro, Philippines'. In: *One health & implementation research* 3, pp. 77–96. ISSN: 2769-6413. DOI: 10.20517/ohir.2023.02. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7615207/ (visited on 07/08/2024).

Takata, Matthew A. et al. (Oct. 2017). 'CG dinucleotide suppression enables antiviral defence targeting non-self RNA'. en. In: *Nature* 550.7674, pp. 124–127. ISSN: 1476-4687. DOI: 10.1038/nature24039. URL: https://www.nature.com/articles/nature24039 (visited on 18/01/2024).

Talbi, Chiraz et al. (Oct. 2010). 'Phylodynamics and Human-Mediated Dispersal of a Zoonotic Virus'. en. In: *PLOS Pathogens* 6.10, e1001166. ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1001166. URL: https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1001166 (visited on 03/05/2021).

Tavaré, Simon (1986). 'Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences'. In: *Lectures on Mathematics in the Life Sciences* 17. URL: https://archive.org/details/someprobabilisticandstatisticalproblemsintheanalysisofdnase
mode/2up.

Thorne, J L, H Kishino and I S Painter (Dec. 1998). 'Estimating the rate of evolution of the rate of molecular evolution.' In: *Molecular Biology and Evolution* 15.12, pp. 1647–1657. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a025892. URL: https://doi.org/10.1093/oxfordjournals.molbev.a025892 (visited on 09/01/2025).

Thoulouze, Maria-Isabel et al. (Sept. 1998). 'The Neural Cell Adhesion Molecule Is a Receptor for Rabies Virus'. In: *Journal of Virology* 72.9, pp. 7181–7190. ISSN: 0022-538X. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC109940/ (visited on 13/12/2023).

Tian, Lin et al. (Sept. 2018). 'The adaptation of codon usage of +ssRNA viruses to their hosts'. In: *Infection, Genetics and Evolution* 63, pp. 175–179. ISSN: 1567-1348. DOI: 10.1016/j.meegid.2018.05.034. URL: https://www.sciencedirect.com/science/article/pii/S1567134818303447 (visited on 07/12/2023).

Tordo, Noël et al. (Aug. 1988). 'Completion of the rabies virus genome sequence determination: Highly conserved domains among the L (polymerase) proteins of unsegmented negative-strand RNA viruses'. In: *Virology* 165.2, pp. 565–576. ISSN: 0042-6822. DOI: 10.1016/0042-6822(88)90600-9. URL: https://www.sciencedirect.com/science/article/pii/0042682288906009 (visited on 13/12/2023).

Townsend, Sunny E. et al. (Aug. 2013a). 'Designing Programs for Eliminating Canine Rabies from Islands: Bali, Indonesia as a Case Study'. en. In: *PLOS Neglected Tropical Diseases* 7.8, e2372. ISSN: 1935-2735. DOI: 10.1371/journal.pntd.0002372. URL: https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0002372 (visited on 24/02/2021).

Townsend, Sunny E. et al. (May 2013b). 'Surveillance guidelines for disease elimination: A case study of canine rabies'. en. In: *Comparative Immunology, Microbiology and Infectious Diseases* 36.3, pp. 249–261. ISSN: 01479571. DOI: 10.1016/j.cimid.2012.10.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S0147957112001221 (visited on 31/05/2023).

Troupin, Cécile et al. (Dec. 2016). 'Large-Scale Phylogenomic Analysis Reveals the Complex Evolutionary History of Rabies Virus in Multiple Carnivore Hosts'. en. In: *PLOS Pathogens* 12.12, e1006041. ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1006041. URL: https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1006041 (visited on 01/02/2023).

Tsiang, Henri (May 1979). 'Evidence for an Intraaxonal Transport of Fixed and Street Rabies Virus'. In: *Journal of Neuropathology & Experimental Neurology* 38.3, pp. 286–296. ISSN: 0022-3069. DOI: 10.1097/00005072-197905000-00008. URL: https://doi.org/10.1097/00005072-197905000-00008 (visited on 13/12/2023).

Tsiang, Henri, Pierre Emmanuel Ceccaldi and Erik Lycke (1991). 'Rabies virus infection and transport in human sensory dorsal root ganglia neurons'. In: *Journal of General Virology* 72.5, pp. 1191–1194. ISSN: 1465-2099. DOI: 10.1099/0022-1317-72-5-1191. URL: https://www.microbiologyresearch.org/content/journal/jgv/10.1099/0022-1317-72-5-1191 (visited on 21/06/2023).

Tuffereau, Christine et al. (Dec. 1998). 'Low-affinity nerve-growth factor receptor (P75NTR) can serve as a receptor for rabies virus'. In: *The EMBO Journal* 17.24, pp. 7250–7259. ISSN: 0261-4189. DOI: 10.1093/emboj/17.24.7250. URL: https://www.embopress.org/doi/full/10.1093/emboj/17.24.7250 (visited on 13/12/2023).

Van Dooren, S., M. Salemi and A.-M. Vandamme (Apr. 2001). 'Dating the Origin of the African Human T-Cell Lymphotropic Virus Type-I (HTLV-I) Subtypes'. In: *Molecular Biology and Evolution* 18.4, pp. 661–671. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a003846. URL: https://doi.org/10.1093/oxfordjournals.molbev.a003846 (visited on 31/01/2023).

Van Zyl, N., W. Markotter and L. H. Nel (June 2010). 'Evolutionary history of African mongoose rabies'. In: *Virus Research* 150.1, pp. 93–102. ISSN: 0168-1702. DOI: 10.1016/j.virusres.2010.02.018. URL: https://www.sciencedirect.com/science/article/pii/S0168170210000869 (visited on 05/08/2024).

Vega, Vinsensius B. et al. (Sept. 2004). 'Mutational dynamics of the SARS coronavirus in cell culture and human populations isolated in 2003'. eng. In: *BMC infectious diseases* 4, p. 32. ISSN: 1471-2334. DOI: 10.1186/1471-2334-4-32.

Wallace, Ryan M. et al. (Oct. 2014). 'Right Place, Wrong Species: A 20-Year Review of Rabies Virus Cross Species Transmission among Terrestrial Mammals in the United States'. en. In: *PLOS ONE* 9.10, e107539. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0107539. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0107539 (visited on 12/07/2024).

Wang, Lina et al. (Aug. 2019). 'Phylodynamic and transmission pattern of rabies virus in China and its neighboring countries'. en. In: *Archives of Virology* 164.8, pp. 2119–2129. ISSN: 1432-8798. DOI: 10.1007/s00705-019-04297-8. URL: https://doi.org/10.1007/s00705-019-04297-8 (visited on 04/07/2023).

Waples, Robin S. (July 2022). 'What Is Ne, Anyway?' eng. In: *The Journal of Heredity* 113.4, pp. 371–379. ISSN: 1465-7333. DOI: 10.1093/jhered/esac023.

World Health Organization (2018). *Rabies vaccines: WHO position paper – April 2018.* en. URL: https://www.who.int/publications/i/item/who-wer9316 (visited on 20/08/2024).

World Health Organization, Food and Agriculture Organization and World Organisation for Animal Health (2018). 'Zero by 30: the global strategic plan to end human deaths from dog-mediated rabies by 2030.' English. In: *Zero by 30: the global strategic plan to end human deaths from dog-mediated rabies by 2030.* URL: https://iris.who.int/bitstream/handle/10665/272756/9789241513838-eng.pdf (visited on 19/05/2021).

Wright, Frank (Mar. 1990). 'The 'effective number of codons' used in a gene'. en. In: *Gene* 87.1, pp. 23–29. ISSN: 0378-1119. DOI: 10.1016/0378-1119(90)90491-9. URL: https://www.sciencedirect.com/science/article/pii/0378111990904919 (visited on 02/08/2023).

Wróbel, Borys et al. (June 2006). 'Analysis of the Overdispersed Clock in the Short-Term Evolution of Hepatitis C Virus: Using the E1/E2 Gene Sequences to Infer Infection Dates in a Single Source Outbreak'. In: *Molecular Biology and Evolution* 23.6, pp. 1242–1253. ISSN: 0737-4038. DOI: 10.1093/molbev/msk012. URL: https://doi.org/10.1093/molbev/msk012 (visited on 24/06/2022).

Yamaoka, Satoko et al. (Nov. 2013). 'Involvement of the Rabies Virus Phosphoprotein Gene in Neuroinvasiveness'. In: *Journal of Virology* 87.22, pp. 12327–12338. DOI: 10.1128/JVI.02132-13. URL: https://journals.asm.org/doi/full/10.1128/JVI.02132-13 (visited on 17/11/2021).

Yu, Guangchuang et al. (2017). 'ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data'. en. In: *Methods in Ecology and Evolution* 8.1, pp. 28–36. ISSN: 2041-210X. DOI: 10.1111/2041-210X.12628. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12628 (visited on 29/01/2024).

Yu, Pengcheng et al. (Nov. 2018). 'A CpG oligodeoxynucleotide enhances the immune response to rabies vaccination in mice'. In: *Virology Journal* 15, p. 174. ISSN: 1743-422X. DOI: 10.1186/s12985-018-1089-1. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6234694/ (visited on 26/09/2024).

Yuson, Mirava et al. (Dec. 2024). 'Combining genomics and epidemiology to investigate a zoonotic outbreak of rabies in Romblon Province, Philippines'. en. In: *Nature Communications* 15.1, p. 10753. ISSN: 2041-1723. DOI: 10.1038/s41467-024-54255-5. URL: https://www.nature.com/articles/s41467-024-54255-5 (visited on 03/01/2025).

Zhang, Sheng et al. (Apr. 2020). 'Estimation of the reproductive number of novel coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A data-driven analysis'. In: *International Journal of Infectious Diseases* 93, pp. 201–204. ISSN: 1201-9712. DOI: 10.1016/j.ijid.2020.02.033. URL: https://www.sciencedirect.com/science/article/pii/S1201971220300916 (visited on 12/12/2023).

Zhang, Shoufeng et al. (May 2013). 'Epidemic and maintenance of rabies in chinese ferret badgers (Melogale moschata) indicated by epidemiology and the molecular signatures of rabies viruses'. In: *Virologica Sinica* 28.3, pp. 146–151. ISSN: 1674-0769. DOI: 10.1007/s12250-013-3316-7. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8208346/ (visited on 11/08/2023).

Zhang, Xu et al. (Aug. 2018). 'Comprehensive Analysis of Codon Usage on Rabies Virus and Other Lyssaviruses'. en. In: *International Journal of Molecular Sciences* 19.8, p. 2397. ISSN: 1422-0067. DOI: 10.3390/ijms19082397. URL: https://www.mdpi.com/1422-0067/19/8/2397 (visited on 02/08/2023).

Zhang, Yuzhen et al. (June 2017). 'Cross-border spread, lineage displacement and evolutionary rate estimation of rabies virus in Yunnan Province, China'. In: *Virology Journal* 14.1, p. 102. ISSN: 1743-422X. DOI: 10.1186/s12985-017-0769-6. URL: https://doi.org/10.1186/s12985-017-0769-6 (visited on 04/07/2023).

Zinsstag, Jakob et al. (Dec. 2017). 'Vaccination of dogs in an African city interrupts rabies transmission and reduces human exposure'. en. In: *Science Translational Medicine* 9.421. ISSN: 1946-6234, 1946-6242. DOI: 10.1126/scitranslmed.aaf6984. URL: https://stm.sciencemag.org/content/9/421/eaaf6984 (visited on 06/04/2021).

Zuckerkandl, Emile and Linus Pauling (Jan. 1965). 'Evolutionary Divergence and Convergence in Proteins'. In: *Evolving Genes and Proteins*. Ed. by Vernon Bryson and Henry J. Vogel. Academic Press, pp. 97–166. ISBN: 978-1-4832-2734-4. URL: https://www.sciencedirect.com/science/article/pii/B9781483227344500176 (visited on 20/12/2024).