Molecular signals of arms race evolution between RNA viruses and their hosts

Spyridon Lytras

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

College of Medical, Veterinary and Life Sciences
University of Glasgow

July 2023

Abstract

Viruses are intracellular parasites that hijack their hosts' cellular machinery to replicate themselves. This creates an evolutionary "arms race" between hosts and viruses, where the former develop mechanisms to restrict viral infection and the latter evolve ways to circumvent these molecular barriers. In this thesis, I explore examples of this virus-host molecular interplay, focusing on events in the evolutionary histories of both viruses and hosts. The thesis begins by examining how recombination, the exchange of genetic material between related viruses, expands the genomic diversity of the *Sarbecovirus* subgenus, which includes SARS-CoV responsible for the 2002 SARS epidemic and SARS-CoV-2 responsible for the COVID-19 pandemic. On the host side, I examine the evolutionary interaction between RNA viruses and two interferon-stimulated genes expressed in hosts. First, I show how the 2′-5′-oligoadenylate synthetase 1 (OAS1) gene of horseshoe bats (Rhinolophoidea), the reservoir host of sarbecoviruses, lost its anti-coronaviral activity at the base of this bat superfamily. By reconstructing the Rhinolophoidea common ancestor OAS1 protein, I first validate the loss of antiviral function and highlight the implications of this event in the virus-host association between sarbecoviruses and horseshoe bat hosts. Second, I focus on the evolution of the human butyrophilin subfamily 3 member A3 (BTN3A3) gene which restricts infection by avian influenza A viruses (IAV). The evolutionary analysis reveals that BTN3A3's anti-IAV function was gained within the primates and that specific amino acid substitutions need to be acquired in IAVs' NP protein to evade the human BTN3A3 activity. Gain of BTN3A3-evasion-conferring substitutions correlate with all major human IAV pandemics and epidemics, making these NP residues key markers for IAV transmissibility potential to humans. In the final part of the thesis, I present a novel approach for evaluating dinucleotide compositional biases in virus genomes. An application of my metric on the *Flaviviridae* virus family uncovers how ancestral host shifts of these viruses correlate with adaptive shifts in their genomes' dinucleotide representation. Collectively, the contents of this thesis extend our understanding of how viruses interact with their hosts along their intertangled evolution and provide insights into virus host switching and pandemic preparedness.

Contents

Tables list

Chapter 5

Appendix A

Appendix C

Figures list

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Appendix A

Appendix B

Appendix D

Acknowledgments

Doing a PhD during a pandemic has not been the easiest task, but it has been an enjoyable and fascinating journey thanks to all the people who stood by my side throughout it and supported me in every single step of the way.

I'm first grateful to my family, especially my parents and grandmother who have always been so supportive of my academic endeavours away from home.

I wouldn't be the scientist I am now without my supervisors, Joseph, Sam and David, who have always kindly given me advice on all kinds of matters, have been there whenever I needed to have a chat and have always trusted and supported my decisions. Along my PhD journey I also met many more researchers I am fortunate to call mentors and to whom I owe much of my scientific expertise, notably Sergei Pond, Darren Martin, Philippe Lemey, Rob Gifford, Massimo Palmarini, Jumpei Ito and Kei Sato.

On a more personal level, I am truly grateful to the dear friends I made since the first day I stepped foot in the CVR, Alex, Anna, Innes and Kasim with who I've experienced so much these past four years and who've always been there for me.

I've had the pleasure to work within two lab groups, the members of which embraced me and made this PhD such a great journey. I'd like to thank everyone in the Robertson lab: Kieran, Sej, Fran, Vandana, Haiting, Dan, Joe, Ben, and in the Wilson lab: Matt, Douglas, Emma, Hollie, Simon, Arthur, Elena, Ulad and Yongtao.

This acknowledgments section would be longer than the chapters of the thesis if I named everyone who I'm grateful to for making these past four years such a wonderful experience. So last but not least, I'd like to thank all the friends I have made at the CVR, chatting for hours in the level 4 office of the Stoker building, bumping into each other in the staircase or meeting up at the Phoenix bar for a drink on a Friday evening, as well as the friends outside the CVR who I met through collaborations and bonded over conference night outs.

In these four years I have learned a lot about being a scientist, but more importantly, in my opinion, I have met many amazing people I can now call my friends.

This thesis is devoted to all the friends I made along the way.

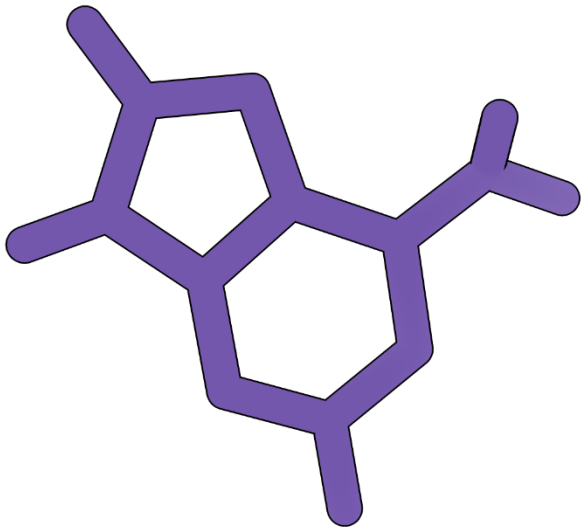
Declaration

I declare that, except where explicit reference is made to the contribution of others, 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.

Spyridon Lytras

July 2023

Chapter 1. General introduction



Cartoon model of adenine.

PDB ligand entry: ADE, visualised with ChimeraX.

“The affinities of all the beings of the same class have sometimes been represented by a great tree. I believe this simile largely speaks the truth.”

Charles Darwin, On the Origins of Species (1859)

1.1 Virus phylogenetics

Viruses are the group of the most diverse genetic entities in our planet, and although they depend on their hosts to replicate, they employ an immense range of mechanisms to interact with these hosts (Wasik and Turner, 2013; Harvey and Holmes, 2022). This unparalleled diversity bears the question of how viruses came to be, but also how they continue to evolve. Virus taxonomy is essential for understanding how these genetic entities relate to one another but also for predicting their phenotype and range of hosts they can infect. In the early years of virus taxonomy, isolates were characterised based on virion morphology, type of nucleic acid and other experimentally determined physical attributes (Simmonds *et al.*, 2023). With the rapid expansion of sequencing technologies and the ubiquitousness of viruses in essentially every biological sample, the mass of known viruses (or at least their genome sequences) has increased exponentially in recent years. This requires new dynamic systems for classifying viruses whether that is broad scale classification or characterising the evolution of circulating viruses in action (Rambaut *et al.*, 2020; Simmonds *et al.*, 2023). Key to modern virus taxonomy and classification is the inference of how different viruses relate to one another based on their genetic sequence.

A phylogeny (derived from the Greek: φῦλον [phûlon], meaning "tribe", and the suffix -γενής [-geny], meaning "producing"; Whitney and Smith, 1911) or phylogenetic tree is a structure representing how organisms or genes are related to one another. These tree-like structures have three key features: i) tips: representing known organisms or genes, ii) branches: representing evolutionary time, and iii) nodes: representing points in time when branches diverged from one another. By using techniques described later on in this chapter, one can infer the phylogeny of groups of viruses based on their genetic sequences. Exploring the evolutionary relatedness of viruses can provide unique insights into their biology and epidemiology. On the fine scale – looking at phylogenies of closely related viruses – we can observe these pathogens' evolution in action, track their spread and monitor genomic changes that may impact their transmissibility and pathogenicity, perfectly exemplified by the molecular epidemiology effort conducted for the COVID-19 pandemic (Oude Munnink, Worp, *et al.*, 2021). On the broader scale – looking at phylogenies of distantly related viruses – we can make inferences about the origins of virus groups and host switches they may have experienced in their past (Koonin, Dolja and

Krupovic, 2015). Throughout this thesis, virus phylogenetics will be used as a tool to explore virus origins, genetic diversity, and their tightly linked interactions with their hosts' evolution.

1.2 Arms-race evolution between viruses and hosts

All kingdoms of life are infected by viruses in what is a truly ancient, entwined interaction between viruses and their hosts (Koonin *et al.*, 2008; Aguado *et al.*, 2017). Throughout this evolutionary history, hosts have been continuously adapting to minimise the impact of viral infection and viruses have been adapting to persist and evade the hosts' barriers. For example, it is estimated that viruses are one of the primary causes of adaptive protein evolution in mammalian genomes (Enard *et al.*, 2016), consistent with the notion that parasitism is a key driver in hosts' adaptive evolution. As a result, both hosts and viruses are constantly engaged in an arms race between their evolutionary landscapes and trajectories, i.e. the available paths they can take in their evolution (Tenthorey, Emerman and Malik, 2022). There are three main levels at which virus-host evolution can take place in: i) Cellular entry: viruses are intracellular parasites and require entry into host cells for their replication. This is normally facilitated by biophysical interactions between viral proteins on the outside of the virion and molecules (receptors) on the outside of the host cell. Both viral entry proteins and host receptors experience adaptive changes that determine which strains of a virus group can infect specific members of a host group (Guo *et al.*, 2020; Fujita *et al.*, 2023). Receptor binding is simply an example of how the viruses initially enter the cells, but certainly the most tangible, binary interaction of cellular entry. Mode of virus transmission and physical barriers of getting to the cells can also considered as part of this first interaction level. ii) Adaptive immunity: hosts can mount antiviral responses specific to the viruses that infect them and retain the memory of infection to fend off future infection by the same virus. For example, vertebrates possess antibody-mediated adaptive immunity where diverse antibodies that target the virus antigen can be produced by B cells undergoing somatic hypermutation (Litman, Rast and Fugmann, 2010; Victora and Nussenzweig, 2012). This adaptive development of immunity against specific viral antigens can match the pace of virus evolution and many well-studied human-circulating pathogens are known to be under constant antigenic evolution as

a response to adaptive immunity (Petrova and Russell, 2017; Eguia *et al.*, 2021). iii) Innate immunity: apart from the more complex adaptive immunity, hosts possess antiviral mechanisms that restrict viral replication or transmission directly upon infection. In this case, innate immunity mechanisms cannot adapt or evolve within an individual, but instead need multiple generations and strong selection by the viral pathogens infecting a given host group (Tenthorey, Emerman and Malik, 2022). Hence, innate immunity usually acts as a barrier against viruses switching between diverse host groups.

During my PhD, I have focused on studying the arms race evolution between viruses and their hosts' innate immunity. More specific examples of such restriction factor interactions will be detailed within the following chapters, but a good illustrative example of such interactions for the purposes of this general introduction is the Apolipoprotein B editing complex 3 (APOBEC3) proteins (Sheehy *et al.*, 2002). The APOBEC3 gene group has experienced a recent expansion through multiple duplication events – primates having seven members – and encode cytidine deaminases that lead to hypermutation in viral genomes (Stavrou and Ross, 2015). APOBEC3G can restrict HIV-1 by inactivating the virus through hypermutation (Armitage *et al.*, 2012). However, HIV-1 and many related lentiviruses possess a gene encoding the virion infectivity factor (Vif) protein which directly counteracts APOBEC3G's antiviral activity (Harris *et al.*, 2003; Gifford, 2012). Deleting Vif from HIV-1 largely decreases infectivity in the presence of human APOBEC3G, although mouse APOBEC3G restricts HIV-1 regardless of Vif being present (Mariani *et al.*, 2003). This suggests a close, host group-specific interaction between Vif and APOBEC3 restriction factors. Functional analysis of the APOBEC3G genes of old world monkeys shows signatures of adaptive variation in the proteins relating to antagonism by Vif (Compton and Emerman, 2013). Furthermore, the primate paralogue of APOBEC3G, APOBEC3F, can also restrict HIV-1 but Vif also counteracts this protein through a distinct mechanism to that of APOBEC3G evasion (Russell and Pathak, 2007). All these findings illustrate an ongoing arms race between primate hosts evolving new ways to restrict retrovirus infection, and retroviruses adapting host-specific mechanisms to evade these.

APOBEC3 restriction factors are an interesting example of virus-host interactions because their hypermutation effect seems to not be completely deleterious for all viruses, as in the case of HIV-1 (Armitage *et al.*, 2012). An outbreak of MPXV

(Monkeypox virus) was detected in early May 2022 in humans, signifying a recent introduction of the virus into humans from a currently unknown animal reservoir (Kraemer *et al.*, 2022; World Health Organization, 2022b). Analysis of available genomes collected during and before the 2022 outbreak showed that the virus was likely introduced to humans around 2016, with this human-circulating clade possessing many more substitutions than would be expected from other MPXV lineages (O'Toole *et al.*, 2023). Interestingly, the pattern of "excess" mutations in this human clade is fully consistent with APOBEC3 DNA deamination in terms of the type of substitutions and context of mutated sites (Forni *et al.*, 2023; O'Toole *et al.*, 2023). Unlike HIV-1's "all or nothing" interaction with these restriction factors, it seems that MPXV – a dsDNA virus with a much larger genome – can tolerate the mutational effect of APOBEC3 deaminases, and this host-specific effect is being "imprinted" onto the virus genome. In this thesis, I will explore examples of both "all or nothing" interactions, where virus adaptations are necessary to infect hosts with unique restriction factors, and tolerant interactions, where host-specific molecules imprint unique signatures on the virus genomes.

1.3 Interferon-stimulated genes (ISG) against virus infection

The interferon (IFN) response is an evolutionarily conserved mechanism of the vertebrate innate immune response, acting as the first, post-entry, defence against intracellular parasites including bacteria, viruses and other parasites (Schneider, Chevillotte and Rice, 2014). Briefly, interferons are secreted from an infected cell to neighbouring cells leading to activation of the Janus kinase signal transducer and activator of transcription (JAK-STAT) pathway and the cells changing their transcriptional state (Stark and Darnell, 2012). The genes that are upregulated through this process are referred to as interferon-stimulated genes (ISGs). These can differ between host species, although many orthologues have conserved ISG function across mammals or even vertebrates and are referred to as 'core mammalian ISGs' and 'core vertebrate ISGs' respectively (Shaw *et al.*, 2017). ISGs can have diverse functions primarily relating to the mounted immune response. Most core ISGs linked to antiviral immunity have functions involving: i) pattern recognition such as RNA-sensing, ii) transcription factors for further gene activation and repression, iii) modulating antigen presentation, ubiquitination and protein

degradation (Shaw *et al.*, 2017). ISG functions, however, can be rather modular depending on the host species or the viruses being targeted (Schneider, Chevillotte and Rice, 2014). Different ISGs can target any part of the virus lifecycle, but three most frequently targeted points are: i) inhibiting virus entry by preventing viral components from reaching their cellular destination, e.g. the IFN-inducible transmembrane (IFITM) gene family; ii) inhibiting virus replication and translation primarily by binding to RNA and degrading it or disrupting translation initiation, e.g., the zinc-finger antiviral protein (ZAP), the IFN-induced protein with tetratricopeptide repeats (IFIT) family and the OAS-RNaseL pathway; and iii) inhibiting viral egress by interfering with virus budding (mainly for enveloped viruses), e.g., tetherin (Stark and Darnell, 2012; Schneider, Chevillotte and Rice, 2014; Shaw *et al.*, 2017). The large diversity of ISG functions is partly due to their rather loose definition as genes induced by any type of IFN molecules, as well as their elusive nature since they are transiently co-opted host genes (Schoggins, 2019). The subset of ISGs relevant to the content of this thesis, their interactions with specific viruses and relevance to virus host switching will be described in detail in the following chapters.

1.4 Zoonotic coronaviruses in humans

The coronaviruses are single-stranded, positive-sense RNA viruses with some of the longest RNA genomes of about 30,000 bases. The *Coronaviridae* family falls under the *Nidovirales* order with the four most important genera for agriculture and public health being the *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus* and *Deltacoronavirus* (Islam *et al.*, 2021; Marchenko *et al.*, 2022). Their genome organisation consists of a long polyprotein gene on the 5’ end of the genome encoding for the polymerase and other non-structural proteins, followed by the structural genes (Spike, Envelope, Matrix and Nucleoprotein) and a variable number of accessory genes on the 3’ end of the genome (Woo *et al.*, 2023). Based on the wide genetic diversity within the group, coronaviruses are an ancient family sharing a common ancestor many millions of years in the past throughout which they have been interacting with and switching between their respective hosts (Wertheim *et al.*, 2013). Alpha and Betacoronaviruses are primarily found in bats, while Gamma and Deltacoronaviruses are found in birds, which are thought to be the ancestral host

groups of the genera respectively (Chan *et al.*, 2013). Currently, members of all four genera are known to circulate across a wide range of vertebrate hosts.

To date, there are at least ten coronaviruses known to have transmitted from an animal reservoir to humans, three of which have a putative rodent origin (HKU1) potentially through a cattle intermediate (OC43 and HECV-4408), five likely originating in bats (229E, NL63, MERS-CoV, SARS-CoV and SARS-CoV-2), one of a canine-feline origin (CCoV-HuPn-2018) and one of porcine origin (Hu-PDCoV) (Zhang *et al.*, 1994; Forni *et al.*, 2017; P. Zhou *et al.*, 2020; Lednicky *et al.*, 2021; Vlasova *et al.*, 2022). These range from old spillovers, now endemic in humans, to recent epidemic and pandemic viruses, to one-off animal-to-human transmission chains. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has had the greatest documented impact on global health, being responsible for the COVID-19 pandemic, and it can be traced to horseshoe bats of the genus *Rhinolophus* (P. Zhou *et al.*, 2020). Since the end of 2019 there have been more than 700 million confirmed cases of COVID-19 and almost 7 million deaths globally, although these are likely underestimates of the true burden this virus has had on global health (World Health Organization, 2023b). Chapters 2 and 3 of this thesis, will focus on the evolution of SARS-related coronaviruses through the process of recombination and how these viruses interact with their reservoir hosts, leading into what these insights mean for the evolution of SARS-CoV-2 in humans.

1.5 Cross-species transmission of influenza A viruses

Influenza A viruses are a genus of segmented, negative-sense, single-stranded RNA viruses that are part of the *Orthomyxoviridae* family (order *Articulavirales*). There are four genera of influenza viruses: *Alphainfluenzavirus*, *Betainfluenzavirus*, *Gammainfluenzavirus* and *Deltainfluenzavirus* their respective main species being Influenza A, B, C and D (abbreviated as IAV, IBV, ICV and IDV respectively). IAV is more closely related to IBV and ICV more closely related to IDV both in evolutionary relatedness and genome organisation (the former group having eight segments and the latter having seven) (Dou *et al.*, 2018; McCauley *et al.*, 2019). Members of all four genera are known to infect vertebrates, but recently, influenza-like viruses have also been sampled from amphibians, fish and jawless vertebrates (Parry *et al.*,

2020). These viruses form sister lineages to each Influenza genus, suggesting a potential association of Influenza viruses with vertebrates since the host group's emergence (Parry *et al.*, 2020). All four genera contain viruses that have had great public health and economic impact. IBV causes seasonal disease in humans, with distinct clades globally circulating and evolving under antigenic pressure in the human population (Langat *et al.*, 2017). Different strains also circulate in marine animals like seals, showing similar recurrent circulation and transmission dynamics (Bodewes *et al.*, 2013). ICV is generally of lesser concern, known to infect humans in early childhood causing cold-like disease, also circulating in pigs, dogs and cattle (Sederdahl and Williams, 2020). IDV infection in cattle is of great agricultural concern, causing Bovine Respiratory Disease (Ruiz *et al.*, 2022). The virus has never been sampled in humans although there is some serological evidence of potential zoonotic infections from cattle (White *et al.*, 2016).

The virus that has had by far the greatest impact on human health is IAV, being the cause of multiple human pandemics and epidemics in the last century (Kaye and Pringle, 2005; Paules and Subbarao, 2017). IAV strains are separated in serotypes based on the sequence similarity of segments four and six encoding for the two surface glycoproteins: hemagglutinin (HA) and neuraminidase (NA) respectively. The first documented human IAV pandemic was in 1918 caused by a H1N1 strain which led to more than 50 million deaths (Centers for Disease Control and Prevention, 2019a). Then followed a H2N2 pandemic first detected in Southeast Asia in 1957 (Centers for Disease Control and Prevention, 2019b) and a H3N2 pandemic in 1968 first detected in the United States (Centers for Disease Control and Prevention, 2019c), both of which led to about one million deaths each. The latter strain is the one globally circulating in the human population until today. All three of these pandemic serotypes had at least some of their segments derived from viruses circulating in wild aquatic birds which are the primary reservoir host of IAV (Smith, Bahl, *et al.*, 2009). From their wild bird hosts IAVs frequently cross to other species, into many animals that interact with humans such as domestic poultry, horses, dogs and swine (Yoon, Webby and Webster, 2014). In fact, the latest IAV pandemic caused by a H1N1 strain in 2009 was transmitted from farmed pigs to humans, causing up to half a million deaths and still circulating in the population (Centers for Disease Control and Prevention, 2019d). Since these viruses are segmented, a lot of the diversity between strains accumulates through reassortment, the process of swapping segments between different strains following

co-infection of the same cell (Wille and Holmes, 2020). The host promiscuity of these viruses allows for frequent reassortment of their segments and constant transmission between host species. This is why understanding how diversity accumulates across IAVs' evolution and how different strains interact with their potential hosts are key to preventing future Influenza pandemics. In Chapter 4, I explore how one primate-specific restriction factor interacts with IAV and how this restriction has shaped the evolution of IAV strains that have jumped into humans.

1.6 The diverse hosts of the *Flaviviridae*

The *Flaviviridae* is a diverse family of positive-sense, single-stranded RNA viruses in the order *Amarillovirales*, most of which have non-segmented genomes of about 10kb and encode a single polyprotein that is proteolytically cleaved into individual peptides (Simmonds *et al.*, 2017). The four genera currently classified by the International Committee on Taxonomy of Viruses (ICTV) are the *Hepacivirus*, *Orthoflavivirus*, *Pegivirus*, and *Pestivirus* groups, although metatranscriptomic studies have recently identified more diverse clades within the family (Mifsud *et al.*, 2023). The Pestiviruses infect mammalian hosts, but are not known to infect humans, being prevalent in cattle and ruminants and causing disease such as bovine viral diarrhoea (BVD) which has a substantial impact on global agriculture and economy (Riitho *et al.*, 2020). The Pegiviruses are a sister group to the Hepaciviruses that also infect mammals, including humans, transmitting through blood, although they do not cause any known human disease and may even exhibit beneficial effects when infecting individuals with other chronic pathogens (Yu *et al.*, 2022). The best-known member of the Hepaciviruses is the hepatitis C virus (HCV) – from which the genus received its name – which establishes chronic infection in humans leading to viral hepatitis disease, with an estimated 58 million people affected globally (Manns and Maasoumy, 2022; World Health Organisation, 2022).

However, the genus whose members have the greatest collective impact on human health is the Orthoflaviviruses. Unlike the aforementioned groups, most Orthoflaviviruses infect both vertebrate and insect hosts, with prevalent human pathogens being arthropod-borne (arboviruses), transmitting primarily through mosquito but also tick vectors (Conway, Colpitts and Fikrig, 2014). These include

Dengue virus (DENV), West Nile virus (WNV), Zika virus (ZIKV), yellow fever virus (YFV), Japanese encephalitis virus (JEV) and tick-borne encephalitis virus (TBEV) all of which circulate in humans and cause neurotropic (e.g., encephalitis), visceral (e.g., hepatitis and haemorrhage) and congenital disease (e.g., infant microcephaly) (Pierson and Diamond, 2020). Other than the vector-borne flaviviruses, there are two distinct clades of flaviviruses exclusively infecting insects (the classical and dual-host insect-specific flaviviruses, cISF and dISF respectively) (Blitvich and Firth, 2015), as well as flaviviruses with no known vectors (NKV) circulating only within their vertebrate hosts (Blitvich and Firth, 2017). The fact that viruses across the *Flaviviridae* family have switched between vertebrate and invertebrate hosts throughout their evolutionary history or can transmit between both distant host groups suggests that they can evade multiple layers of immune responses and replicate in distinct cellular environments (Conway, Colpitts and Fikrig, 2014). Phylogenetic evidence indicates that a potential invertebrate-specific ancestor of the Orthoflaviruses may have switched to also infecting vertebrates consistent with the evolution of hematophagy in ticks' ancestors, subsequently passing onto mosquito vectors (Bamford *et al.*, 2022). Extensive sampling of invertebrate viromes has recently revealed more diverse clades within the *Flaviviridae* family yet to be formally classified by the ICTV. Jigmenviruses are the only known segmented members of the family, circulating in ticks, with few known cases of infection in humans resulting in febrile disease (Qin *et al.*, 2014; Wang *et al.*, 2019). Finally, the "large genome flaviviruses" (LGF) are only known to infect invertebrate hosts and have surprisingly larger genome sizes compared to the rest of the family, of about 25kb (Shi *et al.*, 2016; Mifsud *et al.*, 2023). In Chapter 5, I examine the genetic signatures imprinted on the *Flaviviridae* genomes across their evolutionary history – estimated to be almost 1 billion years old (Bamford *et al.*, 2022) – and how these signatures correlate with switches between diverse host environments.

1.7 Phylogenetic inferences

Modern methods for inferring the evolutionary histories of groups of organisms largely depend on comparing homologous genetic sequences. Sequence homology means that two sequences once shared a common ancestor, i.e. were the same sequence at some point in time, despite having diverged from that ancestral sequence through mutation accumulation. Homologous sequences can be aligned to infer possible homologous positions. A sequence alignment is essentially a hypothesis of homology between two (pairwise alignment) or more sequence (multiple sequence alignment or MSA). Aligning two sequences that show clear homology (e.g., only differ in few sequence positions) is a relatively easy task, but this gets progressively more complicated when the sequences have low similarity, insertions or deletions (indels) or when homology between multiple sequences is being inferred. Most commonly used MSA algorithms involve aligning pairs of the most similar sequences to one another first and progressively adding the other sequences in order of similarity, using heuristic approaches to speed up the process (Thompson, Higgins and Gibson, 1994; Edgar, 2004; Katoh and Standley, 2013). There is further distinction between aligning nucleotide sequences and aligning peptide sequences, since the latter contains many more informative characters (20 possible amino acids). To facilitate alignment inference between peptide sequences most algorithms utilised predefined protein substitution matrices that give each potential amino acid to amino acid substitution a certain weight based on how similar their biochemical properties are (Dayhoff, Schwartz and Orcutt, 1978; Henikoff and Henikoff, 1992).

Once an alignment has been constructed for a set of homologous sequences, a phylogenetic method can be used to infer how the sequences relate to one another based on the nucleotide or amino acid differences between them. Although genetic distance-based phylogenetic approaches exist, such as the neighbour-joining (NJ) method (Saitou and Nei, 1987), this section will focus on the two most sophisticated methods implemented throughout the thesis: maximum likelihood and Bayesian phylogenetics. Maximum likelihood (ML) approaches assess the probability that a given phylogeny produced the observed sequence alignment given a certain evolutionary model (Felsenstein, 1981). Hence, if every possible tree could be assessed for a given set of taxa, the true most likely tree explaining the sequence alignment could be determined. Unfortunately, with increasing sequence sets it

quickly becomes practically impossible to evaluate the likelihood of the entire tree space. Instead, most ML tools begin by constructing a starting tree using a fast distance-based approach (Saitou and Nei, 1987), so that the tree search begins from a representative point in the tree space rather than a random one. Then, different heuristics are used to infer tree topologies in tree space and evaluate the likelihood of different trees, mainly involving tree rearrangement operations (Collienne and Gavryushkin, 2021). Once sufficient tree space has been assessed, the topology with the highest probability for the sequence alignment will be the algorithm's phylogenetic inference. All popular software for ML phylogenetics, such as RAxML, IQ-TREE and PhyML, utilise a combination of methods to optimise their heuristic search through possible tree topologies (Guindon *et al.*, 2010; Kozlov *et al.*, 2019; Minh *et al.*, 2020). The likelihood of all possible topologies cannot be evaluated with these methods, so some pairings in the tree might not be confidently inferred. The "bootstrap" method, which involves resampling the alignment with replacement and seeing if clusters within the tree topology remains the same, is the most commonly used approach for assessing how confident we are for each node in a ML tree (Felsenstein, 1985). Bootstrapping is a fairly time-consuming method, while being sensitive to the lack of phylogenetic information, so faster, probabilistic alternatives have been recently developed (Guindon *et al.*, 2010; Hoang *et al.*, 2018). With expanding sequencing efforts, for example the millions of SARS-CoV-2 genomes that have become available throughout the COVID-19 pandemic, more likelihood-based approaches have been developed that incorporate multiple methods to substantially cut down computational time while improving accuracy (MAPLE, Fasttree) (Price, Dehal and Arkin, 2010; de Maio *et al.*, 2023).

The length of individual branches in a ML phylogeny will represent the genetic change between the tree nodes. This will reflect nucleotide or amino acid substitutions accumulated in the sequences - corrected for saturation with an appropriate substitution model - and not the actual time in which the substitutions took place. Zuckerkandl and Pauling (1965) proposed that if the rate at which sequence substitutions are accumulated is constant, then this rate can be used as a "molecular clock" to infer the units of time that the tree branches represent. Since, many methods have been developed for calibrating the branch lengths of a ML tree to units of time using a molecular clock. This is particularly useful in fast evolving sequences such as these of RNA virus genomes. The best-suited molecular clock of a given virus phylogeny can be inferred by considering the dates when each virus

representing a tip of the tree was sampled; a process referred to as tip-dating (Rambaut, 2000; Sagulenko, Puller and Neher, 2018). This tree transformation is done post-hoc of the ML inference and recent tip-dating software can perform additional functions based on a given tree and sequence alignment such as ancestral sequence reconstruction (Sagulenko, Puller and Neher, 2018); a method for inferring what the sequence was at each internal node of the tree (described in more detail in Chapter 3). Ancestral sequence reconstruction can also be performed in parallel with the ML inference with more advanced methods also assessing the confidence of each inferred nucleotide/residue at each node (Oliva *et al.*, 2019).

Another popular approach for inferring phylogenies from sequence data is Bayesian phylogenetics. Bayesian methods estimate the probability of a tree given the data (as dictated by the Bayes' theorem) which depends on the likelihood of the data (similar to ML) and a prior distribution on the model parameters (i.e., the branch lengths, topology and substitution model) defined by the user (Rannala and Yang, 1996; Nascimento, Reis and Yang, 2017). The interesting aspect of this approach is that if the user has some biological information about the model's parameters (e.g., an approximate evolutionary rate, calibration points, or sampling dates for the sequences) this can be incorporated into the inference as informative prior distributions to aid the tree search. The priors can also be uninformative if an expectation for certain parameters is not known. After priors have been defined, the algorithm determines a posterior probability for the model and each of the parameters, representing the probability that the model is correct given the data (Ronquist, van der Mark and Huelsenbeck, 2009). In this way, Bayesian phylogenetic methods can determine posterior probabilities for topology, node support and many other parameters such as internal node date estimates or ancestral traits, all during the tree inference. The density of posterior distributions for each parameter is constructed by sampling progressive states of a Monte Carlo Markov Chain (MCMC) and combining the sampled estimates (Hastings, 1970; Drummond *et al.*, 2002). MrBayes is one of the earlier tools for performing Bayesian phylogenetics (Huelsenbeck and Ronquist, 2001) but has now been largely superseded by BEAST (Bayesian Evolutionary Analysis Sampling Trees) and BEAST2 (Suchard *et al.*, 2018; Bouckaert *et al.*, 2019). The BEAST framework has had many extensions in recent years, and is generally considered to be more sophisticated and robust for examining complex data than ML approaches. Some examples relating to viruses include reconstructing virus geographic movement

(Dudas *et al.*, 2017), performing continuous phylogeographic inference for pathogen evolution (Dellicour *et al.*, 2021) and handling complex molecular epidemiology models (du Plessis *et al.*, 2021).

1.8 Recombination detection in virus genomes

Genetic recombination in viruses refers to the process where two viruses contribute to the genetic information of new virus progeny, resulting in the progeny having a "mosaic" genome which contains genetic sequences of both parental strains when the parent viruses are divergent. This process can primarily take place through i) "copy-choice" error: where, during virus replication, the polymerase molecule switches between template strands of distinct genomes found in the same cell, or ii) reassortment: a process unique to segmented viruses where segments of two co-infecting viruses are mixed during genome packaging (Simon-Loriere and Holmes, 2011). Detection of reassortment events between divergent genomes is straightforward, since entire segments are being recombined. By calculating the genetic similarity between a putatively reassortant virus segment and the corresponding segment of potentially parental lineages, one can deduce that the reassorted segments will be more similar to sequences of one parent lineage (called the "minor parent" since it has contributed less genetic information to the mosaic genome) compared to the rest of the segments, which will be more similar to the other parent lineage (called the "major parent"). While sequence similarity is a simple proxy for detecting reassortant segments, inferring the phylogeny of each segment can more definitively identify reassortment and infer at which point in the parental strains' evolution the reassortment event took place (Smith, Vijaykrishna, *et al.*, 2009).

In reassortment, by definition, the theoretical "breakpoints" of recombination – the boundaries between recombinant and non-recombinant sequence – will simply be the ends of the reassorted segment. Detection of copy-choice recombination is a much more nuanced process since the recombinant sequence is on the same genomic segment / genome as the non-recombinant backbone and the breakpoints of recombination in the continuous sequence need to be inferred. Breakpoints can be detected by comparing similarity along aligned sequences. Such methods

require comparison of at least three homologous sequences, one being the putative recombinant and the other two representing the major and minor parental lineages. Then, breakpoint inference is done either by comparing polymorphic sites across the sequences' alignment or separating the alignment into blocks and comparing the pairwise similarity between them (Maynard Smith, 1992). Another commonly used similarity approach is the 3SEQ method which also compares triplets of sequences. The method is based on the concept that a hypothetical mosaic sequence between two parental genomes will have excess similarity to the third sequence if that latter is a recombinant of the two parent lineages. The benefit of the approach is that it is based on a nonparametric statistical framework and does not perform a sliding window comparison, making this one of the fastest approaches (Boni, Posada and Feldman, 2007).

Instead of simply comparing similarity between sequences, phylogenetic information can also be incorporated in the process of recombination detection. Early approaches applied in HIV involved assessing phylogenetically informative sites across four related sequences, so that an outgroup sequence is ensured in addition to the two parental and one potentially recombinant sequences (Robertson, Hahn and Sharp, 1995). Additional mutations and indels will also introduce variation between the sequences being compared, so informative sites might not produce discrete phylogenetic signal across sequence regions. Later methods have tried to account for the variation of phylogenetic signals across alignments (Gibbs, Armstrong and Gibbs, 2000). An example of a popular phylogenetic approach is the recombination detection program (RDP) algorithm which compares the phylogenetic placement of three sequences at a time using a sliding window across the sequences (Martin and Rybicki, 2000). All these methods can be complementary to one another, since there is more confidence in breakpoints detected independently by multiple approaches. The current version of the RDP software (RDP5) incorporates a number of the aforementioned and other methods (Salminen *et al.*, 1995; Weiller, 1998; Holmes, Worobey and Rambaut, 1999; Posada and Crandall, 2001; Posada, 2002; Lemey *et al.*, 2009), including accessory function for combining results of different methods (Martin *et al.*, 2021). These methods determine breakpoints on each sequence corresponding to inferred recombination events (Robertson, Hahn and Sharp, 1995; Salminen *et al.*, 1995; Weiller, 1998; Holmes, Worobey and Rambaut, 1999; Gibbs, Armstrong and Gibbs, 2000; Martin and Rybicki, 2000; Posada and Crandall, 2001; Lemey *et al.*, 2009; Martin *et al.*,

2021). If multiple homologous sequences are being examined for breakpoints of recombination one can also assess clustering of independent breakpoints across the sequence (for example a virus genome) and determine whether there are regions where more or less recombination than expected takes place (i.e. genomic hotspots and coldspots of recombination) (Heath *et al.*, 2006; Simon-Loriere *et al.*, 2009).

Unlike most of the aforementioned methods that determine breakpoints unique to each sequence, a different approach is to split the entire alignment to reduce overall recombination signal among trees inferred by different alignment segments. This is especially useful when one needs to perform selection analyses on recombinant sequences, since recombinant segments will introduce excess diversity in these sequences. The Genetic Algorithm for Recombination Detection (GARD) progressively identifies recombination breakpoint across the entire alignment splits it into blocks and reconstructs separate phylogenetic trees for each non-recombinant block. The goodness-of-fit for each model of non-recombinant phylogenies is evaluated each time until the best fit consecutive model is reached (Kosakovsky Pond *et al.*, 2006). The distinction between recombination detection methods implemented in RDP and GARD will be covered in detail, applied on sequence data, in Chapter 2.

1.9 Sequence homology search

As sequencing technologies began to dominate the field of Biological Sciences, databases for depositing and retrieving genetic sequences became integral for comparative genetics and phylogenetics research. Two of the most commonly used are i) the National Center for Biotechnology Information (NCBI) Genbank database (https://www.ncbi.nlm.nih.gov/genbank/) containing the largest number of sequence information including nucleotide and protein sequences, raw reads and genome assemblies (Clark *et al.*, 2016) and ii) the European Bioinformatics Institute (EBI) Ensembl database (https://www.ensembl.org) which focuses on annotated genome data (Cunningham *et al.*, 2022). More databases containing different types of genetic information exist, but the aforementioned two are highlighted since they are repeatedly used in this thesis. The usefulness of sequence databases also depends

on having tools to efficiently query them. Annotation of sequence entries can often be misleading (e.g., orthologous genes can have different names in different organisms) or non-existent. Rather, a more efficient way to query sequence databases is to search for similarity between database entries and an actual genetic sequence of interest. In this way one can: i) find what a novel sequence is, or ii) retrieve related homologous sequences to perform downstream analysis. The Basic Local Alignment Search Tool (BLAST) was developed by NCBI to complement the Genbank database in 1990 (Altschul *et al.*, 1990) and is now heavily used for querying all sequence databases. BLAST can quickly search for similarity between a given query sequence and all entries of a database, providing pairwise alignments as well as "expect" values that statistically assess the likelihood of a true sequence match. The tool's functionalities include searching nucleotide to nucleotide (BLASTn), protein to protein (BLASTp), translated nucleotide to protein (BLASTx) and protein to translated nucleotide (tBLASTn) (Camacho *et al.*, 2009).

Since BLAST's development many additional tools utilising BLAST at the core of their function have been made. One example that has been used in the analysis presented in Chapter 3 is the Database-Integrated Genome Screening (DIGS) tool (http://giffordlabcvr.github.io/DIGS-tool/). DIGS is designed for specifically screening genome assemblies using a set of query sequences, while it retrieves a lot of information on the search results (hits) that can be explored in a local database setting (H. Zhu *et al.*, 2018). The hits can then be easily used for performing downstream comparative genomics and phylogenetic analyses. Usage examples include searching for: i) previously unannotated homologous genes in any genomic assembly, or ii) specific genomic elements such as virus derived sequences in host genomes.

The methods described above depend on pairwise alignment between sequences with detectable similarity. However, when examining very divergent sequences, homology might not be obvious, especially in the nucleotide level that might suffer from mutational saturation or simply high sequence divergence. Similarity at the amino acid level of protein sequences tends to be more conserved and, when trying to resolve the evolutionary history of really divergent genetic entities, protein alignments are more likely to be useful (Wolf *et al.*, 2018). Even then, simple similarity based approaches may fail to detect homology. Proteins with remote homology can be detected using more sensitive similarity search approaches such

as profile hidden Markov models (profile-HMM). Instead of searching for discrete sequence-to-sequence homology, these probabilistic models implemented in software such as HMMER allow for search of a query sequence against a profile of residues with different likelihoods for each sequence position (Johnson, Eddy and Portugaly, 2010). This method substantially increases sensitivity especially when looking for functionally conserved protein domains, although they can lose accuracy when it comes to types of protein sequences with periodic compositional biases (Mistry *et al.*, 2013). The functionality of profile-HMMs has been further extended to looking for remote DNA similarity, for example when it comes to distant repeat sequences such as transposable elements (Wheeler and Eddy, 2013).

1.10 dN/dS based selection detection

The high genetic diversity across all organisms has not come to be through random evolution alone. Instead, some mutations will lead to changes in an organism's genome that will either increase or decrease the organism's fitness, i.e. its ability to replicate/reproduce in a given environment. In this way, beneficial mutations are more likely to be selected for and become fixed in a population (positive selection) while deleterious mutations are more likely to be selected against and disappear from the population's genetic diversity (negative/purifying selection) (Pybus and Shapiro, 2009). This is expected to be directly affected by the organism's effective population size, selection being more efficient in larger populations, since random mutation will have a bigger effect in small populations (Lynch and Gabriel, 1990; Poon and Otto, 2000; Lynch *et al.*, 2020). In the early days of studying the process of molecular evolution, Kimura (1968) proposed the neutral theory of molecular evolution, according to which the primary process of evolution is fixation of stochastic mutations with neutral or nearly neutral effects on the organisms' fitness. Using the neutral theory as a basis, one can detect exceptions where excessive negative or positive selection is taking place in a gene, especially in virus genomes that have much larger population sizes and shorter generation times than multicellular eukaryotes (Gojobori, Moriyama and Kimura, 1990; Frost, Magalis and Kosakovsky Pond, 2018).

One of the most common ways to estimate non-neutral selective pressures in a set of homologous, coding genetic sequences is by comparing the number of synonymous and non-synonymous substitutions accumulated in the sequences (Nei and Gojobori, 1986; Ohta, 1995). Synonymous (or silent) substitutions are changes in the nucleotide level of a coding sequence that will not alter the peptide sequence of the encoded protein, while non-synonymous changes will. Many methods for detecting selection have been developed over the years which statistically compare the number of synonymous substitutions per synonymous site (dN) to that of synonymous substitutions per synonymous site (dS) under an evolutionary model (Goldman and Yang, 1994; Muse and Gaut, 1994). This ratio between dN and dS is also referred to as ω (Yang and Nielsen, 2000). The methods assume that if mutations were only stochastically accumulated in a genetic population then the rate at which synonymous and non-synonymous mutations are accumulated should be the same (ω=1). Assuming that only changes in the peptide sequence affect an organism's fitness, ω values below 1 indicate negative selection acting to purify the deleterious non-synonymous substitutions, while ω values above 1 indicate positive selection fixing beneficial non-synonymous substitutions. Many of the modern state-of-the-art selection detection methods based on dN/dS estimations are implemented in the HyPhy evolutionary hypothesis testing platform (Kosakovsky Pond *et al.*, 2019) and can be used to detect gene-wide (Murrell *et al.*, 2015), site-specific (Kosakovsky Pond and Frost, 2005; Murrell *et al.*, 2012, 2013) and branch-specific (Smith *et al.*, 2015) selection pressures. The accuracy of dN/dS approaches will depend on the model of evolution assumed as well as the sequence alignment provided to calculate the substitution matrix. Recent advancements to improve the false-positive inferences of positive selection with such methods have focused on accounting for the now appreciated importance of variation in synonymous substitution rate among sites (Wisotsky *et al.*, 2020) and accommodating for multi-nucleotide substitutions (Lucaci *et al.*, 2021, 2023). Another limitation of these methods is that they inherently only detect selective pressures on the protein level. In Chapter 5, I discuss how selection on the nucleotide level can also be detected, focusing on dinucleotide representation.

1.11 Thesis summary

All chapters of the thesis will touch on how virus genomic diversity relates to the evolutionary arms race between viruses and their hosts' populations and species.

My PhD coincided with the start of the COVID-19 pandemic, which brought intense interest in the origins of the pandemic and how the SARS-related horseshoe bat coronaviruses jump into humans and other animal hosts. In Chapter 2, I examine the patterns of recombination in SARS-related coronaviruses, how this mechanism is employed to expand genomic diversity and how it relates to the viruses' reservoir hosts' geographical distributions and onwards evolution in humans. The work presented in Chapter 2 has been published as part of the following scientific papers:

- Lytras et al. (2022). Exploring the natural origins of SARS-CoV-2 in the light of recombination. *Genome Biology and Evolution*. 14(2): evac018.

- Tamura et al. (2023). Virological characteristics of the SARS-CoV-2 XBB variant derived from recombination of two Omicron subvariants. *Nature Communications*. 14(1): 2800.

The next two chapters focus on specific host restriction factors and how these interferon-stimulated genes' evolutionary histories relate to viruses' ability to infect specific hosts.

Chapter 3 follows on the theme of SARS-related coronaviruses and horseshoe bats, describing the ancestral loss of the anti-coronaviral form of the OAS1 gene in these bats, potentially explaining their tight-linked interaction with SARS-related coronaviruses. This chapter partly consists of unpublished work (under review at the time of this thesis' submission) and one section of the chapter has been published as part of the following scientific paper:

- Wickenhagen et al. (2021). A prenylated dsRNA sensor protects against severe COVID-19. *Science*. 374(6567): abj3624.

Chapter 4 focuses on the human BTN3A3 gene's ability to restrict avian-circulating, but not human-circulating Influenza A viruses. On the host side, I describe the evolutionary origins of this antiviral function and on the virus side I explore the

distribution of residue changes that can evade BTN3A3's restriction across all influenza A viruses. The majority of the work in this chapter has been published as part of the following scientific paper:

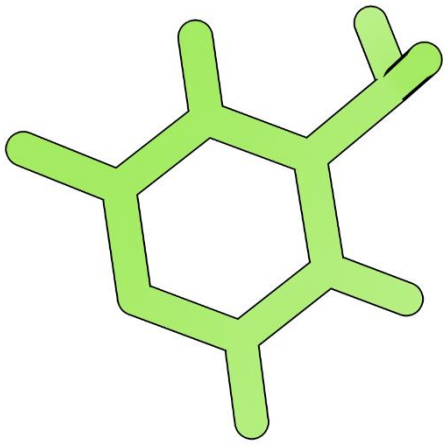
- Pinto et al. (2023). BTN3A3 evasion promotes the zoonotic potential of influenza A viruses. *Nature*. 619: 338.

Chapter 5 takes a more methodological angle, focusing on adaptive changes on the nucleotide rather than the protein level which can be imposed on viral genomes by their respective host environments. I present a novel method for quantifying dinucleotide representation in viral coding sequences, accounting for the intertwined codon usage biases. This metric is then applied on members of the diverse *Flaviviridae* virus family showing how adaptive shifts in the representation of specific dinucleotides coincide with ancestral host switches of these viruses. The majority of the work presented in this chapter is unpublished, except for part of the chapter that has been published in the following scientific paper:

- Lytras and Hughes, (2020). Synonymous dinucleotide usage: a codon-aware metric for quantifying dinucleotide representation in viruses. *Viruses*. 12(4): 462.

The final chapter of the thesis, Chapter 6, outlines three scientific themes overarching the content of all previous chapters, summarising the results' contribution in our understanding of virus-host interactions.

Chapter 2. Exploring the evolution of SARS-related coronaviruses in the light of recombination



Cartoon model of cytosine.

PDB ligand entry: CYT, visualised with ChimeraX.

“You can never put it back together like it was.”

Haruki Murakami, Kafka on the Shore (2002)

The majority of this chapter has been published in *Genome Biology and Evolution* under the title “Exploring the Natural Origins of SARS-CoV-2 in the Light of Recombination” (Lytras *et al.*, [2022]. *GBE*, 14[2]: evac018). I conducted all the phylogenetic analysis and GARD recombination analysis in this work. Co-authors Darren Martin, Phillip Swanepoel, Arné de Klerk and Rentia Lourens contributed to the RDP recombination analysis and recombination hotspot detection presented in Methods’ subsection 2.2.3 “*Recombination hotspot analysis*” and Results’ subsection 2.3.1 *“Hotspots of Recombination”,* and co-authors Joseph Hughes, David L Robertson and Sergei L Kosakovsky Pond supervised the work. The subsection 2.3.4 *“The importance of recombination in human-circulating SARS-CoV-2”* contains my own work published in Tamura *et al.* (2023, *Nature Communications,* 14: 2800) entitled “Virological characteristics of the SARS-CoV-2 XBB variant derived from recombination of two Omicron subvariants”. All other co-authors performed other modelling and experimental analyses for the paper that is not presented in this chapter. The Discussion section has been extended in this chapter compared to the published work.

Aim

The *Coronaviridae* are known to frequently recombine, having complex evolutionary histories and genetic relations among themselves. In this chapter I explore the patterns of recombination in the evolution of SARS-related coronaviruses primarily circulating in Asian horseshoe bats and how these relate to SARS-CoV-2, which recently emerged in humans. I further describe the recombinant origins of the XBB SARS-CoV-2 variant, highlighting the demonstrable importance of recombination in these viruses both in their reservoir and human hosts.

2.1 Introduction

Almost four years since the emergence of SARS-CoV-2, the origins of this new pandemic human coronavirus remain uncertain. First detected in association with an unusual respiratory disease outbreak in December 2019 in Wuhan city, Hubei province, China (Q. Li *et al.*, 2020) no definitive progenitor of animal origin has been identified. The first reports of the initial outbreak were linked to the Huanan animal and seafood market (World Health Organisation, 2021; Worobey, 2021; Worobey *et al.*, 2022) and, while there are some cases with no identifiable association to this location, this is not so surprising given that so many cases are either mild or asymptomatic (Lytras *et al.*, 2021), and it is very possible multiple spillover events at animal markets in Wuhan were involved (Holmes *et al.*, 2021; World Health Organisation, 2021; Pekar *et al.*, 2022). Since the 2020 coronavirus pandemic began, both metagenomic and focused sequencing efforts have uncovered a number of viruses related to SARS-CoV-2, retrieved from locations in China and Southeast Asia (D. Hu *et al.*, 2018; H. Zhou *et al.*, 2020, 2021; P. Zhou *et al.*, 2020; Delaune *et al.*, 2021; Li *et al.*, 2021; Wacharapluesadee *et al.*, 2021). Several of these sarbecoviruses are recombinants necessitating careful analysis as the presence of mosaic genomes violates the assumption of a single evolutionary history, key to reliable phylogenetic inference from mutation patterns in molecular data.

SARS-CoV-2, responsible for COVID-19, and SARS-CoV, the causative agent of the SARS outbreak in 2002-3, are both members of the species *Severe acute respiratory syndrome-related coronavirus* (SARSr-CoV) that forms the sole member of the *Sarbecovirus* subgenus of *Betacoronaviruses* (Gorbalenya *et al.*, 2020) – a group of viruses which have been primarily found in horseshoe bats (family *Rhinolophidae*). Coronaviruses are known to recombine with one another during mixed infections (Graham and Baric, 2010; Boni *et al.*, 2020). Here, we comprehensively characterise the recombinant nature of the SARS-CoV-2-like coronaviruses sampled so far, focusing specifically on the phylogenetic clade of sarbecoviruses that SARS-CoV-2 is a member of; hereafter referred to as the “nCoV” clade (Figure 2.1A) (MacLean *et al.*, 2021). To maintain the focus on this clade from which SARS-CoV-2 emerged, we broadly refer to all other *Sarbecovirus* subclades as ‘non-nCoV’. We present evidence of recombination and several hotspot locations where inferred recombination breakpoints are over-represented.

By comparing the phylogenies inferred for putatively non-recombinant regions of the genome (i.e., best estimates of SARS-CoV-2 and related sarbecoviruses true evolutionary history) with the viruses' sampling locations and their host's geographic range locations, we provide a detailed understanding of the recent evolutionary histories of SARS-CoV-2's closest known relatives including relative divergence times.

In addition to recombination's importance in the evolution of SARSr-CoVs in bats, this process has also recently manifested in SARS-CoV-2's evolution in humans. At the start of the COVID-19 pandemic, SARS-CoV-2 was not under particular positive selection and had accumulated little genetic diversity (MacLean *et al.*, 2021). Once enough antigenic pressure had built up, this was followed by accumulation of multiple adaptive mutations at once creating what was termed variants of concern (VOCs) (World Health Organisation, 2023), starting with the more transmissible Alpha variant, emerging in the autumn of 2020 (Hill *et al.*, 2022). The subsequent evolution of SARS-CoV-2 in humans followed a wave-like pattern where a new VOC with higher relative fitness would quickly displace the previously circulating variant (Markov *et al.*, 2023). For detectable recombination to occur between SARS-CoV-2 variants, distinct genomes need to infect the same individual, similar to what seems to be the case with SARS-related coronaviruses in their reservoir bat populations. Despite the waves of infections where a single variant dominated local or global circulation at a time, some sparse examples of between variant recombination were documented in the first two years of human circulation (Jackson *et al.*, 2021). A significant evolutionary step followed in November 2021, when the Omicron variants emerged harbouring a constellation of non-synonymous mutations not seen before in the virus (Martin *et al.*, 2022; Viana *et al.*, 2022). Omicron's emergence altered the evolutionary landscape of SARS-CoV-2 and, instead of single variant circulation, multiple distinct subvariants of Omicron co-circulated globally by the summer of 2022 (Ito *et al.*, 2023). Most of the prevalent Omicron lineages belong to the phylogenetic clade related to the early BA.2 PANGO lineage. Of these, certain highly diversified BA.2 subvariants, such as BA.2.75 and BJ.1, were first identified in South Asia and are referred to as second-generation BA.2 variants. On the 12th of September 2022 the XBB variant was identified, a recombinant lineage between the second-generation BA.2 variants BJ.1 (BA.2.10.1.1) and BM.1.1.1 (BA.2.75.3.1.1.1; a descendant of BA.2.75) (Roemer, 2022). As I am writing this thesis, descendants of XBB are still the most widely circulating SARS-CoV-2

variants, highlighting the importance of recombination within SARS-CoV-2’s variant pool. The event giving rise to the XBB lineage will be investigated at the end of this chapter.

2.2 Methods

2.2.1 Genome alignment

The whole genome sequences of the 78 *Sarbecovirus* members used in this analysis (Appendix A Table A.1) were aligned and the open reading frames (ORF) of the major protein-coding genes were defined based on SARS-CoV-2 annotation (Wu *et al.*, 2020). Codon-level alignments of the ORFs were created using MAFFT v7.453 (Katoh and Standley, 2013) and PAL2NAL (Suyama, Torrents and Bork, 2006). The intergenic regions were also aligned separately using MAFFT and all alignments were pieced together into the final whole-genome alignment and visually inspected in Bioedit (Hall, 1999).

2.2.2 Genome-specific recombination analysis

We first performed an analysis for detecting unique recombination events specific to individual genome sequences using the RDP (Martin and Rybicki, 2000), GENECONV (Padidam, Sawyer and Fauquet, 1999), BOOTSCAN (Martin *et al.*, 2005), MAXCHI (Maynard Smith, 1992), CHIMAERA (Posada and Crandall, 2001), SISCAN (Gibbs, Armstrong and Gibbs, 2000), and 3SEQ (Boni, Posada and Feldman, 2007) methods implemented in the program RDP5 (Martin *et al.*, 2021). Default settings were used throughout except: i) only potential recombination events detected by three or more of the above methods, coupled with phylogenetic evidence of recombination were considered significant and ii) sequences were treated as linear. We required supporting evidence from three or more of the recombination signal detection methods because none of the methods query the same recombination signals and all have varying power to detect recombination in datasets with different degrees of sequence diversity (Posada and Crandall, 2001; Posada, 2002). The recombinant sequence identification, recombination breakpoint verification and shared recombination event verification steps used are outlined in (Martin *et al.*, 2017), the approximate breakpoint positions and recombinant sequence(s) inferred for every potential recombination event, were manually checked and adjusted where necessary using the phylogenetic and recombination signal analysis features available in RDP5. Breakpoint positions were classified as undetermined if the 95% confidence interval on their location overlapped: i) the 5′

and 3′ ends of the alignment; or ii) the position of a second detected breakpoint within the same sequence that had a lower associated p-value (in such cases it could not be discounted that the actual breakpoint might not have simply been lost due to a more recent recombination event). All of the remaining breakpoint positions were manually checked and adjusted when necessary using the BURT method with the MAXCHI matrix and LARD two breakpoint scan methods (Holmes, Worobey and Rambaut, 1999) used to resolve ties. A putatively non-recombinant version of the original whole-genome alignment was reconstructed by excluding all minor parent sequence segments based on the supervised RDP5 analysis.

2.2.3 Recombination hotspot analysis

The distribution of 236 unambiguously detected breakpoint positions defining 160 unique recombination events based on the RDP5 analysis described above were analysed for evidence of recombination hotspots and coldspots using the permutation-based “recombinant region test” (RRT) (Simon-Loriere *et al.*, 2009) and “breakpoint distribution test” (BDT) (Heath *et al.*, 2006). The RRT accounts for site-to-site variations in the detectability of individual recombination events and examines the distribution of point estimates of pairs of breakpoint locations bounding each of the unique recombination events detected by RDP5. Rather than using point estimates of recombination breakpoint locations, the BDT accounts for underlying uncertainties in the estimation of individual breakpoint locations as determined from the state transition likelihoods yielded by the hidden Markov model-based recombination breakpoint detection method, BURT (described in the RDP5 program manual at http://web.cbio.uct.ac.za/~darren/rdp.html).

The RRT and BDT methods for identifying coldspots and hotspots of recombination have not been previously validated to determine their potential false positivity rates. To verify whether the recombination breakpoint clusters detected with these tests were consistent with the presence of real recombination hotspots, we simulated recombination with SANTA-SIM (Jariani *et al.*, 2019). Four datasets of 100 x 10Kb long sequences that had (i) approximately the same degree of genetic diversity as the analysed sarbecovirus alignment and (ii) approximately the same numbers of detectable recombination events and recombination breakpoints per nucleotide as

those detected in the analysed sarbecovirus alignment. These parameters were empirically chosen and are as follows: population size = 4500, inoculum = all, mutation rate = $3.5x10^{-5}$, rate bias matrix = (0.42, 2.49, 0.29, 1.73, 0.23, 4.73, 6.99, 9.20, 0.60, 1.02, 2.56, 0.88), dual infection probability = 0.1, background recombination probability = 0.06, and generation number = 5000. RRT and BDT consider recombination events with up to two breakpoints between two potential parental sequences and the detected recombinant sequence. Hence, we used a slightly modified version of SANTA-SIM, simulating recombination events with a maximum of two breakpoints, that can be obtained from https://github.com/phillipswanepoel/santa-sim/tree/Recomb_and_align. Within our simulation, we specified that one of the four datasets had no simulated recombination hotspots, and the other three each had a single 100-nucleotide long hotspot between alignment positions 6000 and 6100 wherein recombination frequencies were 4x, 8x or 16x higher than the background level. In this way, we could compare the detected hotspots found by RRT and BDT to the simulated hotspot of the dataset at different frequencies of recombination.

All simulated sequence datasets were analysed for recombination by RDP5 without any supervision and RRT and BDT plots were produced for each dataset (all with the same program settings used to analyse the actual sarbecovirus dataset).

The true positive rate of the BDT was estimated as the proportion of 200-nucleotide windows centred on nucleotides between positions 6000 and 6100, i.e., within the simulated hotspot, that contained a number of breakpoints greater than the upper bound of the 99% confidence interval of the breakpoint clustering distribution expected under random recombination (for example indicated by the light grey areas of the plots in Figure 2.1C). Since a 200-nucleotide sliding window was used for both breakpoint clustering tests, all windows overlapping with the hotspot (positions 5801 to positions 6299) were ignored when determining the BDT and RRT false positive rates. The false positive rate of BDT was calculated as the proportion (across all 100 simulated alignments of each of the four datasets) of the examined 200-nucleotide windows centred on nucleotides outside region 5801 to 6299 that contained a number of breakpoints greater than the upper bound of the 99% confidence interval of the breakpoint clustering distribution expected under random recombination. Since the only true simulated hotspot is between positions 6000 and 6100, any window with breakpoints of recombination above the 99% confidence

interval of the random expectation that is not overlapping these positions would be a false positive hit.

Similarly, the true positive rate of the RRT was estimated as the proportion, across all 100 simulated alignments in a dataset, of 200-nucleotide windows centred on nucleotides between positions 6000 and 6100, i.e., within the simulated hotspot, that had associated breakpoint clustering permutation p-values < 0.01 (for example, indicated by the upper bound of the light grey area of the plot in Figure 2.1B). The RRT false positive rate was calculated as the proportion, across all 100 simulated alignments in a dataset, of the examined 200 nucleotide windows centred on nucleotides outside region 5801 to 6399 that had associated permutation p-values < 0.01.

The true and false positive rates for BDT and RRT with respect to identifying the presence of the simulated recombination hotspots are indicated in Appendix A Table A.2. Note that, due to the nature of the simulations, it was not guaranteed that even with perfect recombination detection power and accuracy (i) the recombination hotspot regions would contain any detectable excess of recombination breakpoints, and (ii) the “normal” genome regions would contain no breakpoint clusters. What these simulations capture is the power of the two clustering tests to indirectly infer the locations of actual recombination hotspot regions that, due to chance during the simulation process, might not even contain any detectable recombination breakpoints. Nevertheless, as expected, the hotspot detection power of both BDT and RRT increases substantially with the intensity of the simulated recombination hotspots: from ~10% for both tests with a 4x increase in simulated breakpoint probabilities within the 100-nucleotide hotspot region to ~60% for a 16x increase in breakpoint probabilities within the hotspot region. It is also noteworthy that the false positive rates for both tests are likely between 1.5 and 2x higher than the expected rate of 0.01 (which is expected given that the windows containing breakpoint clusters exceeding the 99% confidence interval were used to identify breakpoint hotspots). This false positive rate may not seem very high but, for a long alignment such as that examined for the sarbecoviruses that can be broken into ~150 non-overlapping 200-nucleotide windows, it indicates that for such an alignment we might expect to find on average two to three significant clusters of breakpoints that are in fact not associated with any elevation in the underlying recombination rate.

2.2.4 Whole-genome alignment recombination analysis

Next, I sought to conservatively examine the entire genome alignment for the subset of recombination breakpoints that had the largest impacts on the inferred evolutionary relationships between the analysed sarbecoviruses using the Genetic Algorithm for Recombination Detection (GARD) method (Kosakovsky Pond *et al.*, 2006) implemented in Hyphy v2.5.29 (Kosakovsky Pond *et al.*, 2019). Model goodness of fit was evaluated using the small sample Akaike Inference Criterion (c-AIC) (Akaike, 1998). To improve computational efficiency and statistical efficiency (as GARD requires more statistical evidence of recombination for larger phylogenies, and the minimal length of detectable non-recombinant fragments increases with the number of sequences) and focus on the closest relatives of SARS-CoV-2, 22 of the 78 viruses that are closest to SARS-CoV-2 or had preliminary evidence of clustering near detected inter-clade recombinants were included in the GARD analysis (Appendix A Table A.1). Only breakpoints present in more than 2/3 of the 64 GARD consecutive step-up models were retained to produce a final set of 21 likely breakpoints (positions corresponding to the SARS-CoV-2 reference genome Wuhan-Hu-1 in order: 1680, 3093, 3649, 4973, 8208, 11445, 12622, 14401, 15954, 16923, 19965, 20518, 21198, 21411, 22460, 23396, 24144, 24843, 26323, 27388, 27685). Based on these, the whole-genome alignment was split into 22 recombinant breakpoint partitioned (RBP) regions. The position of each region on the alignment and relative to the SARS-CoV-2 genome as well as their length are presented in Appendix A Table A.3.

We further used the GARD recombination analysis to validate the RDP5 recombination hotspot analysis. We performed a permutation test of breakpoint clustering by fixing the number of all inferred breakpoints (64) and the location of the 13,550 variable sites in the alignment. Then defined a sliding window so that each window would have an average of one breakpoint in it (alignment length / 64) producing 474 windows. N = 10,000 replicates were drawn where 64 variable sites were randomly chosen from one of the breakpoints. For each sliding window, we tabulated the distribution of randomly drawn breakpoints in the window. Two hotspots and 17 coldspot windows were identified, presented in Appendix A, Figure A.1. This analysis is not expected to produce results identical to the RDP5-based

hotspot analysis, since the GARD method does not distinguish between potential breakpoints in very near genomic proximity, so this post-hoc test is unlikely to identify clustering of unique breakpoints that are very close to one another (in contrast to the RDP5 approach).

2.2.5 Sarbecoviruses phylogenetic reconstruction

The phylogeny of each RBP alignment region based on the GARD analysis and the non-recombinant whole-genome based on the RDP5 analysis were reconstructed using IQ-TREE version 1.6.12 (Nguyen *et al.*, 2015) under a general time reversible (GTR) substitution model assuming invariable sites and a 4 category Γ distribution. Tree node confidence was determined using 10,000 ultrafast bootstrap replicates (Hoang *et al.*, 2018).

Based on the non-recombinant whole-genome phylogeny, 20 viruses form a monophyletic nCoV clade (Figure 1A). To illustrate the distance of each virus from SARS-CoV-2 for each GARD determined genomic region, I defined the nCoV clade on each phylogeny as the subset of the aforementioned 20 nCoV viruses forming a monophyly with SARS-CoV-2 in each phylogeny. The rest of the viruses were classified as members of the non-nCoV clade for each RBP region. I then used an arbitrary tip distance scale normalised between all phylogenies so distances are comparable between regions. For each maximum likelihood tree, the patristic distance between each tip and SARS-CoV-2 is calculated using ETE 3 (Huerta-Cepas, Serra and Bork, 2016) as d_1 for members of the nCoV clade and d_2 for members of the non-nCoV clade. The distances are then normalised so that for nCoV clade members range between 0.1 and 1.1 (1.1 being SARS-CoV-2 itself and 0.1 being the most distant tip from SARS-CoV-2 within the nCoV clade) and between -0.1 and -1.1 for non-nCoV members (-0.1 being the closest non-nCoV virus to SARS-CoV-2 and -1.1 the most distant), as follows:

$$d'_1 = 1.1 - \frac{d_1}{d_{1,max}} \quad \text{(1:nCoV)}$$

$$d'_2 = -0.1 - \frac{d_2 - d_{2,min}}{d_{2,max} - d_{2,min}} \quad \text{(2:non-nCoV)}$$

With d'_1 and d'_2 being the normalised values for each clade, variables denoted with "min" being the smallest distance and variables denoted with "max" being the largest distance in each given set.

Phylogenies were visualised using FigTree (http://tree.bio.ed.ac.uk/software/figtree/) and ETE 3 (Huerta-Cepas, Serra and Bork, 2016).

2.2.6 Molecular dating

To provide temporal information to the phylogenetic history of the viruses, I performed a Bayesian phylogenetic analysis on RBP region 5, using BEAST v1.10.4 (Suchard *et al.*, 2018). This region was selected due to its length, being one of the two longest non-recombinant regions in the analysis (3,238 bp), and because all 20 nCoV viruses form a monophyly in the respective tree. Based on the observation of an increased evolutionary rate specific to the deepest branch of the nCoV clade reported in MacLean et al. (2021), I adopted the same approach of fitting a separate local clock model to all branches of that clade from the rest of the phylogeny. A normal rate distribution with mean $5x10^{-4}$ and standard deviation $2x10^{-4}$ was used as an informative prior on all other branches. The lineage containing the BtKY72 and BM48-31 bat viruses was constrained as the outgroup to maintain overall topology. Codon positions were partitioned and a GTR+Γ substitution model was specified independently for each partition. The maximum likelihood phylogeny reconstructed previously for RBP region 5 was used as a starting tree (rooted at the BtKY72 and BM48-31 clade). A constant size coalescent model was used for the tree prior and a lognormal prior with a mean of 6 and standard deviation of 0.5 was specified on the population size. Two independent MCMC runs were performed for 500 million states for the dataset. The two chains were inspected for convergence and combined using LogCombiner (Drummond and Rambaut, 2007) using a 10% burn-in for each chain. The effective sample size for all estimated parameters was above 200.

2.2.7 Host range data

All host ranges presented in Figure 4B were retrieved from the IUCN Red List of Threatened Species (https://www.iucnredlist.org/) and the Mammals of China Princeton Pocket Guide (Hoffmann *et al.*, 2013). Geographic visualisation was performed using D3 and JavaScript in Observable (https://observablehq.com/).

2.2.8 XBB recombination analysis

As of October 3, 2022, I retrieved a total of 562 sequences using the following criteria from the GISAID database (https://gisaid.org/): i) human hosts, ii) collected after 2022, iii) with length greater than 28,000 base pairs, and iv) with PANGO lineage designation BJ.1, BM.1, XBB and all their descendants. To ensure that PANGO lineage definitions in the dataset's metadata included the latest circulating lineages, the GISAID metadata were downloaded again on October 15, 2022, and the PANGO lineages of our sequences were updated accordingly. Sequences were aligned to the reference Wuhan-Hu-1 genome (GenBank accession: NC_045512.2) and then converted to a multiple sequence alignment using the "global_profile_alignment.sh" script from the SARS-CoV-2 global phylogeny pipeline (Lanfear, 2020), utilizing MAFFT (Katoh and Standley, 2013). A number of recombination detection methods were performed on the resulting alignment using the Recombination Detection Program (RDP) v.5.21 (Martin *et al.*, 2021), specifically: RDP (Martin and Rybicki, 2000), GENECONV (Padidam, Sawyer and Fauquet, 1999), BOOTSCAN (Martin *et al.*, 2005), MAXCHI (Maynard Smith, 1992), CHIMAERA (Posada and Crandall, 2001), SISCAN (Gibbs, Armstrong and Gibbs, 2000), and 3SEQ (Boni, Posada and Feldman, 2007). Sequences were assumed to be linear in the RDP5 parameters, only recombination events detected consistently by more than 3 independent methods were retrieved and potential false positives were excluded from the final output of RDP5. GISAID acknowledgments for all analysed sequences are available as provided as supplementary information in the published version of this work in Tamura et al. (2023).

2.2.9 XBB phylogenetic analysis

For inferring the phylogenies of each non-recombinant segment of the XBB variant, I first split the alignment used for the recombination analysis above at genome position 22,920 (the breakpoint inferred by RDP5). Due to the lack of many informative sites of the 3' end shorter non-recombinant alignment, two quality filtering steps were implemented: i) the 3' end of the alignment was trimmed up to the position where none of the sequences had 3' end gaps and ii) all sequences with Ns were removed, leading to a reduced alignment of 370 sequences. BA.2 sequence EPI_ISL_10926749 was added to the alignments as an outgroup. IQ-TREE v2.1.3 (Minh *et al.*, 2020) was used for making a phylogeny for each non-recombinant alignment. The TIM2+F+I substitution model was used for both trees as selected by the '-m TEST' (Kalyaanamoorthy *et al.*, 2017) and node support was assessed by performing 1000 ultrafast bootstrap replicates (Hoang *et al.*, 2018).

Both phylogenies were inspected for the presence of temporal signal using TempEst v1.5.3 (Rambaut *et al.*, 2016). The 3' end non-recombinant segment's phylogeny did not have enough substitutions for a root-to-tip regression to be inferred, hence I proceeded with tip-dating analysis only for the 5' end, longer segment. I used BEAST v1.10.4 (Suchard *et al.*, 2018) to infer a time-calibrated Bayesian phylogeny of this genome segment. To avoid missing information affecting the inference I also removed all sequences containing Ns from the alignment, leading to a reduced dataset of 247 sequences. I used a strict molecular clock model with an exponential growth coalescent prior (Griffiths and Tavare, 1994). The HKY substitution model was used, accounting for site heterogeneity with an invariant site and four category Γ distribution model. A clock rate prior with mean of 1×10^{-3} and standard deviation of 1×10^{-4} was provided – consistent with the accepted rate for SARS-CoV-2 (Duchene *et al.*, 2020) – and all XBB sequences were assumed to be monophyletic. Duplicate MCMC chains were run for 100,000,000 states each, sampling every 10,000 states. Convergence was assessed using Tracer v1.7.1 (Rambaut *et al.*, 2018) and maximum clade credibility (MCC) trees were summarized by combining the two chains after removing a 10% burn-in using LogCombiner (https://beast.community/logcombiner) and TreeAnnotator (https://beast.community/treeannotator).

2.2.10 Data availability

The whole-genome alignment and raw phylogenetic tree files associated with this work can be found in the following online repositories: https://github.com/spyros-lytras/SC2_origins_rec, https://github.com/TheSatoLab/XBB/tree/main/Phylogenetic_analysis.

2.3 Results

2.3.1 Hotspots of recombination

For a whole-genome alignment of the set of known complete genomes from 78 members of the *Sarbecovirus* subgenus (including a single representative of SARS-CoV and SARS-CoV-2; Appendix A Table A.1) we performed an initial recombination breakpoint analysis with RDP5 (see Methods subsection 2.2.2) and identified 160 unique recombination events in all the bat and pangolin-derived virus genomes. To infer a reliable phylogeny of the sarbecoviruses, we removed all regions with evidence for a recombination history from the genome alignment. This reconstructed non-recombinant phylogeny (Figure 2.1A) includes a total of 19 non-human viruses that comprise the nCoV clade that SARS-CoV-2 is a member of, a sister lineage to the non-nCoV clade SARS-CoV is part of, first emerged from in 2002.

Using the set of breakpoints inferred by RDP5, we tested for significant clustering of recombination events at specific regions of the genome, suggestive of recombination hot- or coldspots. Two permutation-based recombination breakpoint clustering tests were performed: (i) a “breakpoint distribution test” (BDT) that explicitly accounts for the underlying uncertainties in the positions of identified breakpoint positions (Heath *et al.*, 2006) and (ii) a “recombinant region test” (RRT) that focuses on point estimates of recombination breakpoint pairs that define recombination events and explicitly accounts for region-to-region variations in the detectability of recombination events (Simon-Loriere *et al.*, 2009). Both tests provided support for the presence of several recombination hotspots: seven in the BDT and nine in the RRT analysis, assuming close locations are giving rise to the same peak (Figure 2.1B,C), and recombination refractory regions in the NTD and RBD domains of the Spike gene and within ORF8 (Figure 2.1C).

It is possible that all genomic regions where these breakpoint clusters are detected have elevated recombination rates, linked to the molecular mechanisms likely responsible for recombination (Sola *et al.*, 2015). However, simulations of recombination patterns – in genomes with similar degrees of diversity and numbers of detectable recombination events to the genomes analysed here – indicate that within such a dataset we might expect to find, on average, two to three such clusters

even in the absence of any recombination hotspots (see Methods subsection 2.2.4, Appendix A Table A.2). Therefore, none of the identified breakpoint clusters can be definitively attributed to underlying variations in recombination rates at the genome sites where the clusters are identified. Nonetheless, the distribution of recombination breakpoints is clearly non-uniform across the *Sarbecovirus* genomes, and this non-uniformity is consistent with the presence of recombination hotspots. To independently validate the results of this analysis we also performed a simple permutation test for clustering in the recombination breakpoints inferred by the GARD analysis (see below, Appendix A Figure A.1). Even though this test would not identify potential hotspots in proximal genomic locations (due to the nature of the GARD method which is expected to identify focused recombination hotspots as a single recombination breakpoint), it confirms the recombination hotspots within the Spike ORF (alignment positions 24174 – 24648, Figure A.1- consistent with the BDT results, Figure 2.1C) and at the start of the N ORF (alignment positions 29388 – 29862, Figure A.1, consistent with both RRT and BDT results, Figure 2.1C).

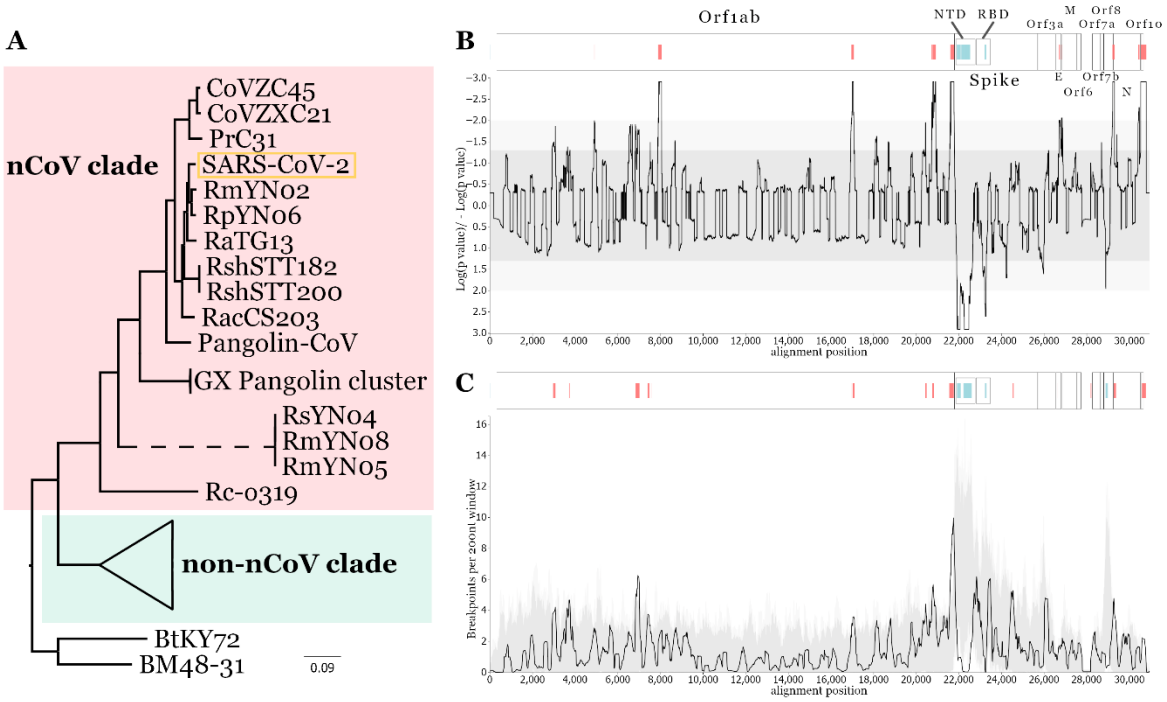


Figure 2.1. Recombination-minimised phylogeny and recombination hot-/coldspots. Maximum likelihood phylogeny inferred from a recombination-free whole genome alignment of the 78 *Sarbecoviruses* **(A)**, see Methods. The non-nCoV/SARS-CoV clade is collapsed for clarity. All nodes presented have bootstrap confidence values above 90%. Distribution of recombination hot- and coldspots across the alignment based on the RRT **(B)** and the BDT **(C)** methods. For both plots light and dark grey represent 95% and 99% confidence intervals of expected recombination breakpoint clustering under random recombination. Peaks above the shaded area represent recombination hotspots and drops below represent coldspots, annotated on the corresponding ORF genome schematic above each plot by vertical red and blue lines respectively. All ORF names and the NTD and RBD encoding regions of Spike are also annotated on the schematics.

Interestingly the pattern of potential hotspots near the Spike ORF has also been noted in previous research (Bobay, O'Donnell and Ochman, 2020). Although selective pressure underlying recombinant regions cannot be assessed in this analysis, antigenic selection – for immune escape – and/or selection associated with switches in host receptor specificity and efficiency - i.e., antigenic shift - are two probable candidate drivers of the observed recombination patterns, consistent with the known immunodominance of the Spike NTD and RBD regions (Walls *et al.*, 2020). It is clearly important to account for these complex recombination patterns when examining the evolutionary history of these pathogens, since multiple evolutionary histories can be inferred from the single whole-genome alignment. As SARS-CoV-2 continues circulating in humans and mutations increase its sequence diversity, identifying SARS-CoV-2 recombination events will become easier and increasingly more important to monitor (Jackson *et al.*, 2021).

2.3.2 Recombination patterns between SARS-CoV-2 relatives

To reconstruct a reliable phylogeny for a set of viruses, sufficient information needs to be present in the underlying sequence alignment. Thus, even though a whole-genome alignment can be split into shorter sub-alignments with the aim of getting rid of all independent recombination events, it is unlikely that all sub-alignments can produce reliable phylogenies. To overcome this trade-off I performed a secondary, more conservative, recombination analysis using GARD and identified the locations of 21 recombination breakpoints that strongly impact the inferred phylogenetic relationships of the analysed sequences when mosaic patterns are ignored (Appendix A Table A.3). In contrast to the RDP5 method used above for assessing breakpoint clustering, the GARD method focuses on extracting recombination signal for the entire alignment, and so is better suited for producing putatively non-recombinant phylogenies. I then determined the phylogenetic relationships of the viral sequences in each of the 22 putatively non-recombinant genome regions bounded by each identifiable breakpoint (Figure 2.3A). The 20 nCoV viruses identified in the non-recombinant whole-genome phylogeny above (Figure 2.1A) were used to inform the clade annotation for the 22 new non-recombinant phylogenies.

The two genetically closest relatives of SARS-CoV-2 that were identified shortly after its emergence were the bat sarbecoviruses, RaTG13 and subsequently RmYN02, both from samples collected in Yunnan (H. Zhou *et al.*, 2020; P. Zhou *et al.*, 2020). We find that RmYN02 shares a common ancestor with SARS-CoV-2 about 40 years ago and RaTG13 – about 50 years ago (Figure 2.4A) consistent with previous estimates (Boni *et al.*, 2020; MacLean *et al.*, 2021; Wang, Pipes and Nielsen, 2021). Although SARS-CoV-2 is most similar to RmYN02 across most of its genome, the region corresponding to the first half of the RmYN02 Spike ORF appears to have been derived through recombination from a parental sequence residing outside the nCoV clade (Figure 2.1A). Two more viruses very recently identified in Yunnan, RpYN06 and PrC31 are most closely related to RmYN02 for part of their genomes (H. Zhou *et al.*, 2021; Li *et al.*, 2021). In the portion of the genome corresponding to recombination breakpoint partitioned (RBP) regions 2 to 5, the three Yunnan viruses (RmYN02, RpYN06 and PrC31) cluster with strong support in a sister clade to SARS-CoV-2 (Figure 2.2A, https://github.com/spyros-lytras/SC2_origins_rec/blob/main/78sarbeco_alltree_fn.json). This pattern

suggests that bat sampling efforts in Yunnan have uncovered a related viral population that has relatively recently shared a common ancestor with SARS-CoV-2's proximal ancestor. Molecular dating of the RBP region 5 phylogeny (corresponding to the C-terminal part of nsp3; Figure 2.4A) indicates that this "Yunnan cluster" shared a common ancestor with SARS-CoV-2 around 1982 (95% HPD: 1970-1994). This analysis further allows us to date the node between PrC31 and RmYN02 to 2005 (95% HPD: 1998-2010), which is one of the most recent nodes in the phylogeny (Figure 2.4A).

The recombination analysis, however, reveals a much more complex evolutionary history for the rest of the PrC31 genome (Li *et al.*, 2021). As seen in the consensus whole-genome phylogeny (Figure 2.1A), most of its genome clusters with viruses CoVZC45 and CoVZXC21 sampled in Zhejiang, a coastal province in East China (Lin *et al.*, 2017; D. Hu *et al.*, 2018). Across the majority of their genomes (excluding segments of Orf1ab and Spike) these viruses are members of the nCoV clade and share a common ancestor with SARS-CoV-2 that existed before 1934 (95% HPD: 1907-1957) according to molecular dating of RBP region 5 (Figure 2.4A). However, in RBP regions 8-12 the sequences of these viruses cluster outside the nCoV clade, and are most closely related to Zhejiang virus Longquan_140 and the HKU3 set of closely related bat sarbecoviruses sampled in Hong Kong (bordering Guangdong province) (Figure 2.2A, see online data Section 2.2.10). The link between SARS-CoV-2's closest relatives and viral populations in the southeast of South China becomes even more apparent in the phylogeny of RBP region 2 where Longquan_140 clusters within the nCoV clade along with CoVZC45 and CoVZXC21 (Figure 2.2A, see online data Methods subsection 2.2.10 - RBP region 2 tree). These relationships indicate ancestral movement of the nCoV viruses across large geographic ranges in China, spanning Yunnan in southwest China and Zhejiang on the east coast (Figure 2.3B).

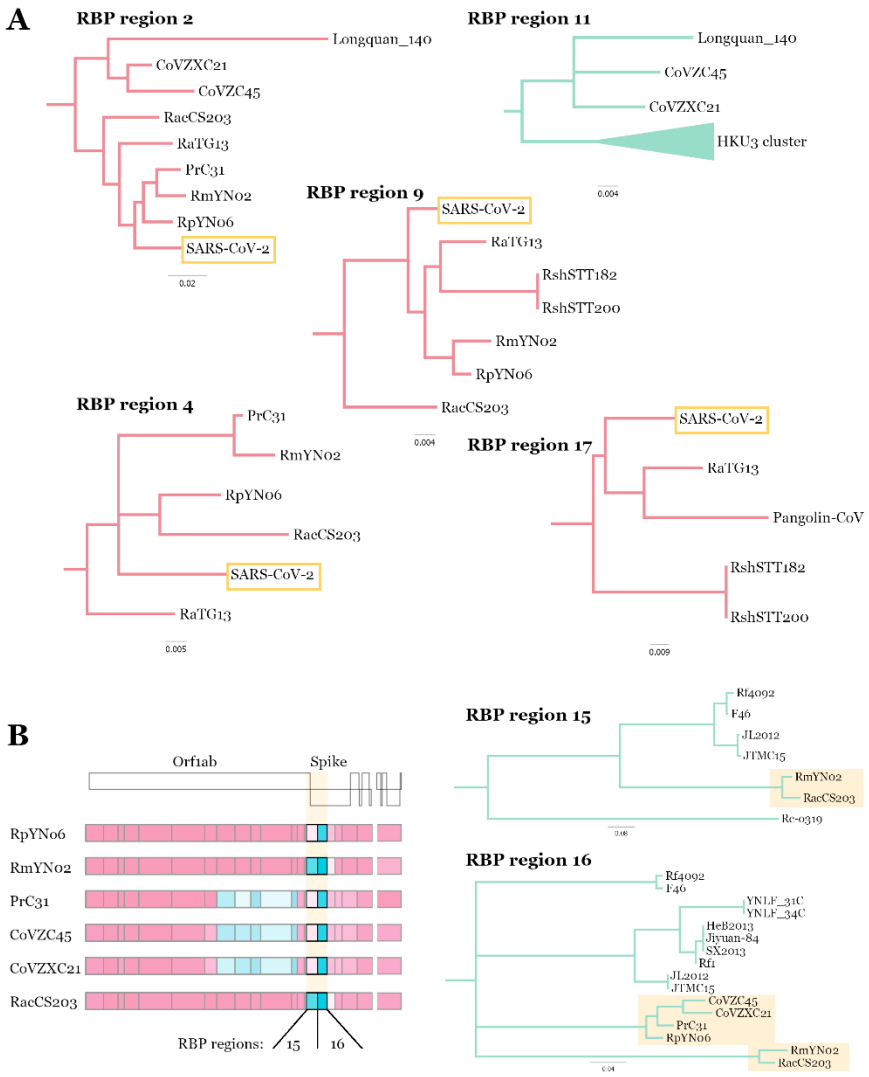


Figure 2.2. Non-recombinant topologies of SARS-CoV-2 relatives. Zoomed in regions of selected RBP region maximum likelihood phylogenies **(A)**. Branches within the nCoV clade are coloured in red and outside the nCoV clade in green. Genome schematics of close SARS-CoV-2 relatives with recombinant Spike regions **(B)**. RBP regions 15 and 16 are highlighted and the non-nCoV subclades of the maximum likelihood phylogenies containing the relevant viruses are presented. The colouring of non-recombinant segments indicate patristic distance to SARS-CoV-2 (see Figure 2.3 legend). Nodes with bootstrap confidence values below 80% have been collapsed.

As more countries initiate wildlife-infecting coronavirus sampling and sequencing efforts, the geographic range of the nCoV clade linked to bat host species will be further refined, evident from the recent reporting of bat sarbecoviruses closely related to SARS-CoV-2 from: (i) two samples collected in Cambodia from *R. shameli* (RShSTT182 and RShSTT200) confirmed by whole-genome analysis (Delaune *et al.*, 2021), and (ii) five bat samples from *R. acuminatus* collected in Thailand with one fully sequenced genome of virus RacCS203 (Wacharapluesadee *et al.*, 2021). These viruses are, after the China sampled CoVs mentioned above, the next closest relatives to SARS-CoV-2 with common ancestor age estimates (using RBP region 5) around 1907 (95% HPD: 1873-1938) and 1883 (95% HPD: 1841-1921),

respectively (Figure 2.4A). Similar to the other nCoV viruses, the recombination analysis uncovers more intricate phylogenetic relations for some parts of the genome. Notably, RShSTT182 and RShSTT200, despite being sampled in Cambodia, cluster with RaTG13 for RBP regions 8 and 9 (Figure 2.2A, see online data subsection 2.2.10), while in RBP region 4 of the genome RacCS203, from Thailand, clusters together with SARS-CoV-2 within the Yunnan clade (Figure 2.2A). This indicates that co-circulation and recombination between these viruses in the last few centuries is responsible for the observed patterns in their inferred evolutionary history, despite their current geographic ranges being at least 2,500km apart. This wide distribution of related viruses, including shared recombination breakpoints, highlights an important feature of bat species: their frequently overlapping/sympatric ranges will provide ample opportunities for transmissions of viral variants from one bat species (or sub-species) to another.

Consistent with the Spike S1 recombination hotspots revealed in the initial analysis (Figure 2.1B,C), the closest relatives of SARS-CoV-2 presented here have non-nCoV derived recombinant sequences at the start of the Spike gene (Figure 2.2B). Despite one collected from Yunnan, China and the other from Thailand, viruses RmYN02 and RacCS203 share a closely related non-nCoV sequence in RBP regions 15 and 16 (encompassing the Spike NTD and RBD respectively; Figure 2.2B) having a distinct RBD compared to that of SARS-CoV-2. On the other hand, viruses RpYN06, PrC31, CoVZC45 and CoVZXC21 cluster within the nCoV clade for region 15 but, similar to the RmYN02 and RacCS203, form a distinct cluster in the non-nCoV clade for region 16 (Figure 2.2B; Wells et al., 2021). We speculate that some of the apparent patterns of recombination-mediated exchange between nCoV and non-nCoV viruses can be partly explained by sequential recombination, i.e., “overprinting” of recombination events involving different ancestral parental viruses. This will occur when an nCoV virus acquires a non-nCoV genomic sequence through ancestral recombination but its progenitors co-circulating with other nCoV viruses incur subsequent recombination events that overlap portions of the original non-nCoV recombinant sequence, producing the more complex “patchy” patterns we see in the currently sampled viruses. Note, overprinting of recombination regions will result in reduced confidence in the breakpoints at deeper nodes in the phylogeny.

The finding that Sunda (also known as Malayan) pangolins, *Manis javanica,* non-native to China, are another mammal species from which nCoV sarbecoviruses have been sampled in Guangxi and Guangdong provinces in South China (Lam *et al.*, 2020; Xiao *et al.*, 2020), indicates these animals are likely being infected in this part of the country (Figure 2.3B). Pangolins are one of the most frequently trafficked animals with multiple smuggling routes leading to southern China (Xu *et al.*, 2016). The most common routes involve moving the animals from Southeast Asia (Myanmar, Malaysia, Laos, Indonesia, Vietnam) to Guangxi, Guangdong, and Yunnan. The most likely scenario that is consistent with both the reported respiratory distress that the sampled pangolins exhibited (Liu, Chen and Chen, 2019; Xiao *et al.*, 2020) and the lack of confirmed CoV infections among Sunda pangolins in Malaysia (Lee *et al.*, 2020), is that the viruses obtained from these animals infected them (presumably from bat sources) after they were trafficked into southern China. Still, serological data of trafficked Sunda pangolins could suggest potential circulation of sarbecoviruses in the animals' wild populations (Wacharapluesadee *et al.*, 2021).

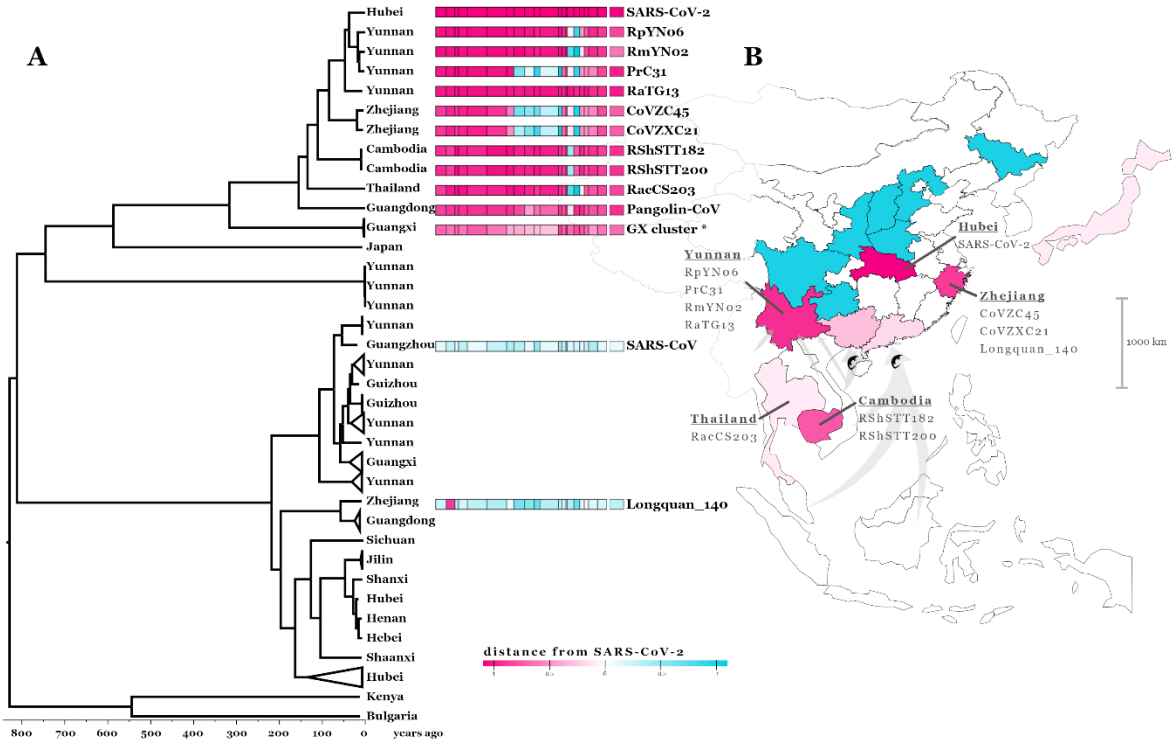


Figure 2.3. Recombination analysis and geographic distribution of *Sarbecoviruses*. Maximum clade credibility (MCC) dated phylogeny of RBP region 5 of 78 *Sarbecoviruses* **(A)**. All tips are annotated with the geographic region the viruses have been sampled in and notable viruses are annotated with genome schematics separated into the 22 inferred RBP regions, each coloured based on phylogenetic distance from SARS-CoV-2 (see scale and Methods). RBP region 21 has been removed from the schematic due to limited phylogenetic information in the alignment. The GX cluster annotated with an asterisk contains the 5 pangolin coronaviruses collected in Guangxi.

Figure 2.3 (cont). Map of East Asia with geographic regions (provinces within China, countries outside China) coloured based on *Sarbecoviruses* sampling **(B)**: blue for regions with only non-nCoV clade samples, pink for regions where nCoV viruses have been sampled. Shading in the nCoV regions corresponds to phylogenetic distance from SARS-CoV-2 (see scale). Notable nCoV viruses and pangolin trafficking routes - adapted from Xu et al. (2016) - are annotated onto the map.

Although the recombination patterns inferred in the pangolin-derived virus genomes seem to be less complex than those of the bat nCoV genomes, the Guangdong Pangolin-CoV has a Spike receptor binding domain that is most similar to that of SARS-CoV-2. This finding was highlighted by (X. Li *et al.*, 2020) and attributed to recombination between the SARS-CoV-2 and Pangolin-CoV proximal ancestors. However, based on the nucleotide divergence between the two viruses in this short Spike segment, the most likely explanation is recombination in RaTG13, making it more divergent than Pangolin-CoV compared to SARS-CoV-2 (Boni *et al.*, 2020) (reflected in region 17 - see online data Methods subsection 2.2.10, Figure 2.2A). The susceptibility of pangolins to an apparently new human coronavirus is not surprising given the well-documented generalist nature of SARS-CoV-2 (Conceicao *et al.*, 2020), which has been found to readily transmit to multiple mammals with similar ACE2 receptors, most notably, on mink farms (Oude Munnink, Sikkema, *et al.*, 2021).

2.3.3 Overlapping horseshoe bat ranges

Considering that almost all sarbecoviruses have been sampled in related horseshoe bat host species, with ranges that span different regions where nCoV clade viruses have been collected (Figure 2.4B), these bat populations should be prioritized for sampling. For example, the least horseshoe bat species, *R. pusillus*, is sufficiently dispersed across China to account for the geographical spread of i) bat sarbecovirus recombinants in the West and East of China, ii) infected imported pangolins in the South, iii) bat sarbecovirus recombinant links to southwest of China, and iv) SARS-CoV-2 emergence towards Hubei in Central China (Figure 2.3B). Strikingly, the ranges of multiple species including *R. affinis*, *R. sinicus* and *R. pusillus* overlap all the regions in China where nCoV members have been collected (Figure 2.4B). Other species known to harbour nCoV viruses have more restricted ranges such as *R. malayanus* found predominantly in the western part of China and countries to the Southwest of China (Myanmar, Thailand, Cambodia, Laos, Viet Nam, and Peninsular Malaysia) (Piraccini, 2016; Bates *et al.*, 2019). On the contrary, the greater horseshoe bat species, *R. ferrumequinum*, is not known to harbour any nCoV viruses and is absent from large parts of South Central China (Figure 2.4B).

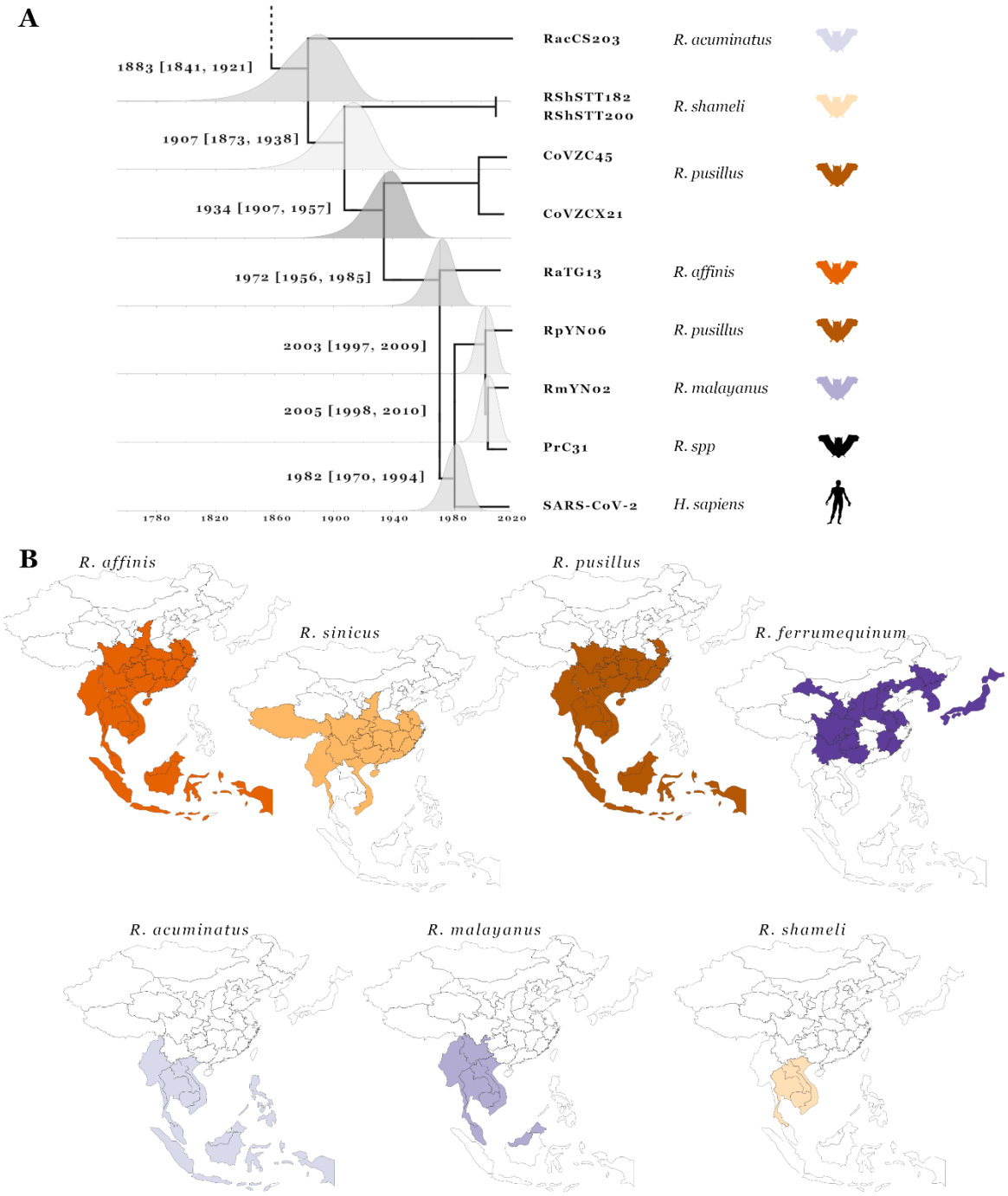


Figure 2.4. Molecular dating and *Rhinolophus* host geographic distributions. Tip-dated Bayesian phylogeny of RBP region 5 showing the 9 closest relatives to SARS-CoV-2 **(A)**. Tree nodes have been adjusted to the mean age estimates and posterior distributions are shown for each node with mean age estimate and 95% HPD confidence intervals presented to their left. Tips are annotated with the host species they were sampled from, bat silhouette colours correspond to panel B. Geographic ranges of *Rhinolophus* species the SARS-CoV-2 closest relatives have been sampled in **(B)**. Maps are restricted to East Asia and separated into province-level within China and country-level outside China.

The wide geographic ranges of *R. pusillus* and *R. affinis* and the fact that two of the closest known relatives of SARS-CoV-2, RpYN06 and RaTG13, have been sampled in these species flags them as prime suspects for the source of the SARS-CoV-2's progenitor in China. Additionally, these two bat species are found in shared roosts with *R. sinicus* and *R. ferrumequinum* in Yunnan and with *R. sinicus* in Guangxi (Luo

et al., 2013), providing opportunities for host switches, co-infections and thus recombination between the sarbecoviruses that these bat species carry. *R. pusillus* and *R. affinis* also link more regions of China with bat species such as *R. shameli*, *R. malayanus* and *R. acuminatus* which are only found in Southeast Asia and southwest of China (Figure 2.4B). Latinne *et al.* (2020) published a large-scale sampling expedition of coronaviruses across bats in China. Despite there only being short RdRp sequence fragments available, the phylogeny for the novel viruses revealed a cluster of seven identical sarbecovirus sequences sampled from *R. affinis* within the nCoV clade (Appends A, Figure A.2). Still, the fact that viruses in the Yunnan clade (consisting of RmYN02, RpYN06 and PrC31) were sampled from three different *Rhinolophus* species supports the hypothesis that these viruses readily infect multiple different horseshoe bat species with overlapping geographical ranges.

Based on the analysis of the sarbecovirus and host data presented here, we propose that to locate the SARS-CoV-2 progenitor sampling should focus on the ranges of horseshoe bat host populations known to harbour nCoV viruses. Specifically, samples should be collected in roosting environments spread across China with care taken both to avoid a further spillover (or reverse zoonosis) and to protect the bat populations (Luo *et al.*, 2013). Sampling strategies will also need to consider the distinct subspecies of *Rhinolophus* as the delineators of genetically meaningful host populations for coronaviruses, for example, there are two *R. affinis* sub-species on mainland China: *himalayanus* and *macrurus* (Mao *et al.*, 2010). Future sampling should also encompass a range of indigenous mammals other than bats that we now know can be infected by these coronaviruses. Although highly endangered, Chinese pangolins, given their susceptibility to infection and their geographical range across southern China (Challender *et al.*, 2019), could be one of the possible candidates for the “missing” intermediate host of the SARS-CoV-2 proximal ancestor (World Health Organisation, 2021).

2.3.4 The importance of recombination in human-circulating SARS-CoV-2

Recombination is a key process in the evolution of SARSr-CoVs, however the sparse sampling of these bat viruses only provides snapshots of the true

recombination events that led to these mosaic genomes. The immense effort of SARS-CoV-2 sequencing during the COVID-19 pandemic allows us to observe recombination in action between variants of the virus, for example in the case of the globally circulating XBB lineage. To trace the recombination event that led to the emergence of the XBB variant, I retrieved all SARS-CoV-2 sequences deposited to GISAID (as of October 3, 2022) with PANGO lineage designation matching BJ.1, BM.1, XBB, and all their descendant lineages (including BM.1.1, BM.1.1.1, and XBB.1). Recombination analysis on the aligned set of sequences, using a number of independent recombination detection methods implemented in RDP5 (Martin *et al.*, 2021) robustly identified a single recombination breakpoint unique to all XBB sequences at genomic position 22,920 (matching the Wuhan-Hu-1 reference genome) (Figure 2.5). No evidence of recombination was found in the BJ.1 and BM.1 sequences in the dataset. Consistent with the result of the RDP5 analysis, visual inspection of the nucleotide differences between the consensus sequences of XBB, BJ.1, and BM.1 (including BM.1.1 and BM.1.1.1) clearly illustrated that the identity of XBB to BJ.1 ends at genome position 22,942, and the identity of XBB to BM.1 starts after position 22,896 (Figure 2.5). Together, the analysis suggests that the recombination breakpoint is between positions 22,897 and 22,941, within the receptor binding domain (RBD) of the Spike protein (corresponding to amino acid positions 445-460) (Figure 2.5).

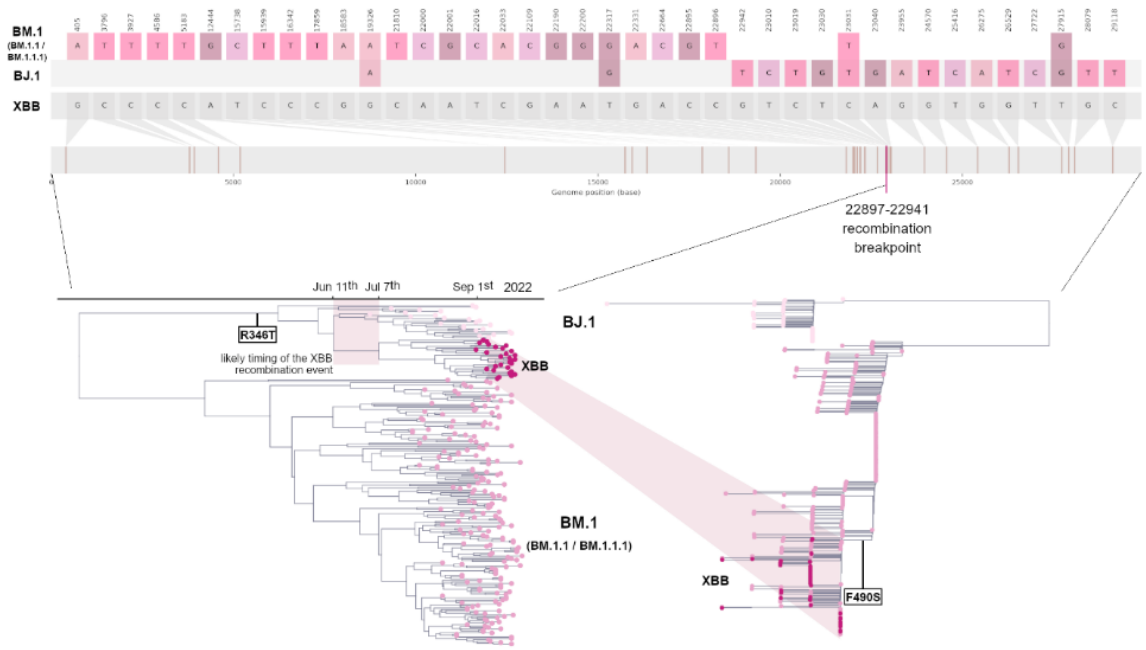


Figure 2.5. Recombination event leading to the SARS-CoV-2 XBB variant. Top: Nucleotide differences between the consensus sequences of the BJ.1, BM.1 (including BM.1.1/BM.1.1.1)

Figure 2.5 (cont). lineages and the XBB (including XBB.1) lineage, visualized with snipit (https://github.com/aineniamh/snipit). Bottom: Maximum clade credibility time-calibrated phylogeny of the 5' non-recombinant segment (1–22,920) of the XBB variant (left) and non-calibrated maximum likelihood phylogeny of the 3' non-recombinant segment (22,920–29,903) (right). The right hand-side tree is rooted on a BA.2 outgroup (not shown).

I then split the sequence alignment at position 22,920 to determine the evolutionary history of each nonrecombinant segment of the XBB genomes. The phylogenetic reconstructions recapitulate the recombination results, with the 5' end major parental sequence being derived from the BJ.1 clade and the 3' end minor parental sequence from the BM.1.1.1 clade (Figure 2.5). Using the longer 5' end non-recombinant part of these genomes, I estimated the emergence date of XBB using Bayesian tip-dated phylogenetic inference (Figure 2.5). The analysis suggests that the XBB clade's most recent common ancestor (MRCA) existed at the start of July 2022 (median posterior date: July 7, 2022; 95% HPD confidence intervals: from June 10, 2022, to July 29, 2022). Furthermore, the MRCA between the XBB and BJ.1 lineages existed at the start of June 2022 (median posterior date: June 11, 2022; 95% HPD intervals: from May 22, 2022, to June 26, 2022) (Figure 2.5). Together, our analyses suggest that XBB emerged through the recombination of two cocirculating lineages, BJ.1 and BM.1.1.1, during the summer of 2022.

2.4 Discussion

The currently available data illustrate a complex reticulate evolutionary history in the lineage of sarbecoviruses SARS-CoV-2 emerged from. This history is influenced by co-circulation of related coronaviruses, over at least the last 100 years, across the bat populations in southern China, and into Southeast Asia with multiple recombination events imprinted on the genomes of these viruses. Considering the high frequency of recombination, it is expected that selection could preferentially favour exchanges of specific genomic regions, in line with our detection of hotspots near the Spike gene (Figure 2.1B,C). The functional implications of selective Spike recombination has recently been corroborated by multiple independent studies, suggesting this might be a mechanism for antigenic shift utilised by the sarbecoviruses or, more broadly, by all coronavirus groups (Bobay, O'Donnell and Ochman, 2020; Nikolaidis *et al.*, 2021; Yang *et al.*, 2021; de Klerk *et al.*, 2022; Goldstein *et al.*, 2022). The analysis further illustrates the importance of accounting for recombination rather than using whole-genome pairwise similarity to determine the shared evolutionary history of these viruses. This is exemplified by RaTG13 which is often described as the "on average" closest sarbecovirus to SARS-CoV-2 despite not being the phylogenetically closest virus once recombination history is accounted for in the other nCoV sarbecoviruses (Figures 2.1A, 2.3A).

The evidence of recombination events between members of the *Sarbecovirus* subgenus sampled in different geographical regions and from different bat hosts, indicates recent extensive movement of the viruses between different regions and species (and sub-species) as a result of the contacts between different bat populations that carry them. Although the closest known relatives of SARS-CoV-2 were sampled in Yunnan, the location of the proximal viral population SARS-CoV-2 emerged from remains unknown. The recombination patterns detected within the nCoV genomes imply the existence of one or several primary reservoir hosts with a geographical range spanning Thailand from the Southwest and Zhejiang to the East, a distribution that is consistent with specific Chinese horseshoe bats acting as the primary reservoir hosts. Our observations are further confirmed by recent report of more bat coronaviruses very closely related to SARS-CoV-2 sampled from *R. pusillus* and *R. malayanus* in Laos (Temmam *et al.*, 2022). Both the sampling location and host species are consistent with expectations based on our analysis, essentially filling in the geographic gap between previous nCoV sampling locations.

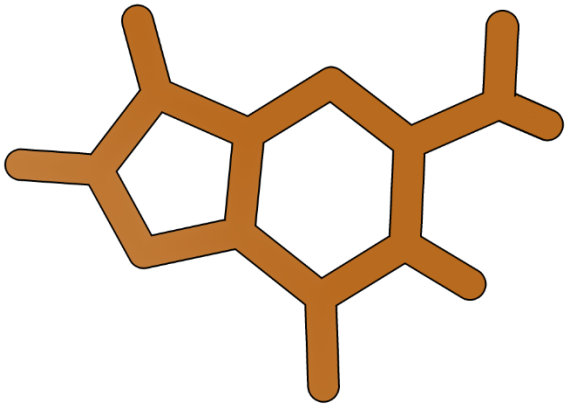
The recombination patterns reported in these newly discovered genomes are also consistent with the extensive recombination reported here (Temmam *et al.*, 2022). Having presented evidence in support of *R. affinis* and *R. pusillus*'s potential significance as the reservoir species, it would be remiss not to note that at least 20 different *Rhinolophus* species are distributed across China (four of which are endemic to China), many of which have not yet been found hosting nCoVs. The generalist nature of *Sarbecoviruses* also means multiple wild or farmed animals (e.g., American mink (*Neovison vison*) both farmed for fur and used as a food source) (World Health Organisation, 2021; Xia *et al.*, 2021; Xiao *et al.*, 2021) could have facilitated transmission of SARS-CoV-2 from bats to humans.

While SARS-like antibodies detected in people from rural communities in China (Wang *et al.*, 2018; Li *et al.*, 2019) indicates an intermediate animal species is potentially not required for transmission to humans, it does seem that emergence in a populated area is required for significant outbreaks to occur. The association of both SARS-CoV and SARS-CoV-2 with animal markets suggests animal trafficking is a key part of this transmission to humans. Human-mediated animal movement increases contact with sarbecovirus infected animals (whether they are susceptible species that have been trapped or farmed in rural locations; (Xia *et al.*, 2021) and subsequently introduces them into city markets (Lytras *et al.*, 2021; World Health Organisation, 2021; Worobey, 2021). An urgent question relating to the prevention of another emergence, is: i) where in China or Southeast Asia is the SARS-CoV-2 progenitor located (our analysis shows this is not necessarily Yunnan); ii) which bat or other animal species are harbouring nCoV sarbecoviruses and iii) what is the risk of future spillover events? There is undoubtedly a virus highly related to SARS-CoV-2 still present somewhere in the wild. The best we can do is maximize the probability that future sampling efforts will uncover that host species or sub-species.

The observation of extensive recombination among SARS-CoV-2's closest known relatives foreshadowed that this process would be eventually involved in SARS-CoV-2's evolution in humans, given enough co-circulating diversity. Indeed, the emergence of XBB confirmed this prediction, with the dominant lineage circulating for almost a year now being a result of recombination. Analysis of the *Sarbecovirus* genomes indicates that recombination takes place around the Spike open reading frame, likely selectively swapping this gene between strains. However, the XBB event seems slightly different, with the breakpoint being within the Spike S1 subunit

encoding region of the genome (Figure 2.5). Experimental work indicates that by combining the Spike genes of its parental strains, XBB acquired an advantageous constellation of Spike substitutions making it more immune evasive and better at ACE2 binding than its progenitors (Tamura *et al.*, 2023; Wang *et al.*, 2023; Yue *et al.*, 2023). Hence, in the case of XBB, recombination acted as a way to quickly accumulate a set of advantageous substitutions, more efficiently than through stepwise mutation. This mechanism is distinct to antigenic shift through entire Spike swapping that likely happens in the horseshoe bat hosts, and probably comes down to the amount of co-circulating diversity of the viruses. The SARSr-CoVs co-circulating in the same bat hosts are much more diverse than SARS-CoV-2 variants co-circulating in the human population. The Delta and Omicron variants of concern did co-circulate for a brief period of time leading to a few detected inter-VOC recombinants (Arora *et al.*, 2022; Colson *et al.*, 2022; Lacek *et al.*, 2022; Wang *et al.*, 2022), but none of these exhibited any observable fitness advantage and became dominant in the population. This suggests that compatibility between co-circulating viruses – the polymerase of one virus being physically capable of replicating the co-infecting virus's genome, or the recombinant genome being viable for further replication – is also key to recombination having an impact in the future evolution of SARS-CoV-2. Given enough diversity, recombination-mediated antigenic shift may play a role in SARS-CoV-2's future evolution whether that is between SARS-CoV-2 variants, or between SARS-CoV-2 and a potential novel zoonotic *Sarbecovirus* introduced to humans. This is why it is necessary to maintain genomic monitoring of the circulating SARS-CoV-2 diversity, tracking future recombination events in this virus's evolution.

Chapter 3. Resurrecting the antiviral activity of the ancient horseshoe bat OAS1 protein



Cartoon model of guanine.

PDB ligand entry: GUN, visualised with ChimeraX.

“A totally blind process can by definition lead to anything.”

Jacques Monod, Chance and Necessity (1970)

The Results subsections 3.3.1 "*An ancient retrotransposition event ablated OAS1 prenylation in horseshoe bats*" and 3.3.2 "*No known Rhinolophoidea-infecting CoVs encode PDEs*" (Figure 3.1) in this chapter are my own work and have been published in Wickenhagen et al. (2021, Science, 374[6567]: abj3624). The published paper includes work conducted by co-authors which is referenced in the text to provide essential context for my own work. The remaining Results subsections of this chapter are currently under review as a separate manuscript. All experimental work presented in subsection 3.3.4 "*Restored anti-SARS-CoV-2 activity in the ancestral OAS1 protein*" and Figure 3.3A,B (cell line modifications, SARS-CoV-2 infections, plaque assays and CPE assays) have been performed by Arthur Wickenhagen, Elena Sugrue and Emma L Davies. The methods for this work can be found in Appendix B Text B.1.

Aim

The human 2′-5′-oligoadenylate synthetase 1 (OAS1) protein has been shown to protect individuals infected with SARS-CoV-2 from severe disease. In this section, I explore the evolutionary history of the horseshoe bat OAS1 orthologue and how an ancient change in this gene might have played a role in enabling the interaction between SARS-related coronaviruses' with horseshoe bat hosts.

3.1 Introduction

3.1.1 Human OAS1 is an antiviral factor against SARS-CoV-2

Multiple barriers prevent viruses from infecting new species and these barriers can also hamper the ability of a virus to thrive in an unfamiliar host. Alternatively, changes in the host genes responsible for these barriers may enable some viruses to diversify in these host lineages. Functional engagement of dependency factors (such as viral receptors) is required for successful spillover, but viruses must also navigate a myriad of host-specific immune defences to sustain infection and transmission in a new host. The main prenylated form of the human 2′-5′-oligoadenylate synthetase 1 (OAS1, isoform p46) protein has been shown to have potent antiviral activity against SARS-CoV-2 and its expression correlates with less severe disease in humans (Soveg *et al.*, 2021; S. Zhou *et al.*, 2021; Wickenhagen *et al.*, 2021; Zeberg and Pääbo, 2021; Huffman *et al.*, 2022). The OASs are interferon-stimulated genes (ISGs) and their expression levels are commonly increased during interferon (IFN) mediated antiviral responses. Most OASs sense double-stranded viral RNA, and this frequently activates the synthesis of 2′-5′-linked oligoadenylates (2-5A). 2-5A induces the dimerization of inactive RNase L, which upon activation mediates the indiscriminate cleavage of viral and host RNAs, leading to inhibition of viral replication (Sadler and Williams, 2008; Hornung *et al.*, 2014). The mammalian OAS family includes three catalytically active members (OAS1, OAS2 and OAS3) (J. Hu *et al.*, 2018) that seem to be under co-adaptive evolutionary pressure with its interacting proteins (e.g., RNase L) in both primates and chiroptera (Mozzi *et al.*, 2015).

For the OAS1 protein to recognise the virus and initiate its inhibition, it needs to be in contact with the viral RNA. Coronaviruses hijack the endoplasmic reticulum of infected cells, creating double-membrane vesicles (DMVs) in which virus replication takes place (Knoops *et al.*, 2008; V'kovski *et al.*, 2020). Similar replication mechanisms are deployed by the majority of positive-sense single-stranded RNA (+ssRNA) viruses (Wolff *et al.*, 2020). This intracellular compartmentalisation of the viral genetic material may be a form of immune evasion and suggests that host antiviral proteins also need to be localised in (or near) the DMVs to target the virus. The p46 isoform of human OAS1 contains a CAAX-box motif at its C-terminal end which acts as a prenylation signal for the protein. The prenylation post-translational

modification targets the protein to membranes in close proximity to where SARS-CoV-2 replicates thereby enabling antiviral activity. In contrast, the non-prenylated p42 isoform does not restrict SARS-CoV-2 replication (Soveg *et al.*, 2021; Wickenhagen *et al.*, 2021). Similar post-translational modification signals seem to be required by other RNA-binding ISG proteins for sensing RNA viruses replicating in cytoplasmic compartments and inhibiting their replication (Kmiec *et al.*, 2021).

3.1.2 Phosphodiesterase-encoding genes in coronaviruses

Phosphodiesterases (PDEs) are a superfamily of enzymes that primarily cleave second messenger molecules bound by phosphodiester bonds, e.g. cyclical adenosine monophosphate (cAMP) (Jeon *et al.*, 2005). As mentioned above, the OAS-Rnase L pathway involves the synthesis of 2-5A molecules which are also bound by phosphodiester bonds between their 2’ and 5’ carbon atoms. Murine hepatitis virus (MHV), a coronavirus widely used as a model system, was the first coronavirus found to encode a 2’-5’ PDE enzyme that can antagonise the OAS system by degrading the 2-5A molecules required for RNase L activation and subsequent virus inhibition (Zhao *et al.*, 2012). This PDE-encoding gene, called NS2, is also found in a number of lineage A *Betacoronaviruses*, including the human seasonal coronavirus OC43 (Goldstein *et al.*, 2017) which is also not inhibited by human OAS1 (Wickenhagen *et al.*, 2021). Functional homology of the coronavirus NS2 2’,5’ PDE domain has also been found in the group A rotavirus VP3 protein and the murine A kinase anchoring protein 7 (AKAP7), presence of either seems to facilitate MHV replication *in vitro* (Zhang *et al.*, 2013; Gusho *et al.*, 2014). This observation suggests that lineage A *Betacoronaviruses* likely acquired their PDE-encoding gene from their host genome through an ancestral insertion, allowing them to antagonise the OAS – RNAse L antiviral pathway. The human coronavirus OC43 likely originated in a murine host (Lau *et al.*, 2015) and entered human populations through a cross-species transmission from cattle (Vijgen *et al.*, 2005). At least one paralogue of both mouse and cow OAS1 orthologues (murine OAS1a and bovine OAS1Y) are prenylated and confer potent anti–SARS-CoV-2 activity (Wickenhagen *et al.*, 2021), confirming that the NS2 gene likely evolved in these viruses in the presence an active antiviral OAS, possibly OAS1.

Interestingly, the Middle East respiratory syndrome-related coronavirus (MERS-CoV), a lineage C *Betacoronavirus* that emerged in humans in 2012 (Zaki *et al.*, 2012), also encodes a PDE (NS4b) capable of antagonising the OAS-RNase L system (Thornbrough *et al.*, 2016). However, the origin of NS4b seems to be independent of the lineage A NS2 gene, despite both coronavirus proteins having a similar role in antagonising host immunity. MERS-CoV entered human populations after transmission from dromedary camels (*Camelus dromedarius*) (Reusken *et al.*, 2013; Briese *et al.*, 2014). The NS4b gene is also found in only a handful of known MERS-CoV-related lineage C *Betacoronaviruses*, including HKU4, HKU5, NeoCoV and PDF-2180, all sampled in bats of the Vespertilionoidea superfamily (*Tylonycteris pachypus*, *Pipistrellus abramus*, *Neoromicia capensis* and *Pipistrellus cf. hesperidus* respectively) (Woo *et al.*, 2006; Corman *et al.*, 2014; Anthony *et al.*, 2017). Accordingly, OAS1 from *Pipistrellus kuhlii* (a Vespertilionoidea bat species) and *C. dromedarius* have CAAX-box motifs and block SARS-CoV-2 replication *in vitro* (Wickenhagen *et al.*, 2021). Hence, it seems that at least two *Betacoronavirus* lineages have independently acquired PDE genes that antagonise the antiviral function of OAS1 through 2-5A degradation. The reservoir hosts of both virus lineages have active OAS–RNAse L pathways that target coronavirus replications, suggesting that PDE acquisition is a means for these viruses to counteract OAS1-dependent antiviral activity.

3.2 Methods

3.2.1 Synteny analysis

The Ensembl web database was used for assessing the OAS1 locus genome synteny between the human genome (GRCh38.p13) and the available *Rhinolophus* species, *R. ferrumequinum*, genome (mRhiFer1_v1.p). The syntenic region between the human OAS1 exon 7 (ENSE00003913305) and the horseshoe bat genome was examined to identify a region in the latter genome sequence starting at position 7,833,728 of scaffold 25 of the mRhiFer1_v1.p primary assembly that lacked synteny to the human genome. Incidentally, the non-syntenic region started in-frame where the p46 “CTIL” encoding human sequence would have been. I extracted the 580 bp *R. ferrumequinum* sequence span up to where synteny resumes to the human genome and used hmmscan (HMMER 3.2.1) (Eddy, 2009) to search against the Dfam database (Hubley *et al.*, 2016) for transposable elements present in the sequence. Two confident matches were identified, one to a partial MER74A-like LTR element at the very start of the non-syntenic sequence and one to a L1-like retrotransposon element at the 3′-end of the sequence.

3.2.2 *In silico* genome screening

To explore how far back in time this LTR insertion at the OAS1 locus took place, I used the Database-Integrated Genome-Screening (DIGS) software (H. Zhu *et al.*, 2018). DIGS uses a nucleotide or amino acid sequence probe to perform a BLAST similarity search through genome assemblies. I collected a set of 44 Chiroptera species genome assemblies to perform three *in silico* screens.

I first used the nucleotide sequence of the syntenic region of *R. ferrumequinum* to human exon 7 (Ensembl) and the adjacent 580 bp region with the detected LTR insertion until homology resumes to the human genome as a probe. The DIGS screen was conducted using a minimum blastn bitscore of 30 and minimum sequence length of 30 nucleotides. Matches were aligned using MAFFT v7.453 (Katoh and Standley, 2013) and inspected for covering all regions of the probe. The second screen used the CAAX terminal amino acid sequence homologous to that encoded by the human exon 7 of 5 previously annotated bat OAS1 proteins holding

a CAAX-box terminus. The minimum tblastn bitscore was set to 60 and the minimum sequence length to 40 nucleotides. The translated sequences of hits were aligned, and only hits with a CAAX domain present were retained. Finally, to cross-validate that sequence hits of the CAAX domain search are part of the OAS1 locus, a screen using the *R. ferrumequinum* OAS1 C-terminal domain amino acid sequence as a probe was performed, with a minimum tblastn bitscore of 100 and minimum sequence length of 100 nucleotides. Only hits of the CAAX search found on scaffolds with a detected OAS1 locus in the third search were maintained in the analysis. It is worth noting that the lack of an OAS1 domain detected on the same scaffold as hits in the CAAX sequence search is most likely a result of low genome assembly quality. Regardless, the hits were excluded for clarity.

3.2.3 PDE analysis

To examine the diversity of PDE proteins encoded by coronaviruses, I first constructed an HMMER protein profile. Two seemingly independently acquired PDEs are encoded by the NS2 of Embecoviruses (Zhao *et al.*, 2012) and NS4b of MERS-like coronaviruses (Thornbrough *et al.*, 2016), respectively. Group A rotavirus (RVA) has also been described to encode a protein with a homologous PDE domain and similar biological function (Zhang *et al.*, 2013). Finally, the AKAP7 mammalian protein holds a PDE domain that has been experimentally shown to complement the function of murine coronaviruses' NS2 activity (Gusho *et al.*, 2014). I aligned the amino acid sequence of the PDE domains of the OC43 NS2 (AAT84352.1), the MERS NS4b (AIA22866.1), and the NS4b proteins of two more bat Merbecoviruses HKU5 (YP_001039965.1) and SC2013 (AHY61340.1), the AKAP7 proteins of *Rattus norvegicus* (NP_001001801.1), *Mus musculus* (NP_001366167.1), and humans (NP_057461.2) (as their homology to CoV PDEs has been previously characterized) and the Rotavirus A VP3 protein (AKD32168.1). The alignment was then manually curated using Bioedit on the basis of the homology described in the literature. The final alignment was used to produce a hidden Markov model (HMM) profile using the HMMER suite (v3.2.1) (Eddy, 2009).

All complete *Coronaviridae* sequences were downloaded from the NCBI virus online database as of the 15th of April 2021 (Hatcher *et al.*, 2017). Only sequences with an

annotated host and length above 25,000 bp were retained and viruses of “severe acute respiratory syndrome-related coronavirus” species with a human host were excluded, producing a dataset of 2042 complete or near-complete coronavirus genomes. The EMBOSS getorf program was used to extract the translated sequences of all methionine starting ORFs with length >100 nucleotides from the filtered virus genome dataset. All putative ORFs were then screened against our custom PDE HMM profile using hmmscan (Eddy, 2009).

3.2.4 Retrieval of bat OAS1 proteins

To find annotated bat OAS1 sequences protein BLAST (Camacho *et al.*, 2009) was used with the *R. ferrumequinum* OAS1 protein (XP_032953023.1) as the query sequence. The search was restricted to the Chiroptera order and after manual examination of the pairwise alignments, the OAS1 protein sequences of 16 bat species (including *R. ferrumequinum*) were retrieved.

From the Rhinolophoidea superfamily, only *R. ferrumequinum* and *Hipposideros armiger* have annotated OAS1 protein sequences available in NCBI Genbank. To increase the phylogenetic resolution of this clade, I retrieved the contigs from the *Megaderma lyra* and *R. sinicus* genomic assemblies (PVJL010007185.1, NW_017739019.1) that are syntenic to the *R. ferrumequinum* OAS1 locus as identified by the DIGS search described in subsection 3.2.2 and used AUGUSTUS (Stanke *et al.*, 2008) to predict the respective OAS1 coding sequences (human version with default transition matrix). Sequence predictions were aligned to the *R. ferrumequinum* OAS1 sequence using MAFFT v7.453 (Katoh and Standley, 2013) and one sequence was selected for each species, based on highest transcript similarity to the *R. ferrumequinum* OAS1 protein (XP_032953023.1).

3.2.5 Ancestral sequence reconstruction

The resulting 18 bat OAS1 protein sequences were aligned with MAFFT (--genafpair option) (Katoh and Standley, 2013) and, in order to avoid low-information sites in the alignment biasing the phylogenetic reconstruction, N- and C-terminal ends not shared by the majority of sequences were trimmed off.

IQ-TREE (version 1.6.1) (Nguyen *et al.*, 2015) was used for the ancestral sequence reconstruction (*-asr*) of the final protein alignment under a LG+I+G4 model, selected by ModelFinder (Kalyaanamoorthy *et al.*, 2017). The reconstructed phylogeny's topology was informed by the species tree of the corresponding bat species retrieved from TimeTree (Kumar *et al.*, 2017), using iqtree's *-te* option.

The IQ-TREE output was used to reconstruct the OAS1 sequence preceding the LTR insertion in the Rhinolophoidea common ancestor, i.e. the node of the phylogeny connecting the Rhinolophoidea and the Pteropodoidea superfamilies. The Rhinolophoidea common ancestor (RhinoCA) sequence was reconstructed using the residue with the highest posterior probability for each site. A second version of the sequence, RhinoCA-T70, was reconstructed by replacing all sites where no residue state had a posterior probability above 0.7 with the corresponding *P. alecto* residue. Since gaps in the alignment provide no information for the site-by-site ancestral reconstruction, the variable indel region corresponding to *R. ferrumequinum* OAS1 positions 159 to 163 was replaced with the *P. alecto* insertion in this region (*P. alecto* OAS1 positions 159-173 – PRSYYSDSQIHEDYR) for both RhinoCA and RhinoCA-T70. Similarly, the *P. alecto* C-terminal end was appended at the C-terminal end of both reconstructed sequences (*P. alecto* OAS1 positions 357-372 – PYDTPHVEEDQWCAIL). Positions in the alignment where residues (rather than gaps) were present in only one out of the 18 bat OAS1 sequences were removed from the reconstructions.

The entropy value for each site in the Chiroptera OAS1 protein alignment shown in Figure 1A was calculated using Shannon's entropy formula with a natural log as implemented in Bioedit (Hall, 1999) ($H(l) = -\Sigma f(a,l)\ln(f(a,l))$; $f(a,l)$ being the frequency of amino acid a at position l).

3.2.6 Selection analysis

The trimmed amino acid alignment of the 18 bat OAS1 proteins was converted to its corresponding coding sequence alignment using PAL2NAL (Suyama, Torrents and Bork, 2006). To exclude potentially non-homologous sites before performing selection analysis, the variable indel region (highlighted in Figure 3.2A) was removed from the alignment. The final codon alignment contained 351 out of the

366 codon sites in the original alignment. The phylogeny was reconstructed again using the gap-free codon alignment in the same way described above (-asr -te) under a GTR+I+F+G4 substitution model using IQ-TREE (Nguyen *et al.*, 2015). The resulting phylogeny and alignment were used for performing a number of selection detection methods of the Hyphy package (v2.5.33) (Kosakovsky Pond *et al.*, 2019). RELAX (Wertheim *et al.*, 2015) was performed to detect potential signals of selection relaxation specific to all branches of the Rhinolophoidea clade and branch leading up to it. The adaptive Branch-Site Random Effects Likelihood (aBSREL) method (Smith *et al.*, 2015) was used to detect branch-specific episodic selection across all branches of the tree. To examine site-specific selection, the Fixed Effects Likelihood (FEL) and Mixed Effects Model of Evolution (MEME) methods (Kosakovsky Pond and Frost, 2005; Murrell *et al.*, 2012) were performed, each with 100 permutations of parametric bootstrapping, separately for the branches of the Rhinolophoidea clade and branch leading up to it and all other branches of the tree. Contrast-FEL (Kosakovsky Pond *et al.*, 2021) was performed to detect sites under different selective pressures in the Rhinolophoidea using the aforementioned branches as test and reference respectively.

3.2.7 Protein structure predictions

ColabFold (https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb) (Mirdita *et al.*, 2022), implementing MMseq2 (Steinegger and Söding, 2017) and AlphaFold2 (Jumper *et al.*, 2021), was used to predict the tertiary structure of the *R. ferrumequinum*, *P. alecto* and RhinoCA OAS1 proteins. ColabFold was performed under default parameters and the best ranked prediction was selected for each protein. The structures were visualised and superimposed onto the RNA-bound human OAS1 crystal structure (pdb: 4IG8) using ChimeraX (version 1.4) (Pettersen *et al.*, 2021).

3.2.8 Data availability

The data and code used for this Chapter are publicly available at the following GitHub repositories:
https://github.com/spyros-lytras/bat_OAS1,
https://github.com/spyros-lytras/ancient_bat_OAS1.

3.3 Results

3.3.1 An ancient retrotransposition event ablated OAS1 prenylation in horseshoe bats

Given the evidence that the prenylated OAS1 protein of humans and other mammals potently inhibits the replication of SARS-CoV-2 *in vitro* (Soveg *et al.*, 2021; Wickenhagen *et al.*, 2021), it is of interest to examine the OAS1 orthologue of horseshoe bats, the primary host of SARS-related coronaviruses. The only horseshoe bat species with an annotated genome in the Ensembl database is that of the greater horseshoe bat, *Rhinolophus ferrumequinum*. Even though there is an annotated OAS1 gene for this species, no transcript or exon encodes the prenylated form of the protein. Indeed, all Ensembl and NCBI database entries specified non-prenylated polypeptides. The prenylation signal (CAAX-box motif) of human OAS1 is encoded by the gene's exon 7. Thus, I examined the genome synteny between human exon 7 and the corresponding region of the *R. ferrumequinum* genome. Interestingly, the CAAX-box encoding region was completely absent from the horseshoe bat genome. Analysis of the syntenic region showed that a long terminal repeat (LTR) retrotransposon sequence was present in this part of the genome instead (Figure 3.1A). LTRs are mobile genetic elements that can integrate in different parts of the genome, similar to retrovirus integration (Bourque *et al.*, 2018). Hence, this LTR-dependent deletion of the CAAX-box encoding region of the *R. ferrumequinum* OAS1 gene should render this OAS1 protein unable to become prenylated.

Experimental work has shown that no *R. ferrumequinum* OAS1 encoded by annotated transcripts restricts SARS-CoV-2 replication when overexpressed *in vitro*, consistent with the hypothesis that prenylation is essential for the antiviral activity (Wickenhagen *et al.*, 2021). To examine how many bat species share this LTR-dependent deletion of the OAS1 prenylation motif, I performed an *in silico* genomic screen of 44 available bat genome sequences, searching for i) the LTR insertion, ii) the CAAX-box encoding sequence and iii) the core OAS1 sequence to ensure synteny. The LTR sequence was identified only in members of the Rhinolophoidea superfamily (including *Rhinolophus*, *Hipposideros*, and *Megaderma* species), indicating that this ancient retrotransposition insertion occurred ~58 to 52 million years ago at the base of this bat superfamily. By contrast, CAAX-box encoding

syntenic sequences could be detected in members of all other bat taxa (Figure 3.1B).

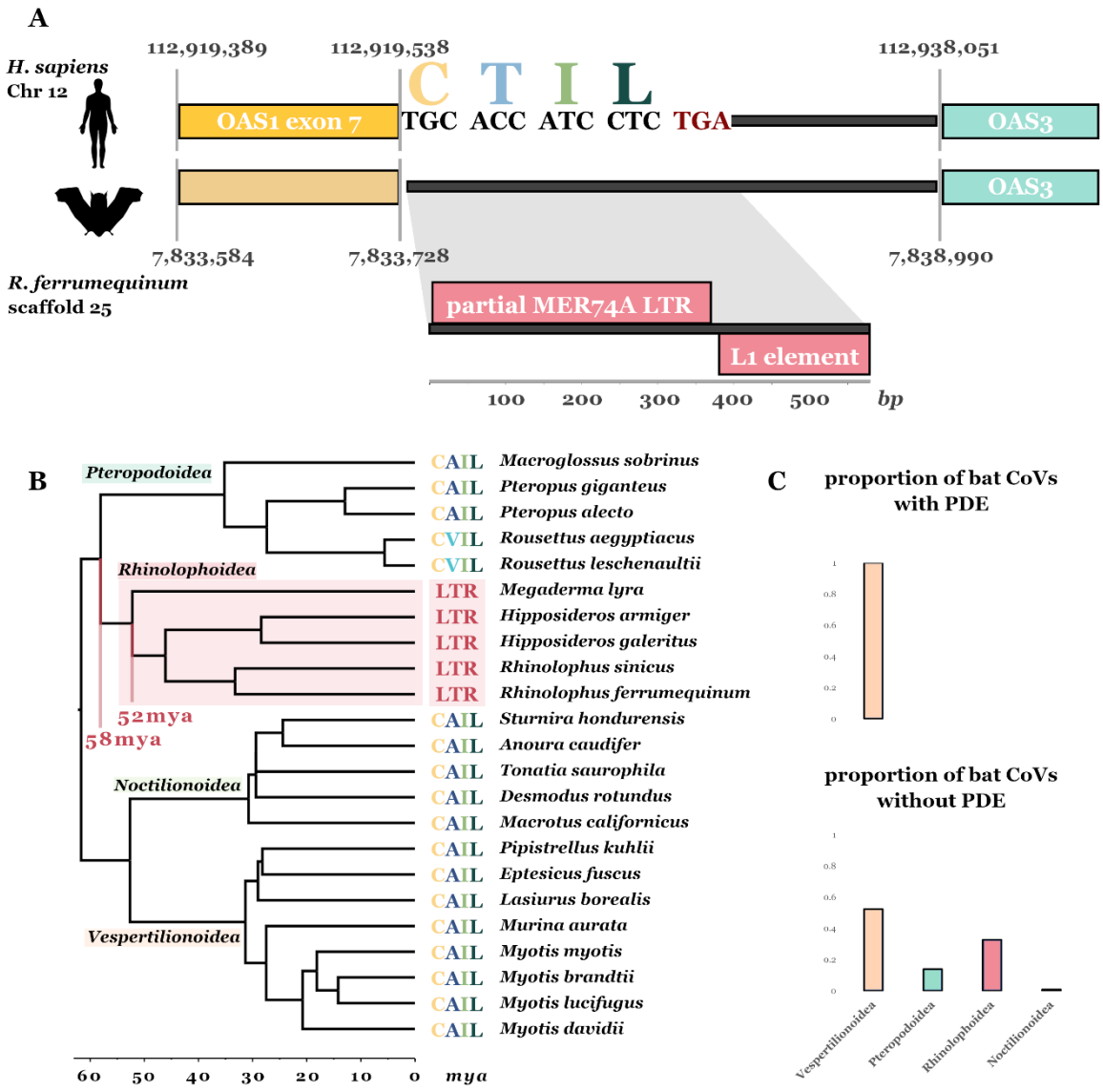


Figure 3.1. Retrotransposition at the OAS1 locus has ablated the CAAX-box prenylation signal in Rhinolophoidea. (A) Schematic of genome synteny between the human OAS1 exon 7 locus (yellow) and the *R. ferrumequinum* genome. The exact syntenic sequence coordinates are annotated for the start of OAS1 exon 7, the start of the CAAX box encoding sequence, and the start of the upstream gene locus, OAS3 (blue). Transposable element hits on the 580 bp non-syntenic region in the R. ferrumequinum genome are shown in the enlarged inset. Noncoding regions are shown in black. Note that the schematic is not to scale. **(B)** Dated phylogeny (retrieved from TimeTree; www.timetree.org) (Kumar *et al.*, 2017) of bat species with a confirmed LTR insertion in the OAS1 locus or a CAAX box–encoding sequence present in the same scaffold as their OAS1 locus. Clades are labelled by superfamily, species names, and CAAX sequence (or LTR) are annotated next to the tree tips. The approximate time period during which the LTR insertion took place is annotated in red. **(C)** Proportion of bat CoV genomes with (top) and without (bottom) PDE-encoding genes, grouped by their host superfamily.

3.3.2 No known Rhinolophoidea-infecting CoVs encode PDEs

Considering the lack of prenylated OAS1 in the Rhinolophoidea and the ability of coronavirus phosphodiesterases to antagonise this pattern recognition pathway, I investigated whether PDE-encoding coronaviruses infect horseshoe bats. Given the variability in coronavirus-encoded PDEs [NS4b in Merbecoviruses and NS2 in Embecoviruses (Zhao *et al.*, 2012; Thornbrough *et al.*, 2016)], I developed a custom HMM protein profile using NS4b, NS2, the mammalian PDE AKAP7 (Gusho *et al.*, 2014), and rotavirus A VP3 (Zhang *et al.*, 2013). I screened for PDEs through all putative open reading frames (ORFs) of all published *Coronaviridae* genomes. This method should capture previously unannotated or undescribed PDEs. Although the available sequence data set is likely biased by sampling, no PDEs could be identified in any known coronaviruses sampled in Rhinolophoidea hosts (Figure 3.1C). In fact, all the bat coronaviruses identified as encoding PDEs were infecting members of the Vespertilionoidea superfamily (in which prenylated OAS1 is intact, Figure 3.1B). It should be mentioned that the majority of virus genomes in this analysis were sampled in Vespertilionoidea hosts, followed by Rhinolophoidea and then Pteropodoidea, with very few sequences from Noctilionoidea-infecting viruses (Figure 3.1C). Although there is an appreciable absence of PDEs in coronaviruses that circulate in horseshoe bats, an absence of PDEs does not necessarily imply an absence of anti–CoV OAS proteins in the relevant host. Many potential strategies exist to evade or antagonize the OAS system (Drappier and Michiels, 2015), and no PDEs were identified in coronaviruses sampled from Pteropodoidea and Noctilionoidea bats.

3.3.3 The Rhinolophoidea common ancestor OAS1 protein

Our current understanding of the deeper phylogenetic relation of bats (order Chiroptera) splits them into two major suborders: the Yinpterochiroptera (including superfamilies Rhinolophoidea and Pteropodoidea) and the Yangochiroptera (including superfamilies Noctilionoidea and Verspertilionoidea) (Teeling *et al.*, 2005). The LTR insertion deleting the OAS1 prenylation signal is shared between all Rhinolophoidea members with available genome sequences. This means that the deletion and putative loss of OAS1 anti-SARS-CoV-2 function took place after the split between the Rhinolophoidea and Pteropodoidea superfamilies and prior to the diversification of the extant Rhinolophoidea species (Figure 3.2A). In addition to OAS1 of all other bat taxa having retained the prenylation signal, SARS-CoV-2 restriction has been confirmed *in vitro* using OAS1 from members of the Pteropodoidea (*Pteropus alecto*) and the Yangochiroptera (*Pipistrellus kuhlii*) as well as humans, camels, cows and mice (Wickenhagen *et al.*, 2021). Hence, the antiviral sensing of CoV dsRNA mediated by prenylated OAS1 is likely the ancestral phenotype, also shared by the deep ancestor of all Rhinolophoidea (Figure 3.2A), prior to the LTR insertion.

I retrieved a set of 18 Chiroptera OAS1 protein sequences – available from NCBI Genbank or reconstructed from whole genome assemblies in this study – and used a phylogenetic approach, informed by the Chiroptera species tree, to predict the sequence of the aforementioned Rhinolophoidea pre-insertion ancestor (RhinoCA) OAS1 protein. This approach requires annotated peptide products instead of genome assemblies, explaining the discrepancy in the number of bat species used here compared to the *in silico* genome screening described above. The method used here provides a posterior probability for each site, indicative of the confidence on the reconstructed state. The majority of sites were confidently predicted with a posterior probability above 0.95. As expected, sites with lower posteriors corresponded to variable positions on the Chiroptera alignment (Figure 3.2A).

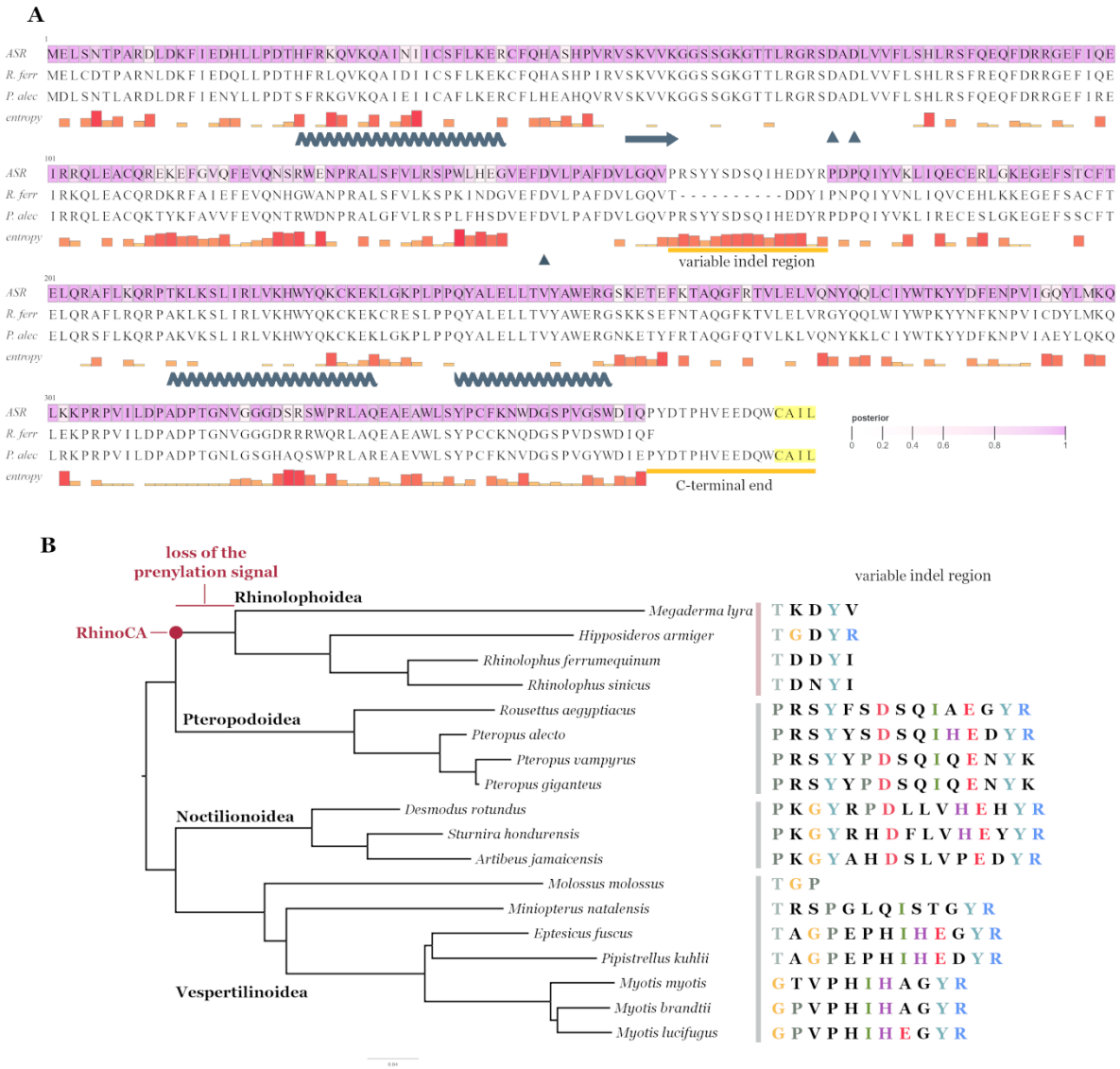


Figure 3.2. Ancestral state reconstruction of the RhinoCA OAS1 sequence. (A) Protein alignment of the RhinoCA (ASR), *R. ferrumequinum* (R. ferr) and *P. alecto* (P. alec) OAS1 sequences. RhinoCA sites are coloured by each predicted state's posterior probability. The bars on the bottom row indicate the Shannon's entropy of each site in the alignment of all 18 Chiroptera OAS1 proteins. Secondary structure alpha helices (zigzag) and beta sheets (arrows) involved in the protein/RNA interface and the active site triad residues D74/D75/D147 are annotated underneath the entropy row as described for the human OAS1 protein in Donovan, Dufner and Korennykh (2013). **(B)** Maximum likelihood phylogeny of the Chiroptera OAS1 proteins informed by their species tree topology. The ancestrally reconstructed (RhinoCA) node, the branch where the prenylation signal was deleted and the clades of each superfamily are annotated on the tree. The variable indel region sequence of OAS1 is shown on the right of each tip. Residues are coloured according to potential homology between the proteins.

Although most ancestral sequence reconstruction (ASR) methods are useful for predicting the state of single informative sites of internal nodes, indel variation in alignments can prove problematic in sequence reconstruction (Vialle, Tamuri and Goldman, 2018). By examining the OAS1 protein alignment a short region with distinct indel variation between bat taxa was identified, corresponding to *R.*

ferrumequinum OAS1 amino acid positions 159 to 163 (Figure 3.2A). Members of the Pteropodoidea and the Noctilionoidea have the longest variable region sharing clear homology, despite the two superfamilies not being monophyletic. This indicates that the longest genotype is likely the ancestral state of all bats, with the Vespirtilionoidea having undergone short deletions in the region (with the exception of the *Molossus molossus* OAS1) while Rhinolophoidea have lost most of this region (Figure 3.2B). To account for this region in the RhinoCA reconstruction I removed the site-by-site predicted segment and replaced it with the corresponding sequence of the *P. alecto* OAS1. This choice was based on: i) the longest region genotype likely being ancestral to all bats, ii) Pteropodoidea being the clade most closely related to the Rhinolophoidea and iii) having confirmed that the *P. alecto* OAS1 restricts SARS-CoV-2 *in vitro* (suggesting this region was unlikely to negatively impact reconstructed OAS1 antiviral activity).

Similarly, the C-terminal end of the RhinoCA OAS1 could not be reconstructed, since most of the region is deleted in the Rhinolophoidea species and there is high indel variation between the rest of the bat OAS1s. To match the variable indel region insertion, the C-terminal end of the *P. alecto* OAS1 (known to initiate a block to SARS-CoV-2) was appended to the reconstructed sequence to complete the RhinoCA OAS1 sequence. Since some sites of the RhinoCA ASR were not confidently predicted (low state posterior), I also implemented an alternative strategy where all sites with a posterior below 0.7 were replaced with the corresponding residues of the *P. alecto* OAS1 sequence. This alternative ancestral sequence reconstruction is referred to as RhinoCA-T70.

3.3.4 Restored anti-SARS-CoV-2 activity in the ancestral OAS1 protein

After reconstructing the RhinoCA and RhinoCA-T70 OAS1 sequences we tested whether exogenous expression of these proteins could initiate a block to SARS-CoV-2 replication. Expression of the ancestral RhinoCA OAS1 in A549-ACE2-TMPRSS2 cells (Rihn *et al.*, 2021) potently inhibited SARS-CoV-2, resulting in more than a 4-log reduction in virus titre (Figure 3.3A) and a 90% reduction of cytopathic-effect-(CPE)-induced well clearance (Rihn *et al.*, 2021) (Figure 3.3B) compared to cells expressing an RFP control. This antiviral phenotype is equivalent to that of the *P. alecto* OAS1 (Figure 3.3A,B) as well as of *P. kuhlii* and human p46 OAS1 proteins (Wickenhagen *et al.*, 2021). On the contrary, overexpressing the *R. ferrumequinum* OAS1 sequence had no effect on virus replication (Wickenhagen *et al.*, 2021). Similarly, ablating the prenylation signal of the RhinoCA OAS1 protein with a single amino acid change (CAAX to AAAX) effectively ablates antiviral activity and rescued virus replication, confirming that prenylation is essential for the antiviral activity (Figure 3.3A,B).

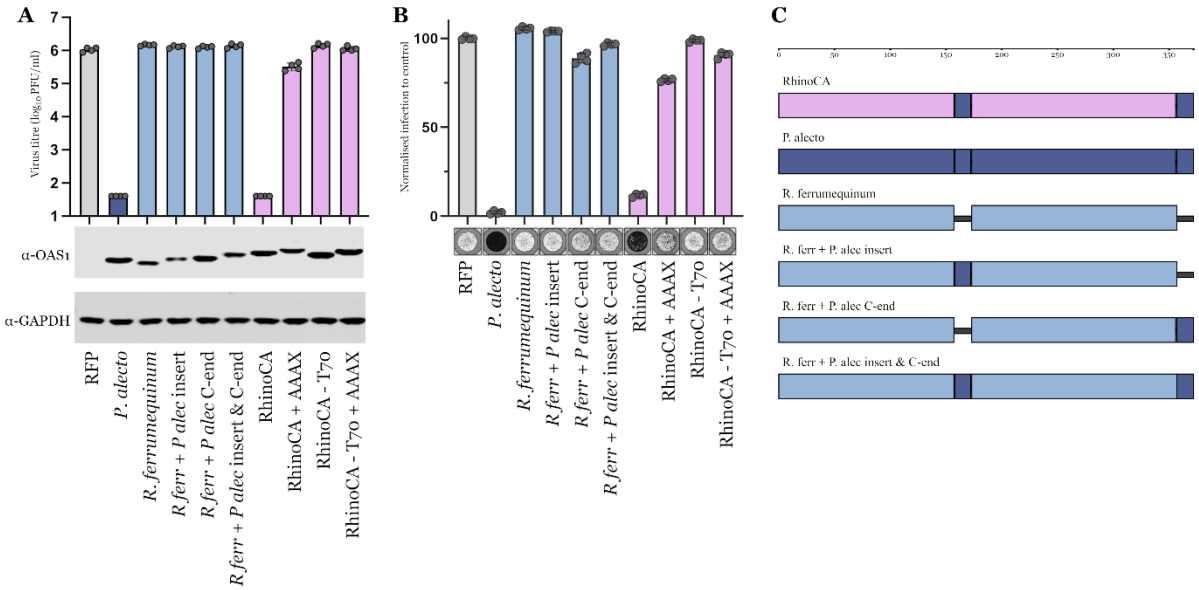


Figure 3.3. RhinoCA restricts SARS-CoV-2 replication. (A) Infectious titers of SARS-CoV-2 (PFU/ml) determined on AAT cells (A549-ACE2-TMPRSS2) modified to expressing bat OAS1 proteins (*P. alecto*, *R. ferrumequinum*), their specified derivatives and the ancestrally reconstructed RhinoCA and RhinoCA-T70 sequences. Controls for the ASR genes include identical sequences with ablated prenylation motifs (CAAX terminal end changed to AAAX). OAS1 expression was monitored in parallel by Western blot. **(B)** SARS-CoV-2 infection on AAT cells expressing exogenous OAS1 constructs as in A (based on well clearance caused by cytopathic effects of virus-replication). Infection normalized to RFP control and a typical picture of virus-induced CPE is shown below each graph. **(C)** Illustration of bat OAS1 recombinant constructs tested in this study.

I questioned whether recapitulation of the antiviral phenotype was simply due to the addition of the *P. alecto* prenylated C-terminal end or the insertion at the gene's internal variable region, rather than the ancestral site reconstruction. To test this, recombinant OAS1 sequences of the *R. ferrumequinum* protein with i) only the prenylated *P. alecto* C-terminal end, ii) only the *P. alecto* variable region insertion or iii) both the insertion and C-terminal end were created (Figure 3.3C). None of the recombinant genotypes restored virus inhibition in the *R. ferrumequinum* protein (Figure 3.3A,B), consistent with the hypothesis that the ancestral loss of the prenylation site was followed by divergence of function over the past 60 million years of Rhinolophoidea diversification. It should be noted that when the CAAX-box sequence is appended to the human p42 OAS1 isoform there is partial rescue of the SARS-CoV-2 restriction phenotype (Wickenhagen *et al.*, 2021). In contrast, none of the *R. ferrumequinum* recombinant proteins initiated a block to SARS-CoV-2 replication, supporting the notion that OAS1 function has changed in the Rhinolophoidea, in a way that is not functionally analogous to the human p42 protein.

Interestingly, RhinoCA-T70 did not restrict SARS-CoV-2 replication (Figure 3.3A,B). This further demonstrates that the anti-coronaviral activity of bat OAS1 proteins is not solely dependent on the presence of a prenylation signal. Instead, amino acid variation also defines the presence or absence of anti-SARS-CoV-2 activity. The RhinoCA and RhinoCA-T70 differ in 16 sites that do not cluster in any specific part of the protein (see online data, Methods subsection 3.2.8).

3.3.5 Unique evolutionary signatures following prenylation loss

Following the loss of OAS1 prenylation at the basal branch of the Rhinolophoidea, I hypothesised that this OAS1 lineage might have taken an evolutionary path distinct from other bat OAS1s; such as lack of conservation of residues needed for the anti-CoV function, or selection for a different function entirely. To assess differences in selective pressures of individual branches across the entire Chiroptera OAS1 phylogeny, I first used the aBSREL method (Smith *et al.*, 2015). Three branches in the tree show evidence of significant episodic diversifying selection: i) the ancestral branch leading to the Yangochiroptera clade ($p = 0.024$) which – considering that both *P. alecto* and *P. kuhlii* OAS1 proteins restrict SARS-CoV-2 replication – is unlikely to have selective changes related to gain or loss of anti-coronaviral function, ii) the terminal branch leading to *M. molossus* ($p = 0.0099$), associated with changes unique to this distant species, and iii) the branch leading to the *Rhinolophus* clade (consisting of *R. ferrumequinum* and *R. sinicus*; $p = 0.018$). Diversification on the latter branch could be associated with divergence of protein function in this non-prenylated group. Still, no episodic selection was detected on the branch where prenylation loss took place, suggesting no major advantageous substitutions happened immediately after the loss of membrane targeting.

Since the *R. ferrumequinum* OAS1 antiviral function cannot be restored simply by appending a prenylation signal at its C-terminal end, subsequent changes to the genome likely occurred that removed this potential function. Similarly, RhinoCA-T70 only has 16 amino acids different to RhinoCA which collectively disrupt anti-SARS-CoV-2 function. Hence, the branches of the Rhinolophoidea clade might have undergone relaxation of potential purifying selection acting on sites required for anti-coronaviral activity in all other bat clades. The RELAX method (Wertheim *et al.*, 2015) showed no evidence of selection relaxation specific to this clade ($K = 0.92$, $p = 0.38$, $LR = 0.77$) compared to the rest of the tree. Consistent with this finding, the Contrast-FEL method (Kosakovsky Pond *et al.*, 2021) found no sites in the alignment to be evolving under a unique selective environment specific to the Rhinolophoidea clade (q value threshold of 0.2). Relaxation or change of selective pressures on this clade could have indicated a lack of significant function of the Rhinolophoidea OAS1s (or progressive pseudogenisation), however that does not seem to be the case. Rather, the nature of selection on the Rhinolophoidea OAS1 genes has not changed substantially following the putative loss of anti-CoV function.

I then sought to understand if the sites under selection are different between the Rhinolophoidea and the other Chiroptera clades. Only testing Rhinolophoidea branches reveals 31 sites under purifying selection using the FEL method (Kosakovsky Pond and Frost, 2005) and 25 sites under diversifying selection using the MEME method (Murrell *et al.*, 2012) (14 of which are also picked up by FEL; p value threshold of 0.1) (see online data, Methods subsection 3.2.8). Testing the remaining branches shows a total of 99 sites under purifying selection (FEL) and 32 sites under diversifying selection (detected with MEME, 8 of which are also picked up by FEL; p value threshold of 0.1). It is notable that although about the same number of positively selected sites are picked up in both sets of branches, only about a third has signal of purifying selection in the Rhinolophoidea branches compared to the rest of the tree. This is not a direct comparison because of differences in the number of branches tested and the amount of diversity between the two sets but could indicate that site-specific purifying selection is weaker in the Rhinolophoidea clade, hence less likely to be detected.

Comparing the identified sites between the two sets revealed 15 sites under diversifying selection unique to the Rhinolophoidea clade (see online data, Methods subsection 3.2.8). These did not seem to cluster in any obvious way on the secondary structure of the protein. To examine potential clustering on the tertiary protein structure I used AlphaFold (Jumper *et al.*, 2021) for predicting structural models of the *R. ferrumequinum*, *P. alecto* and RhinoCA OAS1 sequences. When super-imposed with the human OAS1 protein structure (pdb: 4IG8) there are very few differences between the four structures. The two key distinctions of the *P. alecto* and RhinoCA OAS1s are also obvious in the sequence alignment (Figure 3.2A), namely: i) the variable indel region and ii) the prenylated C-terminal end. The former insertion creates an unresolved loop structure that likely interacts with the dsRNA molecule (Figure 3.4A). Since both OAS1s carrying the insertion restrict SARS-CoV-2 replication, it is unlikely to disrupt RNA-binding, but might modulate binding sensitivity or stability. The latter insertion is also unresolved on the structure, but found on one end of the protein, away from the RNA-binding surface and seems to be easily accessible by enzymes for acquiring post-translational modifications (Figure 3.4B). Finally, I mapped the 15 sites under positive selection unique to the Rhinolophoidea clade onto the tertiary structure of the *R. ferrumequinum* OAS1. Five of these sites (16, 18, 23, 68 and 202) potentially directly interact with the dsRNA helix (Figure 3.4C), suggesting that RNA binding specificity could be under

diversifying selection unique to the Rhinolophoidea. Another six of the sites under selection (264, 292, 296, 297, 329 and 338) cluster on the C-terminal end of the protein (Figure 3.4D). These all seem to face outwards of the core of the protein, exactly where the CAAX box prenylation signal would have been. Deletion of the CAAX end could have resulted in selective changes in sites located structurally near this part of the protein. No apparent function could be speculated for the remaining four positively selected sites (88, 116, 175 and 187).

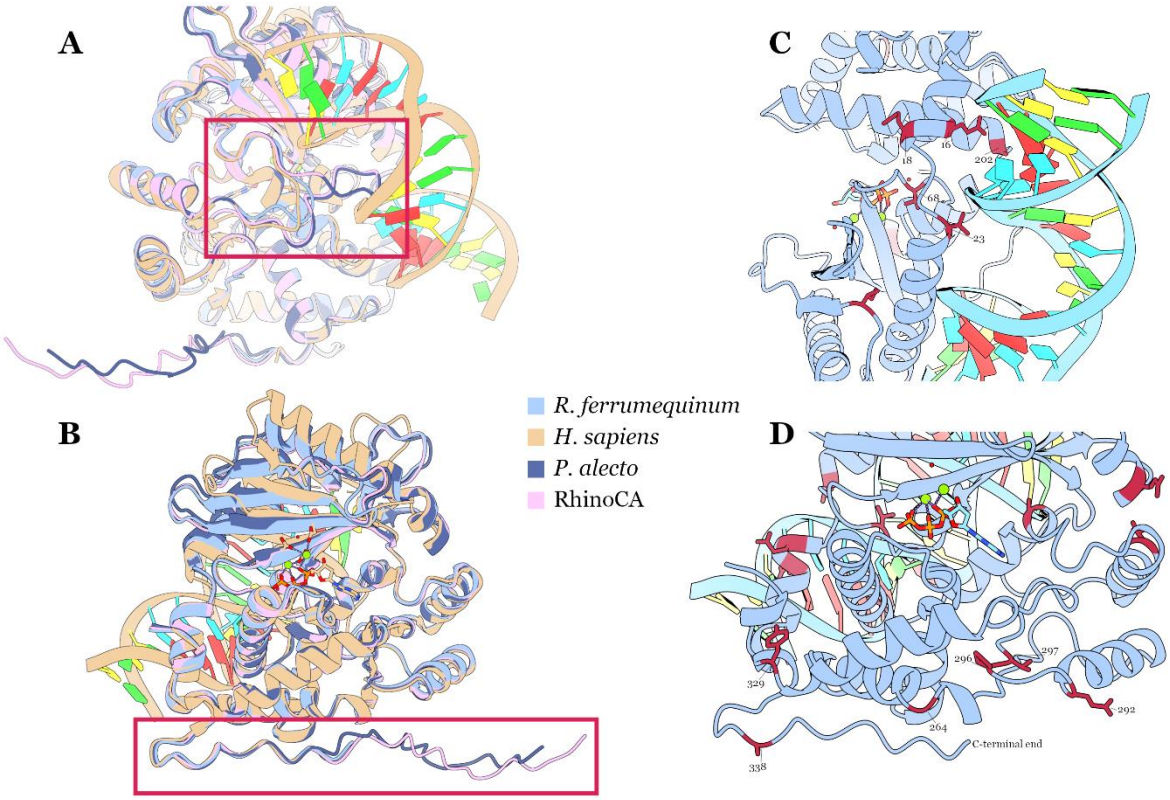


Figure 3.4. Structural comparison and sites under selection unique to the Rhinolophoidea clade. Structural models of the *R. ferrumequinum*, *P. alecto* and RhinoCA protein sequences superimposed onto the human OAS1 structure, highlighting the variable indel region loop **(A)** and the prenylated C-terminal end **(B)**. *R. ferrumequinum* OAS1 structure bound to a dsRNA helix highlighting sites under Rhinolophoidea-specific diversifying selection in red near the RNA molecule **(C)** and near the C-terminal end **(D)**.

3.4 Discussion

The prenylated form of OAS1 is an important defence against SARS-CoV-2 in humans, the virus's novel host. In this chapter, I described the ancient loss of this post-translational modification and subsequent loss of the antiviral activity of the OAS1 orthologue in horseshoe bats (superfamily Rhinolophoidea), the reservoir hosts of SARS-related coronaviruses (SARSr-CoVs). I provide evidence that the loss of antiviral function is unique to this group of bats, only known to be infected by coronaviruses without PDE genes (which would antagonise a functional OAS-RNase L pathway). To look closer into this ancestral change in OAS1 function, I reconstructed the likely sequence of the ancient OAS1 protein found in the Rhinolophoidea common ancestor, before its prenylation signal was lost. The *in vitro* expression of the ancient Rhinolophoidea OAS1 protein in human cells potently inhibits SARS-CoV-2 replication. This anti-CoV function cannot be restored simply by appending the prenylation signal to an extant Rhinolophoidea species' OAS1 protein, suggesting that the ancestral sequence reconstruction based on the bat OAS1 phylogeny, in addition to prenylation, is responsible for restoring this function. This is one of very few examples where antiviral function, lost millions of years ago, is empirically restored by reconstructing the extinct form of a gene (Moraes *et al.*, 2022).

A number of +ssRNA viruses employ cellular compartmentalisation into DMVs most likely as a strategy to avoid targeting from host defences (Roingeard *et al.*, 2022). Viruses are in constant arms race evolution with their hosts, so it is only natural that the adoption of this virus strategy is followed by host defences also localising in the cellular compartments that virus replicates in. The endomembrane targeting of prenylated OAS1 enables the potential sensing of a diverse spectrum of viruses. For example, multiple viruses that use replicative organelles, including hepatitis C virus (Kwon *et al.*, 2013), equine arterivirus (EAV) (Schoggins *et al.*, 2013), and Betaarterivirus suid 1 (PRRSV) (Zhao *et al.*, 2016), are inhibited by OAS1. Membrane localisation of one antiviral gene isoform, in this case human OAS1 p46 instead of the non-prenylated p42 isoform, is not unique to this gene. The Zinc-finger antiviral protein (ZAP) has recently been shown to deploy a similar mechanism, where the longer, ZAP-L, isoform encodes a C-terminal CAAX box that facilitates S-farnesylation and subsequent protein localisation in intracellular compartments (Kmiec *et al.*, 2021). Similar to human OAS1 p46, ZAP-L requires this post-

translational modification signal to potently restrict HIV-1 and SARS-CoV-2 replication, whereas the unmodified ZAP-S protein does not restrict the virus (Kmiec *et al.*, 2021). Thus, it seems that encoding for post-translation signals through gene isoform diversity is a flexible and efficient way for hosts to expand their antiviral targeting while maintaining the genes' core domains, presumably required for foreign RNA recognition.

At least four different human OAS1 isoforms are known (p42, p46, p48, p52), suggesting that there could be unexplored isoform diversity of the bat orthologues. One interesting feature of the bat OAS1 proteins is the variable indel region, that has differing lengths in divergent bat clades (Figure 3.2B). This region forms an unresolved loop in close proximity to the dsRNA molecule bound by OAS1 (Figure 3.4A). Since binding to the viral RNA is required for virus recognition and subsequent restriction of its replication through the RNase L pathway, the elongated variable indel region shared by *P. alecto* and the RhinoCA OAS1 is unlikely to perturb binding. The most likely functional explanation based on the structural predictions is that length variation in this region could modulate RNA binding specificity and potentially allow OAS1 to interact with a different spectrum of dsRNA targets (different viral RNAs or host transcripts). This is also consistent with the Rhinolophoidea losing the elongated genotype along with their prenylated C-terminal end. The start of the bat variable indel region corresponds to the start of exon 3 of the human OAS1 gene. Recently, Banday et al. (2022) showed that another SNP in exon 3 of the human OAS1 gene associated with increased hospitalisation of COVID-19 patients produces isoforms with a shortened exon 3 start. This splicing variation seems to decrease OAS1 expression through nonsense-mediated decay of shorter isoforms, likely explaining its association with more severe disease. If the variable indel region is also near a splicing site in the bat genes, then the length variation we observe across the bats could simply represent modulation of the dominant isoform in each species, most or all producing both long and short isoforms that have not yet been identified. On the same grounds, instead of affecting structural functionality (e.g. RNA binding specificity), the indel variation could affect OAS1 expression in each bat species, although these hypotheses are non-exclusive.

The inability of one of the two versions of RhinoCA to restrict SARS-CoV-2 replication (RhinoCA-T70) can provide some insights into the core OAS1 sites

important for the antiviral activity. The ambiguously predicted sites of RhinoCA-T70 were replaced by *P. alecto* OAS1 residues, hence these residues might not follow the true evolutionary tree expectation. Since the *P. alecto* OAS1 does restrict SARS-CoV-2, its corresponding residues only disrupt antiviral function when placed in the ancestrally predicted backbone, suggestive of epistatic interactions between multiple sites controlling function. The two reconstructions differ at 16 amino acid sites distributed across the length of the protein (Appendix B Figure B.1). Out of these, site 34 seems to directly interact with the dsRNA molecule, being part of a long alpha helix next to the binding site (Appendix B Figure B.1, Figure 3.1A). Residue changes on the nearby site 28 of the human OAS1 protein (site 27 on the RhinoCA protein) are known to be detrimental for RNA binding activity (Donovan, Dufner and Korennykh, 2013). RhinoCA has an asparagine (N) on site 34, which has an uncharged chain, while RhinoCA-T70 has a negatively charged glutamic acid (E) instead (Appendix B Figure B.1). Given that RNA is negatively charged, changing the charge of this site could disrupt RNA binding, making site 34 the most likely single change culprit for RhinoCA-T70's lack of antiviral function.

Showing that OAS1 anti-CoV activity is restored at the base of the Rhinolophoidea superfamily clade, supports loss of this function being due to the ancestral LTR insertion and can provide new insights on the arms race evolution between SARSr-CoVs and these bats. At least two distinct *Betacoronavirus* lineages have independently acquired PDE-encoding genes that counteract OAS1-dependent antiviral activity. Both viral lineages are thought to have ancestrally infected species expressing prenylated OAS1 proteins previously shown to restrict SARS-CoV-2 (Wickenhagen *et al.*, 2021): rodents or cattle for *Betacoronaviruses* in lineage A (Forni *et al.*, 2017) and bats of the Vespertilionidea family for MERS coronaviruses in lineage C (Corman *et al.*, 2014; Yang *et al.*, 2014; Anthony *et al.*, 2017). This suggests that PDE gene acquisition was likely selected for in their distant reservoir hosts. Having lost their OAS1 defence against coronaviruses, the early Rhinolophoidea species would have been an easily accessible niche for non-PDE expressing CoVs, such as the SARSr-CoVs, to establish as their long-term hosts. Thus, OAS1 prenylation loss due to a stochastic LTR insertion about 60 million years ago could be one of the key reasons why SARSr-CoVs circulate in present day horseshoe bats. Previous research has demonstrated how unique evolution of other immune genes in bats has likely led to enhanced 'innate immune tolerance' for these animals (Ahn *et al.*, 2019). This could also be an outcome of OAS1's

evolution in horseshoe bat, explaining the large diversity of SARSr-CoVs that they carry (Wu *et al.*, 2022).

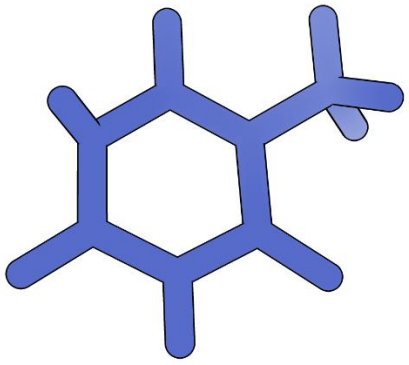
In hosts other than horseshoe bats there likely is firm arms race evolution between host OAS1 antiviral function and virus PDE acquisition. One interesting, previously documented, result of the coronavirus PDE search presented in this chapter is that the Lucheng Rn rat coronavirus is the only example of an *Alphacoronavirus* with a PDE gene. In fact, it seems that this virus has acquired its PDE gene through a recombination event with a lineage A *Betacoronavirus*, acquiring a full coding region resembling the OC43 NS2 gene (Wang *et al.*, 2015). Acquisition of a long insertion between the Orf1b and Spike genes should in principle be deleterious for a coronavirus, however, given that the prenylated rodent OAS1 orthologue restricts SARS-CoV-2 replication (Wickenhagen *et al.*, 2021), it is possible that gain of an NS2-like PDE was largely advantageous for this rat virus instead. The PDE search performed here is sequence based and, given the dissimilarity between NS2 and NS4b PDE genes, unidentified accessory virus genes with PDE function could have been missed. Functional characterisation of more coronavirus accessory genes could improve this search, although other ways to counteract the OAS-RNase L system should be possible, for example changing the dsRNA structure recognised by OAS1.

Although OAS1 prenylation is shared across a number of vertebrate hosts, the Rhinolophoidea might not be the only group where this function has been lost. The *Molossidae* is a family of bats under the Vespertilionoidea superfamily (Teeling *et al.*, 2005). The only species with an annotated OAS1 protein sequence and complete genome assembly used in the presented analyses is *M. molossus.* The NCBI entry of the *M. molossus* OAS1 protein does not have a C-terminal prenylation signal, while the CAAX sequence presence in the *M. molossus* genome assembly could also not be confirmed in the DIGS analysis. Interestingly, the variable indel region of the *M. molossus* OAS1 is shorter than that of the other Vespertilionoidea proteins analysed, being more similar to the Rhinolophoidea version of the region (Figure 3.2B). This suggests that another loss of the OAS1 prenylation signal could have taken place within the *Molossidae* family, independent of that in the Rhinolophoidea. Such a putative loss could also be associated with a diversification in OAS1 function, since the terminal branch leading to *M. molossus* is one of the only three branches with significant evidence of branch-specific positive selection

based on the aBSREL results. A final speculative piece of evidence for a *Molossidae* OAS1 prenylation loss similar to that of the Rhinolophoidea is that *Chaerophon plicatus* (a member of the *Molossidae* family) is the only non-Rhinolophoidea bat species from which a SARSr-CoV has been sampled (Yang *et al.*, 2013); the OAS1 status of this particular species however is not known. The available sequences of these bat species is still very sparse and more data and experimental validation are required to confirm this hypothesis. This observation further raises the question whether OAS1 prenylation could be used as a marker of hosts' susceptibility to SARSr-CoV infection. Although this chapter focuses specifically on bats, being the reservoir hosts of many coronaviruses, independent losses of OAS1 prenylation might have taken place in other animal species.

If the non-prenylated OAS1 isoform has a different function unrelated to innate immunity, it might be advantageous for a host to lose the prenylated form so that more non-prenylated OAS1 is expressed. Despite the ancestral loss of anti-CoV activity, I show that selective pressures have not substantially relaxed on the Rhinolophoidea OAS1 clade. This indicates that the gene is not pseudogenising and probably has biological function(s) that remains conserved within the superfamily. The sites under Rhinolophoidea-specific diversifying selection clustering near the RNA binding surface and C-terminal region (Figure 3.4C,D) considered alongside the *Rhinolophus* branch selection signal, suggest that the horseshoe bat OAS1 has developed a novel function. This could be restricting viruses (where post-translational modification of OAS1 for membrane localisation is not required) or could be a function unrelated to innate immunity. Considering the lack of knowledge of other potential isoforms produced by the bat OAS1 genes, the shift in evolutionary signatures might not be indicative of true novel function, rather changing the evolutionary focus on an existing function performed by a different isoform. The OASs are ancient proteins with extensive retention of duplications in their evolutionary histories (J. Hu *et al.*, 2018) and homology dating back to the animal-insect split (Holleufer *et al.*, 2021; Slavik *et al.*, 2021), so although most research has focused on their immune properties, they could be involved in other cellular functions requiring RNA sensing. Lastly, very few Rhinolophoidea bat genomes have been sequenced so far. Sequencing the OAS1 locus of more species or even acquiring population-level resolution of allele frequencies for these bats would largely enhance our understanding of this functional change in the Rhinolophoidea OAS1.

Chapter 4. Evasion of the human BTN3A3 restriction defines the evolution of zoonotic influenza viruses



Cartoon model of thymine.

PDB ligand entry: TDR, visualised with ChimeraX.

“Among the infinite diversity of singular phenomena science can only look for invariants.”

Jacques Monod, Chance and Necessity (1970)

The work presented in this chapter is part of the paper entitled "*Zoonotic avian influenza viruses evade human BTN3A3 restriction*" published in *Nature* (Pinto *et al.*, 2023. Nature, 619: 338). I have conducted all the genomic, phylogenetic and computational work presented in this chapter. The experimental work performed by co-authors additionally presented in the paper is described in the introduction of this chapter and referenced accordingly across the chapter to provide necessary context.

Aim

Members of the human BTN3 proteins restrict the replication of avian-adapted but not human-adapted influenza A viruses. In this chapter, I characterise the evolutionary timing of the gain of antiviral activity in the BTN3 gene group. The BTN3 restriction of influenza A viruses is dependent on the residues in two sites of the NP protein. I further describe the evolution of the virus NP focussing on these two sites and explore how they can be used as markers for predicting zoonotic potential of influenza A viruses.

4.1 Introduction

4.1.1 Restriction factors against influenza A viruses

Hosts possess multiple barriers against infection by all types of viruses, but restriction factors specifically against influenza A viruses (IAV) have been heavily studied by virologists due to the virus's immense burden on global health. IAVs infect both mammalian and avian hosts and as a result there are distinct barriers specific to the two divergent host environments that the viruses need to overcome for cross-species transmission (Long *et al.*, 2018). The first barrier to infection is entry into the cell, dictated by host cell receptor specificity. The hemagglutinin (HA) protein of IAVs is responsible for cell entry by binding to terminal sialic acid residues at the end of surface glycoconjugates on the surface of vertebrate cells (Dou *et al.*, 2018). Avian IAV tends to favour α(2,3)-linked sialic acid, while human IAVs favour the α(2,6)-linked conformation and, by changing this conformation specificity, different strains can expand their host range (Shinya *et al.*, 2006). Apart from direct receptor specificity, cell entry and onwards transmission also depends on the pH stability required for HA activation (Di Lella, Herrmann and Mair, 2016) as well as the stalk length of the neuraminidase (NA) protein also found in the envelope of the virion along with HA (Blumenkrantz *et al.*, 2013). Both HA and NA are found on the surface of the virion, the first being essential for entry and the latter for virion release, while biochemical balance between the two is important for virus fitness (Wagner, Matrosovich and Klenk, 2002).

Although cell entry and virion release are required for enabling virus transmission, the ability of a virus to infect a host will also depend on other host-imposed mechanisms directly blocking or facilitating virus replication. Focusing on innate immunity, there are many examples of interferon-stimulated genes (ISGs; described in Chapter 1 Section 1.3) known to restrict IAVs. For example, human IFITM3 inhibits IAV by stopping viral fusion of late endosomes and preventing egress of the virus into the cytosol (Feeley *et al.*, 2011; Everitt *et al.*, 2012). Since this antiviral mechanism targets the entry mechanism thought to be employed by all IAVs (and most enveloped viruses), IFITM3 is capable of restricting both mammalian and avian IAV strains. On the contrary, other human barriers to infection seem to only target avian-adapted viruses. This is because many mammalian adaptations on the virus genome relate to evasion of said antiviral mechanisms, especially when virus

restriction requires physical interaction between host proteins and viral RNA or proteins. IAV transcription and replication take place in the nucleus of the infected cell and viral RNA is packaged in the vRNP complex made up of the nucleoprotein (NP) and the three polymerase proteins: PB1, PB2 and PA (Dou *et al.*, 2018). The Acidic Nuclear Phosphoprotein 32 Family Member A (ANP32A) protein has been shown to be essential for viral replication in the nucleus by interacting with the vRNP, specifically PB2 (Long *et al.*, 2016). However, avian ANP32A has a 33 amino acid insertion compared to its mammalian orthologues (ANP32A and ANP32B) which alter how the proteins interact with the virus PB2 (Carrique *et al.*, 2020). As a result, IAVs with avian-adapted PB2 sequences cannot replicate in mammalian cells (Long *et al.*, 2016). The protein interaction can change drastically by a single substitution on PB2 site 627 from the avian glutamic acid (E) to the mammalian lysine (K), enabling efficient replication with mammalian ANP32A. This PB2 adaptation takes place immediately after host switching and is a key factor for bypassing the ANP32A host barrier (Subbarao, London and Murphy, 1993; Long *et al.*, 2018).

While ANP32A can be thought of as having a positive effect for viral replication when the virus strain is compatible with the host version of the protein, other barriers to IAV infection actively restrict specific strains of the virus. Human Mx1 is an ISG that restricts avian IAV but is much less restrictive of human-adapted strains (Dittmann *et al.*, 2008). This effect seems to be dependent on the NP of the virus (Zimmermann *et al.*, 2011). Multiple specific residues in the NP have been proposed to be responsible for Mx1 evasion in human adapted viruses but these differ between the two NP lineages circulating in humans, 1918 pandemic and 2009 swine flu derived strains. The full effect of NP-dependent Mx1 evasion was acquired only when substitutions away from avian strain NPs were made in combination (Mänz *et al.*, 2013). These are few well-studied examples of host barriers against IAV infection which can restrict all strains of IAV (IFITM3) or pose a host species specific restriction that can be efficiently evaded by one (ANP32A) or multiple (Mx1) residue adaptations on the virus proteins. This chapter will focus on a novel host restriction factor and the evolutionary dynamics between this host gene and IAVs' ability to infect different hosts.

4.1.2 Human BTN3A1 and BTN3A3 genes restrict avian IAV

The butyrophilins are a family of proteins first identified to be involved in the stabilisation and production of milk lipid globules in many mammalian species (Heid *et al.*, 1983). It was later revealed that most of the gene family members are primarily involved in adaptive immunity cell differentiation and cellular signalling (Smith *et al.*, 2010; Arnett and Viney, 2014). The butyrophilin proteins have a transmembrane domain and at least one immunoglobulin domain (IgC or IgV). Most members also have a B30.2 or PRYSPRY domain, although there are many cases of duplications and deletions with the family having a complex evolutionary history (Afrache *et al.*, 2012). This particular protein domain comprised of the evolutionary ancient SPRY sub-domain and the more recently incorporated PRY subdomain is found in many gene families involved in immune function (Rhodes, De Bono and Trowsdale, 2005). In humans there are three main subfamilies of butyrophilins genes: BTN1-3 with varying numbers of paralogues under each group (BTN1A1, BTN2A1, BTN2A2, BTN2A3, BTN3A1, BTN3A2, BTN3A3) (Afrache *et al.*, 2012). The human and mouse BTN3 group of the butyrophilins is known to be important for the activation of a specific class of γδ T cells (Vγ9Vδ2 T cells) (Blazquez *et al.*, 2018), proposed to have anti-tumour activity among other functions (Rigau, Uldrich and Behren, 2021).

Apart from their role in T cell activation, the human BTN3 genes have also been shown to be stimulated by interferon, making them ISGs that are likely also involved in innate immunity (Shaw *et al.*, 2017). Following on from this finding, a comprehensive arrayed expression screen of more than 800 *in vitro* overexpressed ISGs against a number of viruses showed that human BTN3A1 and BTN3A3 potently inhibited replication of avian IAV, but not the mammalian lab-adapted or human circulating strains (Pinto *et al.*, 2023). BTN3A3 has a stronger restrictive effect than BTN3A1 and virus growth could be restored by knocking down the constitutive expression of BTN3A3 in primary human cells. BTN3A3 was further shown to be constitutively expressed in the upper and lower respiratory tract of healthy individuals, indicating that this is likely a functional restriction factor against infection by avian IAVs. This inhibitory effect was specific to the Mallard IAV strain out of 24 different viruses tested. A number of BTN3 orthologues and paralogues of humans and other related species were also tested to understand how evolutionarily conserved this anti-avian IAV function is. Interestingly, only old world monkey and

ape butyrophilins restricted avian IAV replication, while none of the genes had any effect on human-adapted IAV. Instead the closest orthologues of species regularly infected by IAVs, chickens, ducks, dogs, horses and pigs had no antiviral effect against any IAV strain. In this chapter, I will explore the evolutionary relations of these genes and determine at which point in their evolutionary history they gained this antiviral function.

4.1.3 Changes in two NP sites independently evade BTN3 restriction

The specificity of the BTN3 antiviral function against only avian IAV infection, suggests that the mammalian-adapted changes in the virus are likely responsible for the evasion of the human restriction factor. Flu reassortment experiments where each of the 8 genome segments were individually swapped between an avian-adapted and a mammalian-adapted strain showed that only the avian segment 5, encoding for the NP, was responsible for the restriction by the BTN3s (Pinto *et al.*, 2023). Conversely, introducing a mammalian-adapted segment 5 into an avian IAV strain was sufficient to evade BTN3 restriction. This means that certain substitutions in the viral NP that are unique to mammalian IAVs are capable of bypassing the antiviral activity. Mutagenesis experiments targeting NP sites that differ between avian and mammalian IAVs showed that a single substitution in either NP site 313 or site 52 independently altered the restriction phenotype. More specifically, changing the avian phenylalanine (F) at site 313 to the human IAV residues, either a tyrosine (Y; present in the 1918 H1N1, 1957 H2N2 and 1968 H3N2 pandemic viruses) or a valine (V; present in the 2009 H1N1 pandemic virus) evades restriction by the BTN3s (Pinto *et al.*, 2023). Virus with a leucine (L) at site 313, found in few avian strains, is susceptible against BTN3, similar to 313F. The F313Y substitution, in combination with other mammalian-adapted NP changes, has been proposed to overcome restriction by Mx1 (described in subsection 4.1.1) (Mänz *et al.*, 2013). However, loss of function experiments have shown that restriction of NP 313F IAVs by BTN3A3 is independent of Mx1's antiviral function (Pinto *et al.*, 2023). The second NP site of interest was uncovered after observing that the H7N9 avian IAV strain that has caused multiple independent bird-to-human epidemics in Southeast Asia in the last decade (W. Zhu *et al.*, 2018) also evades BTN3A3 *in vitro*, despite

having a 313F NP. Further mutagenesis experiments highlighted that substitutions at NP site 52 on a 313F background can independently lead to BTN3A3 evasion. Changes from 52Y (found in most avian strains) to an asparagine (N; present in the H7N9 virus), histidine (H) or glutamine (Q) all lead to bypassing the BTN3A3-dependent barrier (Pinto *et al.*, 2023). This chapter showcases a comprehensive phylogenetic analysis of how these human-adapted NP residues are distributed across IAV diversity and how substitutions from the avian to the human-adapted residues are associated with virtually all recent IAV transmissions into humans.

4.2 Methods

4.2.1 *In silico* identification of BTN3 homologs

To identify proteins expressed by other species, homologous to the human BTN3 genes, I first performed a blastp search (v.2.8.1) against all available members of the *Haplorrhini* suborder in the NCBI blast protein refseq database version 5 (e value cutoff 1e-60, as of the 6th of April 2021) (Camacho *et al.*, 2009). The human BTN3A3 protein sequence (NP_008925.1) was used as a probe for the BLAST search. The isoform with the longest sequence was kept for each protein product annotated with the same name. Similarly, blastp with human BTN3A3 was used for identifying proteins expressed by non-primate species susceptible to IAVs infection: *Gallus gallus*, *Anas platyrhynchos*, *Equus caballus*, and *Sus scrofa* and more distant human paralogues (Appendix C Table C.1).

Protein members of the butyrophilin 3 subfamily retrieved from the BLAST search were manually cross-checked with proteins in the Ensembl database (Cunningham *et al.*, 2022) and if the protein sequences were not identical between the two databases the sequence with highest similarity to the human BTN3A3 protein sequence was retained. A total of 30 proteins were retrieved from the following species: *Pan troglodytes*, *Cebus imitator, Equus caballus, Homo sapiens, Gorilla gorilla gorilla, Chlorocebus sabaeus, Macaca mulatta, Pongo abelii, Carlito syrichta, Mandrillus leucophaeus, Callithrix jacchus, Nomascus leucogenys, Rhinopithecus roxellana*.

A custom set of Pfam hmm profiles were used for identifying the conserved domains in the proteins, comprising the immunoglobulin V-set domain (PF07686), CD80-like C2-set immunoglobulin domain (PF08205), PRY (PF13765) and SPRY (PF00622) domains. All protein sequences were scanned with the profile set using hmmscan (HMMER 3.3) (Mistry *et al.*, 2013). The best hit for each identified domain was extracted from the protein sequence aligned with the respective domain segments using MAFFT (v7.453, --maxiterate 1000 --localpair) (Katoh and Standley, 2013). Protein alignments were converted to codon alignments using PAL2NAL (Suyama, Torrents and Bork, 2006). Phylogenies for each separate domain alignment and concatenated domain sequences were reconstructed using IQ-TREE (version 1.6.12) (Nguyen *et al.*, 2015) with the best suited substitution model selected by the

IQ-TREE '-m TEST' option (Kalyaanamoorthy *et al.*, 2017) and 10,000 ultrafast bootstrap replicates (Hoang *et al.*, 2018). Node support was further assessed using 1,000 nonparametric bootstrap replicates and 10,000 SH-like approximate likelihood ratio test replicates (Guindon *et al.*, 2010) for the individual protein domain phylogenies.

4.2.2 IAVs phylogenetic analysis

A total of 35,477 full-length NP coding sequences unique on the nucleotide level (identical sequences collapsed) were retrieved from the NCBI Flu database (https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database, as of the 8th of June 2021, sampled until the end of 2020), only including type A influenza sequences annotated to have been isolated from avian, canine, equine, human and swine hosts. Sequences with ambiguous nucleotides and internal stop codons were removed, resulting in a dataset of 34,079 sequences. The corresponding protein sequences were aligned using MAFFT (v7.453, --maxiterate 1000 --localpair) (Katoh and Standley, 2013) and then converted to a codon alignment with PAL2NAL (Suyama, Torrents and Bork, 2006). Metadata associated with each sequence accession were retrieved and tabulated. Numbering of NP amino acid residues was assigned based on the PR8 sequence [A/Puerto Rico/8/1934(H1N1), GenBank accession: NP_040982.1]. Therefore, the 6 amino acid N-terminal extension of 2009 pH1N1 viruses were not considered for the residue numbering (i.e. amino acid residue M6 was considered M1).

To reduce oversampling of related sequences the dataset was clustered with a minimum sequence identity of 0.99 using MMseqs2 (--min-seq-id 0.99 --cov-mode 0) (Steinegger and Söding, 2017). One representative was kept from each cluster leading to a filtered dataset of 14,665 sequences. The codon alignment of the filtered set was used to reconstruct a phylogeny with iqtree under a GTR+I+G4 model (selected as the most appropriate model with the '-m TEST' option) (Nguyen *et al.*, 2015). The resulting phylogeny was then time-calibrated using TreeTime (Sagulenko, Puller and Neher, 2018). Eleven sequences with annotated dates inconsistent with the root-to-tip regression were subsequently excluded from the analysis.

To explore the H7N9 epidemic NP clade in more detail, representative sequences from this broader avian NP clade as well as the unfiltered sequences from each representative's corresponding cluster were retrieved (3,150 sequences). The codon alignment of these NP sequences was used to infer a more detailed maximum likelihood phylogenetic reconstruction of this particular clade (IQ-TREE under a GTR+I+F+G4 model with 10,000 Ultrafast bootstrap replicates) (Nguyen *et al.*, 2015; Hoang *et al.*, 2018) and time-calibrated using TreeTime as described above (Sagulenko, Puller and Neher, 2018). All phylogenies were visualized using the ggtree R package (Yu, 2020), unless stated otherwise. Tree statistics were analysed using the ete3 Python package (Huerta-Cepas, Serra and Bork, 2016).

4.2.3 Molecular dating of the NP F313V H1N1 pdm09 change

Based on the tree topology of the MMseqs2 filtered dataset including sequences from all IAV NP clades, representative sequences from the classical swine H1N1 clade as well as the unfiltered sequences from each representative's corresponding cluster were retrieved (9,426 sequences). The codon alignment of all swine clade NP sequences was used to infer a more detailed maximum likelihood phylogenetic reconstruction of this particular clade (IQ-TREE under a GTR+I+F+G4 model with 10,000 Ultrafast bootstrap replicates) (Nguyen *et al.*, 2015; Hoang *et al.*, 2018). The phylogeny was time-calibrated using TreeTime (Sagulenko, Puller and Neher, 2018).

To validate the inferred topology I used an independent Bayesian phylogenetic approach on a subset of the full dataset. From the original dataset (9,426 sequences), the sequences outwith the distinct 313F and 313V subclades and sequences topologically close to the F313V change branch - based on the maximum likelihood phylogeny (for example the swine 313V subclade) - were kept in the subset (263 sequences, see online data Methods subsection 4.2.6). To remove sampling biases from the two remaining subclades I used a subsampling approach where up to 200 sequences were retrieved from each subclade but a maximum of 200 divided by the number of sampling years were kept for each sampling year. For example, if a subclade had sequences from sampling years 2008, 2009, 2010 and 2011 with sequence counts 35, 1000, 800 and 1200 respectively, then all 35

sequences from 2008 would be kept and only 50 (200/4) from each of the other three years. This approach yielded a total of 380 subsampled sequences from the two 313F/V subclades and combined with the 263 non-subsampled isolates produced a total of 643 NP sequences to be analysed using Bayesian phylogenetics.

These sequences were retrieved from the full dataset codon NP alignment and used to infer a BEAST (v1.10.4) phylogeny under a HKY model, accounting for site heterogeneity with a 4 category Γ distribution (Suchard *et al.*, 2018). Codon positions were evaluated separately by the model and sampling years were used for tip-dating. Two independent MCMC chains, 150,000,000 states long each, were performed, sampling every 150,000 states. LogCombiner was used to combine the two independent chains after removing 20% burn-in states from each chain, ensuring chain convergence and an effective sample size >200 for the joint parameters.

4.2.4 Geographical distribution of non-52Y NP sequences

The tips of the filtered dataset phylogeny representing all IAV clades was annotated based on sampling location and NP 52 residue using the ete3 python package (Huerta-Cepas, Serra and Bork, 2016). To determine the number of independent lineages with members predicted to be BTN3A3-resistant based on NP site 52 I retrieved all monophyletic subclades with at least two members, all of which have 52 residues N, H and Q. All countries where IAVs with 52N/H/Q NP sequences have been sampled were annotated by the number of unique BTN3A3-resistant subclades including members sampled in each country. The geographic distribution of the lineages was plotted using JavaScript D3 in an ObservableHQ (https://observablehq.com/) notebook.

4.2.5 GISAID sequence analysis

Protein sequences from atypical, avian-only serotypes (H4-H18) were retrieved from the GISAID database (http://gisaid.org/) for avian, swine and human hosts.

Isolates were filtered for having all 8 segments and having been sampled until the 1st of January 2023. The NP, HA and PB2 proteins were downloaded for each isolate and aligned using MAFFT (v7.453, --maxiterate 1000 --localpair) (Katoh and Standley, 2013). NP and PB2 residues at sites of interest and HA polybasic cleavage site presence (identified as having 3 or more K/R residues at the corresponding region) were summarised along with sequence metadata using Python3. GISAID acknowledgments for all analysed sequences are available as provided as supplementary information in the published version of this work in Pinto et al. (2023).

4.2.6 Data availability

Alignments and raw phylogenetic data related to this Chapter can be found in the following GitHub repository: https://github.com/spyros-lytras/BTN3A3_IAV.

4.3 Results

4.3.1 BTN3 antiviral activity likely evolved after the split between old and new world monkeys

To examine the origin of anti-avian IAV activity in the butyrophilin subfamily 3 (BTN3) gene group, I retrieved the annotated sequences of close orthologues and paralogues of the human BTN3A3 gene. Since the only animals whose BTN3 genes restricted avian IAV *in vitro* were members of the old world monkeys and apes (Pinto *et al.*, 2023), I focused the phylogenetic analysis on the *Haplorrhini* sub-order which encompasses apes, humans, old and new world monkeys as well as the tarsier as the outgroup clade. The phylogeny reconstructed based on all concatenated domains of the proteins indicates that BTN3A1-3 originated through two successive duplications after the split between the new world monkeys lineage (*Platyrrhini*) and the old world monkeys and apes lineage (*Catarrhini*) around 40-44 million years ago (Figure 4.1) (Kumar *et al.*, 2017). Mapping the experimental results of which genes successfully restrict avian IAV onto the phylogeny clearly highlights how this function seems to be specific to the *Catarrhini* lineage and only the BTN3A1 and BTN3A3 gene clades of this group of species. The *Platyrrhini* (*Callithrix jacchus* and *Cebus imitator*) and tarsier (*Carlito syrichta*) orthologues outside the two key gene duplications did not show any antiviral activity. This collectively suggests that the antiviral phenotype of BTN3s was gained after the *Platyrrhini-Catarrhini* split, consistent with the two duplication events. Domain detection analysis showed that the majority of BTN3A1 and BTN3A3 genes have a consistent domain organisation with one set of N-terminal IgV and IgC domains followed by a PRYSPRY domain (Figure 4.1). The BTN3A2 *Catarrhini* genes with known restriction status (human and macaque) showed no antiviral activity (Pinto *et al.*, 2023). This could be explained by the lack of a PRYSPRY domain in this gene group which likely lost the domain during the duplication that produced the group. The only apparent exception is the BTN3A2 of *Nomascus leucogenys* which surprisingly contains a PRYSPRY domain, having the same organisation as BTN3A1 and BTN3A3 (Figure 4.1).

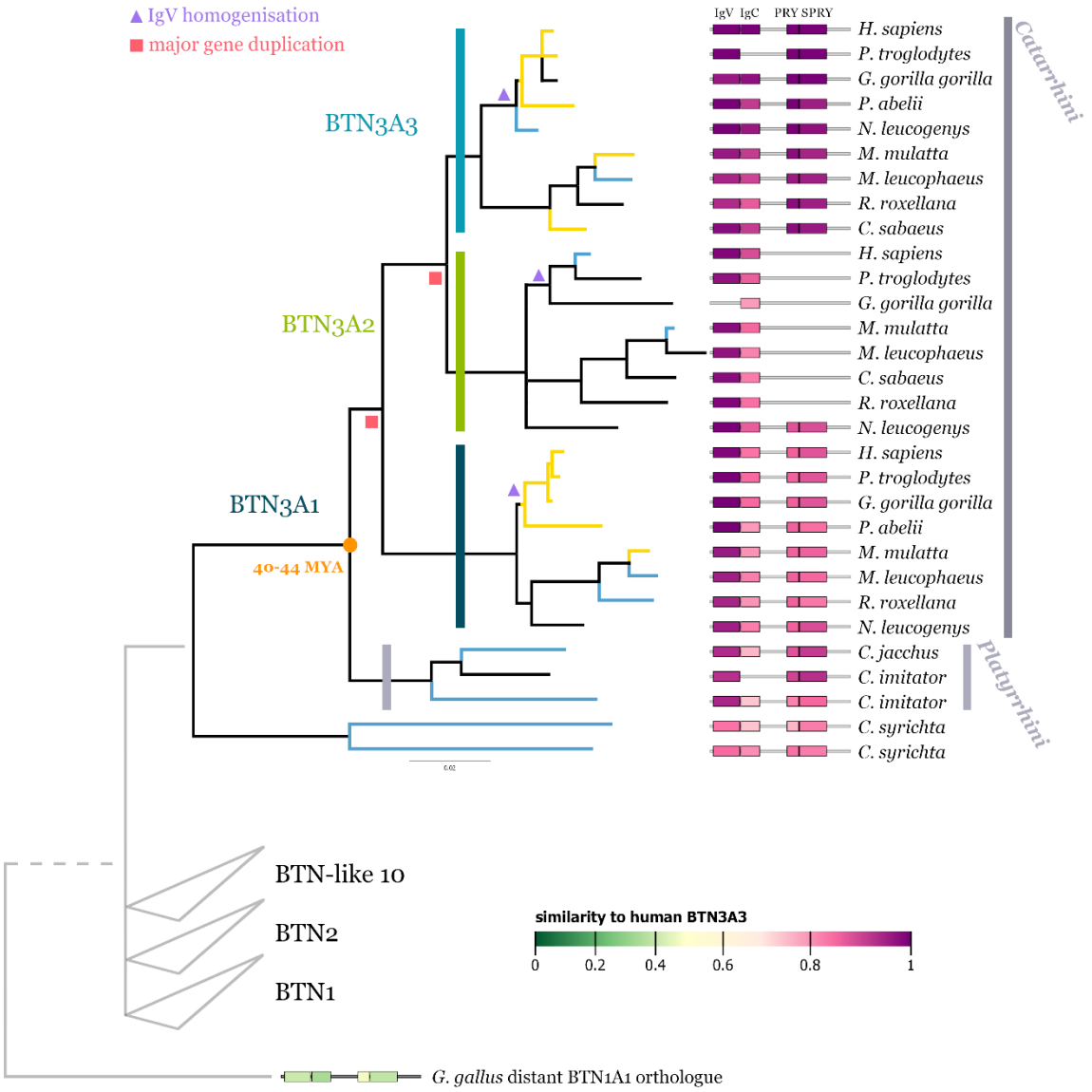


Figure 4.1. Evolution of antiviral activity of BTNs. Maximum likelihood phylogeny of concatenated protein domain coding sequences the Haplorrhini BTN3 genes (K2P+G4 substitution model). Nodes with bootstrap support below 60 have been collapsed. Branches confirmed to have or not have anti-avian IAV activity are highlighted in yellow and blue, respectively. Branches with no available experimental data are shown in black. Relationship to more distant tested homologues and orthologous/paralogous gene families are shown as a schematic in grey. IgV homogenisation events, major gene duplications and gene subfamilies are annotated on the phylogeny. Presence of each of the four protein domains (IgV, IgC, PRY and SPRY) is annotated on the right of each tree tip and coloured by pairwise amino acid similarity to the respective domain of the human BTN3A3. Species names and taxonomic classification is annotated on the right. The median divergence time between *Catarrhini* and *Platyrrhini* was retrieved from TimeTree (Kumar *et al.*, 2017).

Although antiviral function seems to have evolved only in the *Catarrhini* BTN3A1 and BTN3A3 gene groups, not all genes within these groups restrict avian IAV replication, suggesting that the function was not ubiquitously conserved after the initial gain or there were multiple gains of the function within the group. In fact, it has been previously documented that there is a lot of domain diversity within the gene groups, primarily due to reassortment and homogenisation of domains between the paralogues (Afrache *et al.*, 2017). To confirm this, I examined the phylogenetic relatedness of each functional domain (IgV, IgC, PRY and SPRY) independently for the present dataset of *Haplorrhini* BTN3 proteins. The analysis confirms that the IgV domain of the BTN3A2 gene has homogenised in the other two paralogues, within the apes (Figure 4.1). This is evident as the IgV sequences of the old world monkey (*R. roxellana, M. mulatta* and *M. leucophaeus*) form distinct clades for BTN3A1 and BTN3A3, but all ape genes (*H. sapiens, P. troglodytes, G. gorilla gorilla, P. abelii* and *N. leucogenys*) cluster together along with the old world monkey BTN3A2 genes, having almost identical sequences to their paralogues (Figure 4.2A). The only exception to this pattern is the *C. sabaeus* IgV domains. Even though this is a species of old world monkeys, the *C. sabaeus* BTN3A3 and BTN3A2 domains are identical in sequence and both cluster within the overall BTN3A2 clade (Figure 4.2A). This suggests that homogenisation of the BTN3A2 IgV domain into its BTN3A3 paralogue took place independently in this species. It should also be noted that there is no known BTN3A1 gene in *C. sabaeus*, suggesting further genetic reorganisation of the BTN3 locus unique to this species, or genus.

The IgC-derived phylogeny is the most consistent one with the overall concatenated domain topology (Figures 4.1 and 4.2B). In this tree the BTN3A3 and BTN3A1 *Catarrhini* groups cluster closer together, unlike the concatenated domain tree where BTN3A3 and BTN3A2 are closer. It should be noted that this difference could simply be due to insufficient information in the short IgC alignment. The internal branch separating the BTN3A2 group from the BTN3A1/3 groups is very short, suggesting that few IgC-specific mutations may alter the inferred domain tree topology compared to the true gene tree (Figure 4.2B). Similarly, the old world monkey BTN3A3 clade clusters closer to the BTN3A1 clade than to the ape BTN3A3 one, but the corresponding node has relatively low support with all confidence assessment methods (node support: 59/67.4/44).

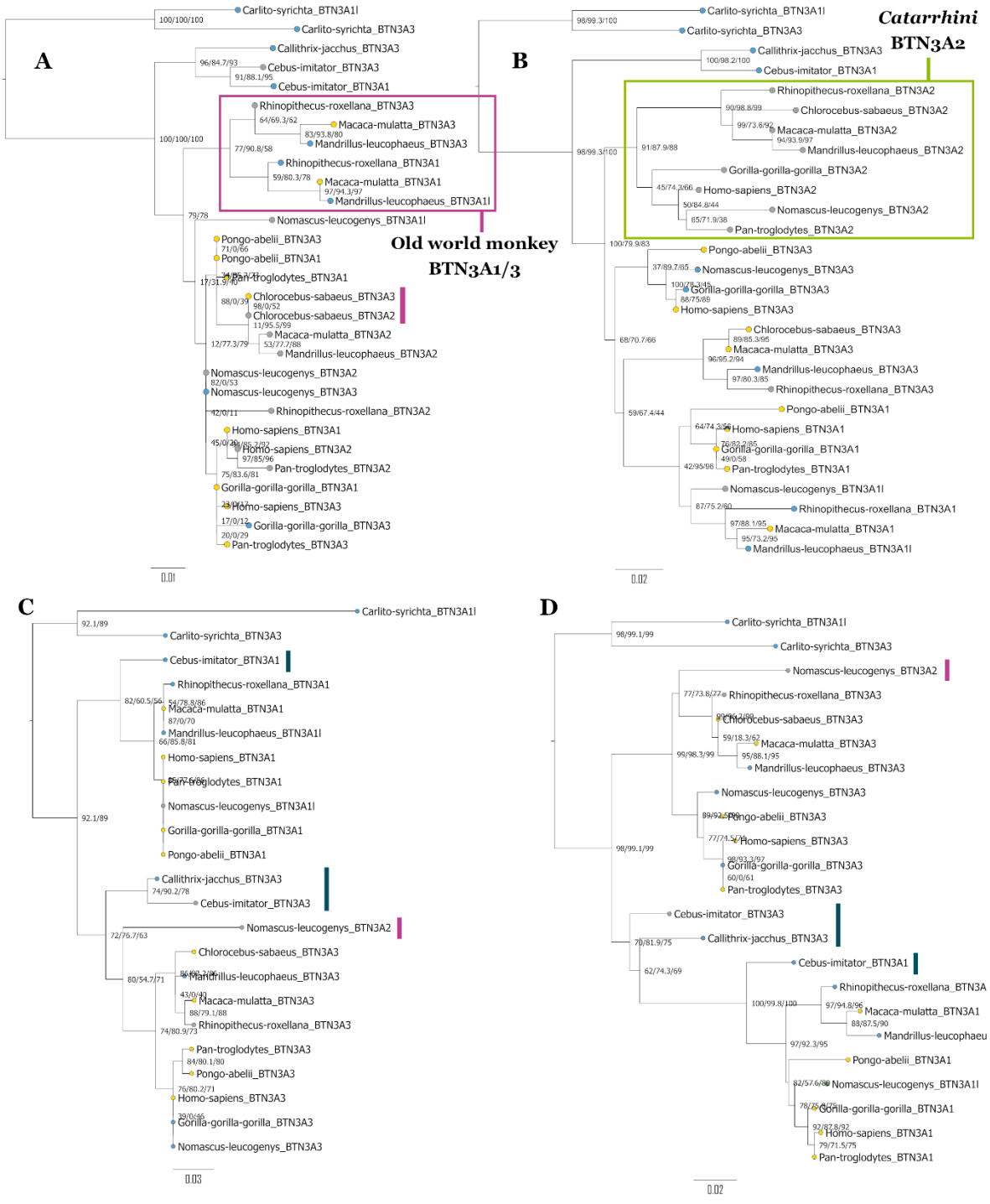


Figure 4.2. Maximum likelihood *Haplorrhini* BTN3 gene coding sequence phylogenies of separate domains: IgV (A), IgC (B), PRY (C), SPRY (D) under a K2P+G4 substitution model. Trees are rooted at the *C. syrichta* branch and node confidence values are annotated on each node (presented as: nonparametric bootstraps / SH-like approximate likelihood ratio test / ultrafast bootstraps). Tip shapes are coloured based on known anti-IAV activity (yellow: restrictive, blue: non-restrictive; as in Figure 4.1) and tips and clades referred to in the text are further annotated. Phylogenies were visualised using FigTree.

Both PRY and SPRY domain phylogenies separate BTN3A1 and BTN3A3 genes in distinct clades, noting that the BTN3A2 genes have lost these domains (Figure 4.2C,D). The only exception, the *N. leucogenys* BTN3A2 PRYSPRY sequences, cluster with the BTN3A3 group for both domains. Based on the PRY domain phylogeny *N. leucogenys* BTN3A2 sits outside all *Catarrhini* BTN3A3 which might mean that, unlike the rest of *Catarrhini*, this species' BTN3A2 PRYSPRY domain was never lost, now forming the sole member of its gene group. On the other hand, the *N. leucogenys* BTN3A2 SPRY domain fairly confidently clusters closer to the old world monkey BTN3A3 gene group (node support: 77/73.8/77). This rather puzzling pattern could be a result of introgression between a *N. leucogenys* ancestor and a now extinct lineage that did not lose the BTN3A2 PRYSPRY domain. Hence, the information in the coding sequences of these genes is not sufficient to delineate the origin of the *N. leucogenys* BTN3A2 PRYSPRY domain, each PRY and SPRY phylogenies supporting an alternate hypothesis.

The phylogenetic placement of the new world monkey genes using the PRY and SPRY domains is also incongruent to that of the other domain trees and the species tree. The *C. imitator* BTN3A1 lacks the IgC domain and clearly clusters with the other two new world monkey genes for the concatenated domain topology and the IgV domain, but consistently clusters closer to the *Catarrhini* BTN3A1 clades for the PRY and SPRY trees (node supports: 82/60.5/56 and 100/99.8/100). The placement of the two BTN3A3 new world monkey (*C. jacchus* and *C. imitator*) genes' PRY and SPRY domains is also incongruent to the other trees and between each other (Figure 4.2C,D). It should be stressed that these are short, conserved domains and it is unlikely that there is sufficient phylogenetic information in the sequences to confidently delineate the evolutionary relationship. Hence, in addition to the duplication and domain homogenisation events that are evident in this gene group, lack of phylogenetic information could also explain some of the tree incongruencies. Finally, examining how the ability to restrict IAV is distributed across the different domain trees does not provide any conclusive patterns as to which domain might be responsible for the antiviral activity. In all four domain sets, there is at least one pair of identical (or almost identical) sequences, where one of the two genes restricts avian IAV but the other does not (Figure 4.2). Thus, gain of anti-IAV activity in the *Catarrhini* BTN3 genes can likely be acquired through a number of substitution combinations in the coding sequence. Another possibility is that changes in the short inter-domain peptide sequences may control the anti-IAV activity, although this is

not clearly apparent from the full protein sequence alignment and the two possibilities are not exclusive e of one another.

4.3.2 Changes at NP residue 313 are key determinants of all known human IAV pandemics

Now focusing on the virus, it has been shown experimentally that the mammalian-adapted residues Y and V at NP site 313 are capable of evading human BTN3A3's antiviral activity (contrary to avian viruses with 313F and L) (Pinto *et al.*, 2023). Hence, it is of interest to examine the distribution of 313 residues across both avian and mammalian IAV strains and infer F313Y and F313V changes across these viruses' evolution, which should lead to viruses better at replicating in humans. I retrieved more than 30,000 IAV NP sequences from the NCBI Flu database and reconstructed a comprehensive time-calibrated maximum likelihood phylogeny for this genome segment (Figure 4.3). Human NP sequences almost exclusively have 313Y or 313V residues while all NP clades circulating in avian hosts, as well as the Eurasian avian-like H1N1 swine clade, are dominated by strains with the conserved 313F residue. Less than 1% of the analysed avian IAV NP sequences contain 313L. Occurrence of the BTN3A3-resistant 313Y is specific to a human NP clade, derived from the H1N1 1918 pandemic, subsequently reassorted into the 1957 and 1968 pandemics and currently circulating as seasonal H3N2. Precise dating of the original F313Y change is difficult due to the small number of 1918 IAV genomes available. However, recently sequenced early pandemic genomes all have a 313Y in NP (Patrono *et al.*, 2022), suggesting that F313Y took place prior to or soon after human emergence of the 1918 H1N1 human strain.

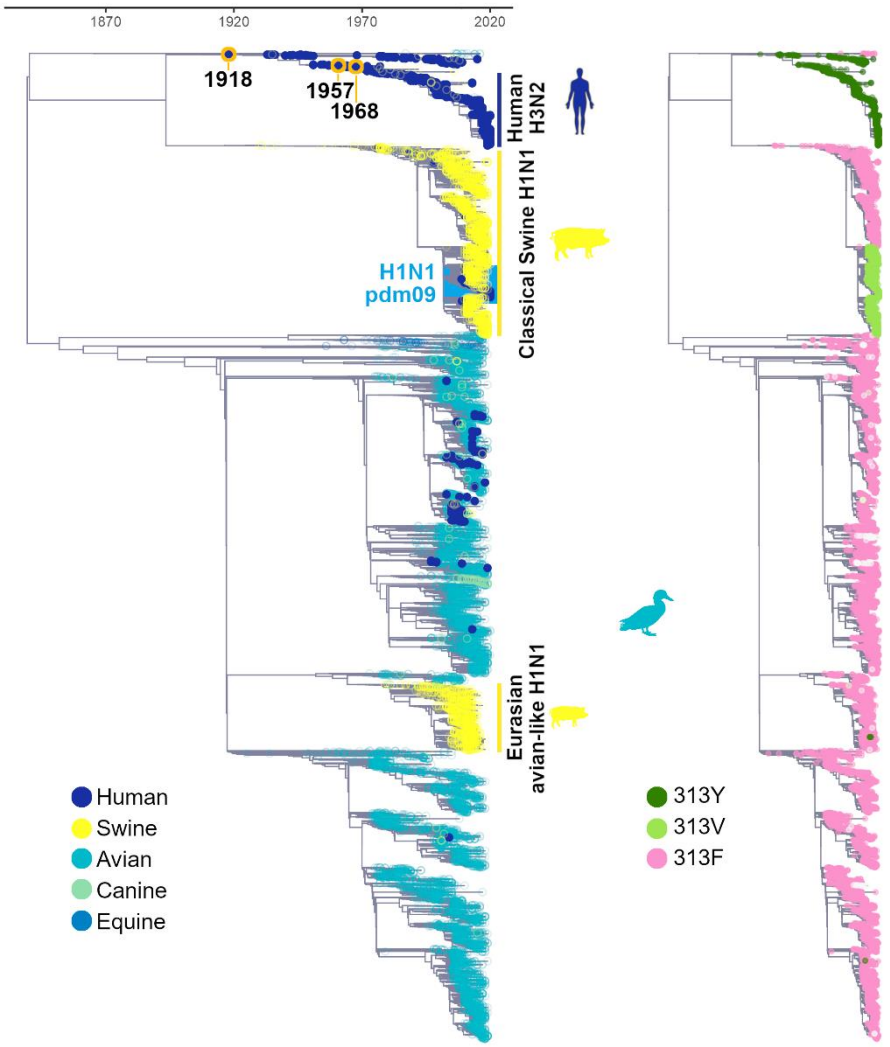


Figure 4.3. Phylogeny of the IAV NP and distribution of site 313 residues. Tip-dated maximum likelihood phylogeny of the filtered IAV NP coding sequence dataset. Tip shapes are annotated by host (left) and 313 residue (right, only residues occurring in more than 5% of the sequences are shown in the legend). Key clades and pandemic isolates are highlighted on the phylogeny.

On the other hand, 313V is specific to the classical swine H1N1 NP subclade, which entered the human population as a result of the 2009 IAV pandemic. A lot more sequences isolated from humans and pigs associated with this IAV introduction to humans are available. To further examine the timing of this F313V NP change, I retrieved all NP sequences that are part of the classical swine H1N1 clade based on the filtered phylogenetic analysis (Figure 4.3, see Methods subsection 4.2.3) and re-inferred the phylogeny for this clade. This tree reveals a clear split between the 313F and 313V clades with a relatively long branch consistent with the F313V change (Figure 4.4A). The divergence date between the two major 313F/V clades is estimated to be July 1997 (1997.52). Upon closer inspection, two isolates:

A/swine/Korea/CY03-12/2012(H3N1) and A/swine/Kansas/12-156064/2012(H1N2) (GenBank accessions: AGE83887 and AHN19642) sampled in South Korea and the US respectively (Noriel *et al.*, 2013; Duff, 2014), are the most basal within the 313V clade. Interestingly, the NP of the latter isolate has a 313F residue. Assuming that this topology is true, the most parsimonious explanation would be that the major F313V change took place after the split between this two-isolate clade and the 313V clade, while an independent F313V change occurred in the A/swine/Korea/CY03-12/2012 lineage. The alternative explanation would require a reversion back to 313F unique to the A/swine/Kansas/12-156064/2012 lineage. This means that we can narrow down the window of time when the F313V change occurred to start from an estimated date of November 2002 (2002.91; Figure 4.4B). Within the 313V clade a distinct subclade consisting of isolates sampled in swine hosts primarily from Mexico between 2011 and 2015 sits as outgroup to the H1N1 swine 2009 pandemic (pdm09) NP sequences. All isolates in this subclade have a 313V residue, supporting a F313V change responsible for human butyrophilin antiviral evasion taking place in the swine host reservoir, prior to human emergence (Figure 4.4B). Furthermore, the 313 residue change should have happened prior to the date of the split between the swine subclade and the pdm09 lineage, estimated to be January 2007 (2007.00).

Inferring phylogenies for large datasets such as this can sometimes lead to erroneous topologies, especially if sampling biases exist between different clades in the tree (for example there are many more sequences from the H1N1 pdm09 lineage compared to any other clade in the tree). Hence, I also performed a Bayesian phylogenetic inference on a targeted subset of the classical H1N1 clade NP sequences aimed to reduce sampling bias that may affect the analysis (Figure 4.4C). Despite low support clustering of a few more 313F sequences within the 313V clade, the aforementioned South Korea and US isolate clade is consistently found within the 313V clade (node posterior = 0.79, Figure 4.4C). The topology of the swine 313V clade from Mexico is also congruent with the maximum likelihood phylogeny. This independent validation of the topology inference strongly supports the above assessment of the F313V change taking place in swine hosts prior to human emergence. The BEAST phylogeny can also provide us with more accurate node dating estimates, placing the earliest date for the F313V change in August 2004 (median: 2004.61, 95% HPD: 2002.57 – 2006.18) and latest date in October 2005 (median: 2005.80, 95% HPD: 2004.68 – 2006.87). These estimates are largely

consistent with the time-calibrated maximum likelihood phylogeny and support that F313V very likely occurred between mid-2002 and the end of 2006.

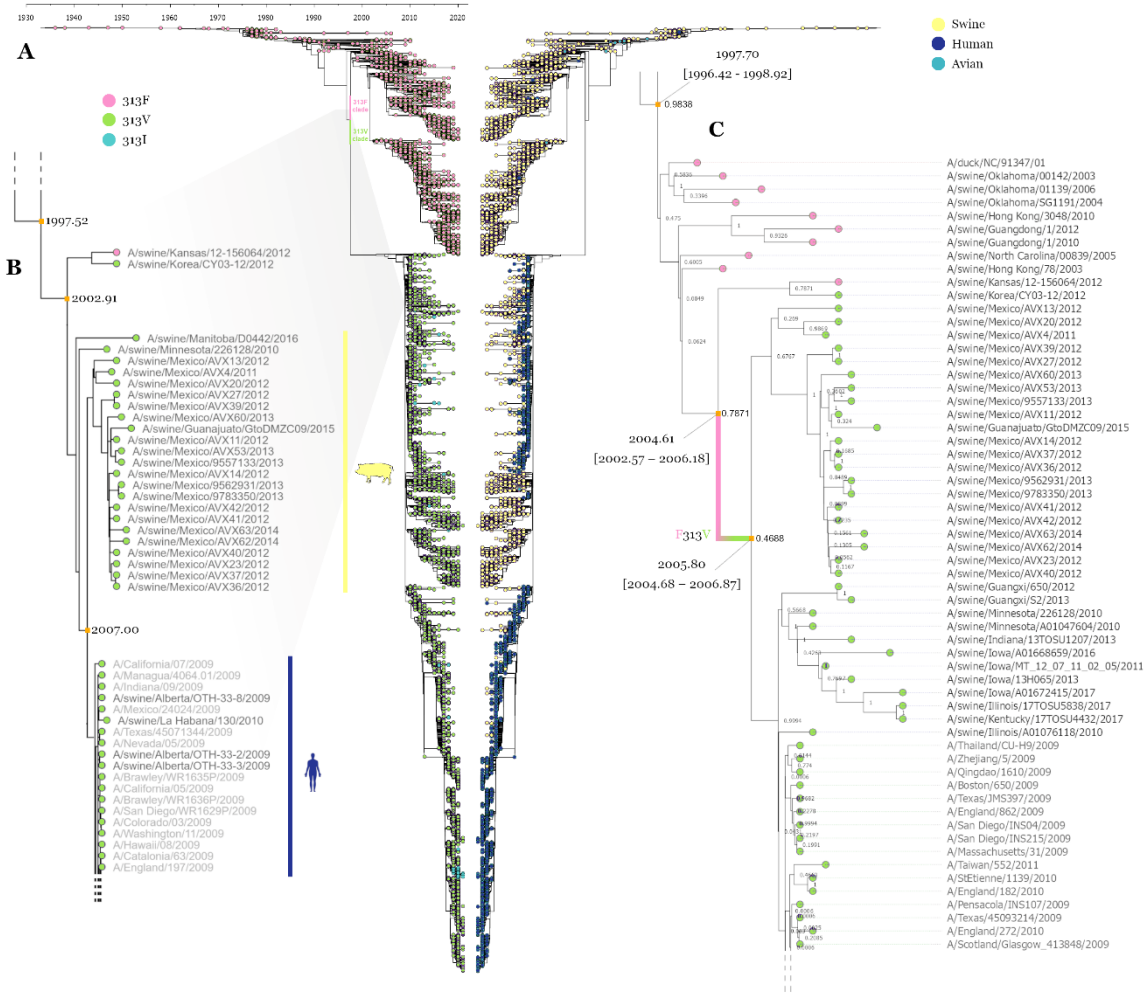


Figure 4.4. Molecular dating of the F313V NP substitution on the classical swine H1N1 lineage. **(A)** Tip-dated maximum likelihood phylogeny of all classical H1N1 lineage NP sequences annotated by position 313 residue (left) and isolation host (mirrored tree, right). **(B)** Zoomed in snippet of the part of the ML phylogeny shown in A where the F313V change has occurred. Tip shapes are coloured by 313 residue, estimated dates for key nodes are annotated, and strain names are shown on the right of the tips. **(C)** Zoomed in snippet of the part of the BEAST maximum clade credibility phylogeny where the F313V change has occurred. Tip shapes are coloured by 313 residue, median node age and 95% highest posterior density confidence intervals are annotated for key nodes, posterior probability values are shown for each node, and strain names are shown on the right of the tips. The branch where F313V is believed to have taken place on is annotated in colour (pink and green). Phylogenies were visualised using FigTree.

4.3.3 NP residue 52 is a key determinant of BTN3A3 resistance associated with avian IAV spillovers into humans

Although all major human-circulating IAV lineages have NPs with residues 313Y or 313V, evading human BTN3A3, a number of bird-to-human spillover viruses have a 313F residue, predicted to be susceptible to BTN3A3. A prime example of this is the H7N9 IAVs that have repeatedly transmitted from birds to humans, causing major outbreaks in Southeast Asia (Liu *et al.*, 2013; Yang *et al.*, 2016; W. Zhu *et al.*, 2018). The H7N9s have a different NP change at site 52, having residue N instead of Y, which bypasses BTN3A3's antiviral activity independent of changes in site 313 (Pinto *et al.*, 2023). Using the filtered dataset NP phylogeny (Figure 4.3), I retrieved all NP sequences matching to the broader avian NP clade where the H7N9 spillover viruses emerged from and re-inferred a time-calibrated phylogeny of the clade to examine the timing of the Y52N change in these viruses. Interestingly, the phylogenetic analysis shows that site 52 changed from Y to N only once in this clade, becoming the dominant residue in the part of the clade where most human infections have come from (Figure 4.5). The change is estimated to have taken place between August 1999 and October 2001 (1999.62 – 2001.81), a bit more than 10 years before members of this clade started emerging in humans, in March 2013 (Gao *et al.*, 2013). It is worth noting that the NPs of this relatively recent 52N clade are primarily found in H9N2 strains based on the analysed sequences (Figure 4.5). Spillovers of H9N2 viruses in humans have been documented, but these are usually linked to close contact with susceptible poultry animals and rarely lead to onward human to human transmission (Zhang *et al.*, 2022). Instead, constellations of mutations that might contribute to transmissibility in humans have been identified in at least five genome segments other than the NP in the re-emerging H7N9 viruses (W. Zhu *et al.*, 2018). Hence, the presented analysis illustrates that the Y52N change stochastically happened in an H9N2 bird virus, creating a pool of BTN3A3 evasion conferring NPs that have been continuously gained by the more "human transmission ready" H7N9 viruses through reassortment in both viruses' reservoir avian hosts.

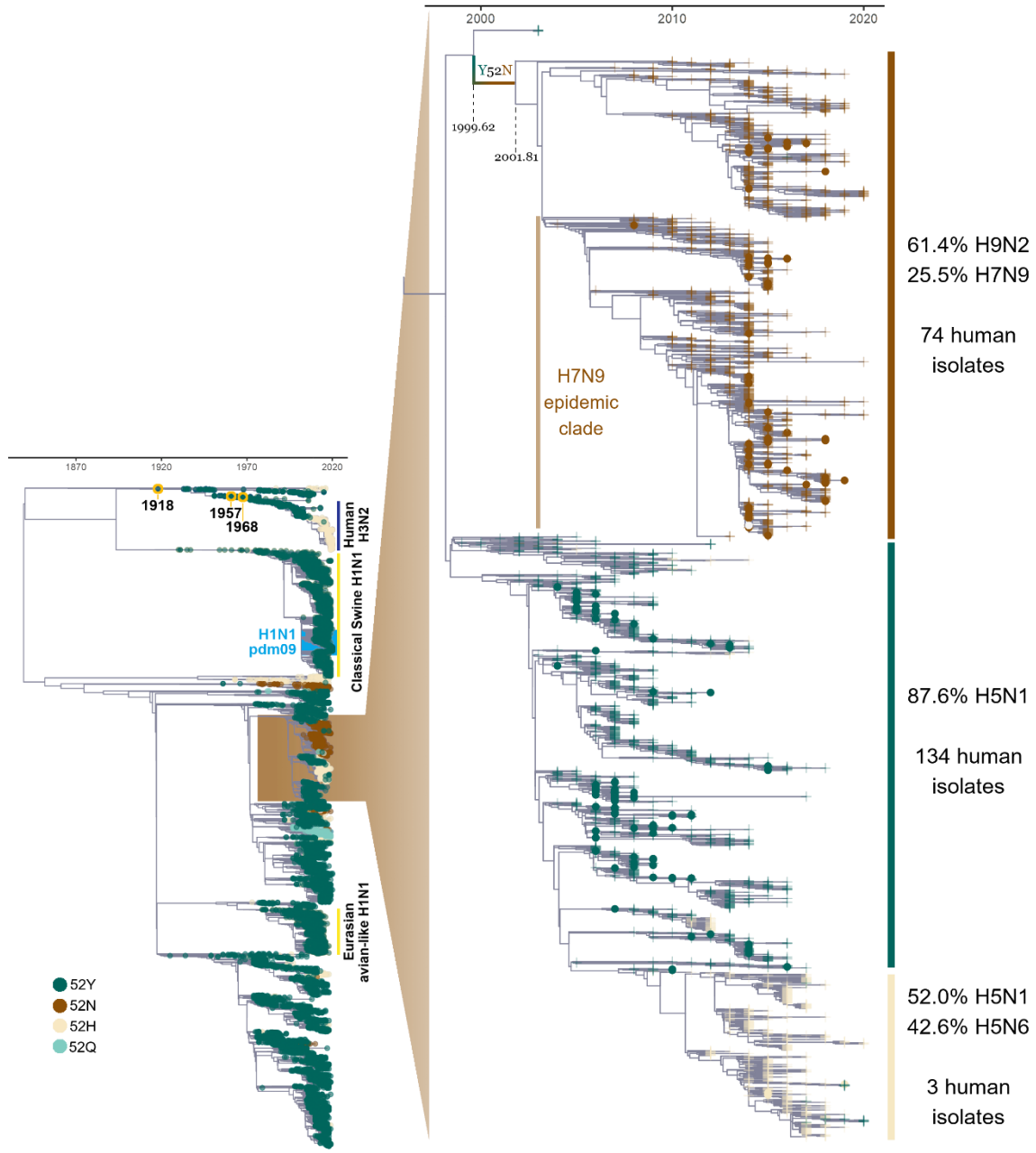


Figure 4.5. Tip-dated maximum likelihood phylogenies of the filtered IAV NP coding sequence dataset (left) and of all sequences clustering within the highlighted avian IAV clade (right). Tip shapes are coloured by site 52 residues (only residues present in more than 1% of the sequences are shown in the legend). Right: human isolates are annotated with circles and non-human isolates with transparent crosses to highlight human transmission. The branch where the Y52N change took place is annotated on the tree. Serotypes present in more than 10% of the sequences in each subclade are shown on the right of each subclade.

The sister NP clade to the 52N clade mainly consists of high pathogenicity H5N1 viruses. Most of these NPs have a 52Y, consistent with the majority of avian strains, however, another change at site 52 this time from Y to H has occurred in this subclade (Figure 4.5). Looking at the comprehensive IAV NP phylogeny annotated by site 52 residues, one can see that there are multiple independent subclades that

have non-52Y residues. The three amino acids that are most prevalent after Y are N, H and Q all of which enable evasion of human BTN3A3, even at the presence of 313F (Pinto *et al.*, 2023). IAV NPs with these three 52 residues are of concern to global health since they can bypass the BTN3A3 restriction factor, likely facilitating transmission to and between humans. To explore whether there is geographical clustering of the non-52Y NP clades I calculated changes from 52Y to N, H or Q across the NP phylogeny that led to clades of at least two sampled members and determined the countries where these viruses were sampled. Among the ~ 13,000 avian IAV NP sequences available in the present filtered dataset, I found a striking 151 independent avian NP lineages which already have a BTN3A3 resistant genotype. Members of these lineages have been sampled across the globe, with hotspots in China and North America correlating with sampling efforts (Figure 4.6). Concerningly, this geographic picture suggests that it is likely impossible to predict where the next BTN3A3 resistant lineage will occur.

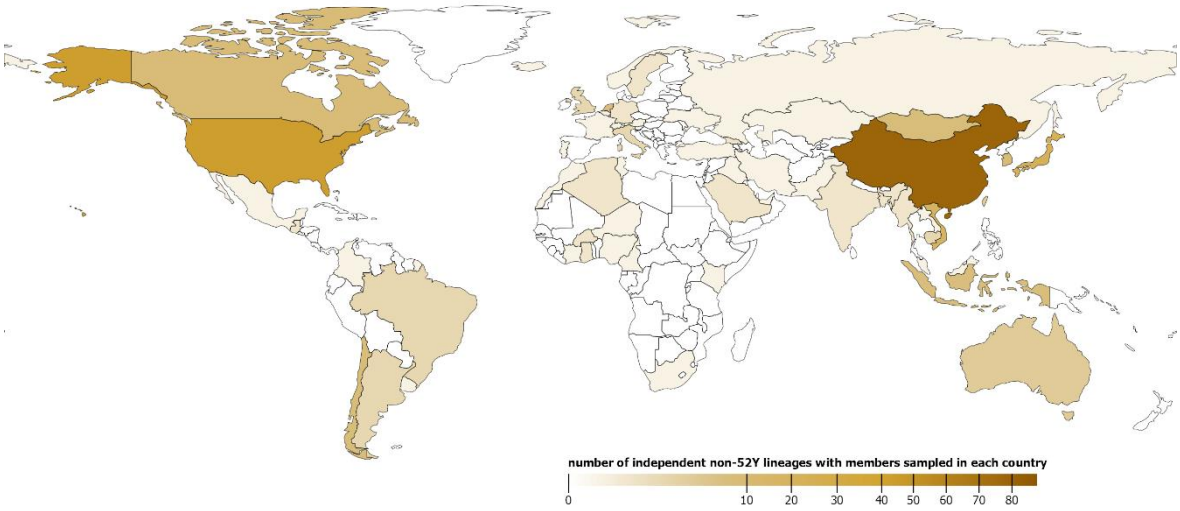


Figure 4.6. Geographical distribution of BTN3A3 resistant avian clade IAV NP independent lineages. Independent changes from the avian BTN3A3 susceptible 52Y genotype to N, H, or Q leading to a clade of at least two members identified in the avian clade of the IAV NP phylogeny. Map shading corresponds to the number of these lineages that each country has sampled at least one isolate from.

4.3.4 Highly pathogenic IAV does not require BTN3A3 evasion for transmission to humans

As shown in the above sections, virtually all IAV strains circulating or having transmitted to humans have changes in their NP site 52 or 313 that confers resistance to human BTN3A3-dependent restriction. There are, however, some groups of viruses isolated in humans that lack any of the known changes required to evade BTN3A3. These are primarily highly pathogenic avian influenza A (HPAI) viruses of the H5N1 serotype (Figure 4.5). The key determinant of the high pathogenicity for these strains is a polybasic cleavage site in the HA protein which expands the viruses' tissue specificity (Luczo *et al.*, 2015). To examine how the presence of this HPAI cleavage site compares to NP-dependent BTN3A3 resistance in non-canonical serotypes (H4-H10) that have spilled into humans, I retrieved a separate, larger dataset of NP and HA proteins from the Global Initiative on Sharing All Influenza Data (GISAID) database. For comparative purposes, I also analysed the composition of PB2 amino acid 627, as viruses harbouring the 627K (or 627V) mammalian adaptation are compatible with the mammalian Acidic Nuclear Phosphoprotein 32 Family Member A (ANP32A), facilitating replication, compared to viruses with the avian 627E (Taft *et al.*, 2015; Long *et al.*, 2016).

Non-canonical serotype genomes (H4-H10) were separated by the host they were isolated in (avian, swine and human) and the proportions of BTN3A3-sensitive/resistant and avian/mammalian ANP32A preference were compared between low (LPAI) and high (HPAI) pathogenicity viruses. Assignment of genomes into LPAI and HPAI was done based on the presence of the high-pathogenicity polybasic cleavage site on the viruses' HA proteins. The vast majority of avian isolates have an avian-like PB2 627 residue, consistent with the PB2-ANP32A compatibility required for efficient replication in each host (Long *et al.*, 2018) (Figure 4.7). On the contrary, avian viruses seem to tolerate the human BTN3A3-evading substitutions, although more than half of the avian isolates have the BTN3A3-resistant NPs, consistent with the phylogenetic reconstruction of the NPs (Figures 4.5 and 4.7). Despite having much fewer sequenced non-canonical serotype isolates from pigs, the proportions of both BTN3A3-resistance and ANP32A compatibility in LPAI and HPAI viruses mirror those of the avian isolates (Figure 4.7). This suggests that the potential barriers between bird-to-swine transmission of these serotypes are independent of all three variables. Hence, the proportions of

PB2 and NP haplotypes in viruses that have infrequently spilled over from birds to pigs simply reflect those in the avian reservoir hosts.

Interestingly, the proportions in the human isolates are strikingly different from these of avian and swine viruses. Starting with ANP32A compatibility, more than 70% of the LPAI genomes have a human-adapted PB2 protein (Figure 4.7). This proportion is much smaller in the HPAI genomes (33%), but still substantially larger than in avian and swine viruses (3% and 5% respectively). The PB2 adaptation to mammalian ANP32A is not crucial for the initial host switch and has been observed to occur once transmission has been established in the new host to improve replication efficiency. Most HPAI virus transmissions into humans are dead-end spillovers, suggesting that the viruses do not have enough time to acquire this PB2 adaptation. This would explain the unique pattern observed in human isolates, where avian PB2 viruses can transmit to humans, and onward transmission or prolonged replications quickly leads to the acquisition of the human ANP32A compatible substitutions. The proportions of human BTN3A3-resistant and BTN3A3-sensitive sequences, however, show a notably different pattern. In HPAI bird-to-human spillovers BTN3A3-resistance proportions seems to match those seen in avian isolates (avian: 31%, human: 26%). On the contrary, only four LPAI bird-to-human spillovers have ever been recorded and sequenced, representing a mere 0.3% of the dataset. This confirms that NP-dependent BTN3A3-resistance is virtually essential for transmission of an LPAI virus into humans. Instead, presence of an HA polybasic cleavage site seems to completely bypass this requirement. In fact, it is of interest to individually inspect the four BTN3A3-sensitive LPAI human isolates. One case is a multi-reassortant H7N4 virus, documented to have acquired human adaptations in its PB2, PB1, HA, NA and M2 proteins, the infected patient being a poultry farm worker who was in close contact with the farmed birds (isolate: A/Jiangsu/1/2018, GISAID accession: EPI_ISL_376123) (Qu *et al.*, 2020). Two of the cases are members of an H7N2 lineage with unique HA adaptations identified to be circulating in domestic cats (Marinova-Petkova *et al.*, 2017). The earlier case sampled in 2003 did not have an identified source (A/New_York/107/2003, GISAID accession: EPI_ISL_16424), but the second case sampled in 2016 was confirmed to have transmitted from a domestic cat (A/New_York/108/2016, GISAID accession: EPI_ISL_253575) (CDC: Morbidity and Mortality Weekly Report, 2004; Marinova-Petkova *et al.*, 2017). Finally, the last case is an H5N1 virus sampled in Vietnam in 2004 which surprisingly lacks the polybasic cleavage site in its HA

(A/Viet_Nam/1203/2004, GISAID accession: EPI_ISL_4156). However, three more H5N1 viruses with almost identical genomes sampled in Vietnam during the same year are all HPAIs (A/Viet_Nam/3062/2004, EPI_ISL_4158; A/Viet_Nam/3046/2004, EPI_ISL_4157; A/Viet_Nam/1194/2004, EPI_ISL_4155), suggesting that this particular genome may have lost the cleavage site after transmission into humans, during culture (it is unclear if the virus had been cultured before sequencing) or might simply be a sequencing error (Hien *et al.*, 2004).

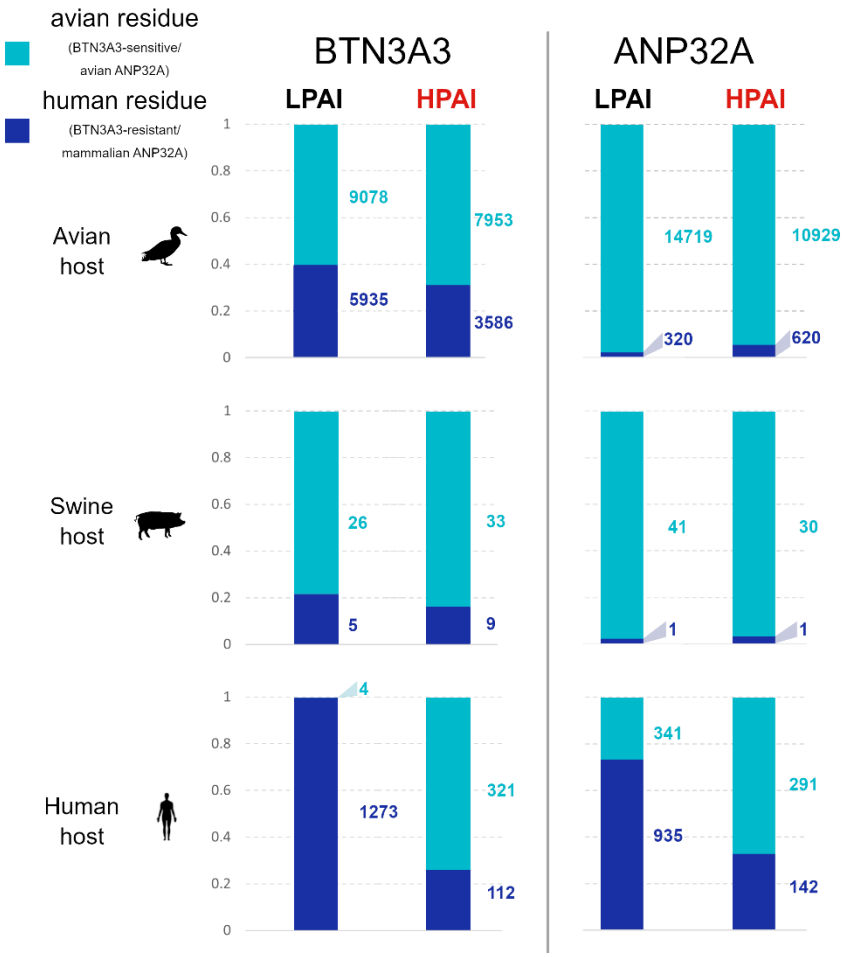


Figure 4.7. Distribution of BTN3A3 and ANP32A human-adapted residues and HA pathogenicity across H4-H10 IAV viruses. BTN3A3 NP-dependent status proportions is shown on the left and ANP32A PB2-dependent proportions shown on the right. Genomes are separated by isolation host (avian, swine and human from top to bottom). Absolute values of genomes for each category are shown on the right of the corresponding proportion bar.

4.4 Discussion

The comprehensive phylogenetic analysis of the *Haplorrhini* BTN3 genes and the IAV NP sequences presented in this chapter, in complement with the experimental results presented in (Pinto *et al.*, 2023), highlight the importance of human BTN3A3 as a restriction factor against avian IAVs. Many genes relating to antiviral immunity experience strong selective pressures and frequently undergo duplications (Duggal and Emerman, 2012). Similarly, all BTN3 paralogues are located on the same locus in the primate genomes which opens up opportunity for further duplications, domain swapping and subsequent diversification of function. Based on the available data, BTN3A3 is the primary human paralogue that confers anti-IAV activity, with BTN3A1 having a more subtle antiviral effect *in vitro* and BTN3A2 missing the PRYSPRY domain and lacking antiviral function altogether. The butyrophilins are known to have many diverse functions (Smith *et al.*, 2010; Afrache *et al.*, 2012; Arnett and Viney, 2014; Rigau, Uldrich and Behren, 2021), this being the first detailed example of an innate antiviral immunity function for the genes. Another recently published ISG screen suggests that the human BTN3A3 gene might have some antiviral activity against Ebola virus (Kuroda *et al.*, 2020), although a potential mechanism for that restriction is still untapped and similar experiments have not shown any specific antiviral activity of BTN3A3 against a large array of other human viruses (Pinto *et al.*, 2023).

Regarding the BTN3 *Catarrhini* genes' restriction of avian IAV, it is intriguing how the topology of no single protein domain can explain the gain of the antiviral activity (Figure 4.2). This means that multiple combinations of substitutions between different domains can independently lead to the function within the specific evolutionary context of the *Catarrhini* BTN3 gene clade. Although it may be tempting to propose that ancestral evolutionary pressures by avian IAV-like viruses have led to the independent gain events of antiviral function across the different BTN3 clades (Figure 4.2), no primates outside of humans are known to host IAVs. Taken together with the knowledge that BTN3 genes have many more unrelated functions that likely impose separate pressures on the genes, I propose that the BTN3 anti-avian IAV activity may be an "accidental" product of sequence changes relating to the genes' other functions. This scenario would be consistent with an initial "accidental" gain of antiviral function specific to the BTN3A1 and BTN3A3 *Catarrhini* clades after their original duplication event and subsequent occasional loss of this function across

different subclades (Figure 4.1). An alternative scenario where gain of antiviral function was selected for through an ancient avian IAV-like virus epidemic in early *Catarrhini* populations 40 million years ago is less likely, in my interpretation of the currently available data, but certainly warrants further investigation. For example, Souilmi et al. (2021) present evidence of a 20,000 year old human coronavirus epidemic specific to ancient East Asian populations, detected by examining ancestral selection in genes known to interact with coronavirus infection. Given enough resolution of primate genomes, a similar approach could potentially be implemented in the future to test if other genes known to interact with IAV infection may have been under unique adaptive selection at the base of the *Catarrhini*. Lastly, an interesting point about the butyrophilin genes is that the chicken one-to-one orthologue of the primate BTN3 genes, BTN1A1 or also referred to as Tvc, acts as the entry receptor of the unrelated Subgroup C Avian Sarcoma and Leukosis Viruses, ASLV(C) (Elleder *et al.*, 2005). This highlights the widely diverse functions of these genes and the convergent interactions of the butyrophilin family with diverse viruses, whether that is restricting virus replication or facilitating virus cellular entry.

Focusing on the virus side, the firm association between BTN3A3-evading NP adaptations and all major IAV human spillovers suggests that the ability to bypass BTN3A3 through this mechanism is a requirement for avian IAVs transmitting to humans. Interestingly, substitutions to the NP site 313 BTN3A3-evading residues 313Y and V, present in the vast majority of human-circulating IAVs (stemming from the 1918 flu pandemic and 2009 swine flu pandemic respectively) is largely infrequent in bird-circulating viruses (Figure 4.3). The strong conservation of NP residue 313F in avian IAVs is suggestive of purifying selection on this site during replication and transmission within avian hosts. Conversely, strains with both 313F and V genotypes circulate in pigs. This indicates relaxed selection for 313F on the virus when in swine hosts, potentially allowing for more frequent changes to BTN3A3-resistant 313 residues. Although there are no sequences from the exact swine population where the F313V substitution took place (estimated to have existed between 2006 and 2009), the phylogenetic analysis of the available sequences confidently supports that the residue change happened in pigs (Figure 4.4). The exact origins of the 1918 pandemic H1N1 virus to the human population remain much more elusive. Current virus sequence data suggest that different segments may have come together through reassortment between multiple co-circulating strains (Smith, Bahl, *et al.*, 2009), while it is believed that all segments

apart from the HA – thought to have come from IAV strains already circulating in humans – likely came from avian IAVs (Worobey, Han and Rambaut, 2014). Furthermore, the earliest available archival H1N1 sequences from 1918 have 313Y NP proteins (Patrono *et al.*, 2022). The data presented in this chapter could have some informative implications regarding the origin of at least the NP segment of the 1918 pandemic H1N1 virus. If 313F is highly conserved across avian-circulating virus, but 313Y or V is required for initial transmission into humans, then one could speculate that the NP of 1918 virus acquired the F313Y substitution while circulating in a non-avian intermediate host (swine or other). A better understanding of what avian mechanism leads to the conservation of NP 313F in these hosts may provide more clues as to potential host environments where the F313Y substitution happened before reassorting with the other 1918 virus segments and spilling into humans.

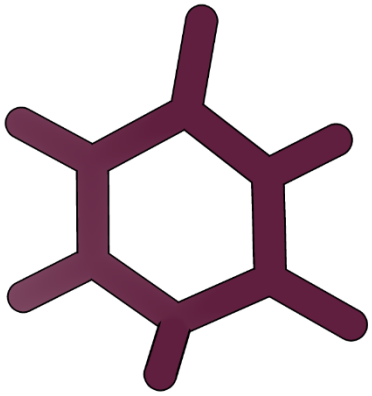
In contrast to the conserved site 313 residues, changes in site 52 of the NP seem to be less constrained during avian circulation of the viruses. Given the global distribution of multiple independent NP clades with BTN3A3-evading site 52 residues (Figures 4.5 and 4.6), viruses with these changes may be expected to be of greater concern to human health. Still, all four recent influenza pandemics (1918 H1N1, 1957 H2N2, 1968 H3N2 and 2009 H1N1) had site 313 BTN3A3-evading residues instead. Although counteracting restriction by BTN3A3 is likely a requirement for bird-to-human transmission, other segments (or other NP changes) could be responsible for how efficient onward human-to-human spread is. The H7N9 52N viruses have caused multiple epidemics in human populations, however the H9N2 viruses that share essentially the same pool of NP segments have only circumstantially spilled into humans without any documented onwards transmission. Hence, perhaps unsurprisingly, a combination of human-adapted changes across all segments of the virus genome control its potential to both cross to and transmit between humans. The substitutions in NP residues, like these at sites 52 and 313, must not simply interact with host proteins, but also the other virus proteins constituting the vRNP complex. An interesting follow up from the data presented here is to investigate whether certain serotypes or segment combinations are more likely to endure (or even prefer) the BTN3A3-evasive residues at sites 52 and 313.

The last important observation is how although virtually all LPAI viruses required a human-adapted NP to transmit to humans, that does not seem to be a requirement

at all for HPAI viruses (Figure 4.7). The defining feature of HPAIs is the polybasic cleavage site in the HA protein relating to more efficient cell entry, while BTN3A3 restricts virus replication post-entry (Pinto *et al.*, 2023); so it is unlikely that the same mechanism is at play allowing HPAIs to transmit to humans. Still, the exact mechanism through which human-adapted NPs counteract BTN3A3 restriction remains unclear and a potential direct or indirect intracellular interaction between HPAI HA peptides and BTN3A3 is not entirely out of the question. So far, documented HPAI human infections (primarily H5N1s) have high mortality rates, but are usually dead-end spillovers, rarely leading to any human-to-human transmission (Krammer and Schultz-Cherry, 2023). The world is currently facing an immense resurgence of H5N1 viruses globally circulating in birds, reassorting their internal segments (Xie *et al.*, 2022) and transmitting to wild mammals, causing mass mortality (Gamarra-Toledo *et al.*, 2023). A recent human infection by these H5N1 viruses was recorded in England in 2022 (Oliver *et al.*, 2022) and, interestingly the virus's NP had a BTN3A3-evasive residue (52H). Also, an outbreak of HPAI H5N1s in mink farms in Spain recorded in 2022 had the NP 52N residue, having acquired this segment from a distant gull H13 clade with 52N NPs (Agüero *et al.*, 2023). It remains unknown whether BTN3A3-evading NP adaptations and HPAI HAs could have an additive effect to the viruses' mortality and transmissibility within humans. What is certain is that BTN3A3-evasive HPAI viruses are encountered more frequently in the animal-to-human interface – either through contact with wild birds or farmed poultry and mammals – and this area warrants monitoring and further research into the molecular mechanisms underlying these effects. Early monitoring of avian IAVs focused on only sequencing the HA and NA segments, responsible for cell entry. It now becomes all the more apparent that robust mutational markers of the viruses' ability to transmit into humans can be found in internal segments, such as the NP. The new approach to monitoring IAVs should be built on a model where combinations and interactions of markers between all segments predict the viruses' potential for human spillover and spread. On a final note, three independent bird-to-human transmissions of H3N8 viruses have been recently documented in China: i) in April 2022 Henan province (A/Henan/4-10CNIC/2022 – EPI_ISL_12322556) (World Health Organization, 2022a), ii) in May 2022 in Hunan province (A/Changsha/1000/2022 - EPI_ISL_12703722) (Centre for Health Protection, 2022), iii) in March 2023 in Guangdong province (A/Guangdong/ZS-23SF005/2023 - EPI_ISL_17464053) (World Health Organization, 2023a) all of

which have residue 52N in their NPs. Based on the data presented in this chapter these H3N8 viruses could be a prime candidate for a future IAV outbreak in humans.

Chapter 5. Quantifying dinucleotide representation in virus genomes



Cartoon model of uracil.

PDB ligand entry: URA, visualised with ChimeraX.

“Never underestimate the ability of the human animal to adapt to its environment.”

Neon Genesis Evangelion (1995)

The original version of the Synonymous Dinucleotide Usage metric has been published in Lytras & Hughes (2020, *Viruses*, 12[4]:462), under the title "*Synonymous Dinucleotide Usage: A Codon-Aware Metric for Quantifying Dinucleotide Representation in Viruses*". I performed all the analysis presented in the paper and co-author Joseph Hughes supervised the work. All other work presented in this chapter, including the extended versions of the metrics and applications on the *Flaviviridae* family, is my own and currently unpublished.

Aim

Most virus genomes have compositional biases manifesting in the dinucleotide level. In this chapter I describe a novel approach for quantifying dinucleotide biases in coding sequences and explore how different nucleotide composition expectations affect the mathematical framework of the method. I apply the approach to genomes of the *Flaviviridae* and *Rhabdoviridae* virus families to investigate the effect that the host environment a virus replicates in has on the genomic dinucleotide composition.

5.1 Introduction

5.1.1 Biases in codon usage

The building blocks of all DNA sequences consist of four nucleotides: guanine (G), cytosine (C), adenine (A) and thymine (T) (or uracil – U – in the case of RNA genomes). The composition of nucleotides is largely non-uniform across genomes. In coding sequences, this is partly due to constraints by the peptide sequence being encoded by the gene. Each amino acid is encoded by specific sets of codons, so conservation on the protein sequence will affect the coding sequence's nucleotide composition. However, the genetic code is degenerate, meaning that some amino acids can be encoded by more than one synonymous codon, allowing some leeway to the nucleotide composition of a coding sequence underlying a conserved protein. We predominantly think of selective pressures acting on the amino acid level, i.e. since the proteins encoded by the genetic code largely control the phenotype, non-synonymous changes in coding sequences (changing the encoded protein sequence) will be under selection. For this reason, the majority of methods used routinely to assess selection on genes are based on comparing the proportion of non-synonymous to synonymous mutations on a given coding sequence (Goldman and Yang, 1994; Muse and Gaut, 1994) (see Chapter 1 Section 1.10). It is now accepted that, other than protein-level selection, changes on the nucleotide sequence (either synonymous changes on coding regions or mutations on non-coding sequences) can also confer phenotypic effects and subsequently be under selective pressure. Focusing on coding sequences, one widely studied area of synonymous selection is on codon usage bias. Selective pressures on codon usage are difficult to assess because of the underlying genetic signatures that could also be biased, either due to selective or mutational pressures.

One of the first approaches developed for studying the non-uniformity of codon usage is the relative synonymous codon usage (RSCU) metric (Sharp, Tuohy and Mosurski, 1986). The RSCU provides a numerical representation for whether specific synonymous codons are more or less abundant than one would expect under equal synonymous codon usage in a sequence. The earliest finding regarding the factors that might influence biases in codon usage is that codon usage correlates with the abundance of tRNA expressing the respective codons (Sharp, Tuohy and Mosurski, 1986). The availability of tRNA molecules present in the cell will affect the

speed and efficiency of mRNA translation. Hence, genes that need to be expressed faster will be selected to have codons corresponding to the tRNAs most abundant in the cell at the time of expression. Similarly, one can predict how highly expressed a gene is based on its codon usage by comparing it to that of reference genes of the organism (Sharp and Li, 1987). In fact, it is now apparent that codon usage selection is much more complex than simply consistently preferring one synonymous codon; instead distinct codon usage patterns can be selected for across different regions of a gene sequence. This relates to improving translational efficiency, but also pacing protein elongation for correct folding of the peptide product (Plotkin and Kudla, 2010; Tuller *et al.*, 2010; Hanson and Coller, 2018). Although selection for translation efficiency and protein folding is an important factor explaining codon usage biases, mutational pressures or even selection on underlying nucleotide patterns (single nucleotide or dinucleotide representation) can also contribute to shaping compositional biases in coding and non-coding sequences (Sharp *et al.*, 1995).

5.1.2 Biases in dinucleotide representation

Other than the biases in the codon usage of genetic entities, distinct patterns have also been extensively described on the underlying dinucleotide composition signatures of genomes. Dinucleotides, two nucleotides adjacent in a sequence bound by a phosphodiester bond, are known to be over- or under-represented across the genomes of living organisms and viruses, creating distinct compositional patterns (Beutler *et al.*, 1989). Shortly after genetic sequencing technologies emerged, researchers observed that distinct dinucleotide patterns - particularly depletion of the CpG dinucleotide - is shared between different vertebrates and across coding and non-coding regions (Russell *et al.*, 1976; Burge, Campbell and Karlin, 1992). The fact that under-representation of CpG dinucleotides was shared between vertebrate and plant genomes, which are methylated, but not methylase-absent organisms like invertebrates, bacteria and fungi, pointed to methylation being potentially relevant to this pattern (Bird, 1980; Karlin and Mrázek, 1997). Indeed, further experimental evidence illustrated that the primary cause for reduced CpG abundance in vertebrate and plant genomes is DNA methyltransferases acting on the nucleotides, frequently converting cytosine to thymine through deamination

following the methylation (Cooper and Krawczak, 1989). Another dinucleotide known to be slightly under-represented across the tree of life, including prokaryotes, is TpA/UpA (Burge, Campbell and Karlin, 1992; Karlin and Burge, 1995). Early speculations for the mechanism behind this compositional bias have suggested that UpA-rich mRNA might be unstable and more prone to degradation by cytoplasmic RNAses, most of which preferentially bind UpA and UpU-rich molecules (Shaw and Kamen, 1986; Beutler *et al.*, 1989; Duan and Antezana, 2003). However, the fact that intronic regions in some organisms also share the TpA under-representation suggests that this might not be the definitive mechanism behind this almost universal dinucleotide signature (Burge, Campbell and Karlin, 1992). These dinucleotide-level biases are likely responsible for overlapping compositional biases in codon usage as well as codon pair usage (Kunec and Osterrieder, 2016).

Interestingly, the dinucleotide signatures observed in eukaryotes and prokaryotes seem to also be shared by virus genomes, often matching their respective hosts' signatures (Subak-Sharpe *et al.*, 1966; Morrison *et al.*, 1967; Russell *et al.*, 1976). Early studies observed that CpG suppression and the weaker UpA suppression was present in small genome vertebrate viruses, but not in these with larger genomes (Karlin, Doerfler and Cardon, 1994). Further analyses of more virus groups confirmed that CpG suppression in both RNA and DNA small genome viruses matches that of their vertebrate hosts, suggesting that these signatures may be caused by host immune pressures that do not affect larger genome viruses (Shackelton, Parrish and Holmes, 2006). Analysis of dinucleotide signatures in all RNA viruses, including retroviruses, suggested that host had a major effect on the viruses' CpG biases (Cheng *et al.*, 2013). More comprehensive modelling of mutational processes showed that CpG and UpA suppression in mammalian viruses are due to selective pressures imposed by the host rather than mutational processes (Simmonds *et al.*, 2013). This led to the speculation that an innate immunity CpG sensor is present in vertebrates or mammals, selecting against virus genomes with this dinucleotide and, as a consequence, producing this mimicry between virus and host CpG levels (Belalov and Lukashev, 2013). These findings were followed by a number of experimental studies validating this prediction, where increasing UpA and CpG levels in RNA viruses led to a decrease in replication and subsequent viral attenuation, while decreasing their abundance has the opposite effect (Atkinson *et al.*, 2014; Tulloch *et al.*, 2014; Gaunt *et al.*, 2016; Witteveldt, Martin-Gans and Simmonds, 2016; Klitting *et al.*, 2018). The likely culprit at least for the CpG

suppression in viral genomes was recently elucidated to be the Zinc-finger Antiviral Protein (ZAP). Takata et al. (2017) demonstrated that ZAP selectively binds to CpG-rich viral RNA, inhibiting virion production of CpG-rich HIV-1 genomes but not of low CpG virus genomes. The discovery of ZAP's selective effect on virus CpG representation has been a breakthrough in our understanding of what causes dinucleotide biases in host and virus genomes. Still, other undiscovered mechanisms likely act on genomic dinucleotide representation, leading to the signatures that will be extensively described in this chapter.

5.1.3 The Zinc-finger Antiviral Protein selects for CpG depletion in virus genomes

The Zinc-finger Antiviral Protein (ZAP) contains a CCCH-type zinc-finger motif and was first described by Gao, Guo and Goff (2002) as an antiviral factor acting through binding of foreign RNA and initiating subsequent viral mRNA degradation (Odon *et al.*, 2019). There are two distinct isoforms of human ZAP, a long form (ZAP-L) that is constitutively expressed in the cell, and a short form (ZAP-S) that is induced by the interferon signalling pathway as an ISG (Schwerk *et al.*, 2019; Kmiec *et al.*, 2021). The ZAP-S isoform seems to be more closely involved to the innate immunity, potentially regulating the RIG-I pathway (Hayakawa *et al.*, 2010). Even prior to the discovery that ZAP proteins preferentially bind to CpG-rich viral RNA (Huang, Wang and Gao, 2010; Takata *et al.*, 2017; Luo *et al.*, 2020), many studies had described its antiviral activity against diverse viruses such as murine leukemia virus (MLV), Semliki virus (Kerns, Emerman and Malik, 2008), HIV-1 (Zhu *et al.*, 2011), hepatitis B virus (HBV) (Mao *et al.*, 2013) and more recently against the model RNA virus Echovirus 7 (Odon *et al.*, 2019), SARS-CoV-2 (Nchioua *et al.*, 2020) and even DNA viruses such as human cytomegalovirus (HCMV) (Lin *et al.*, 2020). On top of ZAP's antiviral activity through RNA binding, a few studies have described direct interactions between ZAP and viral proteins. For example, the H5N1 IAV PA-PB1 complex seems to directly associate with human ZAP-S during infection (Bradel-Tretheway *et al.*, 2011), while the Nsp4 protein of Porcine reproductive and respiratory syndrome virus (PRRSV) specifically cleaves ZAP in what seems to be a virus adaptation to counteract ZAP's activity (Zhao *et al.*, 2020).

The distinction between the long and short ZAP isoforms is not fully understood yet, since both proteins have shown antiviral activity against different viruses. However, ZAP-L has an additional post-translational modification signal at its C-terminal end which mediates S-farnesylation of the protein – the addition of a hydrophobic S-farnesyl group at the protein's end – and subsequent localisation to intracellular membranes (Charron *et al.*, 2013; Kmiec *et al.*, 2021). This means that ZAP-L could target viruses that replicate in intracellular compartments and double-membrane vesicles, such as the *Flaviviridae* and the *Coronaviridae*. Gonçalves-Carneiro et al. (2021) recently explored how ZAP's antiviral function compares between its vertebrate orthologues, showing that the CpG-targeting specificity of ZAP matured in mammals, while bird ZAP does not preferentially bind to a specific nucleotide context. ZAP does not act on its own, with at least two co-factors critical for the antiviral activity having been identified so far: i) TRIM25: which enhances the virus translation inhibition (Li *et al.*, 2017) and has co-evolved with ZAP so that only presence of conspecific TRIM25 can recapitulate the enhancement (Gonçalves-Carneiro *et al.*, 2021), and ii) KHNYN: which provides enzymatic activity along with TRIM25 to complement ZAP's RNA-binding activity and subsequent viral RNA degradation (Ficarelli *et al.*, 2019). Artificially increasing the CpGs in human viruses leads to their attenuation in human cells due to ZAP's antiviral effect (Sharp *et al.*, 2023). A few studies have suggested that synonymously increasing the UpA content of RNA viruses can also lead to – albeit weaker - viral restriction in human cells (Fros *et al.*, 2017; Ibrahim *et al.*, 2019) and this effect could also be due to ZAP recognition (Odon *et al.*, 2019). Although ZAP selectively binding to UpA-containing RNA motifs could explain the consistent under-representation of this dinucleotide in mammalian viruses, the same UpA suppression pattern is seen in non-mammalian viruses, while the known optimal binding motif of ZAP does not contain UpA (Luo *et al.*, 2020). Hence, viral genome UpA suppression is likely caused through an alternative, but potentially related, mechanism. Later in this chapter, I will explore the differences between CpG and UpA signatures in *Flaviviridae* genomes, investigating potential insights about their causes.

5.1.4 Methods for quantifying dinucleotide representation

In addition to exploring and interpreting dinucleotide signatures in viral genomes, this chapter will focus on the development of novel methods for quantifying

dinucleotide representation. One of the earliest approaches for examining dinucleotide signatures in genetic sequences involved estimating equilibrium nucleotide frequencies under mutation rates given by a simple mutation model, then comparing the observed occurrence of each dinucleotide to its expected equilibrium frequency (Sved and Bird, 1990). A simpler method that has been used routinely in the field is comparing the presence of a dinucleotide to that expected based on the frequency of each single nucleotide comprising the sequence (Karlin and Burge, 1995). This metric will be referred to as the relative dinucleotide abundance (RDA) later in the chapter. A later approach developed by Greenbaum et al. (2008) implements a Monte Carlo framework for fixing the amino acid structure and codon usage in a given coding sequence, so these effects are not reflected on the assessment of the dinucleotides' representation. More studies have extended the same approach, focusing on how virus dinucleotide signatures evolve following host switching (Gu *et al.*, 2019). This method is conceptually similar to the synonymous dinucleotide usage framework described in this chapter in that they both account for amino acid abundance and codon usage, as well as provide a statistical assessment of under- or over-representation. More complex, entropy-based approaches have also been developed, borrowing concepts from theoretical physics to estimate the extent of selection acting on the dinucleotide signatures of evolving viral sequences (Greenbaum *et al.*, 2014). The novel approach described in this chapter aims to statistically assess the presence of bias in dinucleotide representation and quantify the extent of the biases. Even though it does not formally examine the selective forces acting on the biases, it can be paired up with existing comparative phylogenetics methods to explore dinucleotide adaptations across phylogenies, which is presented later in the chapter (Results subsection 5.3.9).

5.2 Methods

5.2.1 DinuQ development

The Dinucleotide Quantification (DinuQ) Python3 package is a distributed package I have developed containing functions for calculating the Synonymous Dinucleotide Usage (SDU) and its associated metrics quantifying codon and dinucleotide representation in coding sequences (https://github.com/spyros-lytras/dinuq). The Results section of this chapter contains extensive descriptions of the SDU framework, metrics and the DinuQ package itself.

5.2.2 Testing RDA, SDUc and RSDUc on *Flavivirus* genomes

To test and compare the RDA, SDUc and RSDUc metrics described in this chapter, I retrieved the polyprotein coding sequences of vertebrate- and invertebrate-specific flaviviruses previously analysed in Simón et al. (2017). The analysis focuses on two representatives of the dataset: The insect-specific Aedes flavivirus (AEFV, GenBank accession: AB488408.1) that primarily infects *Aedes* mosquitoes (Blitvich and Firth, 2015), and the Apoi virus (APOIV, GenBank accession: AF160193.1) that has no known insect vector and infects rodents of the *Apodemus* genus (Billoir *et al.*, 2000). To extend the analysis, I retrieved a second dataset of coding sequences from all members of the *Rhabdoviridae* family included in the ICTV Virus Metadata Resource (version November 27, 2019; MSL34) (International Committee on Taxonomy of Viruses, 2019), being labelled as having a vertebrate or invertebrate host. The SDUc, RSDUc, RDA and RSCU values were calculated using the DinuQ Python package. General Linear Models (GLM) for statistical comparisons were performed using the R coding language (R Core Team, 2022).

5.2.3 Detecting adaptive shifts in dinucleotide representation

The set of representative *Flaviviridae* genomes were compiled by combining representative genomes that are part of the Flavi-GLUE database (“*237 reference genome sequences each representing a distinct flavivirid species and linked to*

isolate-associated data") by Bamford et al. (2022) with the recently published diverse genomes from Mifsud et al. (2023). All genomes were aligned and duplicated representatives of the same viruses from the two combined sets were manually excluded from the final set based on the labelled taxonomy and sequence identity. Further filtering involved excluding all sequences without a full RNA dependent RNA polymerase (RdRp or NS5) coding sequence available. First, the coordinates of annotated RdRp/NS5 genes were retrieved from the sequences' NCBI GenBank annotations. To capture potentially unannotated RdRp genes, I proceeded to align the full genomes of each subfamily separately using MAFFT v7.453 (Katoh and Standley, 2013) and manually retrieved the RdRp genes of unannotated genomes based on sequence similarity to the related sequences RdRp. The final set consisted of 350 *Flaviviridae* genomes that should represent all sampled diversity of the virus family (see online data, subsection 5.2.4). The annotated RdRp coding sequences were translated and all 350 amino acid sequences were aligned using MAFFT (--localpair option) (Katoh and Standley, 2013). The resulting alignment was used to reconstruct a phylogeny of these viruses using IQ-TREE v.2.1.3 (Minh *et al.*, 2020) under a Q.pfam+F+I+G4 substitution model selected by ModelFinder (iqtree -m TEST option) (Kalyaanamoorthy *et al.*, 2017). Node confidence was assessed with 1000 replicates of ultrafast bootstrapping implemented in IQ-TREE (Hoang *et al.*, 2018). The full or partial polyprotein coding sequences of all 350 genomes were extracted and SDUc and RSDUc values were calculated for all dinucleotides and all frame positions, assessing the error distribution with 100 replicates per sequence, using the DinuQ python package (described in this chapter, Results subsection 5.3.6).

To infer potential adaptive shifts in the representation of different dinucleotide signatures across the *Flaviviridae* phylogeny I used the PhylogeneticEM R package v.1.6.0 (Bastide, Mariadassou and Robin, 2017; Bastide *et al.*, 2018). This method can identify nodes in a tree representing adaptive shifts of a given multivariate quantitative trait by implementing a modified Ornstein–Uhlenbeck (OU) model of the trait changing across the tree. The RSDUc value of every informative frame position for each polyprotein was used as the quantitative trait representing each dinucleotide's presence. PhylogeneticEM requires a rooted ultrametric tree, hence the maximum likelihood RdRp phylogeny of the 350 *Flaviviridae* was midpoint rooted (separating the Hepaci-Pegi clade from the other *Flaviviridae* subgroups) and transformed using the chronos function of the ape R package v.5.7.1 (Paradis,

Claude and Strimmer, 2004; Paradis, 2013). The relaxed model was used for the transformation (having the largest log-likelihood value) with a lambda value of 1. The resulting ultrametric tree was then used for running PhylogeneticEM under the scalar OU (scOU) model. This method also requires defining a matrix A of values representing the selection strength used in the analysis (Bastide *et al.*, 2018). Matrix A can be calculated based on the length of the given tree using the find_alpha_grid function which, by default, calculates 10 A values. Smaller A values are expected to increase phylogenetic correlation in the model and this can result in mainly detecting shifts on the terminal branches of the tree (*personal observation; Paul Bastide, personal communication*). Since I am interested in detecting adaptive shifts on the deeper nodes of the *Flaviviridae* tree, I selected the five largest values in the automatically computed matrix A. In this way terminal branches are allowed to have relatively more different trait values, increasing shift detection sensitivity in the deeper nodes of the phylogeny. Finally, the K value, representing the maximum number of adaptive shifts that can be detected by the process was set to 10. Subclades of the original maximum likelihood tree were separated by node labels using the phytools R package (Revell, 2012) and were transformed to ultrametric and tested with PhylogeneticEM using the exact parameters described above.

5.2.4 Data availability

All code and data relating to this chapter are available in the following GitHub repositories: https://github.com/spyros-lytras/dinuq, https://github.com/spyros-lytras/flaviviridae_dn.

5.3 Results

5.3.1 The Synonymous Dinucleotide Usage framework

The Relative Dinucleotide Abundance (RDA) is the metric used routinely for quantifying biases in sequence dinucleotide representation. This involves comparing dinucleotide frequencies to those expected based on the single nucleotide composition of the sequence (Karlin and Burge, 1995). Although the RDA is a fitting approach for examining non-coding sequences, the nucleotide composition of protein encoding genes can be constrained by selective pressures for maintaining or removing specific encoded amino acids, especially if these have few synonymous codons (e.g. tryptophan and methionine can only be encoded by a single codon – UGG and AUG respectively). The Synonymous Dinucleotide Usage (SDU) is a novel framework for quantifying dinucleotide representation in coding sequences, while accounting for the potential effects of peptide composition on dinucleotide frequencies. It is based, in principle, on the relative synonymous codon usage (RSCU) metric (Sharp, Tuohy and Mosurski, 1986). In this section, I will describe the calculations behind the SDU and its extensions. Since most of the chapter focuses on coding sequences of RNA genomes, uracil (U) will be used instead of thymine (T) in the sequence notation throughout the chapter.

A coding sequence can have three distinct dinucleotide frame positions. We define the dinucleotide frame position 1 as the first and second nucleotide position of a codon, dinucleotide frame position 2 as the second and third nucleotide position of a codon, and dinucleotide bridge position as the third nucleotide position of a codon and the first position of the downstream codon. Each one of these positions can take a set of different dinucleotides without changing the amino acid (positions 1 and 2) or amino acid pair (bridge position) in the protein sequence, we define these as a set of synonymous dinucleotides. For example, threonine has four synonymous codons ACU, ACC, ACA, ACG. In dinucleotide position 2 of a codon encoding for threonine, there are four synonymous dinucleotides, CpU, CpC, CpA, and CpG. As such, the expected proportion of CpU occurring in position 2 coding for threonine under a null hypothesis of equal synonymous codon usage is: $e_i = 0.25$. The SDU (Table 5.1) compares the observed proportion of a synonymous dinucleotide of interest (o_i) to that expected under equal synonymous codon usage (e_i) for a given dinucleotide frame position. The ratio between o_i and e_i is calculated for each

different amino acid or amino acid pair (for the bridge position) and the SDU is defined as the weighted arithmetic mean of the ratios, weighted by the abundance of each amino acid in the sequence (Equation 5.1).

Table 5.1. Notation used to define SDU and RSDU.

Symbol	Description
i	Amino acid or amino acid pair
j	Dinucleotide
h	Dinucleotide frame position
n_i	Number of occurrences of amino acid or amino acid pair *i* in the sequence
k	Set of informative amino acids or amino acid pairs present in the sequence
$o_{i,j,h}$	Proportion of synonymous dinucleotide *j* in frame position *h* for amino acid or amino acid pair *i* observed in the sequence
$e_{i,j,h}$	Proportion of synonymous dinucleotide *j* in frame position *h* for amino acid or amino acid pair *i* expected under equal synonymous codon usage
N	Total number of amino acids or amino acid pairs present in the sequence

$$SDU_{j,h} = \frac{\sum_{i=1}^{k} n_i \times \frac{o_{i,j,h}}{e_{i,j,h}}}{N}$$

(Equation 5.1)

The set j includes 16 possible dinucleotide combinations. With three frame positions h, the matrix of $SDU_{j,h}$ has 48 possible combinations. Only 3 amino acids can be encoded by different position 1 dinucleotides (arginine, serine, and leucine),

meaning that 11 out of 16 dinucleotides in frame position 1 are non-informative, leaving 37 informative combinations.

The result of the SDU metric directly reflects the overall synonymous dinucleotide representation in each frame position of a given sequence:

- an SDU value of 1 indicates that the representation of the dinucleotide of interest in the given frame position is equal to that expected under the null hypothesis of equal synonymous codon usage;

- an SDU value of 0 indicates that the dinucleotide of interest is completely absent in the given frame position across the sequence;

- an SDU value greater than 1 indicates that the dinucleotide of interest is over-represented in the given frame position, compared to the representation expected under the null hypothesis;

- an SDU value between 0 and 1 indicates that the dinucleotide of interest is under-represented in the given frame position, compared to the representation expected under the null hypothesis.

The number of amino acids or amino acid pairs that can be synonymously encoded by a certain dinucleotide varies between dinucleotides and frame positions. This means that SDU measurements for different positions and dinucleotides can reach different maximum values. Under- (SDU < 1) and over-representation (SDU > 1) can still be consistently interpreted between positions and dinucleotides, since an SDU of 1 always reflects complete agreement with the null hypothesis.

5.3.2 The corrected Synonymous Dinucleotide Usage (SDUc)

In the first iteration of the SDU metric, the expected codon usage was assumed to be equal for all synonymous codons. This assumption facilitates calculations, since a precompiled set of $e_{i,j,h}$ values can be used, corresponding to each combination of dinucleotide, frame position and amino acid or amino acid pair. However, under this assumption, the metric likely misrepresents the true bias in dinucleotide

frequencies for sequences that experience biased synonymous codon usage. The first extension of SDU aims to account for this problem by varying the null hypothesis of expected codon usage. The corrected Synonymous Dinucleotide Usage (SDUc) requires an additional step for calculating expected codon frequencies based on the single nucleotide composition of the sequence of interest. First, nucleotide frequencies are calculated simply as the number of occurrences of each nucleotide over the sequence length (Equation 5.2). Then, the expected frequency for three consecutive nucleotides will be the product of their individual frequencies (Equation 5.3). Since only 61 out of the 64 triplet combinations are coding, we also need to correct the frequency calculations to exclude the 3 stop codons (UAA, UAG and UGA). This can be done by multiplying each codon frequency with the inverse of the sum of calculated frequencies, excluding the stop codons (Equation 5.4). After all corrected expected codon frequencies have been calculated, they can be used for inferring the corrected expected synonymous proportion of a dinucleotide *j* in frame position *h* for amino acid or amino acid pair *i*: $e'_{i,j,h}$. This is shown in Equation 5.5, where set *a* is the set of synonymous codons (or codon pairs) that encode for the informative amino acids (k) and contain the dinucleotide of interest in the relevant coding position and *b* is set of all synonymous codons (or codon pairs) that encode for the informative amino acids (k) regardless of dinucleotide content. To clarify, the expected proportions of bridge position dinucleotides will be the product of the expected proportions for each of the two amino acids in the pair. Other than the ability to vary the null expectation of synonymous dinucleotide proportions, the results of the SDUc (Equation 5.6) can be interpreted in the same way as the SDU.

$$nucleotide\ frequency\ (f_N) = \frac{nucleotide\ count}{sequence\ length}$$

(Equation 5.2)

$$codon\ frequency\ (f_{Cod}) = \prod_{n=1}^{3} f_N$$

(Equation 5.3)

$$f'_{Cod} = f_{Cod} \times \frac{1}{1 - (f_{UAA} + f_{UAG} + f_{UGA})}$$

(Equation 5.4)

$$e'_{i,j,h} = \frac{\sum_{n=1}^{a} f'_{Cod_n}}{\sum_{n=1}^{b} f'_{Cod_n}}$$

(Equation 5.5)

$$SDUc_{j,h} = \frac{\sum_{i=1}^{k} n_i \times \frac{o_{i,j,h}}{e'_{i,j,h}}}{N}$$

(Equation 5.6)

5.3.3 The corrected Relative Synonymous Dinucleotide Usage (RSDUc)

Contrary to the SDU metric, the SDUc metric now accounts for appropriate synonymous codon usage expectations based on each sequence's single nucleotide frequencies. However, similar to SDU, the magnitude of over-representation cannot be compared between different positions and dinucleotides. This is because maximum SDUc values will vary between frame positions and dinucleotides due to amino acids being encoded by different numbers of codons. In order to make this comparison, the SDUc metric can be extended into the corrected Relative Synonymous Dinucleotide Usage (RSDUc). This is simply the calculated SDUc value, normalised by the maximum SDUc for this position and dinucleotide (Equation 5.7). This extension to the metric does not have a consistent scale for comparison to the null hypothesis, and instead allows for comparing relative dinucleotide representation on the same scale for all possible dinucleotide, frame position and sequence parameters.

$$RSDUc_{j,h} = \frac{\sum_{i=1}^{k} n_i \times \frac{o_{i,j,h}}{e'_{i,j,h}}}{\sum_{i=1}^{k} n_i \times \frac{1}{e'_{i,j,h}}}$$

(Equation 5.7)

The results of the RSDUc are as follows:

- an RSDUc of 1 indicates that only the dinucleotide of interest is being used in the sequence, all the other synonymous dinucleotides being absent for the given position;

- an RSDUc of 0, similar to the SDU, indicates that the dinucleotide of interest is completely absent in the given frame position.

With this extension, the magnitude of synonymous dinucleotide representation can be directly compared between dinucleotides, frame positions and codon usage expectations.

5.3.4 SDUc maxima reflect the genetic code's complexity

Since RSDUc values are scaled by the maximum value of the corresponding SDUc calculation, it is of interest to explore how maximum SDUc values can vary depending on different parameters. In this section I will showcase calculations behind the relationship between SDUc maxima and expected single nucleotide frequencies for different dinucleotide-position combinations. This will also hopefully further highlight the mathematical framework of the SDU.

For the purposes of this section, I will make a few assumptions to facilitate the illustration of the metric. First, we can assume equal occurrence of all informative amino acids. Hence, there is no need to weigh the average for the SDU calculation (n_i = 1 for every *i* in set k), and the maximum SDU value is defined as shown in Equation 5.8, where the proportion of synonymous dinucleotides, $o_{i,j,h}$, is always equal to 1.

$$SDU_{j,h}^{MAX} = \frac{\sum_{i=1}^{k} \frac{1}{e_{i,j,h}}}{N}$$

(Equation 5.8)

Given the set of expected synonymous dinucleotide proportions under equal synonymous codon usage ($e_{i,j,h}$), the maxima of all SDU values can be calculated. Dinucleotide-position combinations corresponding to fewer codons / codon pairs tend to have lower maximum SDU values (Figure 5.1), for example position 1 values (with the exception of ApGpos1 – since it corresponds to two out of six codons encoding for arginine). Maximum SDU values will be directly related to the expected single nucleotide frequencies and, intuitively, this relation should be more complex as the set of informative amino acids or amino acid pairs (k) increases. Bridge position calculations have larger sets of k, since there are more pair combinations than single amino acids, and generally have maximum SDU values larger than those of frame position 1 and 2 combinations (Figure 5.1).

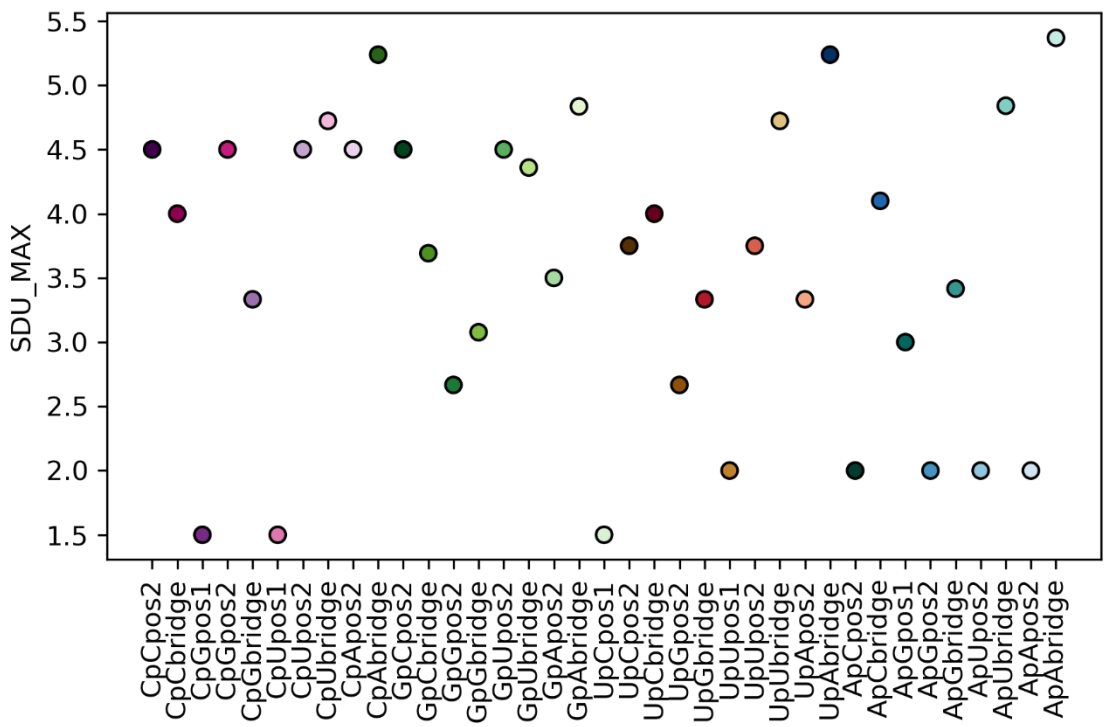


Figure 5.1. SDU maxima. Maximum values of SDU for every informative dinucleotide-position combination, assuming equal synonymous codon usage.

The relation between expected single nucleotide frequencies and SDU can be explored when calculating the SDUc, where the corrected expected proportion of synonymous dinucleotides ($e'_{i,j,h}$) can vary accordingly. Maximum SDUc will be inversely proportional to $e'_{i,j,h}$ (Equation 5.9), hence also to the frequency of

nucleotides contained in the dinucleotide of interest. The exact relation, however, will depend on which amino acids or amino acid pairs can be synonymously encoded by codons containing the relevant dinucleotide.

$$SDU_{j,h}^{MAX} \propto \sum_{i=1}^{k} \frac{1}{e'_{i,j,h}}$$

(Equation 5.9)

The second assumption we will make to test the relation between SDUc maxima and expected single nucleotide composition is that when the frequency of one nucleotide is changed, the frequencies of the other three nucleotides are kept equal to one another. Given these assumptions, I calculated $e'_{i,j,h}$ and maximum SDUc values for every dinucleotide-position combination, varying the frequency of each nucleotide independently (Figure 5.2). The relation between single nucleotide frequency and SDUc maxima for different dinucleotides-positions proved to be non-intuitive, with relation shapes ranging from linear to positive and negative exponential and U-shaped trends. These trends essentially reflect the genetic code's complexity and how single nucleotide composition biases affect expectations for synonymous dinucleotide proportions at different frame positions. To highlight how these relations come to be, I include three examples of explicitly calculating maximum SDUc values below.

Example 1

ApN position 2 SDUc maximum values (N standing for any of the four nucleotides) were always unaffected by changing two out of the four nucleotide frequencies (ApA and ApG when changing frequencies of C and U, ApC and ApU when changing frequencies of A and G, Figure 5.2). This is because all codons containing ApN dinucleotides in their second frame position always only have one other synonymous codon. For example, ApGpos2:

$$SDUc_{ApG,pos2}^{MAX} = \frac{\sum_{i=1}^{k} \frac{1}{e_{i,j,h}}}{N}$$

(Equation 5.10.1)

In this case k will be the set of three amino acids: glutamine (Q), lysine (K) and glutamate (E). We can hence expand the sum in the equation:

$$SDUc_{ApG,pos2}^{MAX} = \frac{\frac{1}{e'_{Q,ApG,pos2}} + \frac{1}{e'_{K,ApG,pos2}} + \frac{1}{e'_{E,ApG,pos2}}}{3}$$

(Equation 5.10.2)

The formula of the e' factor is defined in Equation 5.5. Using this and inversing the fractions in the numerator we get:

$$SDUc_{ApG,pos2}^{MAX} = \frac{\frac{f'_{CAA} + f'_{CAG}}{f'_{CAG}} + \frac{f'_{AAA} + f'_{AAG}}{f'_{AAG}} + \frac{f'_{GAA} + f'_{GAG}}{f'_{GAG}}}{3}$$

(Equation 5.10.3)

Simply looking at Equation 5.10.3 explains why the maximum SDUc value for ApGpos2 might be unaffected by the frequency of Us, since no informative amino acids are encoded by synonymous codons that contain U. Expected codon frequencies are defined as the product of their respective nucleotide frequencies (Equation 5.3), hence the codon frequency terms can be simplified to nucleotide frequencies, some of which cancel out in the fractions:

$$SDUc_{ApG,pos2}^{MAX} = \frac{\frac{f'_A + f'_G}{f'_G} + \frac{f'_A + f'_G}{f'_G} + \frac{f'_A + f'_G}{f'_G}}{3}$$

(Equation 5.10.4)

Assuming equal frequencies of non-varying nucleotides for the purposes of this test means that f'_A is equal to f'_G when varying the frequencies of C and U. In that case:

$$SDUc_{ApG,pos2}^{MAX} = \frac{2 + 2 + 2}{3} = 2$$

(Equation 5.10.5)

This explains why the maximum SDUc for this dinucleotide-position combination remains constant (value of 2) when varying U and C frequencies, as long as A and G frequencies remain equal to one another (Figure 5.2).

Example 2

The maximum SDUc value for CpUpos1 maintains a linear positive relation with nucleotide frequencies of A, G and U when keeping the other frequencies equal (Figure 5.2). Since this is a first frame position dinucleotide measure, set of k includes a single informative amino acid which is leucine (L):

$$SDUc_{CpU,pos1}^{MAX} = \frac{1}{\frac{e'_{L,CpU,pos1}}{1}}$$

(Equation 5.11.1)

Leucine can be encoded by six synonymous codons, four of which contain CpU at frame position 1, hence:

$$SDUc_{CpU,pos1}^{MAX} = \frac{f'_{CUU} + f'_{CUC} + f'_{CUA} + f'_{CUG} + f'_{UUA} + f'_{UUG}}{f'_{CUU} + f'_{CUC} + f'_{CUA} + f'_{CUG}}$$

(Equation 5.11.2)

This can be simplified to:

$$SDUc_{CpU,pos1}^{MAX} = 1 + \frac{f'_{UA} + f'_{UG}}{f'_{CU} + f'_{CC} + f'_{CA} + f'_{CG}}$$

(Equation 5.11.3)

Now we can vary the frequency of U while maintaining the other three frequencies equal to one another, defined as a value f_e (f'_e in its corrected form):

$$SDUc_{CpU,pos1}^{MAX} = 1 + \frac{f'_U \times 2f'_e}{f'_e \times (f'_U + 3f'_e)} = 1 + \frac{2f'_U}{f'_U + 3f'_e}$$

(Equation 5.11.4)

Given that the four nucleotide frequencies should sum up to 1, the denominator can be modified as follows:

$$SDUc_{CpU,pos1}^{MAX} = 1 + \frac{2}{1 - 3f'_e + 3f'_e} f'_U = 1 + 2f'_U$$

(Equation 5.11.5)

Only when varying the frequency of C while keeping the other three nucleotide frequencies equal the relation is exponential instead of linear. This is because f_C is the only frequency not in the numerator.

Example 3

Looking at a more complex relation, GpUpos2 has a U shape trend between the 0 and 1 boundaries when varying each of the four nucleotide frequencies (Figure 5.2). The set k for this dinucleotide-position contains four amino acids: cysteine (C), arginine (R), serine (S) and glycine (G) (Equation 5.12.1).

$$SDUc_{GpU,pos2}^{MAX} = \frac{\frac{1}{e'_{C,GpU,pos2}} + \frac{1}{e'_{R,GpU,pos2}} + \frac{1}{e'_{S,GpU,pos2}} + \frac{1}{e'_{G,GpU,pos2}}}{4}$$

(Equation 5.12.1)

For clarity, we can simplify each part of the numerator's sum separately, in all cases assuming that f_U is varied while $f_A = f_G = f_C$ are all equal with a value f_e (as in Example 2):

$$\frac{1}{e'_{C,GpU,pos2}} = \frac{f'_{UGU} + f'_{UGC}}{f'_{UGU}} = 1 + \frac{f'_e}{f'_U}$$

(Equation 5.12.2)

$$\frac{1}{e'_{R,GpU,pos2}} = \frac{f'_{CGU} + f'_{CGC} + f'_{CGA} + f'_{CGG} + f'_{AGA} + f'_{AGG}}{f'_{CGU}}$$
$$= 1 + \frac{f'_{CC} + f'_{CA} + f'_{CG} + f'_{AA} + f'_{AG}}{f'_{CU}} = 1 + 5\frac{f'_e}{f'_U}$$

(Equation 5.12.3)

$$\frac{1}{e'_{S,GpU,pos2}} = \frac{f'_{UCU} + f'_{UCC} + f'_{UCA} + f'_{UCG} + f'_{AGU} + f'_{AGC}}{f'_{AGU}}$$
$$= 1 + \frac{f'_{UCU} + f'_{UCC} + f'_{UCA} + f'_{UCG} + f'_{AGC}}{f'_{AGU}} = 1 + \frac{f'^2_U + 3f'_U f'_e + f'^2_e}{f'_U f'_e}$$
$$= 4 + \frac{f'^2_U + f'^2_e}{f'_U f'_e}$$

(Equation 5.12.4)

$$\frac{1}{e'_{G,GpU,pos2}} = \frac{f'_{GGU} + f'_{GGC} + f'_{GGA} + f'_{GGG}}{f'_{GGU}} = 1 + 3\frac{f'_e}{f'_U}$$

(Equation 5.12.5)

If we put the simplified terms back in Equation's 5.12.1 numerator, then:

$$SDUc^{MAX}_{GpU,pos2} = \frac{(1 + \frac{f'_e}{f'_U}) + (1 + 5\frac{f'_e}{f'_U}) + (4 + \frac{f'^2_U + f'^2_e}{f'_U f'_e}) + (1 + 3\frac{f'_e}{f'_U})}{4}$$

(Equation 5.12.6)

And simplify as follows:

$$SDUc^{MAX}_{GpU,pos2} = \frac{7 + 9\frac{f'_e}{f'_U} + \frac{f'^2_U + f'^2_e}{f'_U f'_e}}{4}$$

(Equation 5.12.7)

Given that the four nucleotide frequencies should sum up to 1, we can replace all f'_e terms with the corresponding f'_U expression:

$$SDUc_{GpU,pos2}^{MAX} = \frac{7 + 9\frac{\frac{1-f'_U}{3}}{f'_U} + \frac{f'^{\,2}_U + \frac{(1-f'_U)^2}{9}}{f'_U\frac{1-f'_U}{3}}}{4}$$

(Equation 5.12.8)

Which can then be simplified as:

$$SDUc_{GpU,pos2}^{MAX} = \frac{7 + 3\frac{1-f'_U}{f'_U} + \frac{3f'^{\,2}_U + \frac{1}{3}(1 - 2f'_U + f'^{\,2}_U)}{f'_U - f'^{\,2}_U}}{4}$$

$$= \frac{7}{4} + \frac{3}{4}\frac{1-f'_U}{f'_U} + \frac{1}{12}\frac{1 - 2f'_U + 10f'^{\,2}_U}{f'_U - f'^{\,2}_U}$$

(Equation 5.12.9)

This much more complex relation between maximum SDUc values and single nucleotide frequency shown in Equation 5.12.9 (for the limits $0<f'_U<1$) has an asymmetrical U shape, just as calculated for individual datapoints in Figure 5.2.

Overall, the relations between the expected single nucleotide composition of a sequence and the maximum SDUc value that can be calculated for that given sequence vary largely between dinucleotide-position combinations and overall nucleotide composition. As highlighted by the examples above, this variation is explained by the degeneracy of the genetic code, since the expected proportion of synonymous dinucleotides depends on the synonymous codons that contain these dinucleotides in the relevant frame position. To illustrate the metric, equal occurrence of informative amino acids and equal frequencies of three nucleotide frequencies were assumed. The relations between maximum SDUc and nucleotide frequencies presented in Figure 5.2 will change once these assumptions are dropped. Generally, SDUc maxima are similar to these expected by equal synonymous codon usage (Figure 5.1), and only extreme nucleotide composition biases will exponentially inflate the values (Figure 5.2). These extreme compositions (nucleotide frequencies <0.1 or >0.9) are unlikely to be relevant to real (or biologically relevant) coding sequences.

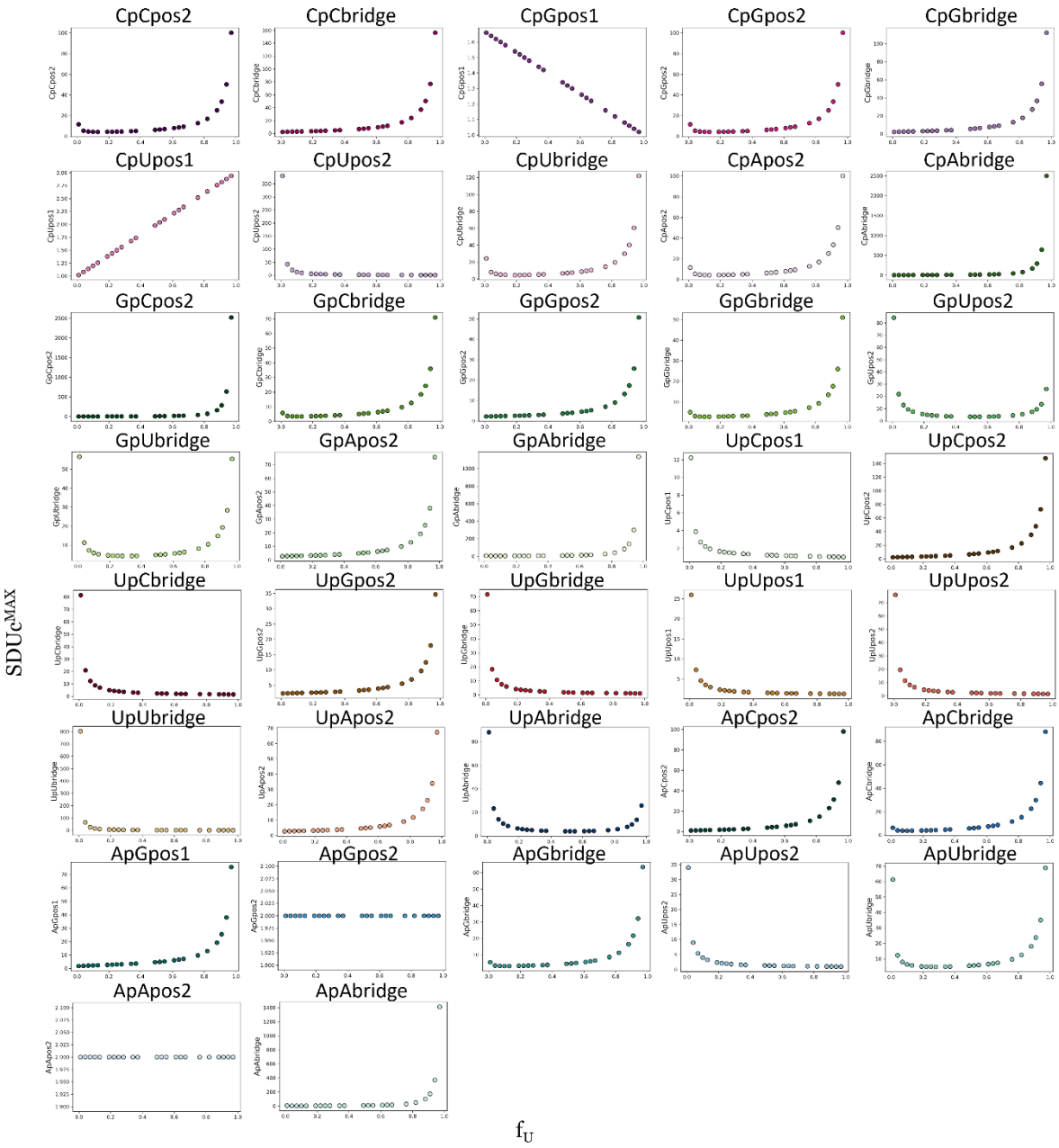


Figure 5.2. Relation between maximum SDUc values and uracil/thymine frequency. Datapoints correspond to varying nucleotide frequency values when the frequency of the other three nucleotides (which is kept equal between them) is a value with less than five nonzero decimals. Relations for all 37 possible dinucleotide-position combinations are presented.

5.3.5 Quantifying error around the null expectation

The SDUc metric equals to exactly 1 when all observed synonymous dinucleotide proportions (o') equal to these expected based on the sequence's single nucleotide composition (e'). However, sequences can theoretically abide to the null expectation of synonymous dinucleotide usage without all o' values being exactly equal to their corresponding e' values. This is simply due to shorter sequences having too limited information to have $o'_{i,j,h}$ being equal to $e'_{i,j,h}$ for every informative amino acid / amino acid pair i. Hence, it is possible to quantify the variability around the SDUc metric's null expectation by calculating the metric for a number of different coding sequences that have codons at the same frequencies as the null expectation and all encode the same protein sequence of equal length.

Given an existing coding sequence for which the SDUc metric has been calculated, the sequence can be translated to protein and then, for each amino acid in the sequence, synonymous codons can be resampled based on their corrected expected codon frequencies (f'_{Cod}, Equation 5.4). In this way, the resampled coding sequence should abide to the null expectation of synonymous dinucleotide representation, given the synonymous codon usage expectation. This can be done for any number of iterations to produce a normal distribution representing the random error of the metric for the given sequence abiding to the null hypothesis. The same process can be done for estimating the null expectation distribution of RSDUc values, where, instead of 1, the distribution's mean is expected to be 1 over the maximum SDUc value for the dinucleotide-position combination and the sequence it is calculated for.

Because the SDU framework splits up the sequence information into many categories (i.e. independent proportions of the dinucleotide of interest for each amino acid or amino acid pair), if the given sequence is short there will be less information in each category and subsequently more variability/error in the calculated value. Therefore, the null distribution's error is expected to vary primarily depending on the length of the sequence that the metric is calculated for. To test this, I randomly simulated 10 amino acid sequences of different lengths sequentially increasing by 10% of the longest sequence's length (from 700 to 7000 amino acids long). The codon resampling approach was then implemented on the amino acid sequences, assuming equal synonymous codon usage, and SDU was calculated

for each random sample. As expected, error inversely correlates with sequence length in a logarithmic fashion (Figure 5.3), with shorter coding sequences (2,100 bp) having a standard deviation of about 0.15, but this sharply drops with increasing length (Figure 5.3). By principle, the standard deviation of the distribution should approach 0 as the sequence length increases. Based on the simulation experiment presented here, the magnitude of error is consistently very low for sequences longer than about 17,000 bp (standard deviation < 0.05).

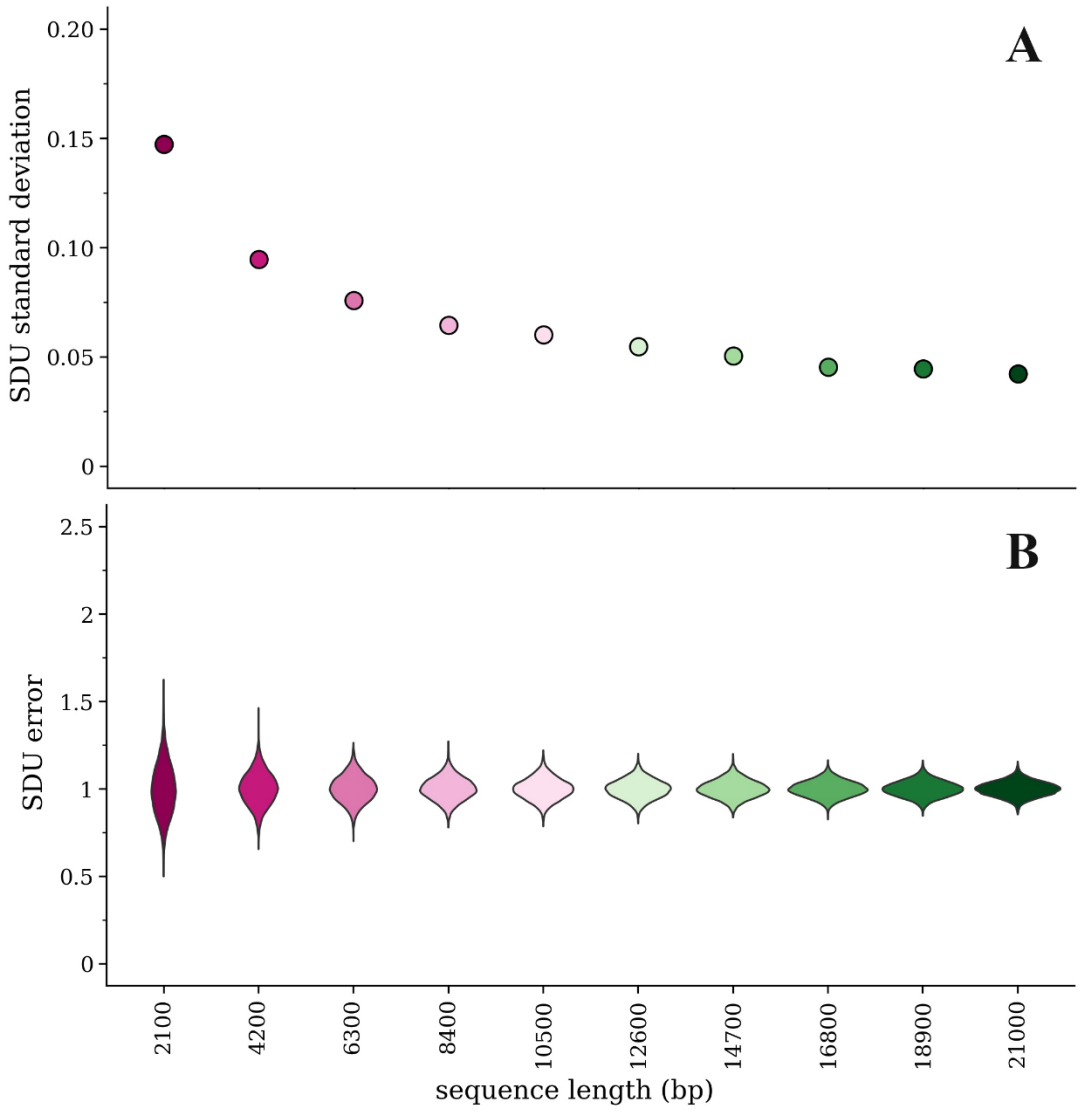


Figure 5.3. Relation of SDU error and sequence length. Comparison of error for the $SDU_{CpGbridge}$ of 10 simulated amino acid sequences of different lengths with 1000 random samples of nucleotide sequences for each amino acid sequence: **(A)** Standard deviation of the mean of the SDU error distributions; **(B)** Violin plots of the error distribution for each simulated sequence.

5.3.6 The DinuQ python package

To implement the SDU framework and allow for other researchers to readily use it, I have developed a Python3 package called DinuQ (Dinucleotide Quantification package). DinuQ is distributed through the python package repository PyPI (https://pypi.org/) and all code for local installation can be found at the relevant GitHub repository (https://github.com/spyros-lytras/dinuq). All sets of changes are monitored with versioned releases (current release is v.1.2.0). The majority of modules performed by DinuQ utilise core python functions, although biopython (https://biopython.org/) (Cock *et al.*, 2009) is an existing package dependency, used to parse sequence files and translate coding sequences to amino acids.

Primary modules calculate the SDUc and RSDUc metrics given three required arguments: i) a set of sequences in FASTA file format, ii) a list of dinucleotides of interest, iii) a list of frame positions of interest. Without any additional arguments only the SDUc/RSDUc values will be calculated for the dinucleotides and frame positions specified. Additionally, the user can provide a number of iterations for the coding sequence simulations that can optionally be performed for estimating the error around the null expectation distribution for each given sequence (as described in Results subsection 5.3.5).

By default, the single nucleotide composition used for calculating the corrected expected proportion of synonymous dinucleotides (e') will be inferred from each sequence for which the metric is being calculated. The *SDUc()* and *RSDUc()* modules include an alternative optional argument for specifying a custom nucleotide composition for the calculations. This could be of use if, for example, the dinucleotide frequencies across all genes in a genome are being compared to the null expectation defined by the overall nucleotide composition of the entire genome.

The package further includes modules for calculating the relative dinucleotide abundance (RDA) – for all positions or separately for frame positions – and the relative synonymous codon usage (RSCU) of coding sequences, so that the user can make comparisons between metrics. All modules include checks and detailed error messages for ensuring that provided sequences are coding (length is multiple of 3) and do not contain internal stop codons. Ambiguous nucleotides are excluded from all calculations as being non-informative, however the user is notified that

multiple ambiguous nucleotides might affect the calculations of bridge dinucleotide representation. Gaps are removed from sequences before initiating calculations, so sequence alignments in FASTA format can also be provided as input files.

Calculated values from all modules are outputted as python objects (dictionaries) to facilitate analysis of results directly in the python environment. Accessory modules are also provided for exporting the result objects into delimiter-separated tables files (comma separated as default). In the output python dictionary, all values of individual simulated sequences for estimating the error around the null expectation are provided. When exporting the results, the user can specify the summary statistic for the null expectation distribution of values to be summarised as. The summary options are: i) minimum and maximum values of the sampled distribution, ii) single standard deviation intervals around the distribution's mean, iii) 95% confidence intervals around the distribution's mean (calculated as 1.96 of the standard deviation, assuming a normal distribution).

An additional accessory module for advanced users with the function name *eprimeall()* can calculate: i) corrected expected dinucleotide proportion for any set of dinucleotide-position combinations (Equation 5.5), ii) corrected expected codon frequencies (Equation 5.4) and iii) the stop codon correction factor (Equation 5.4), all given either a FASTA file with coding sequences, or a set of single nucleotide compositions.

Full documentation and code are available for users online at https://github.com/spyros-lytras/dinuq. To facilitate visualisation of results I have further developed an interactive online ObservableHQ notebook (https://observablehq.com/@spyros-lytras/dinuq-viz) where users can load their DinuQ output for SDUc, RSDUc and RDA values in the table format exported by the python package to create plots of the results.

5.3.7 Applying the SDUc framework on insect- and vertebrate-specific flaviviruses

Previous research (Blitvich and Firth, 2015; Simón *et al.*, 2017) has shown that host environment – mammal-specific, insect-specific, or vector-borne – differently affects the nucleotide, codon and dinucleotide composition in members of the *Flaviviridae* virus family. For example, the Apoi virus (APOIV) and other flaviviruses with no known insect vector show an under-representation of the CpG dinucleotide, while the Aedes flavivirus (AEFV) and other insect-specific flaviviruses do not exhibit this bias. This group of viruses is a fitting example for testing the SDUc framework, since subclades of the *Flaviviridae* are known to show distinct genome composition patterns. The majority of *Flaviviridae* also possess a single polyprotein-encoding coding sequence, facilitating the calculation of SDU framework metrics. RDA, SDUc and RSDUc values for the APOIV and the AEFV polyprotein genes are presented in Figures 5.4 and 5.5 to illustrate their usage. Looking at the SDUc values for the CpG dinucleotide, the clearest difference between the viruses' dinucleotide compositions, the AEFV genome has very little CpG bias. The frame position 1 and 2 values are only marginally above and below the 95% confidence intervals of the null expectation respectively ($SDUc_{CpGpos1}$ = 1.11, 95% CIs: 0.90-1.10; $SDUc_{CpGpos2}$ = 0.84, 95% CIs: 0.88-1.12). Marginal dinucleotide biases in the first two coding positions could simply reflect slight codon preference rather than a specific bias in dinucleotide signatures. A true dinucleotide bias signature should be consistent across all three frame positions, and AEFV's bridge position composition is exactly 1.00 (95% CIs: 0.89-1.12), indicative of no CpG bias in this virus (Figure 5.5). On the contrary, SDUc values for APOIV, the rodent infecting virus, are well below the null expectation confidence intervals consistently for all three frame positions ($SDUc_{CpGpos1}$ = 0.64, 95% CIs: 0.88-1.12; $SDUc_{CpGpos2}$ = 0.28, 95% CIs: 0.87-1.13; $SDUc_{CpGbridge}$ = 0.49, 95% CIs: 0.88-1.13), indicating significant under-representation of CpG in the polyprotein-encoding sequence (Figure 5.4). These observations are in agreement with previous research and the hypothesis of a CpG-targeting antiviral mechanism present in vertebrates and absent in insects.

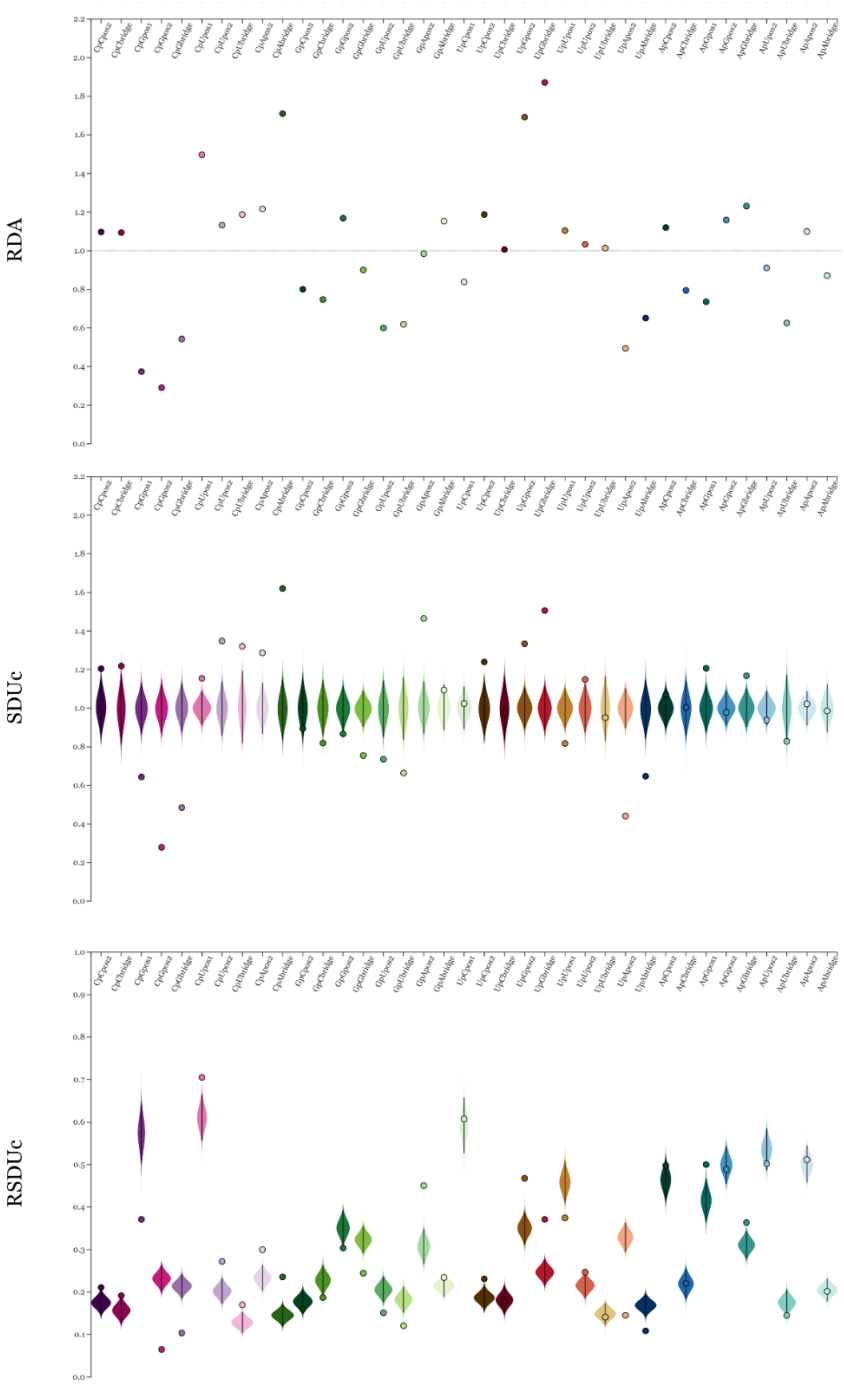


Figure 5.4. Dinucleotide composition of APOIV. RDA (top), SDUc (middle) and RSDUc (bottom) values for all informative dinucleotides and frame positions plotted for the APOIV coding sequence. Dot points indicate observed values and violin plots indicate SDUc/RSDUc error distributions around the null hypothesis (1000 random samples for each value). The grey horizontal line indicates an RDA of 1. Position 1 of dinucleotides CpC, CpA, GpC, GpG, GpU, GpA, UpG, UpA, ApC, ApU, ApA are excluded because they can only produce one amino acid (non-informative).

In contrast to the SDUc, RDA values cannot be compared to a null distribution, so there is no statistical evaluation of over- and under-representation of dinucleotides. For example, AEFV has a CpU frame position 2 RDA value of 0.77, which would be considered as weak under-representation of this dinucleotide (Figure 5.5). The corresponding SDUc value is also below 1 ($SDUc_{CpUpos2}$ = 0.94) but falls well within the null expectation distribution (95% CIs: 0.87-1.13). Thus, by using the SDU framework one can assess how confidently a value reflects true bias or the deviation from 1 is simply due to chance. Another example of differences between using the RDA and SDUc metrics is the evaluation of frame position 2 and bridge for dinucleotides GpU and UpA in the AEFV genome. Based on their RDA values, both dinucleotides are weakly under-represented in both of their informative frame positions ($RDA_{GpUbridge}$ = 0.86, $RDA_{GpUpos2}$ = 0.78, $RDA_{UpAbridge}$ = 0.86, $RDA_{UpApos2}$ = 0.71). Once accounting for the codon table and estimating the error around the unbiased representation expectation with SDUc, the assessment changes for the two dinucleotides, with both GpU values being closer to 1, meeting the null expectation ($SDUc_{GpUpos2}$ = 0.96, 95% CIs: 0.86-1.14; $SDUc_{GpUbridge}$ = 0.98, 95% CIs: 0.83-1.17), and both UpA values falling well below the null distribution's CIs ($SDUc_{UpApos2}$ = 0.66, 95% CIs: 0.89-1.11; $SDUc_{UpAbridge}$ = 0.74, 95% CIs: 0.86-1.14) (Figure 5.5).

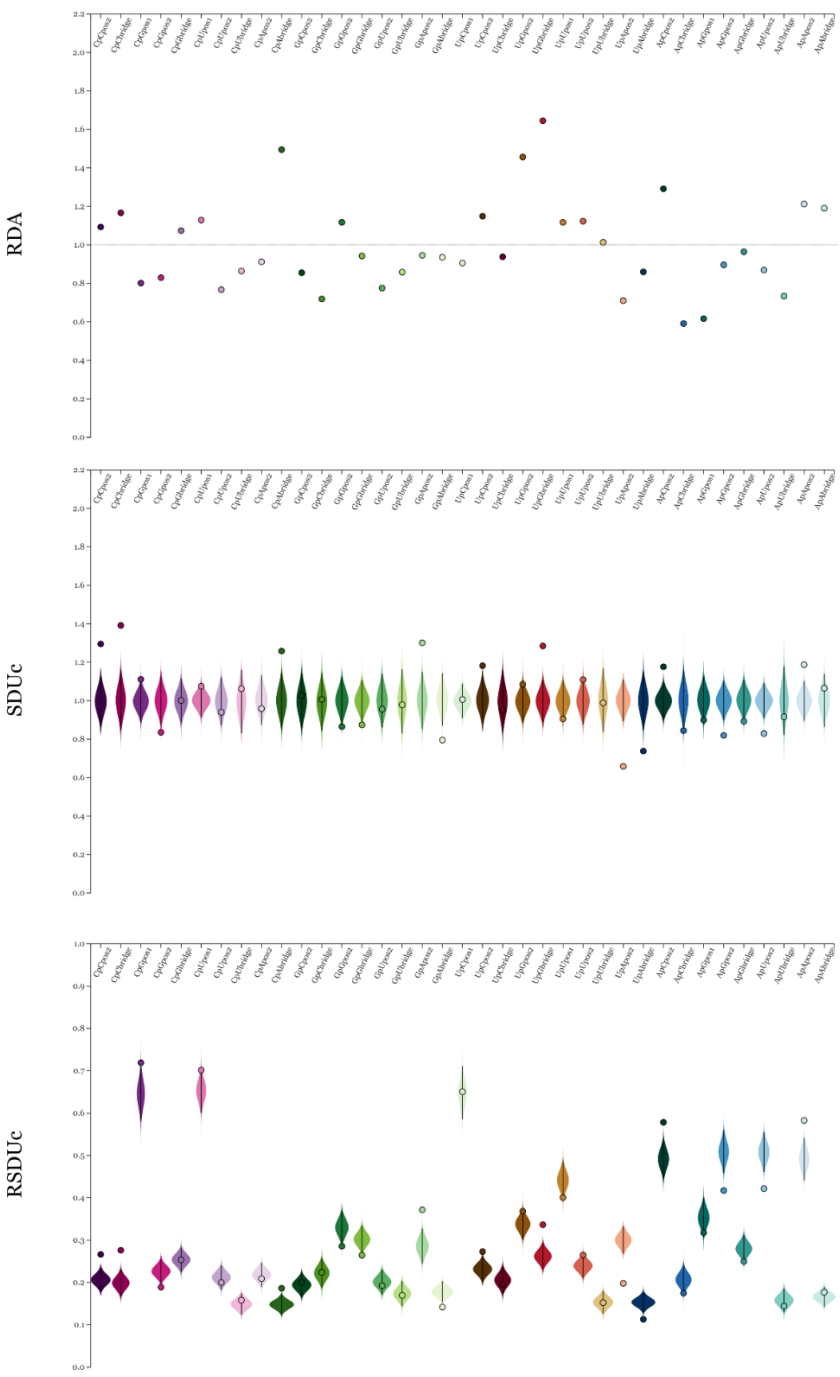


Figure 5.5. Dinucleotide composition of AEFV. RDA (top), SDUc (middle) and RSDUc (bottom) values for all informative dinucleotides and frame positions plotted for the AEFV coding sequence. Dot points indicate observed values and violin plots indicate SDUc/RSDUc error distributions around the null hypothesis (1000 random samples for each value). The grey horizontal line indicates an RDA of 1. Position 1 of dinucleotides CpC, CpA, GpC, GpG, GpU, GpA, UpG, UpA, ApC, ApU, ApA are excluded because they can only produce one amino acid (non-informative).

The RSDUc plots (Figures 5.4 and 5.5) showcase how the relative expected number of occurrences differs between dinucleotides and frame positions. For example, in both genomes UpC representation in frame positions 1 and bridge fall within the null

distribution, however, the expected RSDUc value under the null hypothesis is much larger for UpC position 1 (Figures 5.4 and 5.5). This simply depends on the expected occurrences of codons (or codon pairs for bridge position) that contain this dinucleotide at that frame position under the expected synonymous codon usage, making up the respective SDUc maximum value, discussed above (Results subsection 5.3.4). The RSDUc can be useful when comparing the level of over-representation between two dinucleotide-position combinations. CpU is over-represented in all frame positions of the APOIV coding sequence, but the frame position 1 SDUc value is smaller, and closer to its null expectation, than the position 2 value ($SDUc_{CpUpos1}$ = 1.15, 95% CIs: 0.91-1.08; $SDUc_{CpUpos2}$ = 1.35, 95% CIs: 0.86-1.14) (Figure 5.4). However, the RSDUc value for position 1 is substantially larger than that of position 2, despite both values still being over their null expectation confidence intervals ($RSDUc_{CpUpos1}$ = 0.71, 95% CIs: 0.56-0.66; $RSDUc_{CpUpos2}$ = 0.27, 95% CIs: 0.17-0.23). This discrepancy depends directly on the maximum SDUc values used to normalize the true SDUc values for calculating RSDUc. For this particular example, the maximum values are $SDUcMAX_{CpUpos1}$ = 1.64 and $SDUcMAX_{CpUpos2}$ = 4.95, meaning that, given the synonymous codon usage bias expected by the coding sequence's single nucleotide composition, the expected proportion of CpUs in frame position 1 (corresponding to leucine encoding codons) is much larger than that of CpUs in frame position 2 (corresponding to codons encoding for serine, proline, threonine and alanine). This is because maximum SDUc values are inversely correlated with proportion of expected synonymous dinucleotides (Equation 5.9).

The SDUc of a given dinucleotide should directly reflect the RSCU of the codons that contain it (or in case of the bridge position: the first nucleotide in the third codon position and second nucleotide in the first codon position of the downstream amino acid). To illustrate this relation between the two metrics, the RSCU values for all codons of AEFV and APOIV, calculated using DinuQ, are presented in Table 5.2. ApG in position 1 seems to be over-represented in APOIV ($SDUc_{ApGpos1}$ = 1.21, 95% CIs: 0.89-1.11), which is not the case in AEFV ($SDU_{ApGpos1}$ = 0.90, 95% CIs: 0.87-1.13). This is clearly depicted in the RSCU values of all AG-starting codons being higher in APOIV (Table 5.2), with AGA in particular being highly over-represented in the APOIV genome exclusively (APOIV: RSCU = 2.19; AEFV: RSCU = 1.00). This is also one of the few dinucleotide-position combinations where RDA gives the opposite answer to SDUc ($RDA_{ApGpos1}$ = 0.74). The discrepancy between the two

metrics highlights their different uses, RDA comparing the occurrence of dinucleotides to their single nucleotides' frequencies, while SDUc compares the in-frame occurrence of dinucleotides to that expected by the sequence's synonymous codon usage bias expectation.

Table 5.2. Codon usage bias of APOIV and AEFV. RSCU values for each codon, calculated for the APOIV and AEFV coding sequences. Highlighted in bold are the values mentioned in the text. No values have been calculated for stop codons (UAA, UAG, UGA), since only one coding sequence was used for each virus.

	APOIV	AEFV		APOIV	AEFV		APOIV	AEFV		APOIV	AEFV
UUU	1.11	1.08	**UCU**	1.01	0.75	**UAU**	0.98	1.00	**UGU**	0.95	1.04
UUC	0.89	0.92	**UCC**	0.73	1.13	**UAC**	1.02	1.00	**UGC**	1.05	0.96
UUA	0.38	0.56	**UCA**	1.57	0.96	**UAA**	STOP	STOP	**UGA**	STOP	STOP
UUG	1.39	1.23	**UCG**	**0.34**	**1.06**	**UAG**	STOP	STOP	**UGG**	1.00	1.00
CUU	1.08	0.83	**CCU**	1.19	0.89	**CAU**	1.25	1.05	**CGU**	**0.48**	**1.19**
CUC	1.00	1.42	**CCC**	0.86	1.00	**CAC**	0.75	0.95	**CGC**	**0.39**	**1.19**
CUA	0.67	0.79	**CCA**	1.69	1.32	**CAA**	0.91	1.23	**CGA**	**0.68**	**1.05**
CUG	1.48	1.17	**CCG**	**0.25**	**0.79**	**CAG**	1.09	0.77	**CGG**	**0.68**	**0.88**
AUU	1.10	1.22	**ACU**	1.13	0.98	**AAU**	0.91	0.84	**AGU**	**1.04**	**0.96**
AUC	1.15	1.05	**ACC**	1.27	1.07	**AAC**	1.09	1.16	**AGC**	**1.32**	**1.13**
AUA	0.75	0.73	**ACA**	1.29	1.00	**AAA**	1.01	1.23	**AGA**	**2.19**	**1.00**
AUG	1.00	1.00	**ACG**	**0.32**	**0.95**	**AAG**	0.99	0.77	**AGG**	**1.58**	**0.69**
GUU	1.08	1.29	**GCU**	1.65	1.07	**GAU**	0.97	0.67	**GGU**	0.73	0.74
GUC	1.00	0.98	**GCC**	1.03	1.62	**GAC**	1.03	1.33	**GGC**	0.73	0.73
GUA	0.29	0.62	**GCA**	1.05	0.68	**GAA**	1.09	1.07	**GGA**	1.75	1.57
GUG	1.63	1.11	**GCG**	**0.27**	**0.63**	**GAG**	0.91	0.93	**GGG**	0.80	0.96

5.3.8 SDUc Shows Consistent CpG Differences between Insect- and Vertebrate-Specific Viruses

Since there is evidence for a vertebrate-specific immune response selecting against CpG dinucleotides in viral genomes, I decided to further explore this trend between members of the Flaviviridae family specific to and absent in vertebrate hosts using the SDU framework. First, the RSDUc of CpG for all frame positions was calculated for the two sets of insect-specific and vertebrate-specific (no known insect vector) viruses used by Simón et al. (2017). The genes' overall GC content is intuitively expected to positively correlate with CpG representation, so I fitted a generalised linear model (GLM) with GC content and host group as explanatory factors for CpG RSDUc values ($RSDUc_{CpG}$ ~ GC + Host). Both explanatory variables were significant for all frame positions ($p < 0.05$) with all overall models explaining more than 90% of the variance in dinucleotide representation ($RSDUc_{CpGpos1}$: $F_{2,17} = 220.6$, $p < 0.001$, $R^2 = 0.96$; $RSDUc_{CpGpos2}$: $F_{2,17} = 94.9$, $p < 0.001$, $R^2 = 0.91$; $RSDUc_{CpGbridge}$: $F_{2,17} = 212.4$, $p < 0.001$, $R^2 = 0.96$) (Figure 5.6). Host group had the largest effect on frame position 1 CpGs, vertebrate-specific viruses having RSDUc values that are overall lower by 0.28 compared to the invertebrate-specific viruses (lower by 0.11 for position 2 and 0.08 for bridge position). GC content had the smallest effect on frame position 2 CpG representation (slope of 0.61) and largest effect on frame position 1 (slope of 3.29) with bridge position dinucleotide representation having an intermediate effect (slope of 1.47).

To examine whether this effect is specific to the *Flaviviridae* virus family or can be generalised for other groups, the GLM analysis was replicated for a set of viruses of the *Rhabdoviridae* family (Figure 5.6). Similar to the *Flaviviridae*, both GC content and host group are significant explanatory factors ($p < 0.05$) for CpG representation at all three frame positions, however a lot less variance is explained by the model for the *Rhabdoviridae* set ($RSDUc_{CpGpos1}$: $F_{2,65} = 18.7$, $p < 0.001$, $R^2 = 0.36$; $RSDUc_{CpGpos2}$: $F_{2,65} = 51.5$, $p < 0.001$, $R^2 = 0.60$; $RSDUc_{CpGbridge}$: $F_{2,65} = 101.6$, $p < 0.001$, $R^2 = 0.75$). GC content and host group explain the largest amount of variance ($R^2 = 0.75$) for the bridge position RSDUc values, suggesting that this group of viruses might be under stronger codon usage biases (reflected in frame position 1 and 2 dinucleotide representation) that reduce the relative effect of host and GC content on CpG representation for positions 1 and 2. Vertebrate-specific *Rhabdoviridae* genomes had CpG position 1 RSDUc values that were overall lower

by 0.10 compared to invertebrate-specific viruses of the same family. Host group had the largest effect on this frame position, also the case in the *Flaviviridae* (vertebrate-specific RSDUc values were lower by 0.02 for position 2 and by 0.03 for bridge). The relative effect of GC content on CpG representation of different frame positions followed the same trend as in the *Flaviviridae*, with position 2 having the smallest effect (slope of 0.50), followed by bridge position (slope of 0.96) and position 1 (slope of 1.51).

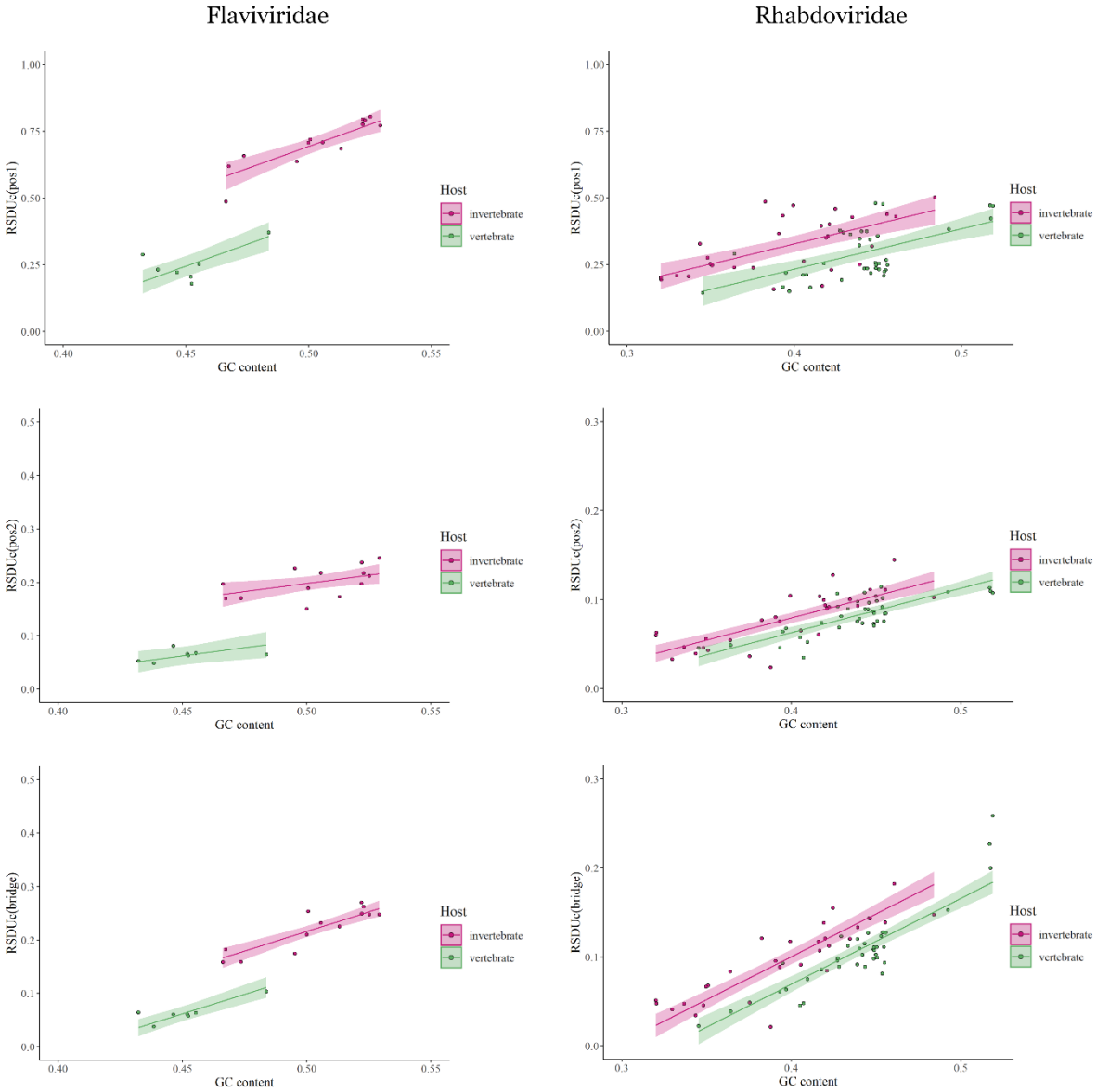


Figure 5.6. Comparison of $RSDUc_{CpG}$ values for each frame position between invertebrate- and vertebrate-specific *Flaviviridae* (left) and *Rhabdoviridae* (right). RSDUc values are plotted against the overall GC content of the coding sequences. The line of the linear regression and 95% confidence intervals of the model are shown along with the datapoints.

5.3.9 Adaptive shifts in CpG and UpA biases across the *Flaviviridae* tree

In addition to simply assessing whether there is dinucleotide bias in each individual coding sequence or genome, the metrics described above can also be implemented in a comparative phylogenetic framework. The RSDUc metric provides a normalised, numerical representation of each dinucleotide's abundance which can be compared between sequences. By using RSDUc values of each sequence in a phylogeny as a quantitative trait and then modelling the expected change in any quantitative trait based on the length of each branch, one can detect nodes in the tree where dinucleotide representation has changed more than expected by chance, i.e. an adaptive shift in dinucleotide representation. This type of analysis can be conducted using the PhylogeneticEM algorithm (Bastide, Mariadassou and Robin, 2017; Bastide *et al.*, 2018) with RSDUc being the quantitative trait being tested. Following on from the above sections, I chose the *Flaviviridae* virus family as an example for applying this approach due to the frequent host switches across their evolution (Bamford *et al.*, 2022) and their long polyprotein genomes that produce SDUc and RSDUc values with narrow error intervals. I collated a comprehensive set of 350 coding sequences including recently published representatives of the entire known *Flaviviridae* family (Mifsud *et al.*, 2023) and used the conserved RNA-dependent RNA polymerase (RdRp) sequences for reconstructing the viruses' phylogeny. The PhylogeneticEM analysis can be performed for all possible dinucleotides, however I will focus on the results for CpG and UpA, the two dinucleotides that have biased representations in most *Flaviviridae* polyproteins (216/350 and 174/350 respectively, Appendix D Figure D.1) and are known to be affected by potential host-driven mechanisms (Simmonds *et al.*, 2013; Takata *et al.*, 2017; Odon *et al.*, 2019).

Firstly, the RdRp phylogeny presented here is consistent with Mifsud et al.'s (2023) reconstruction with Hepaciviruses and Pegiviruses clustering together in a clade more distant from the other groups. Long Genome Flaviviruses (LGF) and Pestiviruses form sister clades which in turn relate to the clade containing Jingmenviruses and Flaviviruses (Figure 5.7). The Tamanavirus group (Bamford *et al.*, 2022) does not form a monophyly in this tree, instead tamana-like viruses form polyphyletic clades that sit as direct outgroups to the Flaviviruses. For this reason, the Tamanavirus group will not be described further in this section. Starting with

CpG representation, inferring adaptive shifts on the entire *Flaviviridae* phylogeny reveals an adaptive increase in CpG at the base of all Pegiviruses, consistent across all three informative frame positions (Figure 5.7). To further inspect shifts that may be missed when testing across the whole family, I also performed the same analysis for subclades in the tree. The shift at the base of the Pegiviruses is robustly inferred when only testing the Hepaci-Pegi clade (Figure 5.7). This group of viruses has been sampled from both birds and mammals, indicating that this adaptive increase in CpGs is unlikely due to a host switch at the base of the Pegiviruses. Instead, it may be a result of a virus-specific adaptation unique to this group (whether that relates to antagonism of host factors targeting CpGs or the viruses' replication machinery). Another adaptive shift, this time representing a marked decrease in CpGs across all frame positions, can be seen in the Pestiviruses. Interestingly, this shift is not at the very base of the group, but the CpG decrease took place after this subclade diverged from the distant pesti-like viruses sampled in fish and ray hosts (Glass knifefish pestivirus, GenBank accession: OX394178; Xiamen fanray pesti-like virus GenBank accession: MG599985; Nanhai dogfish shark pesti-like virus, GenBank accession: MG599984; Wenzhou pesti-like virus, GenBank accession: MG599982). The inner low CpG Pestivirus subclade does not include any fish viruses, suggesting that this adaptive shift could be a result of adaptation to non-fish vertebrate hosts. Other than mammalian viruses, the low CpG clade includes reptile and amphibian viruses (Frog pestivirus, GenBank accession: OX394182; Transcaucasian sand viper pestivirus, GenBank accession: OX394184; Cayenne caecilian pestivirus, GenBank accession: OX394172). If this shift is in fact driven by changes in host environment, this finding would indicate that the immune mechanism responsible for reducing CpGs in Pestiviruses evolved in tetrapods after the split from the fish, and is shared between amphibians, reptiles and mammals. When only testing the Pestivirus clade, the method detects an additional adaptive increase in CpGs unique to the outgroup fish Pestivirus clade, although this is likely due to the lack of CpG representation context outside this virus group.

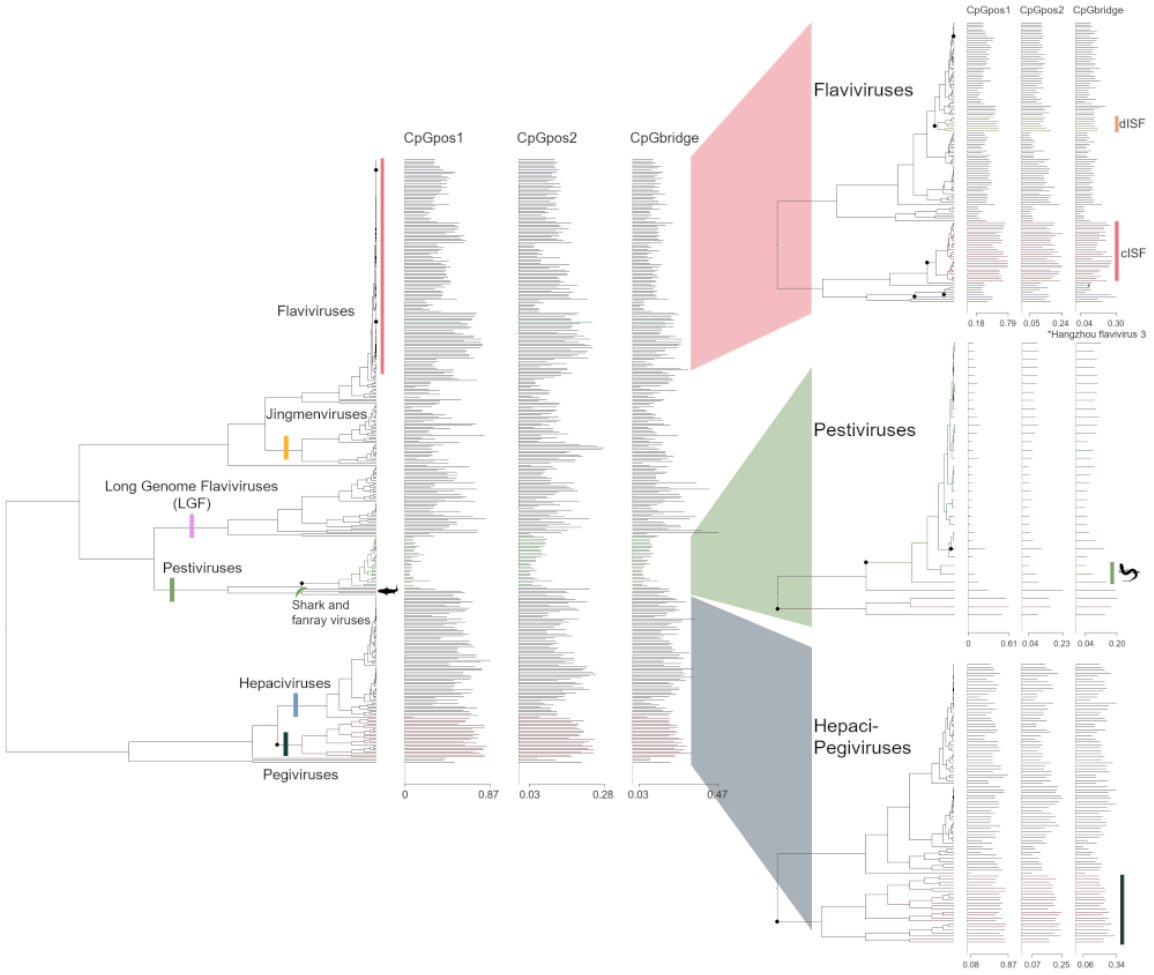


Figure 5.7. Adaptive shifts in CpG representation across the *Flaviviridae* virus family. Left: Ultrametric RdRp phylogeny of 350 representatives of the *Flaviviridae* family. Nodes with adaptive shifts in the genomic CpG representation are denoted with dots on the tree. CpG RSDUc values for all three frame positions are presented on the right of each tip in the tree. **Right**: Subclades of the full phylogeny representing the: i) Flavivirus, ii) Pestivirus and iii) Hepaci-Pegivirus groups. Dots in these trees denote CpG adaptive shifts detected by testing each individual subclade.

Looking at CpG representation within the Flaviviruses, no adaptive shifts were detected in internal branches when testing the full *Flaviviridae* phylogeny. Instead, two terminal branch shifts were picked up within the Flaviviruses, namely branches leading to the Ntaya virus (GenBank accession: NC_018705) and the Menghai flavivirus (GenBank accession: NC_034204). Shifts in terminal branches are more difficult to interpret while neither of these viruses represent any surprising host change. It should also be noted that the CpG changes for these tips are not consistent across frame positions, suggesting that the detection of these shifts may be due to a change in synonymous codon usage rather than the dinucleotide signatures themselves. The height of the Flaviviruses clade is quite shallow in the

full *Flaviviridae* tree which could explain the lack of signal. Only testing the Flavivirus clade improves the detection resolution with two distinct adaptive increases in CpG representation corresponding to internal branches being picked up (Figure 5.7). The first shift is at the base of the classical insect-specific flaviviruses (cISF) and the second at the base of the dual insect-specific flaviviruses (dISF). This is an interesting result since both these clades represent unique switches from a vector-borne to an insect-specific lifestyle for these viruses (Blitvich and Firth, 2015). The CpG levels of both ISF clades have previously been shown to be higher than these of vector-borne flaviviruses (Simón *et al.*, 2017), but recapitulating this finding in the current comparative phylogenetic approach validates its usefulness. By detecting at which exact node in the tree the CpG shift has taken place we can make further predictions about the viruses' host environment. For example, although the shift encapsulates the entire known dISF clade, there are two outer terminal branches in the cISF clade that sit outside the detected CpG shift. These represent the Hangzhou flavivirus 3, sampled from a non-biting midge species (GenBank accession: MZ209680) and the *Tabanus rufidens* flavivirus, sampled in the Japanese horsefly (GenBank accession: LC540441). The two viruses directly related to the main cISF clade have substantially lower CpG levels, comparable to these of vector-borne flaviviruses. Hence, the analysis presented here suggests that these two viruses, despite having no currently known mammalian host, could potentially be vector-borne and the original change from vector-borne to insect-specific took place after the main cISF clade diverged from the *Tabanus rufidens* flavivirus.

Moving on to shifts in UpA representation, there seems to be less overall signal than with CpGs, while the few shifts identified encompass broader virus groups. Testing the full phylogeny only detects two shifts in very deep nodes of the tree: i) one at the base of all Hepaci-Pegiviruses representing a further reduction in UpA representation and ii) one within Pestiviruses, excluding the clade of three shark pesti-like viruses, this time representing a consistent increase in relative UpA representation (Figure 5.8). The latter shift is consistent with the Pestivirus-specific drop in CpGs discussed above, validating that the evolutionary environment of Pestiviruses likely changed notably after the split from their fish-infecting closest relatives. Similar to the CpG analysis, more signal is detected when testing individual groups for shifts. Testing the Flaviviruses first reveals an interesting shift unique to the tick-borne flavivirus (TBFV) clade, denoting further decrease in UpA

levels (Figure 5.8). This could either be a mutational pressure unique to the viruses' replication mechanisms or an adaptation to their tick hosts. Another adaptive shift, towards an increase in UpA, is detected on the Dengue virus (DENV) clade, encompassing all four serotypes (Figure 5.8). These viruses are globally circulating in the human population with transmission mediated by mosquito vectors (Guzman *et al.*, 2010), suggesting that this shift may be specific to human replication. However, looking closer at the UpA representation of the DENV clade, it seems that most of the signal is driven by an increase only in frame position 2 UpAs, suggesting that this could be an adaptation of DENV to human codon usage biases rather than a dinucleotide adaptation. The branch directly outside of the DENV clade leads to Kedougou virus (KEDV; GenBank accession: NC_012533) which has not experienced the UpA shift and does not circulate in humans (Jansen van Vuren *et al.*, 2021).

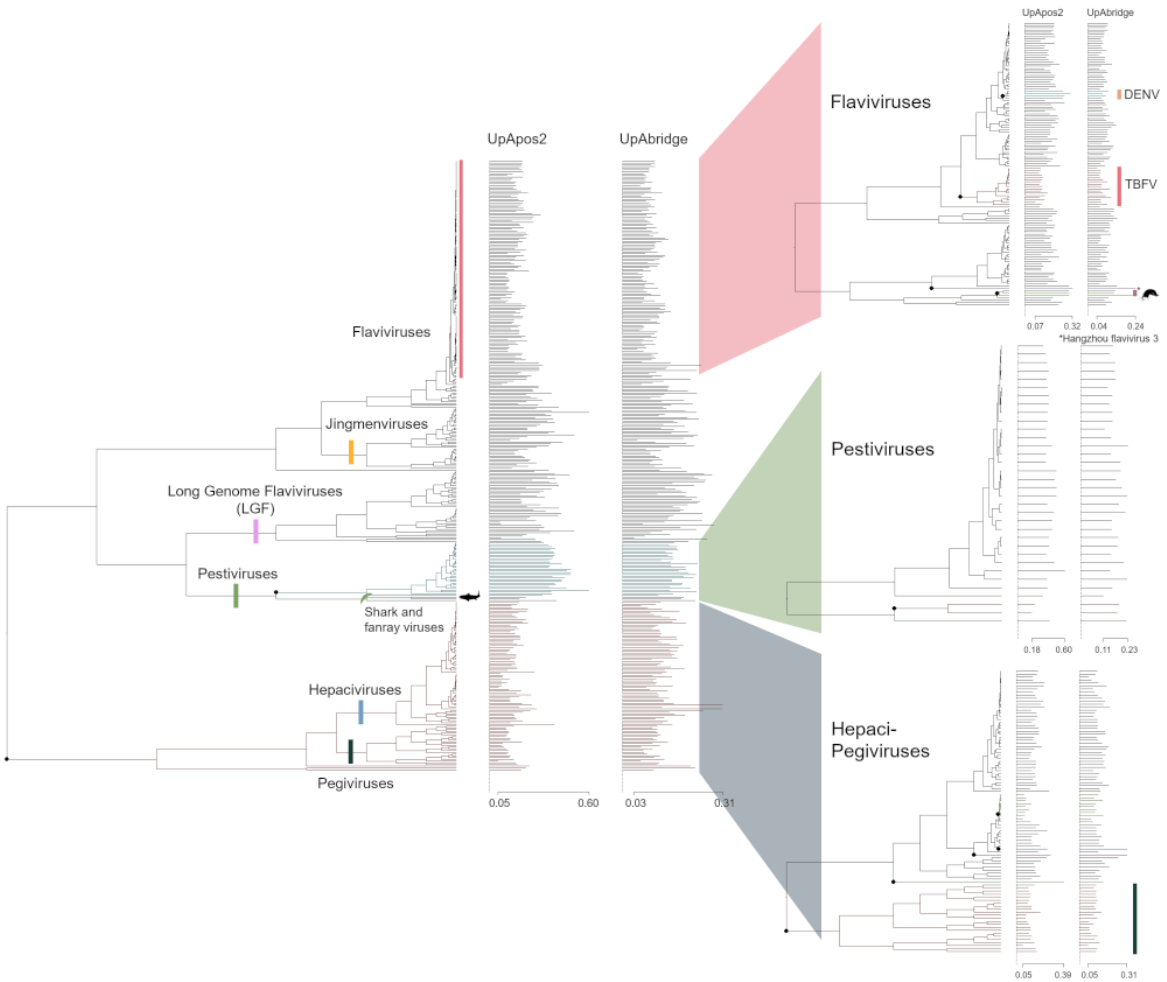


Figure 5.8. Adaptive shifts in UpA representation across the *Flaviviridae* virus family. Left: Ultrametric RdRp phylogeny of 350 representatives of the *Flaviviridae* family. Nodes with adaptive shifts in the genomic UpA representation are denoted with dots on the tree. UpA RSDUc values for the two informative frame positions are presented on the right of each tip in the tree.

Figure 5.8 (cont). Right: Subclades of the full phylogeny representing the: i) Flavivirus, ii) Pestivirus and iii) Hepaci-Pegivirus groups. Dots in these trees denote UpA adaptive shifts detected by testing each individual subclade.

Another UpA increase shift, this time consistent across both frame position 2 and bridge counts, is detected at the base of a group of three crustacean flaviviruses (*Photeros* flavivirus, GenBank: OX394156; Sea-firefly flavivirus, GenBank: OX394161; *Crangon crangon*, GenBank: MK473878). However, this subclade sits within a wider clade of crustacean-infecting flaviviruses, suggesting that this may not be a host-specific adaptation. Little is known about the antiviral mechanisms of crustacean hosts, so it is challenging to speculate about the biological importance of this UpA increase. The final UpA shift detected within the Flaviviruses leads to the terminal branch of the Hangzhou flavivirus 3 (sampled from a non-biting midge species, GenBank: MZ209680). Although terminal branch shifts are more difficult to interpret, the dinucleotide composition of this particular virus is interesting based on both the UpA and CpG results. Hangzhou flavivirus 3 has not experienced the CpG increase associated with its sister clSF clade's switch to insect host specificity, despite not having a known non-insect host. The additional skew in UpA biases specific to this virus implies that it has evolved in a unique host environment and warrants further investigation into the biology of these viruses as well as the ecology of their non-biting midge insect host (Silva and Stur, 2019).

Looking within the Hepaci-Pegivirus group, the adaptive shift at the base of the group detected when testing the full tree is followed by a further adaptive reduction in UpA specific to the Pegiviruses (Figure 5.8). This group seems to have the lowest overall UpA representation across the *Flaviviridae* which seems to be a result of stepwise adaptations across the viruses' deep evolution. Another clade within the Hepaciviruses experiencing a UpA reduction consists of viruses sampled in African cichlids. The outgroups of this group that have not experienced the shift have been sampled in different fish species, making it difficult to interpret the biological relevance of this shift. Similarly, three more shifts detected on terminal branches within the Hepaci-Pegivirus clade cannot be interpreted in terms of biological function. A single UpA shift is detected within the Pestivirus group, denoting a decrease in UpAs at the node leading to the Xiamen fanray pesti-like virus and the Nanhai dogfish shark pesti-like virus (Figure 5.8). Again, this shift may be associated with the viruses' unique host environment, although the decrease is primarily seen in the frame position 2 representation, indicating this could be a codon usage bias

adaptation. Finally, no adaptive shifts were detected in the Jingmenvirus and LGF groups for either CpG or UpA representation, when testing the full *Flaviviridae* phylogeny or their individual groups' trees. Neither of these groups are known to infect vertebrates (or chordates), and likely have not experienced switches between very distant hosts in their recent evolution. This could explain the lack of adaptive signal in the dinucleotide representation of these viral genomes, since the groups have not had recent, substantial changes in their host environments.

5.4 Discussion

Compositional biases in virus genomes can reveal intricate clues about virus biology and complex interactions with the host environments they replicate in. In this chapter, I present a novel framework for assessing biases in the dinucleotide representation of coding sequences. Specifically, the SDUc and RSDUc metrics provide a numerical depiction of whether a dinucleotide is over- or under-represented in a coding sequence while taking into account the codon usage and amino acid abundance of the sequence. Unlike routinely used metrics such as the RDA, by using the SDU framework one can statistically assess the extent to which a dinucleotide's abundance diverges from an unbiased expectation. Previously developed approaches for quantifying coding sequences' dinucleotide representation based on similar principles as the SDU, e.g., accounting for codon usage and amino acid presence (Greenbaum *et al.*, 2008, 2014; Gu *et al.*, 2019), were implemented on standalone scripts and not readily usable software. The SDU metrics can be easily computed using the DinuQ python package which allows for accessible parameterisation of the modules (e.g. changing the nucleotide composition for the null expectation), as well as providing further modules for investigating the genetic composition of any sequence (RSCU, RDA, single nucleotide composition). DinuQ also comes with detailed documentation and a web-based application for visualising results. One of the main caveats of the SDU methods is that they can only be calculated for coding sequences. This is not necessarily a problem when analysing the genomes of RNA viruses where most of the genome is encoding for proteins, as in the applications presented in this chapter. However, a large proportion of most organisms' genomes, including most eukaryotes, is non-coding. There is plenty of potential for using the SDU framework to explore dinucleotide compositions of non-viral genomes, for example comparing the host's gene signatures to those of the infecting virus genomes. Still, these would need to be paired with approaches for examining non-coding sequence biases (such as the RDA) to investigate biases across the entire genome.

Applying these methods to the genomes of the *Flaviviridae* and *Rhabdoviridae* virus families, infecting both vertebrate and invertebrate hosts, shows that the host environment explains a significant amount of the variation in CpG levels of these genomes (Figure 5.6). In fact, machine learning approaches aiming to predict hosts based on features of the virus genomes draw most of their signal from dinucleotide

representation, especially CpG abundance (Babayan, Orton and Streicker, 2018; Young, Rogers and Robertson, 2020; Brierley and Fowler, 2021; Mollentze, Babayan and Streicker, 2021). Although there is an observable effect of the host intra-cellular environment viruses replicate in on their genomes' dinucleotide composition, the exact mechanisms causing these effects are still largely not well understood. The recent discovery of ZAP and its role in restriction of viruses with high CpG levels (Takata *et al.*, 2017) is certainly a major culprit for the pronounced distinction between the CpG representation of vertebrate- and invertebrate-infecting viruses. Still, it is unlikely that ZAP is the sole mechanism affecting CpG biases in viruses, let alone biases in all other dinucleotide combinations. For example, Ficarelli et al. (2020) recently showed that increasing the CpG abundance in the HIV-1 genome also leads to virus inhibition through disruption of pre-mRNA splicing, completely independent of ZAP activity. Unfortunately, simply quantifying compositional biases in the virus genomes cannot fully illuminate the mechanisms underlying these biases and experimental studies are needed to complement the computational approaches. One idea for detecting host factors that interact with viruses' dinucleotide composition is to infect cell lines overexpressing or lacking expression of a panel of known RNA-binding host proteins with strains of the same virus that have synonymously recoded their dinucleotide compositions. Restriction of one strain but not the other in a cell line expressing a specific factor would indicate that this host protein could select against virus with the respective dinucleotide composition in real infection. Other than selection against or for a certain dinucleotide, mutational pressures could also be responsible for the observed biases, the APOBEC proteins being a prime example (Bishop *et al.*, 2004; Milewska *et al.*, 2018; O'Toole *et al.*, 2023) (Chapter 1 Section 1.2). Identifying mutational factors may require long-term passaging of recoded viruses in cells expressing the candidate proteins to properly observe the effects in virus composition. Although these experiments can be very resource- and time-consuming, computational analyses of the virus signatures like the ones presented here could efficiently guide the experimental design both in terms of which dinucleotides are under host pressures and which candidate factors are absent in different hosts.

Interpreting the adaptive dinucleotide shifts on the *Flaviviridae* tree presented in this chapter can also provide some insights into the nature of the mechanisms underlying dinucleotide biases. Odon et al. (2019) suggested that ZAP could be responsible for the under-representation of both CpG and UpA dinucleotides in at

least some RNA virus genomes. Based on my analysis, this is very unlikely to be the case for the *Flaviviridae*, since a common mechanism would also imply congruence between CpG and UpA adaptive shift events across the viruses' evolutionary history, which is certainly not the case (Figures 5.7 and 5.8). The two clades that experience adaptive shifts in both CpG and UpA representation at virtually identical points in the tree are the Pestiviruses and the Pegiviruses. If a common ZAP-dependent mechanism inhibited replication of high CpG and high UpA virus genomes then one would expect ancestral changes in host environments to lead to consistent shifts in both dinucleotide's representation (either increase or decrease). However, in both congruent shifts, the two dinucleotides shift in opposite directions; Pestiviruses experience an adaptive reduction in CpGs but an adaptive increase in UpAs, while the inverse is observed for the Pegiviruses (Figures 5.7 and 5.8). Another hypothesis for the UpA under-representation in virus and other organisms' genomes is that UpA-rich mRNA might be unstable and more prone to degradation by host RNAses (Beutler *et al.*, 1989; Duan and Antezana, 2003). However, RNAses preferentially degrade AU-rich motifs including stretches of uracils (Shaw and Kamen, 1986) implying that, if RNAse targeting is responsible for shifts in UpA representation, these should also be consistent with UpU shifts. This is not the case in the *Flaviviridae*, since performing the same adaptive shift analysis with UpU detects no shifts at all, while UpU is one of the least biased dinucleotides across the *Flaviviridae* genomes used here (RSDUc values being outwith the null expectation in 37 out of 350 genomes; Appendix D Figure D.1). Hence, the mechanism behind the weak UpA under-representation in most viral genomes remains enigmatic, although the fact that UpA shifts in the *Flaviviridae* encompass large groups infecting diverse hosts may indicate that this signature depends more on the virus itself rather than the environment it replicates in (Figure 5.8).

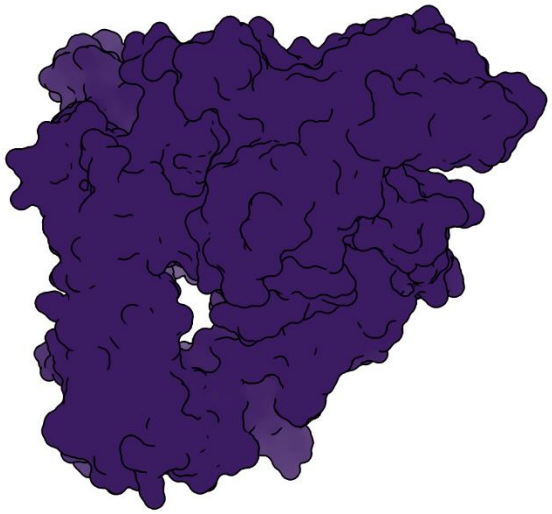
In this chapter, I only present the adaptive shift results for CpG and UpA, but the same analysis can be performed for all dinucleotides. Other than CpG and UpA, the UpG and CpA dinucleotides are also significantly biased for the majority of the *Flaviviridae* (205/350 and 199/350 genomes under bias respectively; Appendix D Figure D.1). In fact, UpG and CpA are almost always over-represented if biased, mirroring the CpG and UpA under-representation. Gu et al. (2019) showed that synonymous C-U transitions was the primary cause of these biases during the recent evolution of influenza A virus in humans (CpGs changing to UpGs and UpAs to CpAs). This pattern has been observed in more studies looking across more virus

groups, although the mirroring biases of these dinucleotides are not consistent in all virus families (di Giallonardo *et al.*, 2017). The current human-circulating IAV strains have only moved to humans from avian hosts up to a century ago or so and the gradual suppression of CpGs can seemingly be observed over the time the viruses have been adapting to the human host environment (Greenbaum *et al.*, 2008). It is important to consider the timing of the host switches that could influence the nucleotide composition of viral genomes. If a virus moves to a new environment where CpGs are selected against and the most efficient way to remove these host recognition motifs is by C to U transition substitutions, then UpG abundance will increase. No true selective force is acting on UpG representation, rather the potential over-representation of this dinucleotide is a biproduct of reducing CpGs. This effect should be observable in the short-term window after the host switch, but maintaining the UpG over-representation is not necessary in the long-term and the bias could wane given enough time in the same host environment. Hence, the presence of mirroring biases that are produced through selection on one dinucleotide is expected to depend on how recently the host environment changed. Given an example of a virus with well-documented host switch events in their evolutionary history this conjecture could be formally tested in the future.

Many of the dinucleotide adaptive host switches detected here coincide with major ancestral host switches (especially for CpGs; Figure 5.7), making it tempting to assume a tight link between the two processes. However, great care should be taken when interpreting these patterns. The deeper the nodes with detected shifts are, the more likely that future sampling of viruses within these clades could change the picture. Discovery of related viruses infecting different hosts may lead to re-evaluating when the host switches took place, while discovery of viruses with divergent dinucleotide signatures could yield different results by the adaptive shift analysis. Furthermore, host switches may not be the only factor that can lead to an adaptive shift in dinucleotide representation. Using the same approach of combining the SDUc framework with PhylogeneticEM in MacLean et al. (2021), we showed that the clade of viruses most closely related to SARS-CoV-2 experienced an adaptive decrease in CpGs after splitting from its sister SARS-CoV-like clade. This could be a result of an ancestral host switch, however the vast majority of known viruses in both clades infect horseshoe bats, while there are recombinant viruses of both CpG backgrounds suggesting that low and high-CpG SARS-related coronaviruses can co-infect the same hosts (MacLean *et al.*, 2021). What could

explain this pattern is a change in the ancestral virus's tissue specificity rather than host. Expression of host factors can drastically differ between tissues, so a change in the tissue where the virus replicates could alter its evolutionary environment – and potentially its dinucleotide composition – without any host change. Finally, virus dinucleotide biases could theoretically be affected by changes in the virus replication machinery itself instead of the environment it replicates in. Most RNA viruses, for example, have an inherent mutational bias towards uracil (Kustin and Stern, 2021; Rice *et al.*, 2021). Adaptations in the polymerase, proofreading mechanisms or mutation-inducing accessory genes may introduce host-independent pressures on the dinucleotide composition of the virus genomes. The intricacies of dinucleotide biases in virus genomes are far from being fully understood and pressures on overlapping signatures such as single nucleotide or codon composition may obstruct these genomic patterns. Further characterisation of the actual molecular mechanisms underlying the compositional biases can greatly help in interpreting the patterns detected through computational analyses like the ones presented in this chapter.

Chapter 6. Concluding remarks



Cartoon model of the SARS-CoV-2 RNA-dependent RNA polymerase.
PDB protein entry: 6m71, visualised with ChimeraX.

“Endless forms most beautiful, most wonderful”
Charles Darwin, On the Origins of Species (1859)

The findings presented in this thesis showcase the long-standing and complex interplay between viruses and their hosts. Better understanding these molecular interactions can translate into lowering the impact of pathogens to global health for humans and animals. In this final chapter of the thesis, I will discuss three overarching themes relating to the work presented across all previous chapters, and touch on future directions of research on virus-host evolution based on the findings.

The first theme is how viruses can generate novel variants by reshuffling genetic diversity through copy-choice recombination or reassortment. In Chapter 2, I present evidence of extensive recombination between SARS-related coronaviruses circulating in horseshoe bats. The distinct hotspots of recombination around the Spike gene and coldspots of recombination within the gene suggest that the antigenically important domains of Spike are likely swapped as a whole between virus populations conferring antigenic shifts (Figure 2.1). These results point to a model where divergent Spike genes can be introduced into a virus population through recombination. We could think of a host population with a relatively homogeneous sarbecovirus population circulating in it. Given infection of enough individuals, we would expect the host population to have some level of immunity against the viruses, primarily driven by antibodies recognising the Spike protein of the circulating virus population. If a distinct sarbecovirus is introduced in this host population by a different individual entering the group (e.g., a bat migrating to a new roost), this new virus could recombine with a virus in the existing sarbecovirus population creating a novel variant better able to transmit in this host population. Consistent with the recombination hotspot analysis (Figure 2.1), a recombinant with the existing population's backbone and the Spike gene of the introduced virus will have a selective advantage compared to other recombinant genomes, since it will evade the host population's existing antigenic immunity. The increased circulation of the recombinant Spike virus will keep recombining with the co-circulating non-recombinant viruses, leading to secondary recombination events. The breakpoints of these subsequent recombination events will not be identical to the original event introducing the new Spike in the existing backbone, resulting in the "overprinting" effect described in Chapter 2 (Figure 2.2). This model is supported by analysing genomes sampled from various locations and by different research groups, providing "snapshots" of how these viruses evolve in their reservoir hosts. Longitudinal studies aiming to systematically sequence the genetic diversity of sarbecoviruses circulating in bat colonies, paired with monitoring of host movement

in and out of these colonies would allow us to confirm this process in action (Cappelle *et al.*, 2021; Giles *et al.*, 2021; Chidoti *et al.*, 2022).

Even though Chapter 2 explicitly focuses on recombination, the results of Chapter 4 also relate to a form of intergenic recombination, reassortment. Reassortment permits efficient swapping of genetic segments in co-infected host cells. IAV is a segmented virus that frequently swaps segments, producing novel variants when this is between strains. Segment 5, encoding the viral NP is the sole known determinant of BTN3A3 evasion, a key requirement for IAV transmission to humans. The substitutions that lead to BTN3A3 evasion occur occasionally in the reservoir hosts of the virus (but much less frequently in NP site 313, Figure 4.3), however reassortment of these BTN3A3 evasive segments – unlocking human infection – has been crucial in the evolution of human-transmissible IAV strains. The current evidence suggests that the 1918 strain had the BTN3A3 evasive 313Y residue before jumping into humans. This same segment 5 lineage – maintaining this residue – reassorted into the antigenically distinct 1957 H2N2 pandemic and 1968 H3N2 pandemic strains, still circulating in the human population (Figure 4.3). This suggests that the virus maintained internal segments essential for human infection, while swapping the glycoprotein segments and evading the existing antigenic immunity – reminiscent of the sarbecovirus Spike recombination described above. On the other hand, Figure 4.5 shows how the Y52N NP substitution in the NP lineage responsible for all recent avian H7N9 human epidemics happened in an H9N2 virus. H9N2 strains have had little to no success in causing onward human-to-human transmission, despite occasional bird-to-human spillovers. This means that other segments encode more factors determining transmissibility and infectiousness in humans, and reassortment is the primary mechanism for bringing these together to create a human transmissible strain. The current situation with high pathogenicity H5 viruses transmitting in birds and mammals highlights the importance of reassortment and how mixing with diverse host populations – and subsequently virus populations – expands the potential reassortant combinations (Xie *et al.*, 2022). Geographic movement, either through roost migration in horseshoe bats transmitting sarbecoviruses (Figure 2.4) or intercontinental migration of wild birds transmitting IAVs (Figure 4.6) can aid the effectiveness of the mechanisms of recombination and reassortment (Breed *et al.*, 2010; Gass *et al.*, 2023). This bares the question whether host generalism of a virus, i.e. the ability to readily infect and transmit in different host species, as well as preference for

infecting hosts that can migrate through long distances, correlates with the ability of certain virus groups to recombine. These conjectures could be formally tested for virus groups where mechanisms of recombination are well understood, sampling across hosts is consistent, and metadata are well annotated.

The second theme is the co-evolutionary dynamics between viruses and host ISG evolutionary pressures. There are multiple levels at which viruses interact with their host defences, outlined in Chapter 1 (Section 1.2). One of the most important of these is innate immunity, comprising primarily of the interferon response. Chapters 3 and 4 describe the interaction between two virus groups and two ISGs expressed by the virus's vertebrate hosts. First, the prenylated form of OAS1 restricts SARS-CoV-2 but this isoform is absent in horseshoe bats, the reservoir host of SARS-related coronaviruses (Wickenhagen *et al.*, 2021), and second, the human BTN3A3 restricts avian IAV strains preventing transmission to humans without the necessary evasive substitutions in the virus genomes (Pinto *et al.*, 2023). Additionally, Chapter 5 describes adaptive shifts in the CpG dinucleotide representation of *Flaviviridae* genomes, which are likely a result of another ISG, the mammalian ZAP protein (Takata *et al.*, 2017). In order for viruses to jump species, they need to traverse a complex evolutionary landscape where certain genomic adaptations can bypass these host-specific antiviral mechanisms. The way the hosts gain or lose these mechanisms, however, may be more stochastic than one would intuitively think with a fair amount of turnover of ISGs on different host lineages (Shaw *et al.*, 2017). OAS1 in horseshoe bats seems to have lost its antiviral function by chance through a random LTR insertion (Figure 3.1). Similarly, BTN3 antiviral function in primates seems to have appeared "accidentally", with no apparent evidence of selection for restricting flu-like viruses at the likely time of the function's appearance (Figure 4.1). Hence, at least based on the examples presented in this thesis, and as expected from the "blind" nature of evolution, hosts will gain and lose genes with potential ISG functions with retention/fixation in a population biased towards genes that are helpful for counteracting specific virus groups. The antiviral property can then be selected for in the host population if a restricted virus infects the host. Following the stochastic loss of potential antiviral activity, a virus could initiate an association with this host species or population, readily spreading in the now "vulnerable" population. Most ISGs are co-opted host genes so have multiple functions (with the prime example of butyrophilins genes; Chapter 4 subsection 4.1.2), exemplifying how small changes in the genes' coding sequences could easily control pleiotropy. Another

example of the evolutionary flexibility of these functions is the frequent utilisation of isoform diversity for readily evolving novel functions of these host genes. Both human OAS1 and ZAP have at least one alternative isoform each (OAS1 p46 and ZAP-L) that are post-translationally modified to allow cellular compartmentalisation and subsequent targeting of specific viruses (Chapters 3 and 5 subsections 3.1.1 and 5.1.3). Both coronaviruses and flaviviruses replicate in membrane-associated replication complexes, away from cytosolic proteins, while influenza viruses replicate in the cell's nucleus. The compartmentalisation of virus presence within the cell – presumably evolved by viruses to evade host immunity – has likely resulted in this flexibility of function for ISGs. The stochastic readiness of a host group to restrict (or sustain) infection by specific viruses is expected to shape its virome, as well as the landscape of potential zoonotic pathogens that can spill into the host group. A comprehensive understanding of more of these host species specific antiviral mechanisms could ultimately guide predictions of which viruses can spill over into humans and inform pandemic preparedness or yield novel intervention strategies.

The third and final theme is how viruses can switch host species and start circulating in a new host, in the context of host-specific ISG defences and vertebrate RNA viruses. This theme is of direct relevance to human health, with examples of two major viral threats to humans outlined in Chapters 2 and 4: the recent COVID-19 pandemic, and the multiple IAV pandemics in the past century. There is a myriad of factors affecting the ability of a virus to switch to a new host species. Part – if not most – of these will relate to the ecological and epidemiological opportunity of the viruses to come into contact with a different host population. This side of virus-host interactions is essential for our complete understanding of viral infection and transmission, however it falls outside the scope of this thesis. Instead, my PhD work focused on the molecular barriers against virus infection. For this final theme I would like to present a simplified model of virus cross-species transmission, directly relating to the work presented in this thesis. There is genomic diversity within virus groups circulating in a host reservoir and only some of these virus haplotypes can successfully transmit to a new host. The diversity within the virus population is built up through mutations and, in the case of viruses like the sarbecoviruses (Chapter 2) and influenza viruses (Chapter 4), genetic recombination/reassortment. Even if there is an interface for transmission between the reservoir host species and a different, unrelated host, much of the haplotypes in the virus population will not be able to infect the latter host. Two key host-specific barriers to infection that the virus

will need to circumvent in order to establish infection in the new host are: gaining cell entry and evading intracellular immunity. In the case of the SARS-related coronaviruses, described in Chapter 2, the vast majority of bat SARSr-CoVs sampled so far cannot bind human ACE2 and infect human cells (Starr *et al.*, 2022). Only the, relatively under-sampled, viruses with SARS-CoV-like and SARS-CoV-2-like Spike genes are expected to be able to bind human ACE2. In fact, many of these viruses have diverse genomes to one another outside of the Spike gene, again highlighting the importance of recombination in increasing within virus population diversity. Hence, at least part of the virus population should be able to bind cellular receptors of both the reservoir and the secondary host species for cross-species transmission to occur.

Similarly with immune evasion, genomic adaptations that bypass host-specific immune mechanisms will usually need to be present in the virus before it crosses to a new host species. This seems to be the case with BTN3A3-evading NP substitutions for the 1918 pandemic derived NP lineage (Figure 4.3), the 2009 swine pandemic NP (Figure 4.4) and the H7N9 epidemics in Asia (Figure 4.5). Avoiding host immunity can also come from the host rather than the virus. In Chapter 3 I show how the reconstructed Rhinolophoidea common ancestor OAS1 protein restricts SARS-CoV-2 replication *in vitro*, unlike its extant Rhinolophoidea OAS1 proteins. This means that OAS1 anti-coronaviral function was ablated before the expansion of the Rhinolophoidea, allowing infection by these viruses and potentially explaining why horseshoe bats are the dominant reservoir host for these viruses.

Based on these examples, virus or host genomic changes are required to “unlock” the ability of a virus to switch host species and these changes will be expected to take place prior to the host switch. The interaction between virus and host does not stop there. Following the jump to a new host, the virus population is expected to be under evolutionary pressure for gradual adaptation to its new evolutionary environment, particularly as that environment becomes less naïve to the new infections. One example of this is the dinucleotide shifts described in Chapter 5. Representation of the CpG dinucleotide has experienced multiple detectable adaptive shifts across the evolutionary history of the *Flaviviridae*, correlating with ancestral host switches (Figure 5.7). These changes in the virus may represent gradual evasion of less detrimental antiviral mechanisms, such as ZAP, or adaptation to the new host’s unique codon usage and replication machinery. All in

all, virus host switches are governed by a complex interplay between viral genetic diversity, unique changes in the hosts' immune systems and further viral adaptation to novel host evolutionary environments.

In the closing sentence of his book *On the Origin of Species* (1859), Charles Darwin finishes with the phrase *"Endless forms most beautiful and most wonderful"*, referring to the vast diversity of living organisms that evolved from what once was a single common ancestor. Even though the words "beautiful" and "wonderful" have a contradictory connotation when referring to pathogenic viruses, I think Darwin's phrase is still quite fitting for describing the everchanging, *"endless forms"* of interactions between viruses and their hosts.

Bibliography

Afrache, H. *et al.* (2012) ‘The butyrophilin (BTN) gene family: From milk fat to the regulation of the immune response’, *Immunogenetics*, 64(11), pp. 781–794. Available at: https://doi.org/10.1007/s00251-012-0619-z.

Afrache, H. *et al.* (2017) ‘Evolutionary and polymorphism analyses reveal the central role of BTN3A2 in the concerted evolution of the BTN3 gene family’, *Immunogenetics*, 69(6), pp. 379–390. Available at: https://doi.org/10.1007/s00251-017-0980-z.

Aguado, L.C. *et al.* (2017) ‘RNase III nucleases from diverse kingdoms serve as antiviral effectors’, *Nature*, 547(7661), pp. 114–117. Available at: https://doi.org/10.1038/nature22990.

Agüero, M. *et al.* (2023) ‘Highly pathogenic avian influenza A(H5N1) virus infection in farmed minks, Spain, October 2022’, *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*, 28(3), p. 2300001. Available at: https://doi.org/10.2807/1560-7917.ES.2023.28.3.2300001.

Ahn, M. *et al.* (2019) ‘Dampened NLRP3-mediated inflammation in bats and implications for a special viral reservoir host’, *Nature Microbiology*, 4(5), pp. 789–799. Available at: https://doi.org/10.1038/s41564-019-0371-3.

Akaike, H. (1998) ‘Information Theory and an Extension of the Maximum Likelihood Principle’, in *Selected Papers of Hirotugu Akaike*. Springer, New York, NY, pp. 199–213. Available at: https://doi.org/10.1007/978-1-4612-1694-0_15.

Altschul, S.F. *et al.* (1990) ‘Basic local alignment search tool’, *Journal of Molecular Biology*, 215(3), pp. 403–410. Available at: https://doi.org/10.1016/S0022-2836(05)80360-2.

Anthony, S.J. *et al.* (2017) ‘Further evidence for bats as the evolutionary source of middle east respiratory syndrome coronavirus’, *mBio*, 8(2). Available at: https://doi.org/10.1128/MBIO.00373-17.

Armitage, A.E. *et al.* (2012) 'APOBEC3G-Induced Hypermutation of Human Immunodeficiency Virus Type-1 Is Typically a Discrete "All or Nothing" Phenomenon', *PLOS Genetics*, 8(3), p. e1002550. Available at: https://doi.org/10.1371/JOURNAL.PGEN.1002550.

Arnett, H.A. and Viney, J.L. (2014) 'Immune modulation by butyrophilins', *Nature Reviews Immunology*, 14(8), pp. 559–569. Available at: https://doi.org/10.1038/nri3715.

Arora, P. *et al.* (2022) 'The SARS-CoV-2 Delta-Omicron Recombinant Lineage (XD) Exhibits Immune-Escape Properties Similar to the Omicron (BA.1) Variant', *International Journal of Molecular Sciences*, 23(22), p. 14057. Available at: https://doi.org/10.3390/ijms232214057.

Atkinson, N.J. *et al.* (2014) 'The influence of CpG and UpA dinucleotide frequencies on RNA virus replication and characterization of the innate cellular pathways underlying virus attenuation and enhanced replication', *Nucleic Acids Research*, 42(7), pp. 4527–4545. Available at: https://doi.org/10.1093/nar/gku075.

Babayan, S.A., Orton, R.J. and Streicker, D.G. (2018) 'Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes', *Science*, 362(6414), pp. 577–580. Available at: https://doi.org/10.1126/SCIENCE.AAP9072.

Bamford, C.G.G. *et al.* (2022) 'Comparative analysis of genome-encoded viral sequences reveals the evolutionary history of flavivirids (family Flaviviridae)', *Virus Evolution*, 8(2). Available at: https://doi.org/10.1093/VE/VEAC085.

Banday, A.R. *et al.* (2022) 'Genetic regulation of OAS1 nonsense-mediated decay underlies association with COVID-19 hospitalization in patients of European and African ancestries', *Nature Genetics*, 54(8), pp. 1103–1116. Available at: https://doi.org/10.1038/s41588-022-01113-z.

Bastide, P. *et al.* (2018) 'Inference of Adaptive Shifts for Multivariate Correlated Traits', *Syst. Biol*, 67(4), pp. 662–680. Available at: https://doi.org/10.1093/sysbio/syy005.

Bastide, P., Mariadassou, M. and Robin, S. (2017) 'Detection of Adaptive Shifts on Phylogenies by using Shifted Stochastic Processes on a Tree', *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4), pp. 1067–1093. Available at: https://doi.org/10.1111/RSSB.12206.

Bates, P. *et al.* (2019) *Rhinolophus malayanus*, *The IUCN Red List of Threatened Species 2019*. Available at: https://doi.org/10.2305/IUCN.UK.2019-3.RLTS.T19551A21978424.en.

Belalov, I.S. and Lukashev, A.N. (2013) 'Causes and Implications of Codon Usage Bias in RNA Viruses', *PLOS ONE*, 8(2), p. e56642. Available at: https://doi.org/10.1371/JOURNAL.PONE.0056642.

Beutler, E. *et al.* (1989) 'Evolution of the genome and the genetic code: Selection at the dinucleotide level by methylation and polyribonucleotide cleavage', *Proceedings of the National Academy of Sciences*, 86(1), pp. 192–196. Available at: https://doi.org/10.1073/pnas.86.1.192.

Billoir, F. *et al.* (2000) 'Phylogeny of the genus Flavivirus using complete coding sequences of arthropod-borne viruses and viruses with no known vector', *Journal of General Virology*, 81(3), pp. 781–790. Available at: https://doi.org/10.1099/0022-1317-81-3-781.

Bird, A.P. (1980) 'DNA methylation and the frequency of CpG in animal DNA', *Nucleic Acids Research*, 8(7), pp. 1499–1504. Available at: https://doi.org/10.1093/nar/8.7.1499.

Bishop, K.N. *et al.* (2004) 'APOBEC-mediated editing of viral RNA', *Science*, 305(5684), p. 645. Available at: https://doi.org/10.1126/science.1100658.

Blazquez, J.L. *et al.* (2018) 'New insights into the regulation of γδ T cells by BTN3A and other BTN/BTNL in tumor immunity', *Frontiers in Immunology*, 9(JUL), p. 1. Available at: https://doi.org/10.3389/fimmu.2018.01601.

Blitvich, B. and Firth, A. (2017) 'A Review of Flaviviruses that Have No Known Arthropod Vector', *Viruses*, 9(6), p. 154. Available at: https://doi.org/10.3390/v9060154.

Blitvich, B.J. and Firth, A.E. (2015) 'Insect-Specific Flaviviruses: A Systematic Review of Their Discovery, Host Range, Mode of Transmission, Superinfection Exclusion Potential and Genomic Organization', *Viruses*, 7(4), pp. 1927–1959. Available at: https://doi.org/10.3390/V7041927.

Blumenkrantz, D. *et al.* (2013) 'The Short Stalk Length of Highly Pathogenic Avian Influenza H5N1 Virus Neuraminidase Limits Transmission of Pandemic H1N1 Virus in Ferrets', *Journal of Virology*, 87(19), pp. 10539–10551. Available at: https://doi.org/10.1128/JVI.00967-13.

Bobay, L.M., O'Donnell, A.C. and Ochman, H. (2020) 'Recombination events are concentrated in the spike protein region of Betacoronaviruses', *PLoS Genetics*, 16(12), p. e1009272. Available at: https://doi.org/10.1371/JOURNAL.PGEN.1009272.

Bodewes, R. *et al.* (2013) 'Recurring Influenza B Virus Infections in Seals', *Emerging Infectious Diseases*, 19(3), pp. 511–512. Available at: https://doi.org/10.3201/eid1903.120965.

Boni, M.F. *et al.* (2020) 'Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic', *Nature Microbiology*, 5, pp. 1408–1417. Available at: https://doi.org/10.1038/s41564-020-0771-4.

Boni, M.F., Posada, D. and Feldman, M.W. (2007) 'An exact nonparametric method for inferring mosaic structure in sequence triplets', *Genetics*, 176(2), pp. 1035–1047. Available at: https://doi.org/10.1534/genetics.106.068874.

Bouckaert, R. *et al.* (2019) 'BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis', *PLOS Computational Biology*, 15(4), p. e1006650. Available at: https://doi.org/10.1371/JOURNAL.PCBI.1006650.

Bourque, G. *et al.* (2018) 'Ten things you should know about transposable elements', *Genome Biology*, 19(1), pp. 1–12. Available at: https://doi.org/10.1186/S13059-018-1577-Z.

Bradel-Tretheway, B.G. *et al.* (2011) 'Comprehensive Proteomic Analysis of Influenza Virus Polymerase Complex Reveals a Novel Association with

Mitochondrial Proteins and RNA Polymerase Accessory Factors', *Journal of Virology*, 85(17), pp. 8569–8581. Available at: https://doi.org/10.1128/JVI.00496-11.

Breed, A.C. *et al.* (2010) 'Bats Without Borders: Long-Distance Movements and Implications for Disease Risk Management', *EcoHealth*, 7(2), pp. 204–212. Available at: https://doi.org/10.1007/s10393-010-0332-z.

Brierley, L. and Fowler, A. (2021) 'Predicting the animal hosts of coronaviruses from compositional biases of spike protein and whole genome sequences through machine learning', *PLOS Pathogens*, 17(4), p. e1009149. Available at: https://doi.org/10.1371/journal.ppat.1009149.

Briese, T. *et al.* (2014) 'Middle east respiratory syndrome coronavirus quasispecies that include homologues of human isolates revealed through whole- genome analysis and virus cultured from dromedary camels in Saudi Arabia', *mBio*, 5(3). Available at: https://doi.org/10.1128/MBIO.01146-14.

Burge, C., Campbell, A.M. and Karlin, S. (1992) 'Over- and under-representation of short oligonucleotides in DNA sequences.', *Proceedings of the National Academy of Sciences*, 89(4), pp. 1358–1362. Available at: https://doi.org/10.1073/PNAS.89.4.1358.

Camacho, C. *et al.* (2009) 'BLAST plus : architecture and applications', *Bmc Bioinformatics*, 10, p. 9. Available at: https://doi.org/10.1186/1471-2105-10-421.

Cappelle, J. *et al.* (2021) 'Longitudinal monitoring in Cambodia suggests higher circulation of alpha and betacoronaviruses in juvenile and immature bats of three species', *Scientific Reports*, 11(1), p. 24145. Available at: https://doi.org/10.1038/s41598-021-03169-z.

Carrique, L. *et al.* (2020) 'Host ANP32A mediates the assembly of the influenza virus replicase', *Nature*, 587(7835), pp. 638–643. Available at: https://doi.org/10.1038/s41586-020-2927-z.

CDC: Morbidity and Mortality Weekly Report (2004) *Update: Influenza Activity --- United States and Worldwide, 2003--04 Season, and Composition of the 2004--05*

Influenza Vaccine. Available at: https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5325a1.htm (Accessed: 25 April 2023).

Centers for Disease Control and Prevention (2019a) *1918 Pandemic (H1N1 virus)*, *www.cdc.gov*. Available at: https://www.cdc.gov/flu/pandemic-resources/1918-pandemic-h1n1.html.

Centers for Disease Control and Prevention (2019b) *1957-1958 Pandemic (H2N2 virus)*, *www.cdc.gov*. Available at: https://www.cdc.gov/flu/pandemic-resources/1957-1958-pandemic.html (Accessed: 4 June 2023).

Centers for Disease Control and Prevention (2019c) *1968 Pandemic (H3N2 virus)*, *www.cdc.gov*. Available at: https://www.cdc.gov/flu/pandemic-resources/1968-pandemic.html (Accessed: 4 June 2023).

Centers for Disease Control and Prevention (2019d) *2009 H1N1 Pandemic (H1N1pdm09 virus)*, *www.cdc.gov*. Available at: https://www.cdc.gov/flu/pandemic-resources/2009-h1n1-pandemic.html (Accessed: 4 June 2023).

Centre for Health Protection (2022) *CHP notified of human case of avian influenza A(H3N8) in Mainland*. Available at: https://www.info.gov.hk/gia/general/202205/30/P2022053000587.htm (Accessed: 10 April 2023).

Challender, D. *et al.* (2019) *Manis pentadactyla*, *The IUCN Red List of Threatened Species 2019*. Available at: https://doi.org/10.2305/IUCN.UK.2019-3.RLTS.T12764A168392151.en.

Chan, J.F.W. *et al.* (2013) 'Interspecies transmission and emergence of novel viruses: lessons from bats and birds', *Trends in Microbiology*, 21(10), pp. 544–555. Available at: https://doi.org/10.1016/J.TIM.2013.05.005.

Charron, G. *et al.* (2013) 'Prenylome profiling reveals S-farnesylation is crucial for membrane targeting and antiviral activity of ZAP long-isoform', *Proceedings of the National Academy of Sciences*, 110(27), pp. 11085–11090. Available at: https://doi.org/10.1073/PNAS.1302564110.

Cheng, X. *et al.* (2013) 'CpG Usage in RNA Viruses: Data and Hypotheses', *PLoS ONE*, 8(9). Available at: https://doi.org/10.1371/journal.pone.0074109.

Chidoti, V. *et al.* (2022) 'Longitudinal Survey of Coronavirus Circulation and Diversity in Insectivorous Bat Colonies in Zimbabwe', *Viruses*, 14(4), p. 781. Available at: https://doi.org/10.3390/v14040781.

Clark, K. *et al.* (2016) 'GenBank', *Nucleic Acids Research*, 44(Database issue), p. D67. Available at: https://doi.org/10.1093/NAR/GKV1276.

Cock, P.J.A. *et al.* (2009) 'Biopython: Freely available Python tools for computational molecular biology and bioinformatics', *Bioinformatics*, 25(11), pp. 1422–1423. Available at: https://doi.org/10.1093/bioinformatics/btp163.

Collienne, L. and Gavryushkin, A. (2021) 'Computing nearest neighbour interchange distances between ranked phylogenetic trees', *Journal of Mathematical Biology*, 82(1). Available at: https://doi.org/10.1007/S00285-021-01567-5.

Colson, P. *et al.* (2022) 'Culture and identification of a "Deltamicron" SARS-CoV-2 in a three cases cluster in southern France', *Journal of Medical Virology*, 94(8), pp. 3739–3749. Available at: https://doi.org/10.1002/jmv.27789.

Compton, A.A. and Emerman, M. (2013) 'Convergence and Divergence in the Evolution of the APOBEC3G-Vif Interaction Reveal Ancient Origins of Simian Immunodeficiency Viruses', *PLoS Pathogens*, 9(1), p. e1003135. Available at: https://doi.org/10.1371/journal.ppat.1003135.

Conceicao, C. *et al.* (2020) 'The SARS-CoV-2 Spike protein has a broad tropism for mammalian ACE2 proteins', *PLoS Biology*, 18(12 December), p. e3001016. Available at: https://doi.org/10.1371/journal.pbio.3001016.

Conway, M.J., Colpitts, T.M. and Fikrig, E. (2014) 'Role of the Vector in Arbovirus Transmission', *Annual Review of Virology*, 1(1), pp. 71–88. Available at: https://doi.org/10.1146/annurev-virology-031413-085513.

Cooper, D.N. and Krawczak, M. (1989) ‘Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes.’, *Human genetics*, 83(2), pp. 181–8. Available at: https://doi.org/10.1007/bf00286715.

Corman, V.M. *et al.* (2014) ‘Rooting the phylogenetic tree of middle East respiratory syndrome coronavirus by characterization of a conspecific virus from an African bat’, *Journal of virology*, 88(19), pp. 11297–11303. Available at: https://doi.org/10.1128/JVI.01498-14.

Cunningham, F. *et al.* (2022) ‘Ensembl 2022’, *Nucleic Acids Research*, 50(D1), pp. D988–D995. Available at: https://doi.org/10.1093/NAR/GKAB1049.

Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978) ‘A model of evolutionary change in proteins’, in M.O. Dayhoff (ed.) *Atlas of Protein Sequence and Structure*. 3rd edn. Washington DC: National Biomedical Research Foundation, pp. 345–352.

Delaune, D. *et al.* (2021) ‘A novel SARS-CoV-2 related coronavirus in bats from Cambodia’, *Nature Communications*, 12(1). Available at: https://doi.org/10.1038/s41467-021-26809-4.

Dellicour, S. *et al.* (2021) ‘Relax, keep walking—a practical guide to continuous phylogeographic inference with BEAST’, *Molecular Biology and Evolution*. Edited by R. Nielsen, 38(8), pp. 3486–3493. Available at: https://doi.org/10.1093/molbev/msab031.

Dittmann, J. *et al.* (2008) ‘Influenza A virus strains differ in sensitivity to the antiviral action of Mx-GTPase’, *Journal of virology*, 82(7), pp. 3624–3631. Available at: https://doi.org/10.1128/JVI.01753-07.

Donovan, J., Dufner, M. and Korennykh, A. (2013) ‘Structural basis for cytosolic double-stranded RNA surveillance by human oligoadenylate synthetase 1’, *Proceedings of the National Academy of Sciences of the United States of America*, 110(5), pp. 1652–1657. Available at: https://doi.org/10.1073/PNAS.1218528110.

Dou, D. *et al.* (2018) ‘Influenza A virus cell entry, replication, virion assembly and movement’, *Frontiers in Immunology*, 9(JUL), p. 1581. Available at: https://doi.org/10.3389/FIMMU.2018.01581/BIBTEX.

Drappier, M. and Michiels, T. (2015) ‘Inhibition of the OAS/RNase L pathway by viruses’, *Current Opinion in Virology*, 15, pp. 19–26. Available at: https://doi.org/10.1016/J.COVIRO.2015.07.002.

Drummond, A.J. *et al.* (2002) ‘Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data’, *Genetics*, 161(3), pp. 1307–1320. Available at: https://doi.org/10.1093/GENETICS/161.3.1307.

Drummond, A.J. and Rambaut, A. (2007) ‘BEAST: Bayesian evolutionary analysis by sampling trees’, *BMC Evolutionary Biology*, 7(1), pp. 1–8. Available at: https://doi.org/10.1186/1471-2148-7-214.

Duan, J. and Antezana, M.A. (2003) ‘Mammalian Mutation Pressure, Synonymous Codon Choice, and mRNA Degradation’, *Journal of Molecular Evolution*, 57(6), pp. 694–701. Available at: https://doi.org/10.1007/s00239-003-2519-1.

Duchene, S. *et al.* (2020) ‘Temporal signal and the phylodynamic threshold of SARS-CoV-2’, *Virus Evolution*, 6(2). Available at: https://doi.org/10.1093/ve/veaa061.

Dudas, G. *et al.* (2017) ‘Virus genomes reveal factors that spread and sustained the Ebola epidemic’, *Nature*, 544(7650), pp. 309–315. Available at: https://doi.org/10.1038/nature22040.

Duff, M.A. (2014) *CHARACTERIZATION OF H1N2 VARIANT INFLUENZA VIRUSES IN PIGS*. Kansas State University, Manhattan, Kansas.

Duggal, N.K. and Emerman, M. (2012) ‘Evolutionary conflicts between viruses and restriction factors shape immunity’, *Nature Reviews Immunology*, 12(10), pp. 687–695. Available at: https://doi.org/10.1038/nri3295.

Eddy, S.R. (2009) ‘A new generation of homology search tools based on probabilistic inference.’, *Genome informatics. International Conference on Genome Informatics*, 23(1), pp. 205–211. Available at: https://doi.org/10.1142/9781848165632_0019.

Edgar, R.C. (2004) ‘MUSCLE: A multiple sequence alignment method with reduced time and space complexity’, *BMC Bioinformatics*, 5(1), pp. 1–19. Available at: https://doi.org/10.1186/1471-2105-5-113.

Eguia, R.T. *et al.* (2021) ‘A human coronavirus evolves antigenically to escape antibody immunity’, *PLOS Pathogens*, 17(4), p. e1009453. Available at: https://doi.org/10.1371/JOURNAL.PPAT.1009453.

Elleder, D. *et al.* (2005) ‘The Receptor for the Subgroup C Avian Sarcoma and Leukosis Viruses, Tvc, Is Related to Mammalian Butyrophilins, Members of the Immunoglobulin Superfamily’, *Journal of Virology*, 79(16), pp. 10408–10419. Available at: https://doi.org/10.1128/JVI.79.16.10408-10419.2005.

Enard, D. *et al.* (2016) ‘Viruses are a dominant driver of protein adaptation in mammals’, *eLife*, 5. Available at: https://doi.org/10.7554/ELIFE.12469.

Everitt, A.R. *et al.* (2012) ‘IFITM3 restricts the morbidity and mortality associated with influenza’, *Nature*, 484(7395), pp. 519–523. Available at: https://doi.org/10.1038/nature10921.

Feeley, E.M. *et al.* (2011) ‘IFITM3 Inhibits Influenza A Virus Infection by Preventing Cytosolic Entry’, *PLOS Pathogens*, 7(10), p. e1002337. Available at: https://doi.org/10.1371/JOURNAL.PPAT.1002337.

Felsenstein, J. (1981) ‘Evolutionary trees from DNA sequences: A maximum likelihood approach’, *Journal of Molecular Evolution*, 17(6), pp. 368–376. Available at: https://doi.org/10.1007/BF01734359.

Felsenstein, J. (1985) ‘CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP’, *Evolution; international journal of organic evolution*, 39(4), pp. 783–791. Available at: https://doi.org/10.1111/J.1558-5646.1985.TB00420.X.

Ficarelli, M. *et al.* (2019) ‘KHNYN is essential for the zinc finger antiviral protein (ZAP) to restrict HIV-1 containing clustered CpG dinucleotides’, *eLife*, 8. Available at: https://doi.org/10.7554/eLife.46767.

Ficarelli, M. *et al.* (2020) ‘CpG Dinucleotides Inhibit HIV-1 Replication through Zinc Finger Antiviral Protein (ZAP)-Dependent and -Independent Mechanisms’, *Journal of Virology*, 94(6). Available at: https://doi.org/10.1128/JVI.01337-19.

Forni, D. *et al.* (2017) ‘Molecular Evolution of Human Coronavirus Genomes’, *Trends in Microbiology*, 25(1), pp. 35–48. Available at: https://doi.org/10.1016/j.tim.2016.09.001.

Forni, D. *et al.* (2023) ‘An APOBEC3 Mutational Signature in the Genomes of Human-Infecting Orthopoxviruses’, *mSphere*, 8(2). Available at: https://doi.org/10.1128/MSPHERE.00062-23.

Fros, J.J. *et al.* (2017) ‘CpG and upA dinucleotides in both coding and non-coding regions of echovirus 7 inhibit replication initiation post-entry’, *eLife*, 6. Available at: https://doi.org/10.7554/eLife.29112.

Frost, S.D.W., Magalis, B.R. and Kosakovsky Pond, S.L. (2018) ‘Neutral Theory and Rapidly Evolving Viral Pathogens’, *Molecular Biology and Evolution*, 35(6), p. 1348. Available at: https://doi.org/10.1093/MOLBEV/MSY088.

Fujita, S. *et al.* (2023) ‘Determination of the factors responsible for host tropism of SARS-CoV-2-related bat coronaviruses’, *bioRxiv*, p. 2023.04.13.536832. Available at: https://doi.org/10.1101/2023.04.13.536832.

Gamarra-Toledo, V. *et al.* (2023) ‘Mass Mortality of Marine Mammals Associated to Highly Pathogenic Influenza Virus (H5N1) in South America’, *bioRxiv*, p. 2023.02.08.527769. Available at: https://doi.org/10.1101/2023.02.08.527769.

Gao, G., Guo, X. and Goff, S.P. (2002) ‘Inhibition of retroviral RNA production by ZAP, a CCCH-type zinc finger protein’, *Science*, 297(5587), pp. 1703–1706. Available at: https://doi.org/10.1126/science.1074276.

Gao, Rongbao *et al.* (2013) ‘Human Infection with a Novel Avian-Origin Influenza A (H7N9) Virus’, *New England Journal of Medicine*, 368(20), pp. 1888–1897. Available at: https://doi.org/10.1056/NEJMOA1304459.

Gass, J.D. *et al.* (2023) ‘Global dissemination of influenza A virus is driven by wild bird migration through arctic and subarctic zones’, *Molecular Ecology*, 32(1), pp. 198–213. Available at: https://doi.org/10.1111/mec.16738.

Gaunt, E. *et al.* (2016) ‘Elevation of CpG frequencies in influenza a genome attenuates pathogenicity but enhances host response to infection’, *eLife*, 5. Available at: https://doi.org/10.7554/eLife.12735.

di Giallonardo, F. *et al.* (2017) ‘Dinucleotide Composition in Animal RNA Viruses Is Shaped More by Virus Family than by Host Species’, *Journal of Virology*, 91(8). Available at: https://doi.org/10.1128/jvi.02381-16.

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

Gifford, R.J. (2012) ‘Viral evolution in deep time: Lentiviruses and mammals’, *Trends in Genetics*, 28(2), pp. 89–100. Available at: https://doi.org/10.1016/j.tig.2011.11.003.

Giles, J.R. *et al.* (2021) ‘Optimizing noninvasive sampling of a zoonotic bat virus’, *Ecology and Evolution*, 11(18), pp. 12307–12321. Available at: https://doi.org/10.1002/ece3.7830.

Gojobori, T., Moriyama, E.N. and Kimura, M. (1990) ‘Molecular clock of viral evolution, and the neutral theory.’, *Proceedings of the National Academy of Sciences*, 87(24), pp. 10015–10018. Available at: https://doi.org/10.1073/PNAS.87.24.10015.

Goldman, N. and Yang, Z. (1994) ‘A codon-based model of nucleotide substitution for protein-coding DNA sequences.’, *Molecular Biology and Evolution*, 11(5), pp. 725–736. Available at: https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A040153.

Goldstein, S.A. *et al.* (2017) ‘Lineage A Betacoronavirus NS2 Proteins and the Homologous Torovirus Berne pp1a Carboxy-Terminal Domain Are

Phosphodiesterases That Antagonize Activation of RNase L', *Journal of Virology*, 91(5), pp. 2201–2217. Available at: https://doi.org/10.1128/jvi.02201-16.

Goldstein, S.A. *et al.* (2022) 'Extensive Recombination-driven Coronavirus Diversification Expands the Pool of Potential Pandemic Pathogens', *Genome Biology and Evolution*, 14(12). Available at: https://doi.org/10.1093/gbe/evac161.

Gonçalves-Carneiro, D. *et al.* (2021) 'Origin and evolution of the zinc finger antiviral protein', *PLOS Pathogens*, 17(4), p. e1009545. Available at: https://doi.org/10.1371/JOURNAL.PPAT.1009545.

Gorbalenya, A.E. *et al.* (2020) 'The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2', *Nature Microbiology*, 5(4), pp. 536–544. Available at: https://doi.org/10.1038/s41564-020-0695-z.

Graham, R.L. and Baric, R.S. (2010) 'Recombination, Reservoirs, and the Modular Spike: Mechanisms of Coronavirus Cross-Species Transmission', *Journal of Virology*, 84(7), pp. 3134–3146. Available at: https://doi.org/10.1128/jvi.01394-09.

Greenbaum, B.D. *et al.* (2008) 'Patterns of Evolution and Host Gene Mimicry in Influenza and Other RNA Viruses', *PLoS Pathogens*, 4(6), p. e1000079. Available at: https://doi.org/10.1371/journal.ppat.1000079.

Greenbaum, B.D. *et al.* (2014) 'Quantitative theory of entropic forces acting on constrained nucleotide sequences applied to viruses', *Proceedings of the National Academy of Sciences*, 111(13), pp. 5054–5059. Available at: https://doi.org/10.1073/pnas.1402285111.

Griffiths, R.C. and Tavare, S. (1994) 'Sampling theory for neutral alleles in a varying environment', *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1310), pp. 403–410. Available at: https://doi.org/10.1098/rstb.1994.0079.

Gu, H. *et al.* (2019) 'Dinucleotide evolutionary dynamics in influenza A virus', *Virus Evolution*, 5(2). Available at: https://doi.org/10.1093/VE/VEZ038.

Guindon, S. *et al.* (2010) ‘New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0’, *Systematic Biology*, 59(3), pp. 307–321. Available at: https://doi.org/10.1093/SYSBIO/SYQ010.

Guo, H. *et al.* (2020) ‘Evolutionary Arms Race between Virus and Host Drives Genetic Diversity in Bat Severe Acute Respiratory Syndrome-Related Coronavirus Spike Genes’, *Journal of Virology*, 94(20). Available at: https://doi.org/10.1128/JVI.00902-20.

Gusho, E. *et al.* (2014) ‘Murine AKAP7 has a 2',5'-phosphodiesterase domain that can complement an inactive murine coronavirus ns2 gene’, *mBio*, 5(4). Available at: https://doi.org/10.1128/mBio.01312-14.

Guzman, M.G. *et al.* (2010) ‘Dengue: a continuing global threat’, *Nature Reviews Microbiology*, 8(S12). Available at: https://doi.org/10.1038/nrmicro2460.

Hall, T.A. (1999) ‘BioEdit A User-Friendly Biological Sequence Alignment Editor and Analysis Program for Windows 95/98/NT’, in *Nucleic Acids Symposium Series 41*, pp. 95–98. Available at: https://thalljiscience.github.io/ (Accessed: 22 July 2023).

Hanson, G. and Coller, J. (2018) ‘Codon optimality, bias and usage in translation and mRNA decay’, *Nature Reviews Molecular Cell Biology*, 19(1), pp. 20–30. Available at: https://doi.org/10.1038/nrm.2017.91.

Harris, R.S. *et al.* (2003) ‘DNA Deamination Mediates Innate Immunity to Retroviral Infection’, *Cell*, 113, pp. 803–809.

Harvey, E. and Holmes, E.C. (2022) ‘Diversity and evolution of the animal virome’, *Nature Reviews Microbiology*, 20(6), pp. 321–334. Available at: https://doi.org/10.1038/s41579-021-00665-x.

Hastings, W.K. (1970) ‘Monte Carlo sampling methods using Markov chains and their applications’, *Biometrika*, 57(1), pp. 97–109. Available at: https://doi.org/10.1093/BIOMET/57.1.97.

Hatcher, E.L. *et al.* (2017) 'Virus Variation Resource – improved response to emergent viral outbreaks', *Nucleic Acids Research*, 45(D1), pp. D482–D490. Available at: https://doi.org/10.1093/NAR/GKW1065.

Hayakawa, S. *et al.* (2010) 'ZAPS is a potent stimulator of signaling mediated by the RNA helicase RIG-I during antiviral responses', *Nature Immunology*, 12(1), pp. 37–44. Available at: https://doi.org/10.1038/ni.1963.

Heath, L. *et al.* (2006) 'Recombination Patterns in Aphthoviruses Mirror Those Found in Other Picornaviruses', *Journal of Virology*, 80(23), pp. 11827–11832. Available at: https://doi.org/10.1128/jvi.01100-06.

Heid, H.W. *et al.* (1983) 'Butyrophilin, an apical plasma membrane-associated glycoprotein characteristic of lactating mammary glands of diverse species', *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 728(2), pp. 228–238. Available at: https://doi.org/10.1016/0005-2736(83)90476-5.

Henikoff, S. and Henikoff, J.G. (1992) 'Amino acid substitution matrices from protein blocks.', *Proceedings of the National Academy of Sciences*, 89(22), p. 10915. Available at: https://doi.org/10.1073/PNAS.89.22.10915.

Hien, T.T. *et al.* (2004) 'Avian Influenza A (H5N1) in 10 Patients in Vietnam', *New England Journal of Medicine*, 350(12), pp. 1179–1188. Available at: https://doi.org/10.1056/NEJMOA040419.

Hill, V. *et al.* (2022) 'The origins and molecular evolution of SARS-CoV-2 lineage B.1.1.7 in the UK', *Virus Evolution*, 8(2). Available at: https://doi.org/10.1093/ve/veac080.

Hoang, D.T. *et al.* (2018) 'UFBoot2: Improving the ultrafast bootstrap approximation', *Molecular Biology and Evolution*, 35(2), pp. 518–522. Available at: https://doi.org/10.1093/molbev/msx281.

Hoffmann, R.S. *et al.* (2013) *Princeton Pocket Guides: Mammals of China*. Edited by A.T. Smith and Y. Xie. Oxfordshire: Princeton University Press.

Holleufer, A. *et al.* (2021) ‘Two cGAS-like receptors induce antiviral immunity in Drosophila’, *Nature*, 597, pp. 114–118. Available at: https://doi.org/10.1038/s41586-021-03800-z.

Holmes, E.C. *et al.* (2021) ‘The origins of SARS-CoV-2: A critical review’, *Cell*, 184(19), pp. 4848–4856. Available at: https://doi.org/10.1016/j.cell.2021.08.017.

Holmes, E.C., Worobey, M. and Rambaut, A. (1999) ‘Phylogenetic evidence for recombination in dengue virus.’, *Molecular Biology and Evolution*, 16(3), pp. 405–409. Available at: https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A026121.

Hornung, V. *et al.* (2014) ‘OAS proteins and cGAS: unifying concepts in sensing and responding to cytosolic nucleic acids’, *Nature Reviews Immunology*, 14(8), pp. 521–528. Available at: https://doi.org/10.1038/nri3719.

Hu, D. *et al.* (2018) ‘Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats’, *Emerging Microbes and Infections*, 7(1). Available at: https://doi.org/10.1038/s41426-018-0155-5.

Hu, J. *et al.* (2018) ‘Origin and development of oligoadenylate synthetase immune system’, *BMC Evolutionary Biology*, 18(1), p. 201. Available at: https://doi.org/10.1186/S12862-018-1315-X/FIGURES/7.

Huang, Z., Wang, X. and Gao, G. (2010) ‘Analyses of SELEX-derived ZAP-binding RNA aptamers suggest that the binding specificity is determined by both structure and sequence of the RNA’, *Protein & Cell*, 1(8), pp. 752–759. Available at: https://doi.org/10.1007/S13238-010-0096-9.

Hubley, R. *et al.* (2016) ‘The Dfam database of repetitive DNA families’, *Nucleic Acids Research*, 44(D1), pp. D81–D89. Available at: https://doi.org/10.1093/nar/gkv1272.

Huelsenbeck, J.P. and Ronquist, F. (2001) ‘MRBAYES: Bayesian inference of phylogenetic trees’, *Bioinformatics*, 17(8), pp. 754–755. Available at: https://doi.org/10.1093/BIOINFORMATICS/17.8.754.

Huerta-Cepas, J., Serra, F. and Bork, P. (2016) 'ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data', *Molecular Biology and Evolution*, 33(6), pp. 1635–1638. Available at: https://doi.org/10.1093/molbev/msw046.

Huffman, J.E. *et al.* (2022) 'Multi-ancestry fine mapping implicates OAS1 splicing in risk of severe COVID-19', *Nature Genetics*, 54(2), pp. 125–127. Available at: https://doi.org/10.1038/s41588-021-00996-8.

Ibrahim, A. *et al.* (2019) 'A functional investigation of the suppression of CpG and UpA dinucleotide frequencies in plant RNA virus genomes', *Scientific Reports*, 9(1), pp. 1–14. Available at: https://doi.org/10.1038/s41598-019-54853-0.

International Committee on Taxonomy of Viruses (2019) *Virus Metadata Resource (VMR) MSL34*. Available at: https://ictv.global/vmr (Accessed: 29 April 2023).

Islam, A. *et al.* (2021) 'Evolutionary Dynamics and Epidemiology of Endemic and Emerging Coronaviruses in Humans, Domestic Animals, and Wildlife', *Viruses*, 13(10), p. 1908. Available at: https://doi.org/10.3390/V13101908.

Ito, J. *et al.* (2023) 'Convergent evolution of SARS-CoV-2 Omicron subvariants leading to the emergence of BQ.1.1 variant', *Nature Communications*, 14(1), p. 2671. Available at: https://doi.org/10.1038/s41467-023-38188-z.

Jackson, B. *et al.* (2021) 'Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic', *Cell*, 184(20), pp. 5179-5188.e8. Available at: https://doi.org/10.1016/j.cell.2021.08.014.

Jansen van Vuren, P. *et al.* (2021) 'A 1958 Isolate of Kedougou Virus (KEDV) from Ndumu, South Africa, Expands the Geographic and Temporal Range of KEDV in Africa', *Viruses*, 13(7), p. 1368. Available at: https://doi.org/10.3390/v13071368.

Jariani, A. *et al.* (2019) 'SANTA-SIM: Simulating viral sequence evolution dynamics under selection and recombination', *Virus Evolution*, 5(1). Available at: https://doi.org/10.1093/ve/vez003.

Jeon, Y.H. *et al.* (2005) 'Phosphodiesterase: Overview of protein structures, potential therapeutic applications and recent progress in drug development',

Cellular and Molecular Life Sciences, 62(11), pp. 1198–1220. Available at: https://doi.org/10.1007/S00018-005-4533-5.

Johnson, L.S., Eddy, S.R. and Portugaly, E. (2010) 'Hidden Markov model speed heuristic and iterative HMM search procedure', *BMC Bioinformatics*, 11(1), pp. 1–8. Available at: https://doi.org/10.1186/1471-2105-11-431.

Jumper, J. *et al.* (2021) 'Highly accurate protein structure prediction with AlphaFold', *Nature*, 596(7873), pp. 583–589. Available at: https://doi.org/10.1038/s41586-021-03819-2.

Kalyaanamoorthy, S. *et al.* (2017) 'ModelFinder: fast model selection for accurate phylogenetic estimates', *Nature Methods*, 14(6), pp. 587–589. Available at: https://doi.org/10.1038/nmeth.4285.

Karlin, S. and Burge, C. (1995) 'Dinucleotide relative abundance extremes: a genomic signature', *Trends in Genetics*, 11(7), pp. 283–290. Available at: https://doi.org/10.1016/S0168-9525(00)89076-9.

Karlin, S., Doerfler, W. and Cardon, L.R. (1994) 'Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses?', *Journal of virology*, 68(5), pp. 2889–2897.

Karlin, S. and Mrázek, J. (1997) 'Compositional differences within and between eukaryotic genomes', *Proceedings of the National Academy of Sciences*, 94(19), pp. 10227–10232. Available at: https://doi.org/10.1073/PNAS.94.19.10227.

Katoh, K. and Standley, D.M. (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30(4), pp. 772–780. Available at: https://doi.org/10.1093/molbev/mst010.

Kaye, D. and Pringle, C.R. (2005) 'Avian Influenza Viruses and their Implication for Human Health', *Clinical Infectious Diseases*, 40(1), pp. 108–112. Available at: https://doi.org/10.1086/427236.

Kerns, J.A., Emerman, M. and Malik, H.S. (2008) ‘Positive Selection and Increased Antiviral Activity Associated with the PARP-Containing Isoform of Human Zinc-Finger Antiviral Protein’, *PLOS Genetics*, 4(1), p. e21. Available at: https://doi.org/10.1371/JOURNAL.PGEN.0040021.

Kimura, M. (1968) ‘Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles’, *Genetical research*, 11(3), pp. 247–270. Available at: https://doi.org/10.1017/S0016672300011459.

de Klerk, A. *et al.* (2022) ‘Conserved recombination patterns across coronavirus subgenera’, *Virus Evolution*, 8(2). Available at: https://doi.org/10.1093/ve/veac054.

Klitting, R. *et al.* (2018) ‘Exploratory re-encoding of yellow fever virus genome: new insights for the design of live-attenuated viruses.’, *Virus evolution*, 4(2), p. vey021. Available at: https://doi.org/10.1093/ve/vey021.

Kmiec, D. *et al.* (2021) ‘S-farnesylation is essential for antiviral activity of the long ZAP isoform against RNA viruses with diverse replication strategies’, *PLOS Pathogens*, 17(10), p. e1009726. Available at: https://doi.org/10.1371/JOURNAL.PPAT.1009726.

Knoops, K. *et al.* (2008) ‘SARS-Coronavirus Replication Is Supported by a Reticulovesicular Network of Modified Endoplasmic Reticulum’, *PLOS Biology*, 6(9), p. e226. Available at: https://doi.org/10.1371/JOURNAL.PBIO.0060226.

Koonin, E. V. *et al.* (2008) ‘The Big Bang of picorna-like virus evolution antedates the radiation of eukaryotic supergroups’, *Nat. Rev. Microbiol.*, 6(12), pp. 925–939. Available at: https://doi.org/10.1038/nrmicro2030.

Koonin, E. V., Dolja, V. V. and Krupovic, M. (2015) ‘Origins and evolution of viruses of eukaryotes: The ultimate modularity’, *Virology*, 479–480, pp. 2–25. Available at: https://doi.org/10.1016/j.virol.2015.02.039.

Kosakovsky Pond, S.L. *et al.* (2006) ‘GARD: a genetic algorithm for recombination detection’, *BIOINFORMATICS APPLICATIONS NOTE*, 22(24), pp. 3096–3098. Available at: https://doi.org/10.1093/bioinformatics/btl474.

Kosakovsky Pond, S.L. *et al.* (2019) ‘HyPhy 2.5-A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies’, *Molecular Biology and Evolution*, 37(1), pp. 295–299. Available at: https://doi.org/10.1093/molbev/msz197.

Kosakovsky Pond, S.L. *et al.* (2021) ‘Contrast-FEL—A Test for Differences in Selective Pressures at Individual Sites among Clades and Sets of Branches’, *Molecular Biology and Evolution*, 38(3), pp. 1184–1198. Available at: https://doi.org/10.1093/MOLBEV/MSAA263.

Kosakovsky Pond, S.L. and Frost, S.D.W. (2005) ‘Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection’, *Molecular Biology and Evolution*, 22(5), pp. 1208–1222. Available at: https://doi.org/10.1093/molbev/msi105.

Kozlov, A.M. *et al.* (2019) ‘RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference’, *Bioinformatics*, 35(21), pp. 4453–4455. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTZ305.

Kraemer, M.U.G. *et al.* (2022) ‘Tracking the 2022 monkeypox outbreak with epidemiological data in real-time’, *The Lancet Infectious Diseases*, 22(7), pp. 941–942. Available at: https://doi.org/10.1016/S1473-3099(22)00359-0.

Krammer, F. and Schultz-Cherry, S. (2023) ‘We need to keep an eye on avian influenza’, *Nature Reviews Immunology 2023 23:5*, 23(5), pp. 267–268. Available at: https://doi.org/10.1038/s41577-023-00868-8.

Kumar, S. *et al.* (2017) ‘TimeTree: A Resource for Timelines, Timetrees, and Divergence Times’, *Molecular biology and evolution*, 34(7), pp. 1812–1819. Available at: https://doi.org/10.1093/molbev/msx116.

Kunec, D. and Osterrieder, N. (2016) ‘Codon Pair Bias Is a Direct Consequence of Dinucleotide Bias’, *Cell Reports*, 14(1), pp. 55–67. Available at: https://doi.org/10.1016/J.CELREP.2015.12.011.

Kuroda, M. *et al.* (2020) ‘Identification of interferon-stimulated genes that attenuate Ebola virus infection’, *Nature Communications*, 11(1). Available at: https://doi.org/10.1038/s41467-020-16768-7.

Kustin, T. and Stern, A. (2021) ‘Biased Mutation and Selection in RNA Viruses’, *Molecular Biology and Evolution*, 38(2), pp. 575–588. Available at: https://doi.org/10.1093/molbev/msaa247.

Kwon, Y.C. *et al.* (2013) ‘The ribonuclease l-dependent antiviral roles of human 2′,5′-oligoadenylate synthetase family members against hepatitis C virus’, *FEBS Letters*, 587(2), pp. 156–164. Available at: https://doi.org/10.1016/J.FEBSLET.2012.11.010.

Lacek, K.A. *et al.* (2022) ‘SARS-CoV-2 Delta–Omicron Recombinant Viruses, United States’, *Emerging Infectious Diseases*, 28(7), pp. 1442–1445. Available at: https://doi.org/10.3201/eid2807.220526.

Lam, T.T.Y. *et al.* (2020) ‘Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins’, *Nature*, 583, pp. 282–285. Available at: https://doi.org/10.1038/s41586-020-2169-0.

Lanfear, R. (2020) ‘A global phylogeny of SARS-CoV-2 sequences from GISAID’. Zenodo. Available at: https://doi.org/10.5281/zenodo.3958883.

Langat, P. *et al.* (2017) ‘Genome-wide evolutionary dynamics of influenza B viruses on a global scale’, *PLOS Pathogens*, 13(12), p. e1006749. Available at: https://doi.org/10.1371/journal.ppat.1006749.

Latinne, A. *et al.* (2020) ‘Origin and cross-species transmission of bat coronaviruses in China’, *Nature Communications*, 11(1), pp. 1–15. Available at: https://doi.org/10.1038/s41467-020-17687-3.

Lau, S.K.P. *et al.* (2015) ‘Discovery of a Novel Coronavirus, China Rattus Coronavirus HKU24, from Norway Rats Supports the Murine Origin of Betacoronavirus 1 and Has Implications for the Ancestor of Betacoronavirus Lineage A’, *Journal of Virology*, 89(6), pp. 3076–3092. Available at: https://doi.org/10.1128/JVI.02420-14.

Lednicky, J.A. *et al.* (2021) ‘Independent infections of porcine deltacoronavirus among Haitian children’, *Nature 2021 600:7887*, 600(7887), pp. 133–137. Available at: https://doi.org/10.1038/s41586-021-04111-z.

Lee, J. *et al.* (2020) 'No Evidence of Coronaviruses or Other Potentially Zoonotic Viruses in Sunda pangolins (Manis javanica) Entering the Wildlife Trade via Malaysia', *EcoHealth*, 17(3), pp. 406–418. Available at: https://doi.org/10.1007/s10393-020-01503-x.

Di Lella, S., Herrmann, A. and Mair, C.M. (2016) 'Modulation of the pH Stability of Influenza Virus Hemagglutinin: A Host Cell Adaptation Strategy', *Biophysical journal*, 110(11), pp. 2293–2301. Available at: https://doi.org/10.1016/J.BPJ.2016.04.035.

Lemey, P. *et al.* (2009) 'Identifying recombinants in human and primate immunodeficiency virus sequence alignments using quartet scanning', *BMC Bioinformatics*, 10(1). Available at: https://doi.org/10.1186/1471-2105-10-126.

Li, H. *et al.* (2019) 'Human-animal interactions and bat coronavirus spillover potential among rural residents in Southern China', *Biosafety and Health*, 1(2), pp. 84–90. Available at: https://doi.org/10.1016/j.bsheal.2019.10.004.

Li, L. *et al.* (2021) 'A novel SARS-CoV-2 related coronavirus with complex recombination isolated from bats in Yunnan province, China', *Emerging Microbes & Infections*, 10(1), pp. 1683–1690. Available at: https://doi.org/10.1080/22221751.2021.1964925.

Li, M.M.H. *et al.* (2017) 'TRIM25 Enhances the Antiviral Action of Zinc-Finger Antiviral Protein (ZAP)', *PLOS Pathogens*, 13(1), p. e1006145. Available at: https://doi.org/10.1371/JOURNAL.PPAT.1006145.

Li, Q. *et al.* (2020) 'Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia', *New England Journal of Medicine*, 382(13), pp. 1199–1207. Available at: https://doi.org/10.1056/NEJMoa2001316.

Li, X. *et al.* (2020) 'Emergence of SARS-CoV-2 through recombination and strong purifying selection', *Science Advances*, 6, p. eabb9153. Available at: https://doi.org/10.1101/2020.03.20.000885.

Lin, X.D. *et al.* (2017) 'Extensive diversity of coronaviruses in bats from China', *Virology*, 507, pp. 1–10. Available at: https://doi.org/10.1016/j.virol.2017.03.019.

Lin, Y.T. *et al.* (2020) ‘Human cytomegalovirus evades ZAP detection by suppressing CpG dinucleotides in the major immediate early 1 gene’, *PLOS Pathogens*, 16(9), p. e1008844. Available at: https://doi.org/10.1371/JOURNAL.PPAT.1008844.

Litman, G.W., Rast, J.P. and Fugmann, S.D. (2010) ‘The origins of vertebrate adaptive immunity’, *Nature Reviews Immunology*, 10(8), pp. 543–553. Available at: https://doi.org/10.1038/nri2807.

Liu, D. *et al.* (2013) ‘Origin and diversity of novel avian influenza A H7N9 viruses causing human infection: phylogenetic, structural, and coalescent analyses’, *The Lancet*, 381(9881), pp. 1926–1932. Available at: https://doi.org/10.1016/S0140-6736(13)60938-1.

Liu, P., Chen, W. and Chen, J.-P. (2019) ‘Viral Metagenomics Revealed Sendai Virus and Coronavirus Infection of Malayan Pangolins (Manis javanica)’, *Viruses*, 11(11), p. 979. Available at: https://doi.org/10.3390/v11110979.

Long, J.S. *et al.* (2016) ‘Species difference in ANP32A underlies influenza A virus polymerase host restriction’, *Nature*, 529(7584), pp. 101–104. Available at: https://doi.org/10.1038/nature16474.

Long, J.S. *et al.* (2018) ‘Host and viral determinants of influenza A virus species specificity’, *Nature Reviews Microbiology*, 17(2), pp. 67–81. Available at: https://doi.org/10.1038/s41579-018-0115-z.

Lucaci, A.G. *et al.* (2021) ‘Extra base hits: Widespread empirical support for instantaneous multiple-nucleotide changes’, *PLOS ONE*, 16(3), p. e0248337. Available at: https://doi.org/10.1371/JOURNAL.PONE.0248337.

Lucaci, A.G. *et al.* (2023) ‘Evolutionary Shortcuts via Multinucleotide Substitutions and Their Impact on Natural Selection Analyses’, *Molecular Biology and Evolution*, 40(7). Available at: https://doi.org/10.1093/molbev/msad150.

Luczo, J.M. *et al.* (2015) ‘Molecular pathogenesis of H5 highly pathogenic avian influenza: the role of the haemagglutinin cleavage site motif’, *Reviews in Medical Virology*, 25(6), pp. 406–430. Available at: https://doi.org/10.1002/RMV.1846.

Luo, J. *et al.* (2013) ‘Bat conservation in China: Should protection of subterranean habitats be a priority?’, *ORYX*, 47(4), pp. 526–531. Available at: https://doi.org/10.1017/S0030605311001505.

Luo, X. *et al.* (2020) ‘Molecular Mechanism of RNA Recognition by Zinc-Finger Antiviral Protein’, *Cell Reports*, 30(1), pp. 46-52.e4. Available at: https://doi.org/10.1016/j.celrep.2019.11.116.

Lynch, M. *et al.* (2020) ‘Inference of Historical Population-Size Changes with Allele-Frequency Data’, *G3 Genes|Genomes|Genetics*, 10(1), pp. 211–223. Available at: https://doi.org/10.1534/g3.119.400854.

Lynch, M. and Gabriel, W. (1990) ‘MUTATION LOAD AND THE SURVIVAL OF SMALL POPULATIONS’, *Evolution*, 44(7), pp. 1725–1737. Available at: https://doi.org/10.1111/j.1558-5646.1990.tb05244.x.

Lytras, S. *et al.* (2021) ‘The animal origin of SARS-CoV-2’, *Science*, 373(6558), pp. 968–970. Available at: https://doi.org/10.1126/SCIENCE.ABH0117.

Lytras, S. *et al.* (2022) ‘Exploring the Natural Origins of SARS-CoV-2 in the Light of Recombination’, *Genome Biology and Evolution*, 14(2). Available at: https://doi.org/10.1093/GBE/EVAC018.

Lytras, S. and Hughes, J. (2020) ‘Synonymous Dinucleotide Usage: A Codon-Aware Metric for Quantifying Dinucleotide Representation in Viruses’, *Viruses*, 12(4), p. 462. Available at: https://doi.org/10.3390/v12040462.

MacLean, O.A. *et al.* (2021) ‘Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen’, *PLOS Biology*. Edited by D. Tully, 19(3), p. e3001115. Available at: https://doi.org/10.1371/journal.pbio.3001115.

de Maio, N. *et al.* (2023) ‘Maximum likelihood pandemic-scale phylogenetics’, *Nature Genetics*, 55, pp. 746–752. Available at: https://doi.org/10.1038/s41588-023-01368-0.

Manns, M.P. and Maasoumy, B. (2022) ‘Breakthroughs in hepatitis C research: from discovery to cure’, *Nature Reviews Gastroenterology & Hepatology*, 19(8), pp. 533–550. Available at: https://doi.org/10.1038/s41575-022-00608-8.

Mänz, B. *et al.* (2013) ‘Pandemic Influenza A Viruses Escape from Restriction by Human MxA through Adaptive Mutations in the Nucleoprotein’, *PLoS Pathogens*, 9(3). Available at: https://doi.org/10.1371/journal.ppat.1003279.

Mao, R. *et al.* (2013) ‘Inhibition of Hepatitis B Virus Replication by the Host Zinc Finger Antiviral Protein’, *PLoS Pathogens*, 9(7). Available at: https://doi.org/10.1371/JOURNAL.PPAT.1003494.

Mao, X.G. *et al.* (2010) ‘Pleistocene climatic cycling drives intra-specific diversification in the intermediate horseshoe bat (Rhinolophus affinis) in Southern China’, *Molecular Ecology*, 19(13), pp. 2754–2769. Available at: https://doi.org/10.1111/j.1365-294X.2010.04704.x.

Marchenko, V. *et al.* (2022) ‘Diversity of gammacoronaviruses and deltacoronaviruses in wild birds and poultry in Russia’, *Scientific Reports*, 12(1). Available at: https://doi.org/10.1038/s41598-022-23925-z.

Mariani, R. *et al.* (2003) ‘Species-specific exclusion of APOBEC3G from HIV-1 virions by Vif’, *Cell*, 114(1), pp. 21–31. Available at: https://doi.org/10.1016/S0092-8674(03)00515-4.

Marinova-Petkova, A. *et al.* (2017) ‘Avian Influenza A(H7N2) Virus in Human Exposed to Sick Cats, New York, USA, 2016’, *Emerging Infectious Diseases*, 23(12), pp. 2046–2049. Available at: https://doi.org/10.3201/EID2312.170798.

Markov, P. V. *et al.* (2023) ‘The evolution of SARS-CoV-2’, *Nature Reviews Microbiology*, 21(6), pp. 361–379. Available at: https://doi.org/10.1038/s41579-023-00878-2.

Martin, D. and Rybicki, E. (2000) ‘RDP: detection of recombination amongst aligned sequences’, *BIOINFORMATICS APPLICATIONS NOTE*, 16(6), pp. 562–563. Available at: https://doi.org/10.1093/bioinformatics/16.6.562.

Martin, D.P. *et al.* (2005) ‘A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints’, *AIDS Research and Human Retroviruses*, 21(1), pp. 98–102. Available at: https://doi.org/10.1089/aid.2005.21.98.

Martin, D.P. *et al.* (2017) ‘Detecting and analyzing genetic recombination using RDP4’, in *Methods in Molecular Biology*. Humana Press Inc., pp. 433–460. Available at: https://doi.org/10.1007/978-1-4939-6622-6_17.

Martin, D.P. *et al.* (2021) ‘RDP5: A computer program for analysing recombination in, and removing signals of recombination from, nucleotide sequence datasets’, *Virus Evolution*, 7(1), p. 87. Available at: https://doi.org/10.1093/ve/veaa087.

Martin, D.P. *et al.* (2022) ‘Selection Analysis Identifies Clusters of Unusual Mutational Changes in Omicron Lineage BA.1 That Likely Impact Spike Function’, *Molecular Biology and Evolution*, 39(4). Available at: https://doi.org/10.1093/molbev/msac061.

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

McCauley, J.W. *et al.* (2019) *International Committee on the Taxonomy of Viruses: 2019; Negative Sense RNA Viruses: Orthomyxoviridae*.

Mifsud, J.C.O. *et al.* (2023) ‘Transcriptome mining extends the host range of the Flaviviridae to non-bilaterians’, *Virus Evolution*, 9(1). Available at: https://doi.org/10.1093/VE/VEAC124.

Milewska, A. *et al.* (2018) ‘APOBEC3-mediated restriction of RNA virus replication’, *Scientific Reports*, 8, p. 5960. Available at: https://doi.org/10.1038/s41598-018-24448-2.

Minh, B.Q. *et al.* (2020) ‘IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era’, *Molecular Biology and Evolution*, 37(5), pp. 1530–1534. Available at: https://doi.org/10.1093/MOLBEV/MSAA015.

Mirdita, M. *et al.* (2022) ‘ColabFold: making protein folding accessible to all’, *Nature Methods 2022 19:6*, 19(6), pp. 679–682. Available at: https://doi.org/10.1038/s41592-022-01488-1.

Mistry, J. *et al.* (2013) ‘Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions’, *Nucleic acids research*, 41(12). Available at: https://doi.org/10.1093/NAR/GKT263.

Mollentze, N., Babayan, S.A. and Streicker, D.G. (2021) ‘Identifying and prioritizing potential human-infecting viruses from their genome sequences’, *PLOS Biology*, 19(9), p. e3001390. Available at: https://doi.org/10.1371/JOURNAL.PBIO.3001390.

Moraes, S.N. *et al.* (2022) ‘Evidence linking APOBEC3B genesis and evolution of innate immune antagonism by gamma-herpesvirus ribonucleotide reductases’, *eLife*, 11. Available at: https://doi.org/10.7554/eLife.83893.

Morrison, J.M. *et al.* (1967) ‘Nearest neighbour base sequence analysis of the deoxyribonucleic acids of a further three mammalian viruses: Simian virus 40, human papilloma virus and adenovirus type 2.’, *The Journal of general virology*, 1(1), pp. 101–108. Available at: https://doi.org/10.1099/0022-1317-1-1-101.

Mozzi, A. *et al.* (2015) ‘OASes and STING: Adaptive Evolution in Concert’, *Genome Biology and Evolution*, 7(4), pp. 1016–1032. Available at: https://doi.org/10.1093/GBE/EVV046.

Murrell, B. *et al.* (2012) ‘Detecting individual sites subject to episodic diversifying selection’, *PLoS Genetics*, 8(7), p. e1002764. Available at: https://doi.org/10.1371/journal.pgen.1002764.

Murrell, B. *et al.* (2013) ‘FUBAR: A fast, unconstrained bayesian AppRoximation for inferring selection’, *Molecular Biology and Evolution*, 30(5), pp. 1196–1205. Available at: https://doi.org/10.1093/molbev/mst030.

Murrell, B. *et al.* (2015) ‘Gene-Wide Identification of Episodic Selection’, *Molecular Biology and Evolution*, 32(5), pp. 1365–1371. Available at: https://doi.org/10.1093/MOLBEV/MSV035.

Muse, S. V. and Gaut, B.S. (1994) ‘A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome’, *Molecular biology and evolution*, 11(5), pp. 715–724. Available at: https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A040152.

Nascimento, F.F., Reis, M. Dos and Yang, Z. (2017) ‘A biologist’s guide to Bayesian phylogenetic analysis’, *Nature Ecology & Evolution 2017 1:10*, 1(10), pp. 1446–1454. Available at: https://doi.org/10.1038/s41559-017-0280-x.

Nchioua, R. *et al.* (2020) ‘SARS-CoV-2 Is Restricted by Zinc Finger Antiviral Protein despite Preadaptation to the Low-CpG Environment in Humans’, *mBio*, 11(5). Available at: https://doi.org/10.1128/mBio.01930-20.

Nei, M. and Gojobori, T. (1986) ‘Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.’, *Molecular Biology and Evolution*, 3(5), pp. 418–426. Available at: https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A040410.

Nguyen, L.T. *et al.* (2015) ‘IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies’, *Molecular Biology and Evolution*, 32(1), pp. 268–274. Available at: https://doi.org/10.1093/molbev/msu300.

Nikolaidis, M. *et al.* (2021) ‘The Neighborhood of the Spike Gene Is a Hotspot for Modular Intertypic Homologous and Nonhomologous Recombination in Coronavirus Genomes’, *Molecular Biology and Evolution*, 39(1). Available at: https://doi.org/10.1093/MOLBEV/MSAB292.

Noriel, P. *et al.* (2013) ‘Emergence of H3N2pM-like and novel reassortant H3N1 swine viruses possessing segments derived from the A (H1N1)pdm09 influenza virus, Korea’, *Influenza and other respiratory viruses*, 7(6), pp. 1283–1291. Available at: https://doi.org/10.1111/irv.12154.

Odon, V. *et al.* (2019) ‘The role of ZAP and OAS3/RNAseL pathways in the attenuation of an RNA virus with elevated frequencies of CpG and UpA dinucleotides’, *Nucleic Acids Research*, 47(15), pp. 8061–8083. Available at: https://doi.org/10.1093/nar/gkz581.

Ohta, T. (1995) ‘Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory’, *Journal of molecular evolution*, 40(1), pp. 56–63. Available at: https://doi.org/10.1007/BF00166595.

Oliva, A. *et al.* (2019) ‘Accounting for ambiguity in ancestral sequence reconstruction’, *Bioinformatics*, 35(21), pp. 4290–4297. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTZ249.

Oliver, I. *et al.* (2022) ‘A case of avian influenza A(H5N1) in England, January 2022’, *Eurosurveillance*, 27(5), p. 2200061. Available at: https://doi.org/10.2807/1560-7917.ES.2022.27.5.2200061.

O’Toole, Á. *et al.* (2023) ‘Putative APOBEC3 deaminase editing in MPXV as evidence for sustained human transmission since at least 2016’, *bioRxiv*, p. 2023.01.23.525187. Available at: https://doi.org/10.1101/2023.01.23.525187.

Oude Munnink, B.B., Worp, N., *et al.* (2021) ‘The next phase of SARS-CoV-2 surveillance: real-time molecular epidemiology’, *Nature Medicine*, 27(9), pp. 1518–1524. Available at: https://doi.org/10.1038/s41591-021-01472-w.

Oude Munnink, B.B., Sikkema, R.S., *et al.* (2021) ‘Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans’, *Science*, 371(6525), pp. 172–177. Available at: https://doi.org/10.1126/science.abe5901.

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

Paradis, E. (2013) ‘Molecular dating of phylogenies by likelihood methods: A comparison of models and a new information criterion’, *Molecular Phylogenetics and Evolution*, 67(2), pp. 436–444. Available at: https://doi.org/10.1016/J.YMPEV.2013.02.008.

Paradis, E., Claude, J. and Strimmer, K. (2004) ‘APE: Analyses of Phylogenetics and Evolution in R language’, *Bioinformatics*, 20(2), pp. 289–290. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTG412.

Parry, R. *et al.* (2020) 'Divergent Influenza-Like Viruses of Amphibians and Fish Support an Ancient Evolutionary Association', *Viruses*, 12(9), p. 1042. Available at: https://doi.org/10.3390/v12091042.

Patrono, L. V *et al.* (2022) 'Archival influenza virus genomes from Europe reveal genomic variability during the 1918 pandemic', *Nature Communications*, 13(1). Available at: https://doi.org/10.1038/s41467-022-29614-9.

Paules, C. and Subbarao, K. (2017) 'Influenza', *The Lancet*, 390(10095), pp. 697–708. Available at: https://doi.org/10.1016/S0140-6736(17)30129-0.

Pekar, J.E. *et al.* (2022) 'The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2', *Science*, 377(6609), pp. 960–966. Available at: https://doi.org/10.1126/science.abp8337.

Petrova, V.N. and Russell, C.A. (2017) 'The evolution of seasonal influenza viruses', *Nature Reviews Microbiology*, 16(1), pp. 47–60. Available at: https://doi.org/10.1038/nrmicro.2017.118.

Pettersen, E.F. *et al.* (2021) 'UCSF ChimeraX: Structure visualization for researchers, educators, and developers', *Protein science : a publication of the Protein Society*, 30(1), pp. 70–82. Available at: https://doi.org/10.1002/PRO.3943.

Pierson, T.C. and Diamond, M.S. (2020) 'The continued threat of emerging flaviviruses', *Nature Microbiology*, 5(6), pp. 796–812. Available at: https://doi.org/10.1038/s41564-020-0714-0.

Pinto, R.M. *et al.* (2023) 'BTN3A3 evasion promotes the zoonotic potential of influenza A viruses', *Nature*, 619, pp. 338–347. Available at: https://doi.org/10.1038/s41586-023-06261-8.

Piraccini, R. (2016) *Rhinolophus ferrumequinum*, *The IUCN Red List of Threatened Species 2016*. Available at: https://doi.org/https://dx.doi.org/10.2305/IUCN.UK.2016-2.RLTS.T19517A21973253.en.

du Plessis, L. *et al.* (2021) ‘Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK’, *Science*, 371(6530), pp. 708–712. Available at: https://doi.org/10.1126/SCIENCE.ABF2946.

Plotkin, J.B. and Kudla, G. (2010) ‘Synonymous but not the same: the causes and consequences of codon bias’, *Nature Reviews Genetics 2011 12:1*, 12(1), pp. 32–42. Available at: https://doi.org/10.1038/nrg2899.

Poon, A. and Otto, S.P. (2000) ‘COMPENSATING FOR OUR LOAD OF MUTATIONS: FREEZING THE MELTDOWN OF SMALL POPULATIONS’, *Evolution*, 54(5), pp. 1467–1479. Available at: https://doi.org/10.1111/j.0014-3820.2000.tb00693.x.

Posada, D. (2002) ‘Evaluation of methods for detecting recombination from DNA sequences: Empirical data’, *Molecular Biology and Evolution*, 19(5), pp. 708–717. Available at: https://doi.org/10.1093/oxfordjournals.molbev.a004129.

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

Price, M.N., Dehal, P.S. and Arkin, A.P. (2010) ‘FastTree 2 - Approximately maximum-likelihood trees for large alignments’, *PLoS ONE*, 5(3), p. e9490. Available at: https://doi.org/10.1371/journal.pone.0009490.

Pybus, O.G. and Shapiro, B. (2009) ‘Natural selection and adaptation of molecular sequences’, in P. Lemey, M. Salemi, and A.-M. Vandamme (eds) *The Phylogenetic Handbook*. 2nd edn. Cambridge: Cambridge University Press, pp. 407–418.

Qin, X.-C. *et al.* (2014) ‘A tick-borne segmented RNA virus contains genome segments derived from unsegmented viral ancestors’, *Proceedings of the National Academy of Sciences*, 111(18), pp. 6744–6749. Available at: https://doi.org/10.1073/pnas.1324194111.

Qu, B. *et al.* (2020) 'Reassortment and adaptive mutations of an emerging avian influenza virus H7N4 subtype in China', *PLOS ONE*, 15(1), p. e0227597. Available at: https://doi.org/10.1371/JOURNAL.PONE.0227597.

R Core Team (2022) *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/ (Accessed: 29 April 2023).

Rambaut, A. (2000) 'Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies', *Bioinformatics*, 16(4), pp. 395–399. Available at: https://doi.org/10.1093/BIOINFORMATICS/16.4.395.

Rambaut, A. *et al.* (2016) 'Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen)', *Virus Evolution*, 2(1), p. vew007. Available at: https://doi.org/10.1093/ve/vew007.

Rambaut, A. *et al.* (2018) 'Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7', *Syst. Biol*, 67(5). Available at: https://doi.org/10.1093/sysbio/syy032.

Rambaut, A. *et al.* (2020) 'A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology', *Nature Microbiology*, 5(11), pp. 1403–1407. Available at: https://doi.org/10.1038/s41564-020-0770-5.

Rannala, B. and Yang, Z. (1996) 'Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference', *Journal of Molecular Evolution*, 43(3), pp. 304–311. Available at: https://doi.org/10.1007/BF02338839/METRICS.

Reusken, C.B.E.M. *et al.* (2013) 'Middle East respiratory syndrome coronavirus neutralising serum antibodies in dromedary camels: A comparative serological study', *The Lancet Infectious Diseases*, 13(10), pp. 859–866. Available at: https://doi.org/10.1016/S1473-3099(13)70164-6.

Revell, L.J. (2012) 'phytools: an R package for phylogenetic comparative biology (and other things)', *Methods in Ecology and Evolution*, 3(2), pp. 217–223. Available at: https://doi.org/10.1111/J.2041-210X.2011.00169.X.

Rhodes, D.A., De Bono, B. and Trowsdale, J. (2005) 'Relationship between SPRY and B30.2 protein domains. Evolution of a component of immune defence?', *Immunology*, 116(4), pp. 411–417. Available at: https://doi.org/10.1111/J.1365-2567.2005.02248.X.

Rice, A.M. *et al.* (2021) 'Evidence for Strong Mutation Bias toward, and Selection against, U Content in SARS-CoV-2: Implications for Vaccine Design', *Molecular Biology and Evolution*, 38(1), pp. 67–83. Available at: https://doi.org/10.1093/molbev/msaa188.

Rigau, M., Uldrich, A.P. and Behren, A. (2021) 'Targeting butyrophilins for cancer immunotherapy', *Trends in Immunology*, 42(8), pp. 670–680. Available at: https://doi.org/10.1016/J.IT.2021.06.002.

Rihn, S.J. *et al.* (2021) 'A plasmid DNA-launched SARS-CoV-2 reverse genetics system and coronavirus toolkit for COVID-19 research', *PLOS Biology*, 19(2), p. e3001091. Available at: https://doi.org/10.1371/JOURNAL.PBIO.3001091.

Riitho, V. *et al.* (2020) 'Bovine Pestivirus Heterogeneity and Its Potential Impact on Vaccination and Diagnosis', *Viruses*, 12(10), p. 1134. Available at: https://doi.org/10.3390/v12101134.

Robertson, D.L., Hahn, B.H. and Sharp, P.M. (1995) 'Recombination in AIDS viruses', *Journal of Molecular Evolution*, 40(3), pp. 249–259. Available at: https://doi.org/10.1007/BF00163230.

Roemer, C. (2022) *BJ.1/BM.1.1.1 (=BA.2.75.3.1.1.1) recombinant with breakpoint in S1 [>=5 sequences, 3x Singapore, 2x US as of 2022-09-12]*, *GitHub*. Available at: https://github.com/cov-lineages/pango-designation/issues/1058 (Accessed: 2 June 2023).

Roingeard, P. *et al.* (2022) 'The double-membrane vesicle (DMV): a virus-induced organelle dedicated to the replication of SARS-CoV-2 and other positive-sense single-stranded RNA viruses', 79, p. 425. Available at: https://doi.org/10.1007/s00018-022-04469-x.

Ronquist, F., van der Mark, P. and Huelsenbeck, J.P. (2009) 'Bayesian phylogenetic analysis using MRBAYES', in P. Lemey, M. Salemi, and A.-M. Vandamme (eds) *The Phylogenetic Handbook*. 2nd edn. Cambridge: Cambridge University Press, pp. 210–266.

Ruiz, M. *et al.* (2022) 'Influenza D Virus: A Review and Update of Its Role in Bovine Respiratory Syndrome', *Viruses*, 14(12), p. 2717. Available at: https://doi.org/10.3390/v14122717.

Russell, G.J. *et al.* (1976) 'Doublet frequency analysis of fractionated vertebrate nuclear DNA', *Journal of Molecular Biology*, 108(1), pp. 1–20. Available at: https://doi.org/10.1016/S0022-2836(76)80090-3.

Russell, R.A. and Pathak, V.K. (2007) 'Identification of Two Distinct Human Immunodeficiency Virus Type 1 Vif Determinants Critical for Interactions with Human APOBEC3G and APOBEC3F', *Journal of Virology*, 81(15), pp. 8201–8210. Available at: https://doi.org/10.1128/JVI.00395-07.

Sadler, A.J. and Williams, B.R.G. (2008) 'Interferon-inducible antiviral effectors', *Nature Reviews. Immunology*, 8(7), p. 559. Available at: https://doi.org/10.1038/NRI2314.

Sagulenko, P., Puller, V. and Neher, R.A. (2018) 'TreeTime: Maximum-likelihood phylodynamic analysis', *Virus Evolution*, 4(1). Available at: https://doi.org/10.1093/VE/VEX042.

Saitou, N. and Nei, M. (1987) 'The neighbor-joining method: a new method for reconstructing phylogenetic trees', *Molecular biology and evolution*, 4(4), pp. 406–425. Available at: https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A040454.

Salminen, M.O. *et al.* (1995) 'Identification of Breakpoints in Intergenotypic Recombinants of HIV Type 1 by Bootscanning', *AIDS Research and Human Retroviruses*, 11(11), pp. 1423–1425. Available at: https://doi.org/10.1089/aid.1995.11.1423.

Schneider, W.M., Chevillotte, M.D. and Rice, C.M. (2014) 'Interferon-Stimulated Genes: A Complex Web of Host Defenses', *Annual Review of Immunology*, 32, pp.

513–545. Available at: https://doi.org/10.1146/ANNUREV-IMMUNOL-032713-120231.

Schoggins, J.W. *et al.* (2013) 'Pan-viral specificity of IFN-induced genes reveals new roles for cGAS in innate immunity', *Nature 2013 505:7485*, 505(7485), pp. 691–695. Available at: https://doi.org/10.1038/nature12862.

Schoggins, J.W. (2019) 'Interferon-Stimulated Genes: What Do They All Do?', *Annual Review of Virology*, 6(1), pp. 567–584. Available at: https://doi.org/10.1146/ANNUREV-VIROLOGY-092818-015756.

Schwerk, J. *et al.* (2019) 'RNA-binding protein isoforms ZAP-S and ZAP-L have distinct antiviral and immune resolution functions', *Nature Immunology*, 20(12), pp. 1610–1620. Available at: https://doi.org/10.1038/s41590-019-0527-6.

Sederdahl, B.K. and Williams, J. V. (2020) 'Epidemiology and Clinical Characteristics of Influenza C Virus', *Viruses*, 12(1), p. 89. Available at: https://doi.org/10.3390/v12010089.

Shackelton, L.A., Parrish, C.R. and Holmes, E.C. (2006) 'Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses', *Journal of Molecular Evolution*, 62(5), pp. 551–563. Available at: https://doi.org/10.1007/S00239-005-0221-1.

Sharp, C.P. *et al.* (2023) 'CpG dinucleotide enrichment in the influenza A virus genome as a live attenuated vaccine development strategy', *PLOS Pathogens*, 19(5), p. e1011357. Available at: https://doi.org/10.1371/journal.ppat.1011357.

Sharp, P.M. *et al.* (1995) 'DNA sequence evolution: the sounds of silence', *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 349(1329), pp. 241–247. Available at: https://doi.org/10.1098/RSTB.1995.0108.

Sharp, P.M. and Li, W.H. (1987) 'The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications', *Nucleic Acids Research*, 15(3), pp. 1281–1295. Available at: https://doi.org/10.1093/NAR/15.3.1281.

Sharp, P.M., Tuohy, T.M.F. and Mosurski, K.R. (1986) 'Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes', *Nucleic Acids Research*, 14(13), pp. 5125–5143. Available at: https://doi.org/10.1093/NAR/14.13.5125.

Shaw, A.E. *et al.* (2017) 'Fundamental properties of the mammalian innate immune system revealed by multispecies comparison of type I interferon responses', *PLoS Biology*, 15(12), p. e2004086. Available at: https://doi.org/10.1371/journal.pbio.2004086.

Shaw, G. and Kamen, R. (1986) 'A conserved AU sequence from the 3′ untranslated region of GM-CSF mRNA mediates selective mRNA degradation', *Cell*, 46(5), pp. 659–667. Available at: https://doi.org/10.1016/0092-8674(86)90341-7.

Sheehy, A.M. *et al.* (2002) 'Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein', *Nature*, 418(6898), pp. 646–650. Available at: https://doi.org/10.1038/nature00939.

Shi, M. *et al.* (2016) 'Divergent Viruses Discovered in Arthropods and Vertebrates Revise the Evolutionary History of the Flaviviridae and Related Viruses', *Journal of Virology*, 90(2), pp. 659–669. Available at: https://doi.org/10.1128/JVI.02036-15.

Shinya, K. *et al.* (2006) 'Influenza virus receptors in the human airway', *Nature*, 440(7083), pp. 435–436. Available at: https://doi.org/10.1038/440435a.

Silva, F.L. da and Stur, E. (2019) 'Pentaneurella katterjokki Fittkau & Murray (Chironomidae, Tanypodinae): redescription and phylogenetic position', *ZooKeys*, 833, pp. 107–119. Available at: https://doi.org/10.3897/zookeys.833.30936.

Simmonds, P. *et al.* (2013) 'Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla -selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses', *BMC Genomics*, 14(1). Available at: https://doi.org/10.1186/1471-2164-14-610.

Simmonds, P. *et al.* (2017) 'ICTV Virus Taxonomy Profile: Flaviviridae', *Journal of General Virology*, 98(1), pp. 2–3. Available at: https://doi.org/10.1099/jgv.0.000672.

Simmonds, P. *et al.* (2023) ‘Four principles to establish a universal virus taxonomy’, *PLOS Biology*, 21(2), p. e3001922. Available at: https://doi.org/10.1371/journal.pbio.3001922.

Simón, D. *et al.* (2017) ‘Host influence in the genomic composition of flaviviruses: A multivariate approach’, *Biochemical and Biophysical Research Communications*, 492(4), pp. 572–578. Available at: https://doi.org/10.1016/j.bbrc.2017.06.088.

Simon-Loriere, E. *et al.* (2009) ‘Molecular Mechanisms of Recombination Restriction in the Envelope Gene of the Human Immunodeficiency Virus’, *PLoS Pathogens*. Edited by E.C. Holmes, 5(5), p. e1000418. Available at: https://doi.org/10.1371/journal.ppat.1000418.

Simon-Loriere, E. and Holmes, E.C. (2011) ‘Why do RNA viruses recombine?’, *Nature Reviews Microbiology*, 9(8), pp. 617–626. Available at: https://doi.org/10.1038/nrmicro2614.

Slavik, K.M. *et al.* (2021) ‘cGAS-like receptors sense RNA and control 3′2′-cGAMP signaling in Drosophila’, *Nature*, 597, pp. 109–113. Available at: https://doi.org/10.1038/s41586-021-03743-5.

Smith, G.J., Bahl, J., *et al.* (2009) ‘Dating the emergence of pandemic influenza viruses’, *Proceedings of the National Academy of Sciences*, 106(28), pp. 11709–11712. Available at: https://doi.org/10.1073/PNAS.0904991106.

Smith, G.J., Vijaykrishna, D., *et al.* (2009) ‘Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic’, *Nature*, 459(7250), pp. 1122–1125. Available at: https://doi.org/10.1038/nature08182.

Smith, I.A. *et al.* (2010) ‘BTN1A1, the Mammary Gland Butyrophilin, and BTN2A2 Are Both Inhibitors of T Cell Activation’, *The Journal of Immunology*, 184(7), pp. 3514–3525. Available at: https://doi.org/10.4049/jimmunol.0900416.

Smith, M.D. *et al.* (2015) ‘Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection’, *Molecular Biology and Evolution*, 32(5), pp. 1342–1353. Available at: https://doi.org/10.1093/molbev/msv022.

Sola, I. *et al.* (2015) ‘Continuous and Discontinuous RNA Synthesis in Coronaviruses’, *Annual Review of Virology*, 2(1), pp. 265–288. Available at: https://doi.org/10.1146/annurev-virology-100114-055218.

Souilmi, Y. *et al.* (2021) ‘An ancient viral epidemic involving host coronavirus interacting genes more than 20,000 years ago in East Asia’, *Current Biology*, 31, pp. 3504-3514.e9. Available at: https://doi.org/10.1016/j.cub.2021.05.067.

Soveg, F.W. *et al.* (2021) ‘Endomembrane targeting of human OAS1 p46 augments antiviral activity’, *eLife*, 10. Available at: https://doi.org/10.7554/ELIFE.71047.

Stanke, M. *et al.* (2008) ‘Using native and syntenically mapped cDNA alignments to improve de novo gene finding’, *Bioinformatics*, 24(5), pp. 637–644. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTN013.

Stark, G.R. and Darnell, J.E. (2012) ‘The JAK-STAT Pathway at Twenty’, *Immunity*, 36(4), pp. 503–514. Available at: https://doi.org/10.1016/J.IMMUNI.2012.03.013.

Starr, T.N. *et al.* (2022) ‘ACE2 binding is an ancestral and evolvable trait of sarbecoviruses’, *Nature*, 603(7903), pp. 913–918. Available at: https://doi.org/10.1038/s41586-022-04464-z.

Stavrou, S. and Ross, S.R. (2015) ‘APOBEC3 Proteins in Viral Immunity’, *The Journal of Immunology*, 195(10), pp. 4565–4570. Available at: https://doi.org/10.4049/JIMMUNOL.1501504.

Steinegger, M. and Söding, J. (2017) ‘MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets’, *Nature Biotechnology*, 35(11), pp. 1026–1028. Available at: https://doi.org/10.1038/nbt.3988.

Subak-Sharpe, H. *et al.* (1966) ‘An approach to evolutionary relationships of mammalian DNA viruses through analysis of the pattern of nearest neighbor base sequences.’, *Cold Spring Harbor symposia on quantitative biology*, 31, pp. 737–748. Available at: https://doi.org/10.1101/SQB.1966.031.01.094.

Subbarao, K.E., London, W. and Murphy, B.R. (1993) ‘A Single Amino Acid in the PB2 Gene of Influenza A Virus Is a Determinant of Host Range’, *JOURNAL OF*

VIROLOGY, 67(4), pp. 1761–1764. Available at: https://doi.org/10.1128/JVI.67.4.1761-1764.1993.

Suchard, M.A. *et al.* (2018) 'Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10', *Virus Evolution*, 4(1). Available at: https://doi.org/10.1093/VE/VEY016.

Suyama, M., Torrents, D. and Bork, P. (2006) 'PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments', *Nucleic Acids Research*, 34(suppl_2), pp. W609–W612. Available at: https://doi.org/10.1093/nar/gkl315.

Sved, J. and Bird, A. (1990) 'The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model', *Proceedings of the National Academy of Sciences*, 87(12), pp. 4692–4696. Available at: https://doi.org/10.1073/pnas.87.12.4692.

Taft, A.S. *et al.* (2015) 'Identification of mammalian-adapting mutations in the polymerase complex of an avian H5N1 influenza virus', *Nature Communications*, 6(1). Available at: https://doi.org/10.1038/ncomms8491.

Takata, M.A. *et al.* (2017) 'CG dinucleotide suppression enables antiviral defence targeting non-self RNA', *Nature*, 550(7674), pp. 124–127. Available at: https://doi.org/10.1038/nature24039.

Tamura, T. *et al.* (2023) 'Virological characteristics of the SARS-CoV-2 XBB variant derived from recombination of two Omicron subvariants', *Nature Communications*, 14(1), p. 2800. Available at: https://doi.org/10.1038/s41467-023-38435-3.

Teeling, E.C. *et al.* (2005) 'A molecular phylogeny for bats illuminates biogeography and the fossil record', *Science*, 307(5709), pp. 580–584. Available at: https://doi.org/10.1126/SCIENCE.1105113.

Temmam, S. *et al.* (2022) 'Bat coronaviruses related to SARS-CoV-2 and infectious for human cells', *Nature*, 604(7905), pp. 330–336. Available at: https://doi.org/10.1038/s41586-022-04532-4.

Tenthorey, J.L., Emerman, M. and Malik, H.S. (2022) 'Evolutionary Landscapes of Host-Virus Arms Races', *Annual Review of Immunology*, 40, pp. 271–294. Available at: https://doi.org/10.1146/ANNUREV-IMMUNOL-072621-084422.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) 'CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.', *Nucleic Acids Research*, 22(22), p. 4673. Available at: https://doi.org/10.1093/NAR/22.22.4673.

Thornbrough, J.M. *et al.* (2016) 'Middle east respiratory syndrome coronavirus NS4b protein inhibits host RNase L activation', *mBio*, 7(2). Available at: https://doi.org/10.1128/mBio.00258-16.

Tuller, T. *et al.* (2010) 'Translation efficiency is determined by both codon bias and folding energy', *Proceedings of the National Academy of Sciences*, 107(8), pp. 3645–3650. Available at: https://doi.org/10.1073/PNAS.0909910107.

Tulloch, F. *et al.* (2014) 'RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies', *eLife*, 3, p. e04531. Available at: https://doi.org/10.7554/eLife.04531.

Vialle, R.A., Tamuri, A.U. and Goldman, N. (2018) 'Alignment Modulates Ancestral Sequence Reconstruction Accuracy', *Molecular Biology and Evolution*, 35(7), p. 1783. Available at: https://doi.org/10.1093/MOLBEV/MSY055.

Viana, R. *et al.* (2022) 'Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa', *Nature*, 603(7902), pp. 679–686. Available at: https://doi.org/10.1038/s41586-022-04411-y.

Victora, G.D. and Nussenzweig, M.C. (2012) 'Germinal Centers', *Annual Review of Immunology*, 30, pp. 429–457. Available at: https://doi.org/10.1146/ANNUREV-IMMUNOL-020711-075032.

Vijgen, L. *et al.* (2005) 'Complete Genomic Sequence of Human Coronavirus OC43: Molecular Clock Analysis Suggests a Relatively Recent Zoonotic Coronavirus Transmission Event', *Journal of Virology*, 79(3), pp. 1595–1604. Available at: https://doi.org/10.1128/JVI.79.3.1595-1604.2005.

V'kovski, P. *et al.* (2020) 'Coronavirus biology and replication: implications for SARS-CoV-2', *Nature Reviews Microbiology*, 19(3), pp. 155–170. Available at: https://doi.org/10.1038/s41579-020-00468-6.

Vlasova, A.N. *et al.* (2022) 'Novel Canine Coronavirus Isolated from a Hospitalized Patient With Pneumonia in East Malaysia', *Clinical Infectious Diseases*, 74(3), pp. 446–454. Available at: https://doi.org/10.1093/CID/CIAB456.

Wacharapluesadee, S. *et al.* (2021) 'Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia.', *Nature communications*, 12(1), p. 972. Available at: https://doi.org/10.1038/s41467-021-21240-1.

Wagner, R., Matrosovich, M. and Klenk, H.D. (2002) 'Functional balance between haemagglutinin and neuraminidase in influenza virus infections', *Reviews in Medical Virology*, 12(3), pp. 159–166. Available at: https://doi.org/10.1002/RMV.352.

Walls, A.C. *et al.* (2020) 'Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein', *Cell*, 181(2), pp. 281-292.e6. Available at: https://doi.org/10.1016/j.cell.2020.02.058.

Wang, H., Pipes, L. and Nielsen, R. (2021) 'Synonymous mutations and the molecular evolution of SARS-CoV-2 origins', *Virus Evolution*, 7(1). Available at: https://doi.org/10.1093/ve/veaa098.

Wang, L. *et al.* (2022) 'Potential intervariant and intravariant recombination of Delta and Omicron variants', *Journal of Medical Virology*, 94(10), pp. 4830–4838. Available at: https://doi.org/10.1002/jmv.27939.

Wang, N. *et al.* (2018) 'Serological Evidence of Bat SARS-Related Coronavirus Infection in Humans, China', *Virologica Sinica*, 33(1), pp. 104–107. Available at: https://doi.org/10.1007/s12250-018-0012-7.

Wang, Q. *et al.* (2023) 'Alarming antibody evasion properties of rising SARS-CoV-2 BQ and XBB subvariants', *Cell*, 186(2), pp. 279-286.e8. Available at: https://doi.org/10.1016/j.cell.2022.12.018.

Wang, W. *et al.* (2015) ‘Discovery, diversity and evolution of novel coronaviruses sampled from rodents in China’, *Virology*, 474, pp. 19–27. Available at: https://doi.org/10.1016/J.VIROL.2014.10.017.

Wang, Z.-D. *et al.* (2019) ‘A New Segmented Virus Associated with Human Febrile Illness in China’, *New England Journal of Medicine*, 380(22), pp. 2116–2125. Available at: https://doi.org/10.1056/NEJMoa1805068.

Wasik, B.R. and Turner, P.E. (2013) ‘On the Biological Success of Viruses’, *Annual Review of Microbiology*, 67(1), pp. 519–541. Available at: https://doi.org/10.1146/annurev-micro-090110-102833.

Weiller, G.F. (1998) ‘Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences.’, *Molecular Biology and Evolution*, 15(3), pp. 326–335. Available at: https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A025929.

Wells, H.L. *et al.* (2021) ‘The evolutionary history of ACE2 usage within the coronavirus subgenus Sarbecovirus’, *Virus Evolution*, 7(1), p. 7. Available at: https://doi.org/10.1093/VE/VEAB007.

Wertheim, J.O. *et al.* (2013) ‘A Case for the Ancient Origin of Coronaviruses’, *Journal of Virology*, 87(12), pp. 7039–7045. Available at: https://doi.org/10.1128/JVI.03273-12.

Wertheim, J.O. *et al.* (2015) ‘RELAX: Detecting Relaxed Selection in a Phylogenetic Framework’, *Molecular Biology and Evolution*, 32(3), pp. 820–832. Available at: https://doi.org/10.1093/MOLBEV/MSU400.

Wheeler, T.J. and Eddy, S.R. (2013) ‘nhmmer: DNA homology search with profile HMMs’, *Bioinformatics*, 29(19), pp. 2487–2489. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTT403.

White, S.K. *et al.* (2016) ‘Serologic evidence of exposure to influenza D virus among persons with occupational contact with cattle’, *Journal of Clinical Virology*, 81, pp. 31–33. Available at: https://doi.org/10.1016/j.jcv.2016.05.017.

Whitney, W.D. and Smith, B.E. (1911) *The Century dictionary and cyclopedia : with a new atlas of the world ; a work of general reference in all departments of knowledge in twelve volumes*. New York: Century Co. Available at: http://triggs.djvu.org/century-dictionary.com/ (Accessed: 16 June 2023).

Wickenhagen, A. *et al.* (2021) 'A prenylated dsRNA sensor protects against severe COVID-19', *Science*, 374(6567). Available at: https://doi.org/10.1126/SCIENCE.ABJ3624.

Wille, M. and Holmes, E.C. (2020) 'The Ecology and Evolution of Influenza Viruses', *Cold Spring Harbor Perspectives in Medicine*, 10(7), p. a038489. Available at: https://doi.org/10.1101/CSHPERSPECT.A038489.

Wisotsky, S.R. *et al.* (2020) 'Synonymous Site-to-Site Substitution Rate Variation Dramatically Inflates False Positive Rates of Selection Analyses: Ignore at Your Own Peril', *Molecular biology and evolution*, 37(8), pp. 2430–2439. Available at: https://doi.org/10.1093/molbev/msaa037.

Witteveldt, J., Martin-Gans, M. and Simmonds, P. (2016) 'Enhancement of the replication of hepatitis C virus replicons of genotypes 1 to 4 by manipulation of CpG and UpA dinucleotide frequencies and use of cell lines expressing SECL14L2 for antiviral resistance testing', *Antimicrobial Agents and Chemotherapy*, 60(5), pp. 2981–2992. Available at: https://doi.org/10.1128/AAC.02932-15.

Wolf, Y.I. *et al.* (2018) 'Origins and evolution of the global RNA virome', *mBio*, 9(6). Available at: https://doi.org/10.1128/MBIO.02329-18.

Wolff, G. *et al.* (2020) 'Double-Membrane Vesicles as Platforms for Viral Replication', *Trends in Microbiology*, 28(12), pp. 1022–1033. Available at: https://doi.org/10.1016/J.TIM.2020.05.009.

Woo, P.C.Y. *et al.* (2006) 'Molecular diversity of coronaviruses in bats', *Virology*, 351(1), pp. 180–187. Available at: https://doi.org/10.1016/J.VIROL.2006.02.041.

Woo, P.C.Y. *et al.* (2023) 'ICTV Virus Taxonomy Profile: Coronaviridae 2023', *Journal of General Virology*, 104(4). Available at: https://doi.org/10.1099/jgv.0.001843.

World Health Organisation (2021) *WHO-convened global study of origins of SARS-CoV-2*, *WHO*. Available at: https://www.who.int/publications/i/item/who-convened-global-study-of-origins-of-sars-cov-2-china-part (Accessed: 23 May 2021).

World Health Organisation (2022) *Hepatitis C*. Available at: https://www.who.int/news-room/fact-sheets/detail/hepatitis-c (Accessed: 13 June 2023).

World Health Organisation (2023) *Tracking SARS-CoV-2 variants*. Available at: https://www.who.int/activities/tracking-SARS-CoV-2-variants (Accessed: 2 June 2023).

World Health Organization (2022a) *Disease Outbreak News; Avian Influenza A(H3N8) - China*. Available at: https://www.who.int/emergencies/disease-outbreak-news/item/2022-DON378 (Accessed: 1 May 2023).

World Health Organization (2022b) *Mpox (monkeypox) outbreak 2022*. Available at: https://www.who.int/emergencies/situations/monkeypox-oubreak-2022 (Accessed: 26 April 2023).

World Health Organization (2023a) *Disease Outbreak News; Avian Influenza A(H3N8) - China*. Available at: https://www.who.int/emergencies/disease-outbreak-news/item/2023-DON456 (Accessed: 1 May 2023).

World Health Organization (2023b) *WHO Coronavirus (COVID-19) Dashboard*. Available at: https://covid19.who.int/ (Accessed: 2 June 2023).

Worobey, M. (2021) 'Dissecting the early COVID-19 cases in Wuhan', *Science*, 374(6572), pp. 1202–1204. Available at: https://doi.org/10.1126/SCIENCE.ABM4454.

Worobey, M. *et al.* (2022) 'The Huanan Seafood Wholesale Market in Wuhan was the early epicenter of the COVID-19 pandemic', *Science*, 377(6609), pp. 951–959. Available at: https://doi.org/10.1126/science.abp8715.

Worobey, M., Han, G.Z. and Rambaut, A. (2014) 'Genesis and pathogenesis of the 1918 pandemic H1N1 influenza a virus', *Proceedings of the National Academy of*

Sciences of the United States of America, 111(22), pp. 8107–8112. Available at: https://doi.org/10.1073/PNAS.1324197111.

Wu, F. *et al.* (2020) 'A new coronavirus associated with human respiratory disease in China', *Nature*, 579(7798), pp. 265–269. Available at: https://doi.org/10.1038/s41586-020-2008-3.

Wu, Z. *et al.* (2022) 'A comprehensive survey of bat sarbecoviruses across China in relation to the origins of SARS-CoV and SARS-CoV-2', *National Science Review*, 10(6). Available at: https://doi.org/10.1093/NSR/NWAC213.

Xia, W. *et al.* (2021) 'How one pandemic led to another: ASFV, the disruption contributing to SARS-CoV-2 emergence in Wuhan', *Preprints* [Preprint]. Available at: https://doi.org/10.20944/preprints202102.0590.v1.

Xiao, K. *et al.* (2020) 'Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins', *Nature*, 583(7815), pp. 286–289. Available at: https://doi.org/10.1038/s41586-020-2313-x.

Xiao, X. *et al.* (2021) 'Animal sales from Wuhan wet markets immediately prior to the COVID-19 pandemic', *Scientific Reports*, 11(1), pp. 1–7. Available at: https://doi.org/10.1038/s41598-021-91470-2.

Xie, R. *et al.* (2022) 'The episodic resurgence of highly pathogenic avian influenza H5 virus', *bioRxiv*, p. 2022.12.18.520670. Available at: https://doi.org/10.1101/2022.12.18.520670.

Xu, L. *et al.* (2016) *An overview of pangolin trade in China - Wildlife Trade Report from TRAFFIC*, *TRAFFIC*. Available at: https://www.traffic.org/publications/reports/pangolin-trade-in-china/ (Accessed: 23 May 2021).

Yang, L. *et al.* (2013) 'Novel SARS-like Betacoronaviruses in Bats, China, 2011 - Volume 19, Number 6—June 2013 - Emerging Infectious Diseases journal - CDC', *Emerging Infectious Diseases*, 19(6), pp. 989–991. Available at: https://doi.org/10.3201/EID1906.121648.

Yang, L. *et al.* (2014) 'MERS–Related Betacoronavirus in Vespertilio superans Bats, China', *Emerging Infectious Diseases*, 20(7), p. 1260. Available at: https://doi.org/10.3201/EID2007.140318.

Yang, P.F. *et al.* (2016) 'Characterization of Avian Influenza A (H7N9) Virus Prevalence in Humans and Poultry in Huai'an, China: Molecular Epidemiology, Phylogenetic, and Dynamics Analyses', *Biomedical and environmental sciences : BES*, 29(10), pp. 742–753. Available at: https://doi.org/10.3967/BES2016.099.

Yang, Y. *et al.* (2021) 'Characterizing Transcriptional Regulatory Sequences in Coronaviruses and Their Role in Recombination', *Molecular Biology and Evolution*, 38(4), pp. 1241–1248. Available at: https://doi.org/10.1093/MOLBEV/MSAA281.

Yang, Z. and Nielsen, R. (2000) 'Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models', *Molecular Biology and Evolution*, 17(1), pp. 32–43. Available at: https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A026236.

Yoon, S.W., Webby, R.J. and Webster, R.G. (2014) 'Evolution and ecology of influenza a viruses', *Current Topics in Microbiology and Immunology*, 385, pp. 359–375. Available at: https://doi.org/10.1007/82_2014_396.

Young, F., Rogers, S. and Robertson, D.L. (2020) 'Predicting host taxonomic information from viral genomes: A comparison of feature representations', *PLOS Computational Biology*, 16(5), p. e1007894. Available at: https://doi.org/10.1371/journal.pcbi.1007894.

Yu, G. (2020) 'Using ggtree to Visualize Data on Tree-Like Structures', *Current Protocols in Bioinformatics*, 69(1), p. e96. Available at: https://doi.org/10.1002/CPBI.96.

Yu, Y. *et al.* (2022) 'Review of human pegivirus: Prevalence, transmission, pathogenesis, and clinical implication', *Virulence*, 13(1), pp. 323–340. Available at: https://doi.org/10.1080/21505594.2022.2029328.

Yue, C. *et al.* (2023) ‘ACE2 binding and antibody evasion in enhanced transmissibility of XBB.1.5’, *The Lancet Infectious Diseases*, 23(3), pp. 278–280. Available at: https://doi.org/10.1016/S1473-3099(23)00010-5.

Zaki, A.M. *et al.* (2012) ‘Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia’, *New England Journal of Medicine*, 367(19), pp. 1814–1820. Available at: https://doi.org/10.1056/NEJMOA1211721.

Zeberg, H. and Pääbo, S. (2021) ‘A genomic region associated with protection against severe COVID-19 is inherited from Neandertals’, *Proceedings of the National Academy of Sciences*, 118(9), p. e2026309118. Available at: https://doi.org/10.1073/PNAS.2026309118.

Zhang, G. *et al.* (2022) ‘A H9N2 Human Case and Surveillance of Avian Influenza Viruses in Live Poultry Markets — Huizhou City, Guangdong Province, China, 2021’, *China CDC Weekly*, 4(1), p. 8. Available at: https://doi.org/10.46234/CCDCW2021.273.

Zhang, R. *et al.* (2013) ‘Homologous 2’,5’-phosphodiesterases from disparate RNA viruses antagonize antiviral innate immunity’, *Proceedings of the National Academy of Sciences*, 110(32), pp. 13114–13119. Available at: https://doi.org/10.1073/pnas.1306917110.

Zhang, X.M. *et al.* (1994) ‘Biological and genetic characterization of a hemagglutinating coronavirus isolated from a diarrhoeic child’, *Journal of Medical Virology*, 44(2), pp. 152–161. Available at: https://doi.org/10.1002/JMV.1890440207.

Zhao, J. *et al.* (2016) ‘2′,5′-Oligoadenylate synthetase 1(OAS1) inhibits PRRSV replication in Marc-145 cells’, *Antiviral Research*, 132, pp. 268–273. Available at: https://doi.org/10.1016/J.ANTIVIRAL.2016.07.001.

Zhao, L. *et al.* (2012) ‘Antagonism of the interferon-induced OAS-RNase L pathway by murine coronavirus ns2 protein is required for virus replication and liver pathology’, *Cell host & microbe*, 11(6), pp. 607–616. Available at: https://doi.org/10.1016/J.CHOM.2012.04.011.

Zhao, Y. *et al.* (2020) 'Porcine reproductive and respiratory syndrome virus Nsp4 cleaves ZAP to antagonize its antiviral activity', *Veterinary Microbiology*, 250, p. 108863. Available at: https://doi.org/10.1016/J.VETMIC.2020.108863.

Zhou, H. *et al.* (2020) 'A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein', *Current Biology*, 30. Available at: https://doi.org/10.1016/j.cub.2020.05.023.

Zhou, H. *et al.* (2021) 'Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses', *Cell*, 184, pp. 1–12. Available at: https://doi.org/10.1016/j.cell.2021.06.008.

Zhou, P. *et al.* (2020) 'A pneumonia outbreak associated with a new coronavirus of probable bat origin', *Nature*, 579(7798), pp. 270–273. Available at: https://doi.org/10.1038/s41586-020-2012-7.

Zhou, S. *et al.* (2021) 'A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity', *Nature Medicine*, 27(4), pp. 659–667. Available at: https://doi.org/10.1038/s41591-021-01281-1.

Zhu, H. *et al.* (2018) 'Database-integrated genome screening (DIGS): exploring genomes heuristically using sequence similarity search tools and a relational database', *bioRxiv*, p. 10.1101/246835. Available at: https://doi.org/10.1101/246835.

Zhu, W. *et al.* (2018) 'A Gene Constellation in Avian Influenza A (H7N9) Viruses May Have Facilitated the Fifth Wave Outbreak in China', *Cell Reports*, 23(3), pp. 909–917. Available at: https://doi.org/10.1016/J.CELREP.2018.03.081.

Zhu, Y. *et al.* (2011) 'Zinc-finger antiviral protein inhibits HIV-1 infection by selectively targeting multiply spliced viral mRNAs for degradation', *Proceedings of the National Academy of Sciences*, 108(38), pp. 15834–15839. Available at: https://doi.org/10.1073/PNAS.1101676108.

Zimmermann, P. *et al.* (2011) 'The Viral Nucleoprotein Determines Mx Sensitivity of Influenza A Viruses', *Journal of Virology*, 85(16), p. 8133. Available at: https://doi.org/10.1128/JVI.00712-11.

Zuckerkandl, E. and Pauling, L. (1965) ‘Evolutionary Divergence and Convergence in Proteins’, *Evolving Genes and Proteins*, pp. 97–166. Available at: https://doi.org/10.1016/B978-1-4832-2734-4.50017-6.

Appendix A. *Sarbecovirus* recombination additional material

Table A.1. Accessions, metadata and GISAID acknowledgments for all 78 sarbecovirus genomes used in this analysis.

Virus name	Host	Location	Accession	GARD analysis
Longquan_140	*Rhinolophus_monoceros*	Zhejiang	KF294457	Y
LYRa11	*Rhinolophus_affinis*	Yunnan	KF569996	Y
Rs3367	*Rhinolophus_sinicus*	Yunnan	KC881006	
WIV1	*Rhinolophus_sinicus*	Yunnan	KF367457	
279_2005	*Rhinolophus_macrotis*	Hubei	DQ648857	
Pangolin-CoV*	*Manis_javanica*	Guangdong	EPI_ISL_410721	Y
SARS-CoV-2	*Homo_sapiens*	Hubei	MN908947	Y
RmYN02**	*Rhinolophus_malayanus*	Yunnan	EPI_ISL_412977	Y
RaTG13	*Rhinolophus_affinis*	Yunnan	MN996532	Y
CoVZC45	*Rhinolophus_pusillus*	Zhejiang	MG772933	
CoVZXC21	*Rhinolophus_pusillus*	Zhejiang	MG772934	Y
P4L	*Manis_javanica*	Guangxi	MT040333	
P1E	*Manis_javanica*	Guangxi	MT040334	Y
P5L	*Manis_javanica*	Guangxi	MT040335	
P5E	*Manis_javanica*	Guangxi	MT040336	
P2V	*Manis_javanica*	Guangxi	MT072864	
BM48-31	*Rhinolophus_blasii*	Bulgaria	NC_014470	
BtKY72	*Rhinolophus_spp*	Kenya	KY352407	
RpShaanxi2011	*Rhinolophus_pusillus*	Shaanxi	JX993987	Y
F46	*Rhinolophus_pusillus*	Yunnan	KU973692	Y
Yunnan2011	*Chaerephon_plicata*	Yunnan	JX993988	Y
Rp3	*Rhinolophus_pearsoni*	Guangxi	DQ071615	
Rs672	*Rhinolophus_sinicus*	Guizhou	FJ588686	
HSZ-Cc|SC1	*Homo_sapiens*	Guangdong	AY394995	
Rs4237	*Rhinolophus_sinicus*	Yunnan	KY417147	Y
YN2013	*Rhinolophus_sinicus*	Yunnan	KJ473816	
YN2018D	*Rhinolophus_affinis*	Yunnan	MK211378	
YN2018A	*Rhinolophus_affinis*	Yunnan	MK211375	
YN2018B	*Rhinolophus_affinis*	Yunnan	MK211376	
Rs4874	*Rhinolophus_sinicus*	Yunnan	KY417150	
WIV16	*Rhinolophus_sinicus*	Yunnan	KT444582	
Rs4081	*Rhinolophus_sinicus*	Yunnan	KY417143	
RsSHC014	*Rhinolophus_sinicus*	Yunnan	KC881005	
Anlong-103	*Rhinolophus_sinicus*	Guizhou	KY770858	
Anlong-112	*Rhinolophus_sinicus*	Guizhou	KY770859	
Rs4247	*Rhinolophus_sinicus*	Yunnan	KY417148	
Rs4231	*Rhinolophus_sinicus*	Yunnan	KY417146	
Rs4255	*Rhinolophus_sinicus*	Yunnan	KY417149	
Rs7327	*Rhinolophus_sinicus*	Yunnan	KY417151	
As6526	*Aselliscus_stoliczkanus*	Yunnan	KY417142	
Rs4084	*Rhinolophus_sinicus*	Yunnan	KY417144	
Rs9401	*Rhinolophus_sinicus*	Yunnan	KY417152	
Rf4092	*Rhinolophus_ferrumequinum*	Yunnan	KY417145	Y
YN2018C	*Rhinolophus_affinis*	Yunnan	MK211377	

Table A.1. *(continued)*.

Virus name	Host	Location	Accession	GARD analysis
HKU3-9	*Rhinolophus_sinicus*	Guangdong	GQ153544	
HKU3-10	*Rhinolophus_sinicus*	Guangdong	GQ153545	
HKU3-13	*Rhinolophus_sinicus*	Guangdong	GQ153548	
HKU3-11	*Rhinolophus_sinicus*	Guangdong	GQ153546	
HKU3-5	*Rhinolophus_sinicus*	Guangdong	GQ153540	
HKU3-7	*Rhinolophus_sinicus*	Guangdong	GQ153542	
HKU3-2	*Rhinolophus_sinicus*	Guangdong	DQ084199	
HKU3-12	*Rhinolophus_sinicus*	Guangdong	GQ153547	
HKU3-4	*Rhinolophus_sinicus*	Guangdong	GQ153539	
HKU3-6	*Rhinolophus_sinicus*	Guangdong	GQ153541	
HKU3-1	*Rhinolophus_sinicus*	Guangdong	DQ022305	Y
HKU3-3	*Rhinolophus_sinicus*	Guangdong	DQ084200	
HKU3-8	*Rhinolophus_sinicus*	Guangdong	GQ153543	
YNLF_31C	*Rhinolophus_ferrumequinum*	Yunnan	KP886808	
HuB2013	*Rhinolophus_sinicus*	Hubei	KJ473814	
GX2013	*Rhinolophus_sinicus*	Guangxi	KJ473815	
SC2018	*Rhinolophus_spp*	Sichuan	MK211374	
HeB2013	*Rhinolophus_ferrumequinum*	Hebei	KJ473812	
SX2013	*Rhinolophus_ferrumequinum*	Shanxi	KJ473813	
Rf1	*Rhinolophus_ferrumequinum*	Hubei	DQ412042	
Jiyuan-84	*Rhinolophus_ferrumequinum*	Henan	KY770860	
YNLF_34C	*Rhinolophus_ferrumequinum*	Yunnan	KP886809	
JTMC15	*Rhinolophus_ferrumequinum*	Jilin	KU182964	Y
Rm1	*Rhinolophus_macrotis*	Hubei	DQ412043	
JL2012	*Rhinolophus_ferrumequinum*	Jilin	KJ473811	Y
RshSTT182* **	*Rhinolophus_shameli*	Cambodia	EPI_ISL_852604	Y
RshSTT200* **	*Rhinolophus_shameli*	Cambodia	EPI_ISL_852605	
RacCS203	*Rhinolophus_acuminatus*	Thailand	MW251308	Y
Rc-o319	*Rhinolophus_cornutus*	Japan	LC556375	Y
RsYN04	*Rhinolophus_sinicus*	Yunnan	MZ081380	Y
RmYN08	*Rhinolophus_malayanus*	Yunnan	MZ081378	
RmYN05	*Rhinolophus_malayanus*	Yunnan	MZ081376	
RpYN06	*Rhinolophus_pusillus*	Yunnan	MZ081381	Y
PrC31	*Rhinolophus_spp*	Yunnan	MW703458	Y

*GISAID acknowledgment: Yongyi Shen; Lihua Xiao; Wu Chen.

**GISAID acknowledgment: Weifeng Shi; Tao Hu; Hong Zhou; Juan Li; Xing Chen; Alice Catherine Hughes; Yuhai Bi.

***GISAID acknowledgment: Vibol Hul; Deborah Delaune; Erik A Karlsson; Ou Tey Putita; Alexandre Hassanin; Artem Baidaliuk; Fabiana Gambaro; Vuong Tan Tu; Lucy Keatts; Jonna Mazet; Christine Johnson; Philippe Buchy; Philippe Dussart; Tracey Goldstein; Etienne Simon-Loriere; Veasna Duong.

Table A.2. False and true positivity rates of the hotspot detection methods BDT and RRT based on simulated datasets.

Hotspot intensity	BDT true positive rate	BDT false positive rate	RRT true positive rate	RRT false positive rate
None	-	0.021	-	0.019
4x	0.097	0.016	0.104	0.015
8x	0.29	0.018	0.254	0.017
16x	0.59	0.014	0.576	0.014

Table A.3. Sequence length and start and end nucleotide positions of each RBP region on the whole-genome alignment and in relation to the reference SARS-CoV-2 genome.

RBP_region	al_start	al_end	length	SC2_start	SC2_end
1	1	1732	1732	1	1680
2	1733	3154	1421	1681	3093
3	3155	3761	606	3094	3649
4	3762	5193	1431	3650	4973
5	5194	8431	3237	4974	8208
6	8432	11671	3239	8209	11445
7	11672	12854	1182	11446	12622
8	12855	14633	1778	12623	14401
9	14634	16186	1552	14402	15954
10	16187	17155	968	15955	16923
11	17156	20197	3041	16924	19965
12	20198	20750	552	19966	20518
13	20751	21430	679	20519	21198
14	21431	21643	212	21199	21411
15	21644	22747	1103	21412	22460
16	22748	23692	944	22461	23396
17	23693	24452	759	23397	24144
18	24453	25151	698	24145	24843
19	25152	26638	1486	24844	26323
20	26639	28165	1526	26324	27388
21	28166	28571	405	27389	27685
22	28572	30956	2384	27686	29903

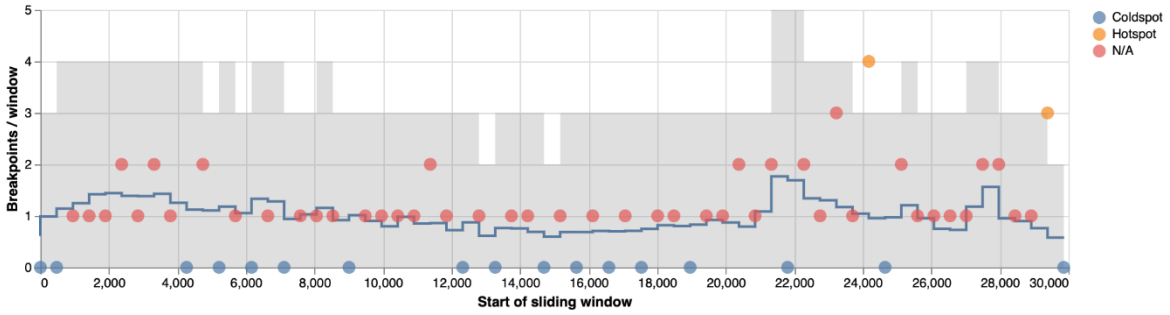


Figure A.1. Permutation test for assessing potential clustering of the recombination breakpoints inferred by GARD. The blue line represents the mean of the number of breakpoints in the window (proportional to the density of variable sites). Grey shading shows the 2.5% - 97.5% intervals of breakpoints in each window. The number of inferred breakpoints in each window is shown in dots, in red if they fall within the permutation intervals (N/A), blue if they represent recombination coldspots (in the left tail of the permutation distribution) and orange if they represent recombination hotspots (in the right tail of the distribution.

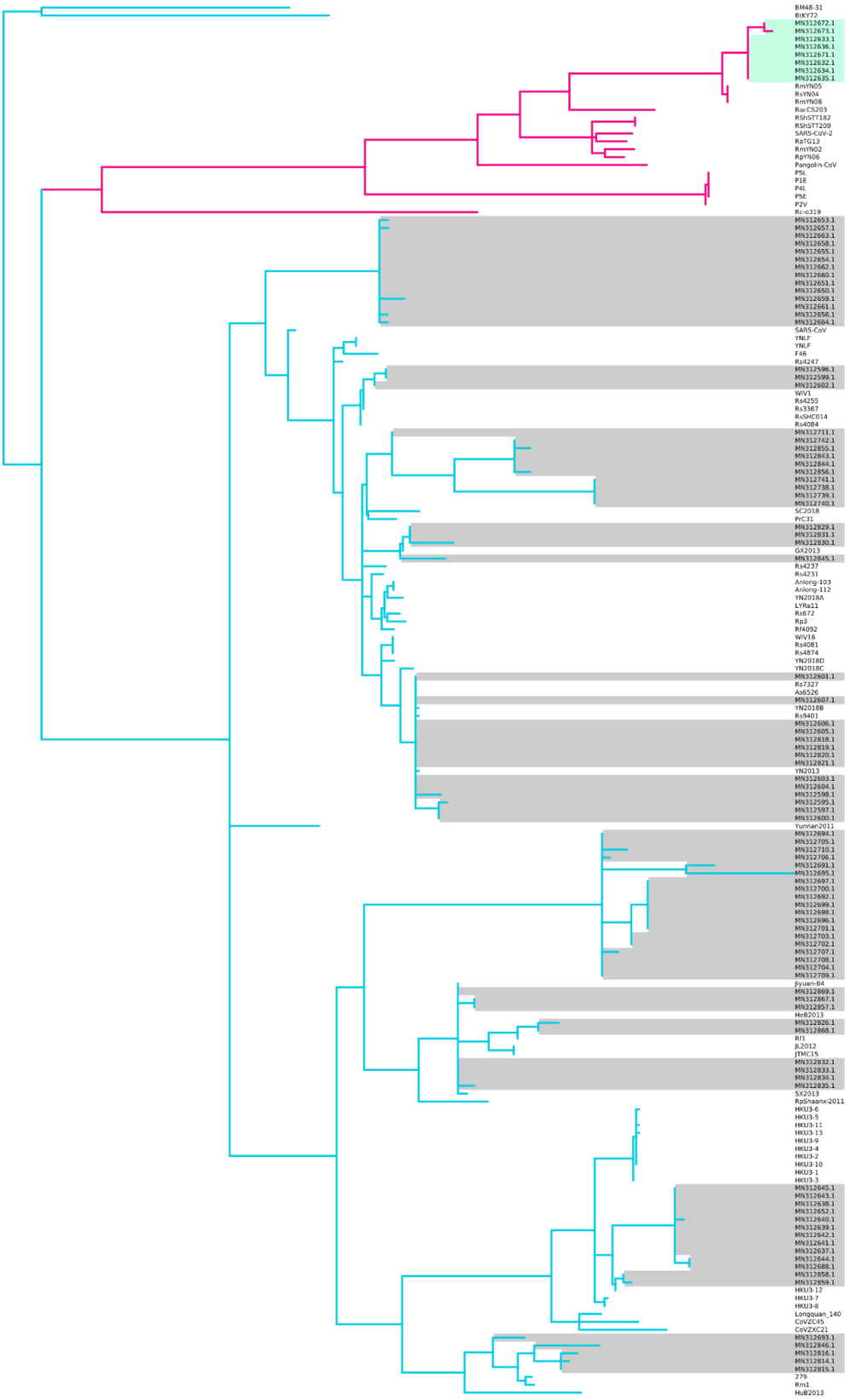


Figure A.2. Maximum likelihood phylogeny reconstructed using IQ-TREE (GTR+I+Γ4) of all 78 sarbecoviruses used throughout the analysis, including the short RdRp fragments of related sarbecoviruses reported in (Latinne *et al.*, 2020). The genomic region used for the alignment corresponds to the SARS-CoV-2 reference genome's Wuhan-Hu-1 coordinates 15280 - 16282. Nodes with bootstrap support (10,000 replicates) below 80 have been collapsed. The nCoV clade is annotated in pink and the non-nCoV clade in blue. SARS-CoV-2 and SARS-CoV are highlighted in pink and blue respectively. Viruses from Latinne et al. are highlighted in grey, apart from the 7 sequences that cluster within the nCoV clade which are highlighted in green. Out of this cluster of sequences MN312634.1 has been collected from a confirmed *R. affinis* bat species.

Appendix B. Rhinolophoidea OAS1 additional material

Text B.1 Virus infections and titrations

A549-ACE2-TMPRSS2 (‘AAT’) cells [described before in Wickenhagen et al. (2021) and Rihn et al. (2021)] were maintained in Dulbecco’s modified Eagle’s medium (DMEM) supplemented with 9% fetal calf serum (FCS) and 10 µg/ml gentamicin. The SARS-CoV-2 isolate CVR-GLA-1 was used for all SARS-CoV-2 infections under appropriate biosafety conditions and has been described previously (Rihn et al., 2021).

Overexpression of genes corresponding to the cDNA of open reading frames for: P. alecto OAS1 (NP_001277091.1), R. ferrumequinum OAS1 (XP_032953023.1) and the ancestrally reconstructed constructs shown in Figure 2 (online supplementary) were synthesised as gene blocks with flanking SfiI sites (IDT DNA) and subcloned into the lentiviral vector pLV-EF1a-IRES-Puro-SfiI-TagRFP (Wickenhagen et al. 2021). Successful expression of the gene products in AAT cells was confirmed by Western blot analysis. Briefly, cells were seeded at 106 cells/well in six-well plates the day before harvest. Cells were washed once with PBS, harvested in SDS sample buffer [12.5% glycerol, 175 mM Tris-HCl (pH 8.5), 2.5% SDS, 70 mM 2-mercaptoethanol, and 0.5% bromophenol blue] and then heated for 10 min at 70°C and sonicated. After protein separation on NuPage 4-12% Bis-Tris polyacrylamide gels and transfer onto nitrocellulose membranes, proteins were detected using OAS1 (rabbit polyclonal 14955-1-AP, Proteintech) or GAPDH (mouse monoclonal 60004-1-Ig, Proteintech) antibodies. Goat anti-rabbit IgG (Thermo Fisher Scientific, 35568) and goat anti-mouse IgG (Thermo Fisher Scientific, SA5-10176) fluorescently labelled secondary antibodies were used for detection on a LiCor Odyssey scanner.

Infection assays with SARS-CoV-2 (plaque assay and CPE induced well-clearance assays) have been described before (Rihn et al., 2021; Wickenhagen et al., 2021). For plaque assays, 12-well plates were seeded with 3x105 cells/well of AAT derivative cells overnight. The next day cells were inoculated with 10-fold logarithmic dilutions of virus stock and absorbed for 1 hour at 37C. Cells were subsequently overlaid with 0.6% Avicel in MEM and incubated for 72 hours. Followed by fixation in 8% formaldehyde and stained with a Coomassie blue solution for plaque visualization. Well-clearance assays were seeded in 96-well plates at 1.25x104 cells/well and infected the following day with titrated threefold dilutions. After 72

hours cells were fixed in 8% formaldehyde and cell monolayers were stained with Coomassie blue. The assay quantifies transmitted light (Celigo, Nexcelom) that penetrates stained cell monolayers with CPE cleared wells transmitting more light than intact monolayers of protected or uninfected cells.

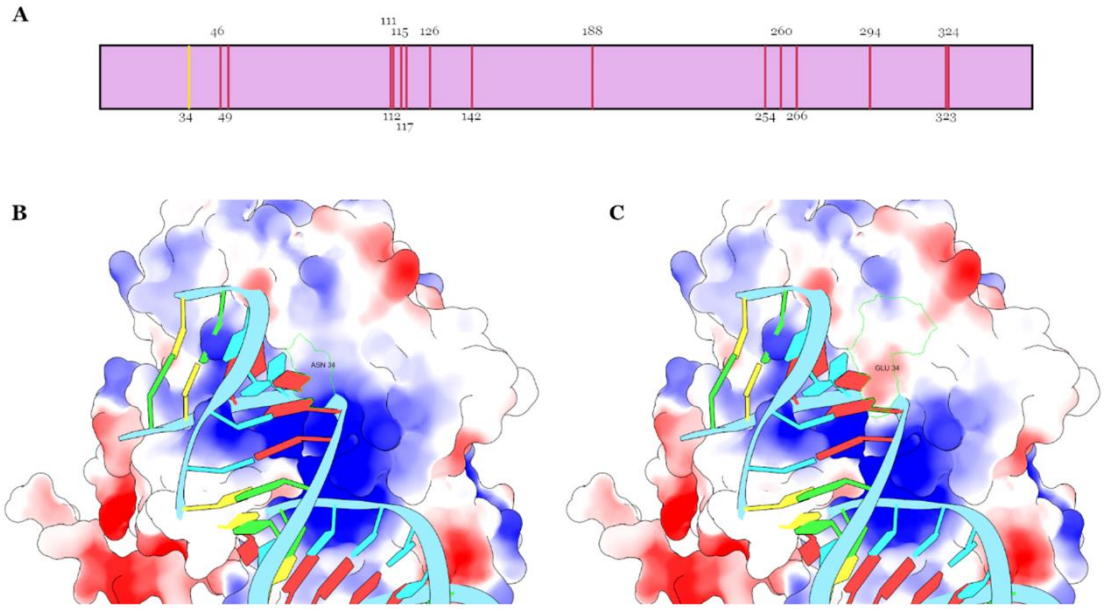


Figure B.1. Amino acid differences between the RhinoCA and RhinoCA-T70 sequence reconstructions. (A) Schematic of amino acid differences between RhinoCA and RhinoCA-T70 on the secondary sequence structure. Site 34 is highlighted in yellow. Electrostatic potential prediction calculated with ChimeraX (Pettersen *et al.*, 2021) on the RhinoCA structural protein model with an asparagine residue **(B)** and a glutamic acid residue **(C)** on site 34.

Appendix C. BTN3 gene evolution additional material

Table C.1. NCBI accessions of all BTN3 homologues presented in Chapter 4.

accession	gene	species	taxonomy
XP_005248890.1	BTN3A1	*Homo sapiens*	Catarrhini
XP_006715042.1	BTN3A2	*Homo sapiens*	Catarrhini
NP_008925.1	BTN3A3	*Homo sapiens*	Catarrhini
ENSCSAT00000005954.1	BTN3A2	*Chlorocebus sabaeus*	Catarrhini
XP_037861528.1	BTN3A3	*Chlorocebus sabaeus*	Catarrhini
ENSGGOT00000053672.1	BTN3A1	*Gorilla gorilla gorilla*	Catarrhini
ENSGGOT00000064133.1	BTN3A2	*Gorilla gorilla gorilla*	Catarrhini
ENSGGOT00000012801.3	BTN3A3	*Gorilla gorilla gorilla*	Catarrhini
XP_014991222.2	BTN3A1	*Macaca mulatta*	Catarrhini
XP_014991234.2	BTN3A2	*Macaca mulatta*	Catarrhini
XP_001091527.2	BTN3A3	*Macaca mulatta*	Catarrhini
XP_011821812.1	BTN3A1l	*Mandrillus leucophaeus*	Catarrhini
XP_011821809.1	BTN3A2	*Mandrillus leucophaeus*	Catarrhini
XP_011821814.1	BTN3A3	*Mandrillus leucophaeus*	Catarrhini
XP_030673288.1	BTN3A1l	*Nomascus leucogenys*	Catarrhini
XP_030674240.1	BTN3A2	*Nomascus leucogenys*	Catarrhini
ENSNLET00000004096.2	BTN3A3	*Nomascus leucogenys*	Catarrhini
ENSPTRT00000032930.4	BTN3A1	*Pan troglodytes*	Catarrhini
XP_016810535.1	BTN3A2	*Pan troglodytes*	Catarrhini
ENSPTRT00000032935.4	BTN3A3	*Pan troglodytes*	Catarrhini
XP_024103851.1	BTN3A1	*Pongo abelii*	Catarrhini
ENSPPYT00000018987.1	BTN3A3	*Pongo abelii*	Catarrhini
ENSRROT00000058440.1	BTN3A1	*Rhinopithecus roxellana*	Catarrhini
XP_030784330.1	BTN3A2	*Rhinopithecus roxellana*	Catarrhini
XP_030784311.1	BTN3A3	*Rhinopithecus roxellana*	Catarrhini
XP_017827069.1	BTN3A3	*Callithrix jacchus*	Platyrrhini
XP_037591290.1	BTN3A1	*Cebus imitator*	Platyrrhini
XP_037591286.1	BTN3A3	*Cebus imitator*	Platyrrhini
XP_021567778.1	BTN3A1l	*Carlito syrichta*	Tarsiiformes
XP_021567779.1	BTN3A3	*Carlito syrichta*	Tarsiiformes
NP_001723.2	BTN1A1	*Homo sapiens*	Catarrhini
NP_008980.1	BTN2A1	*Homo sapiens*	Catarrhini
NP_001184166.1	BTN2A2	*Homo sapiens*	Catarrhini
NP_001291490.1	BTNL2	*Homo sapiens*	Catarrhini
NP_932079.1	BTNL3	*Homo sapiens*	Catarrhini
NP_001035552.1	BTNL8	*Homo sapiens*	Catarrhini
XP_024310148.1	BTNL9	*Homo sapiens*	Catarrhini
XP_011542317.1	BTNL10	*Homo sapiens*	Catarrhini
NP_001017922.1	ERMAP	*Homo sapiens*	Catarrhini
NP_001350539.1	MOG	*Homo sapiens*	Catarrhini
NP_001316557.1	CD276	*Homo sapiens*	Catarrhini

Table C.1. *(continued)*.

accession	gene	species	taxonomy
XP_023480325.1	BTN3A1	*Equus caballus*	Equine
XP_014589734.1	BTN3A3	*Equus caballus*	Equine
XP_020940718.1	BTNL10	*Sus scrofa*	Porcine
XP_021133907.2	BTN1A1l	*Anas platyrhynchos*	Anatine
XP_027326214.1	BTN3A1	*Anas platyrhynchos*	Anatine
XP_027326300.1	BTN3A3	*Anas platyrhynchos*	Anatine
XP_027326231.1	BTN3A2	*Anas platyrhynchos*	Anatine
NP_001029989.1	BTN1A1	*Gallus gallus*	Galline
XP_015156030.1	BTN2A1	*Gallus gallus*	Galline
XP_004949822.2	BTN3A2	*Gallus gallus*	Galline
XP_022282885.1	BTNL10	*Canis lupus familiaris*	Canine

Appendix D. Dinucleotide representation shifts in *Flaviviridae* genomes additional material

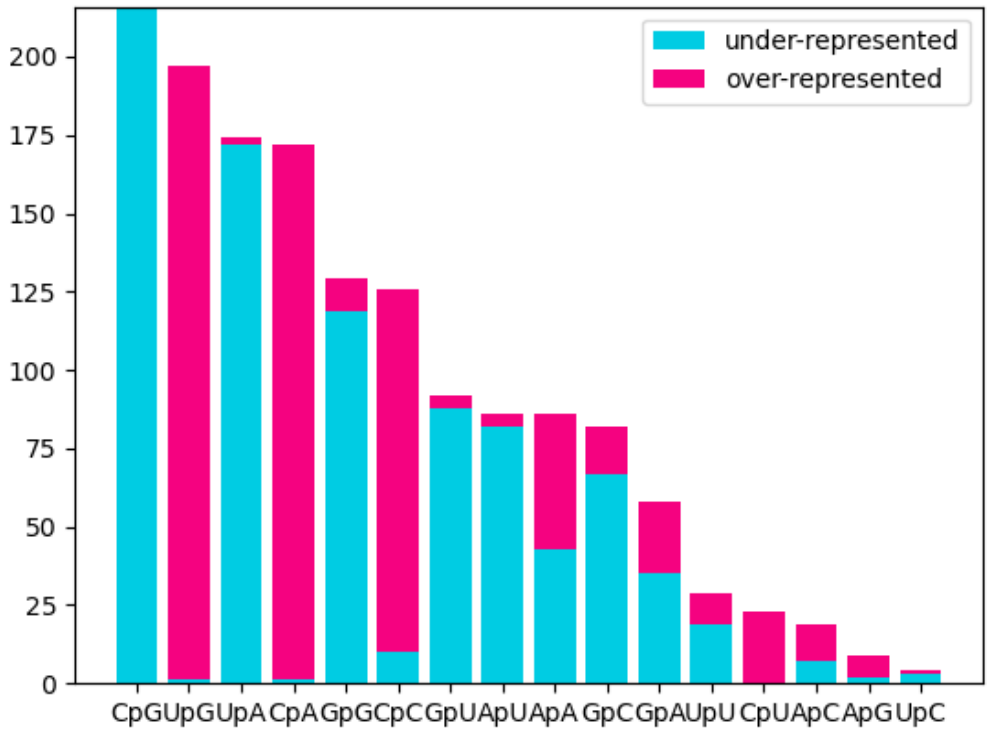


Figure D.1. Dinucleotide biases across the *Flaviviridae*. Stacked bar plot showing the number of the 350 Flaviviridae genomes with RSDUc values falling above (over-represented) or below (under-represented) of the null expectation's 95% confidence intervals across all informative frame positions for all dinucleotides.