Global genomic diversity of a major wildlife pathogen: *Ranavirus*, past and present

Christopher J. Owen

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

December, 2021

UCL Genetics Institute, University College London
Institute of Zoology, Zoological Society of London

Le biologiste passe, la grenouille reste.
The biologist passes, the frog remains.
— *Jean Rostand, 1967*

Acknowledgments

First and foremost, I extend my utmost gratitude and appreciation to all those who provided samples, either in the spirit of collaboration or sharing of published data, from which the genomic sequences scrutinised in this thesis were derived. Thank you all, most sincerely, may our work continue. I would also like to kindly thank the Natural Environment Research Council (NERC) for providing my Studentship, and making this work possible.

My supervisors, Prof. François Balloux and Prof. Trenton Garner. Trent, thank you for seeing in me which I did not myself. In the face of my incessant stubbornness, I will always be grateful for every instance you proved me wrong. François, thank you for challenging me in all the right ways. You have provided me with a set of set tools with which I hope to carve a career, but which I am sure will also prove invaluable in dealing with all manner of life's complexities.

The force of nature that is Lucy van Dorp, where would I be without you? You have not only been my analytical and computational guru, but also the most supportive of friends. Between juggling countless research projects, overseeing too many students, and the never-ending funding applications, you always found time for me and my quires, no matter how innocuous. I cannot express my gratitude enough.

A massive thank you to Stephen Price, who introduced me to ranaviruses, and in many ways passed on the baton. Beyond all else, my remaining hope is to have done justice to the fruits of your immense labour. Which of course extends to Will Leung, my friend and colleague, who not only physically generated half the data analysed in this thesis, but also taught me everything I know of virology outside a computer. Thank you mate.

It is always a great pleasure to chat to Prof. Richard Nichols, and I thank him dearly for never failing to impart a deep wisdom in our discussions. Likewise, I would like to thank Prof. Kate Jones for playing such a crucial role early in my academic training, and for our enjoyable discussions since – particularly during my Upgrade Viva(s!).

To Mislav, Cedric, Liam, Damien, and Dave C., and everyone else within our group and office at UCL who have come and gone. You all provided the fuel to

my inquisitiveness through our many chats, journal clubs, and sessions down the Huntley or IoE bar. In no small way did our time together impact the work presented in this thesis. And a special mention to the true source of our productivity – Magnifica, our beloved, if not leaky, coffee machine.

To all within the extended herp group at IoZ, who have been so supportive of my obscure genetic fascinations, always eager to listen to my over-technical talks, and provide me with valuable, grounding insights beyond my virus-centric views. Bryony, Joice, Lola, Phil, Charlotte, Gonçalo, Sarah, and Chris S., thank you.

To my bestest DTP buddies, Renée, Mike, and Lucy B. I feel extremely fortunate to have shared the many highs and lows of this journey alongside you all, every step of the way. You are some of my most favorite humans!

Last, but in no way least, my undying gratitude to Connie and my family, my parents and my brothers, for their illimitable love and support. Together you anchored me and kept me sane (and probably prevented my institutionalisation) over these last few challenging years. It feels cliché to say I could not have done this without you, but then again, I most certainly could not have.

Statement of Originality

I, Christopher Owen, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis. I further acknowledge the helpful guidance and support of my supervisors, Prof. François Balloux and Prof. Trenton Garner.

Signed:

Christopher J. Owen,
London, December 2021

Abstract

Ranavirus is a genus of large double-stranded DNA viruses (family *Iridoviridae*) that parasitise three taxonomic classes of poikilothermic vertebrates. They are important wildlife pathogens of conservation and economic concern, posing significant threat to amphibian biodiversity and aquaculture commerce. Despite substantial advances since their discovery in the 1960s, the evolutionary history of ranaviruses remains poorly characterised.

The aim of this thesis is to advance the characterisation of *Ranavirus* evolutionary dynamics to contemporary standards. A large whole-genome dataset was collated and scrutinised, combining all publicly available material with a novel collection of isolate genomes. Cutting-edge microbial genomics tools were applied to gain insight into ranavirus genetic diversity, phylogeography, and genome evolution. Delineation of the *Ranavirus* pan-genome served as a foundation to conduct phylogenetic and population genetic analyses. Where the limitations of alignment-based methodologies were met, alignment-free techniques were employed to make full use of all genomic information.

Phylogenetic reconstructions uncovered unique genetic diversity incompatible with current taxonomic demarcations amongst several lineages of *Ranavirus*. Pervasive genetic recombination was detected across the genus, and certain lineages contained a high degree of ancestral polyphyly. Recurrent patterns linked to animal trade and aquaculture were detected. Extensively polyphyletic viruses were isolated from captive animals, and population genetic analysis revealed ancestry components shared by ranaviruses isolated from farmed animals on separate continents. Finally, phylodynamic analysis suggests human-mediated translocation of FV3-like ranaviruses began more than a century before present.

The inadequacies of current *Ranavirus* taxonomy are highlighted by this work, and suggests a substantial diversity remains to be characterised. The processes by which ranaviral genetic diversity is generated appears particularly dynamic, with significant contributions made via recombination between distinct linages. Altogether, this thesis underscores the vital impact trade and captive rearing of fish and herpetofauna have had on the global spread of ranaviruses and their processes of genetic diversification. Finally, these results suggest that anthropogenic influences commenced decades earlier than previously thought, likely upon the acceleration of modern globalisation.

Impact Statement

Characterising viral evolutionary dynamics can be challenging, but contemporary genomic technologies provide unprecedented avenues for advancement. In this thesis, whole-genome sequence data of the viral genus *Ranavirus* were used to conduct a variety of genomic analyses centering on phylogenetic reconstruction with spatiotemporal components, together with population genetic approaches. The results of these analyses provide several routes for academic advancement. First, the novel genetic diversity uncovered amongst certain lineages can contribute to improved *Ranavirus* taxonomic classification, such as the reassignment of species. A further consideration in this vein that is not currently taken concerns the high rates of ranavirus recombination, which have contributed to the generation of novel lineage diversity; the results conclusively demonstrate that it is not appropriate to assess mosaic species relationships on a single phylogenetic tree, particularly with restricted marker information. Next, a benefit to current and future endeavours within the ranavirus field also stems from the genetic distance characterisations of all individual ranavirus core genes described in this thesis. A separate NERC-funded project has already utilised these genetic distances to assess which genes best reconstruct ranavirus diversity according to the core phylogeny. PCR amplification protocols targeting the identified genes have been developed for use in diagnostic typing applications. Outside of academia, the findings of this work have the greatest potential impact on considerations for policy relating to animal trade and aquacultural practices. The high rates of ranavirus recombination demonstrated in this thesis elevates our understanding of the risk associated to novel strain emergence in captive settings. Evidence suggests that recombinant ranaviruses can be more virulent than wild type strains, and captivity-associated ranaviruses frequently escape into wild host populations. As such, the insights imparted by this thesis suggest tighter restrictions/mitigations are warranted at local and international scales of animal trade and aquaculture commerce.

Contents

List of Figures

Chapter Figures

Figure 4.3. Phylogenetically inferred dates of FV3-like ranavirus introductions and phylogeographic emergences in the UK.

Figure 4.4. Ancestral population structure of 58 FV3-like ranavirus isolates.

Appended Figures

Figure S1. *Ranavirus* pan-genome presence/absence of orthologous ORFs clustered at 80% amino acid homology.

Figure S2. Isolate gene histories of ALRV clades.

Figure S3. Proportion of polyphyletic genes associated with *Ranavirus* isolates from captive and wild sources.

Figure S4. Root-to-tip regression of FV3-like isolates with known sample dates.

Figure S5. Comparison of overlapping HPD intervals between phylodynamic reconstructions of FV3-like ranaviruses.

Figure S6. Phylogeographic mapping of FV3-like ranaviruses from the USA and the UK.

Figure S7. A*DMIXTURE* cross-validation error scores giving support to the optimal value of *K* ancestral populations for FV3-like ranavirus isolates.

List of Tables

Chapter Tables

Appended Tables

1

Introduction to Thesis

Despite nearly six decades of research, the genus *Ranavirus* (family *Iridoviridae*) remains relatively poorly characterised in many aspects of their biology, at least compared to some of their viral relatives. They belong to the viral phylum *Nucleocytoviricota* (realm *Varidnaviria*), which are a diverse but monophyletic group of large double-stranded DNA (dsDNA) viruses notably encompassing poxviruses (*Poxviridae*), African swine fever viruses (*Asfarviridae*), and the more recently discovered giant viruses (*Mimiviridae*) (Rodrigues et al., 2021). Ranaviruses changed our view of the *Iridoviridae*, which were thought to be largely non-pathogenic parasites of fish and invertebrates (Weissenberg, 1965; Xeros, 1954). In fact, ranaviruses themselves were initially considered similarly benign for several decades following their accidental discovery by Allen Granoff in the mid-1960s (Granoff et al., 1966).

1.1 *Ranavirus*, a Historical View

Granoff and colleagues were attempting to establish cell lines with the hope to support the growth of *Ranid herpesvirus 1* (RaHV-1; also known as *Lucké tumor herpesvirus*) for the study of oncovirus biology, as at the time it was an unproven but suspected aetiological agent of renal adenocarcinoma in the northern leopard frog (*Rana pipeins*). To their surprise, the leopard frog renal tumour cells they isolated underwent spontaneous cytopathic effect, despite no obvious sign of a viral infection, except possibly the adenocarcinoma itself (Granoff et al., 1965). Subsequent investigation of the virus showed its growth was uncharacteristic of RaHV-1, and later turned out not to be oncogenic (Tweedell & Granoff, 1968). In lieu of further characterisation, Granoff and colleagues gave their novel viral cultures placeholder names. Though, the name of their third isolate was destined to stay, and little did they know at the time that their *Frog virus 3* (FV3) was a new *Iridovirus* species, and the first of its genus.

In the 20 years that followed its discovery, Granoff and colleagues laid an impressive groundwork in the molecular characterisation of *Frog virus 3*. The newly discovered virus was used a model system to gain insights into a number of aspects *Iridovirus* biology, but also that of their eukaryotic hosts (Chinchar, 2002). Fundamental discoveries during these years included the linear double-stranded DNA structure of the ranavirus genome, with its terminal redundancy and circularly permuted genomic organisation (Goorha & Murti, 1982). The FV3 genome is also heavy methylenated (Willis & Granoff, 1980), which has since been discovered as common to ranaviruses, but lacking in some of the fish-infecting relatives (Song et al., 2004). Given the role of methylation in detection of foreign DNA (Jeudy et al., 2020; Murray, 2002), this points towards important differences in the antiviral immune response of iridovirid host taxa. Other significant characteristics related to the structure and function of the icosahedral capsid, such as its uncoating (Gendrault et al., 1981) and the fact that it remains infective whether enveloped or naked (Braunwald et al., 1979). The progression of key proteins expressed upon infection was also elucidated (Willis et al., 1985), which contributed to the understanding that replication takes place in both nuclear and cytoplasmic locations of the host cell – one of the defining features of the *Nucleocytoviricota* (Koonin & Yutin, 2012). However, following the initial flurry of productive molecular research, enthusiasm for the new system eventually waned during the 1980s. Given the apparent benign effect on its amphibian host (Tweedell & Granoff, 1968), let alone the seeming lack of any medical or economic impact, the FV3 ranavirus fell into relative obscurity amongst the virological community (Gray & Chinchar, 2015).

A second era of ranavirus research was ushered in upon severe outbreaks of mystery iridovirids amongst farmed fish. Langdon et al. (1986) first reported a virial agent isolated following a large mortality event of farmed juvenile redfin perch (*Perca fluviatilis*) in south-eastern Australia. Ascertained through electron microscopy, they deemed the culprit to be *Iridovirus*-like based on the size and shape of the virion capsids (~150 nm icosahedron) which had the presence of an internal spherical layer of nucleoproteins characteristic of the family. However, the pathology caused by the virus did not progress exactly like *lymphocystis disease virus* (LCDV), a then well-known group of iridovirid fish pathogens and potential agent (Weissenberg, 1965; Wolf, 1988). As such, based on the host pathology, the new virus was named *Epizootic haematopoietic necrosis virus* (EHNV; Langdon & Humphrey 1987), which would later become the second characterised species of the genus *Ranavirus*.

Although not known at the time, an earlier disease outbreak of a fish-associated ranavirus had occurred amongst European farmed cod (*Gadus morhua*) in Denmark, which was erroneously thought to be caused by a rhabdovirus (Jensen et al., 1979). Future genetic characterisation of the only isolate secured from the outbreak retrospectively recharacterised it as a ranavirus (Ariel et al., 2010). Nevertheless, over the course of a decade, several reports emerged of mass mortality events amongst farmed fish – both freshwater and marine – caused by apparently similar and novel iridoviruses. The timing and global distribution of these reports were particularly striking, ranging across Australia (Langdon et al., 1986), Germany (Ahne et al., 1989), France (Pozet et al., 1992), Denmark (Jensen et al., 1979), Finland (Tapiovaara et al., 1998), the USA (Plumb et al., 1996), Japan (Inouye et al., 1992), and South East Asia (Chua et al., 1994). The relationship between these viruses, and particularly to the obscure frog pathogen isolated approximately two decades before, was unclear at the time. Though, what was clear, was that a novel group of iridoviruses were wreaking havoc on global aquaculture commerce, commonly causing between 30% and 100% mortality in affected fish stocks.

During the period of the iridoviral outbreaks amongst farmed fish, the late 1980s and 1990s concurrently saw an explosion of global ranavirus detection associated with amphibians and reptiles. The second-ever report of an amphibian-associated iridovirus came from a die-off of captive ornate burrowing frogs (*Limnodynastes ornatus*) in north-eastern Australia (Speare & Smith, 1992). The isolate became known as Bohle iridovirus (BIV), named after the town it was discovered in. Several reports from around the globe followed the detection of BIV (see later), but amongst them, two were particularly distinct and important, as they described the first large-scale mortality amongst *wild* amphibian populations. The first was the detection of an iridovirid in the UK, following bouts of unusual common frog (*Rana temporaria*) mortality across the country (Cunningham et al., 1996). The second described periodic mass mortality events of an endangered tiger salamander subspecies (*Ambystoma tigrinum stebbinsi*) that had been occurring since the mid 1980s in south-western USA (Jancovich et al., 1997). The latter agent became known as the species *Ambystoma tigrinum virus* (ATV), and was associated with mortality events of tiger salamander populations up to their most northern geographic range in Canada (Bollinger et al., 1999). Despite minor differences in disease presentation between animals, such as presence or absence of skin lesions, the mass mortality events of captive and wild animals were all attributed to severe

systemic haemorrhagic syndromes, which became the defining feature of amphibian-associated ranavirosis. Together with the economic impacts being incurred in farmed fish, these virulent new disease emergences in amphibians triggered a reinvigorated interest in iridovirid research.

Two landmark investigations were soon conducted following the global emergences of fish and amphibian iridoviruses, which were the first large-scale studies of *Ranavirus* relations. Both included many of the piscine viruses mentioned above, together with a collection of the newly described amphibian-associated iridoviruses (except for ATV). First, Mao et al. (1997) focused primarily on typing a panel of nine isolates by developing a protocol that used the conserved Major Capsid Protein (MCP) as a taxonomic marker. They found that all the iridoviruses they characterised were much more closely related to FV3 than to the well-known LCDV fish-associated viruses, concluding that they indeed belonged to *Ranavirus*, a sister genus to *Lymphocystivirus* within the *Iridoviridae*. Then, Hyatt et al. (2000) employed a series of molecular characterisations including restriction fragment length polymorphism (RFLP), DNA hybridisation and additional sequencing of the MCP gene. Their results not only supported the findings of Mao et al. (1997), but further provided several particularly important insights. They delineated six distinctive groups, describing candidate *Ranavirus* species, including three that were fish-associated and three associated with amphibians and reptiles. The next significant finding was that the FV3 group contained both amphibian and reptile hosts, and importantly, the isolates from the UK. This provided the first glimpse that FV3 could infect more than one taxonomic class of host animal, but also suggested that the lineage had been introduced into the UK from North America. Finally, both these studies provided the first evidence that the amphibian/reptile viruses were the derived state, as the fish-associated isolates were ancestrally positioned in their phylogenetic reconstructions of the MCP gene. An image was beginning to emerge that the expanding genus *Ranavirus* was unique amongst the *Iridoviridae* in their capability of infecting a broad range of host species, spanning three classes of poikilothermic vertebrates.

In addition to world-wide die-offs occurring in farmed fish, the concurrent emergences in amphibians began to reveal a trend of ranaviral disease associated with captive animals. At around the same time that the novel ranaviruses in the UK and the USA were reported, disease outbreaks were taking place amongst farmed soft-shelled turtles (*Trionyx sinensis*) and pig frogs (*Rana grylio*) at aquaculture facilities in China (Chen et al., 1999; Zhang et al., 2001). The soft-

shelled turtle iridovirus (STIV) and *Rana grylio* virus (RGV) isolates, as they became known, would later be confirmed as FV3-like (Huang et al., 2009; Lei et al., 2012). Not long after these Asian FV3 emergences, yet another ranavirus was implicated in the mass mortality of farmed edible tiger frogs (*Hoplobatrachus rugulosus*) in southern China (Weng et al., 2002), and amongst captive Asian bullfrogs in Thailand (*Hoplobatrachus tigerinus;* Kanchanakhan 1998). The aetiological agent was named Tiger frog virus (TFV), and in the same year of its reported outbreak in China, Chan and colleagues sequenced its full genome (He et al., 2002). Incidentally, TFV was the first *Ranavirus* to have its whole genome characterised, and rudimentary phylogenetic analysis suggested it was closely related to FV3. It was speculated that these Asian ranaviruses could have also been introduced, given the association of FV3 to North America, and its precedence for translocation into the UK (Hyatt et al., 2000; Zhang et al., 2001).

With the firm establishment of multiple *Ranavirus* variants, each linked to mass morbidity and mortality of poikilothermic vertebrates, the view of iridoviruses as benign pathogens changed. This shift, together with the increasing incidence in novel host types and widening geographic distributions, prompted *Ranavirus* to be designated a group of emerging pathogens (Daszak et al., 1999, 2000). Mounting concern was shared amongst stake holders in aquaculture economies, conservationists and members of the general public alike (Williams et al., 2005), given the apparent links to farmed animals heavily implicated in global trade and the continued reports and characterisations of die-offs amongst amphibian communities in the USA (Jancovich et al., 2003; Storfer et al., 2007) and the UK (Cunningham et al., 2007; Daszak et al., 2001). Following in 2008, *Ranavirus* was listed by the World Organisation of Animal health (OIE) as notifiable pathogens in the Aquatic Animal Health Code (Schloegel et al., 2010). The listing is upheld to this day, currently containing two separate mandates legally obligating member states to report disease detected from both infection by any *Ranavirus* species (Aquatic Code Chapter 2.1.2), and infection specifically by the EHNV species (Aquatic Code Chapter 2.3.1; OIE 2019).

The importance of global States to draw mitigative policy against *Ranavirus* spread became clear almost immediately following the decision of the OIE to list the genus as notifiable pathogens. Amidst the ongoing epizootic outbreaks in common frogs of the UK, Teacher et al. (2010) provided the first evidence of wild amphibian populations experiencing persistent ranavirus-mediated decline. They did so by measuring frog populations in ponds with known ranavirus presence

and absence, respectively, over a decade-long period. Ponds with ranavirus presence demonstrated a median population decline of 81% compared to ponds without, and the change was most significantly associated with disease status as opposed to habitat-related factors. Then, a mainland European amphibian-associated ranavirus emerged for the first time on the Iberian Peninsula in northern Spain, which expressed a phenotype hitherto unseen. Balseiro et al. (2009) reported what would become a new ranavirus species, naming it *Common midwife toad virus* (CMTV), as it was the agent of a mass mortality event amongst *Alytes obstetricans* larva in the Picos de Europa (PNPE) National Park.

After five years of monitoring amphibian populations in the PNPE following the detection of CMTV, Price et al. (2014) reported a worrying trend. At several sites, not only were common midwife toad populations suffering, but so to were populations of common toads (*Bufo bufo*) and alpine newts (*Mesotriton alpestris*), which together showed persistent declines in ponds positive for CMTV presence over the monitoring period. What is more, disease was detected in the entire amphibian species assemblage, as well as a natricine water snake (*Natrix maura*), demonstrating the extreme host breadth and virulence of this novel virus. During this period and the years that followed, CMTV was detected and implicated in die-offs of wild amphibians in neighbouring Portugal (Rosa et al., 2017), the Netherlands (Kik et al., 2011), Belgium (Sharifian-Fard et al., 2011), and France (Miaud et al., 2016). Unfortunately, CMTV-like viruses were quickly detected in cultured animals far from their region of emergence in Europe, including farmed giant Chinese salamanders *(Andrias davidianus*) in China (Chen et al., 2013) and American bullfrogs (*Lithobates catesbeianus*) in the USA (Majji et al., 2006).

These examples demonstrate a serious and realised risk of highly pathogenic viruses – capable of causing population extinctions – passing into naïve host populations through animal trade. The risk becomes particularly concerning considering that amphibian disease is a contributing factor to approximately 50% of all amphibian species across their three extant taxonomic orders being threatened with extinction (González-del-Pliego et al., 2019). The listing of ranaviruses as notifiable pathogens by OIE represented a significant policy decision aimed to mitigate trade-related risks. However, it stood merely as the first small step in the direction of the legislation and collective action required to prevent further ranavirus spread to both wild and captive host animal populations across the globe.

1.2 Present *Ranavirus* Diversity

Over approximately 30 years, *Ranavirus* rose from a viral genus of relative benign obscurity to a diverse group of important pathogens carrying serious threats to amphibian biodiversity and aquaculture commerce. New ranaviruses are still being detected, the most recent being the marine *European North Atlantic ranavirus* (ENARV), a clade of ranaviruses found in wild and farmed lumpfish (*Cyclopterus lumpus*) off the shores of the Republic of Ireland, Scotland, the Faroe Isles and Iceland (Stagg et al., 2020). The ENARV marks the seventh formal species of *Ranavirus* currently classified by the ICTV (https://talk.ictvonline.org/ictv-reports/ictv_online_report/dsdna-viruses/w/iridoviridae/616/genus-ranavirus, last accessed December 2021; Chinchar et al., 2017a). It joins the exclusively fish-associated ranaviruses, which are considered ancestral, and are globally distributed across marine and freshwater habitats (Price et al., 2017a). These comprise the Southeast Asian *Singapore grouper iridovirus* (SGIV; Qin et al., 2003), the North American *Santee-Cooper ranavirus* (often abbreviated to LMBV for the exemplar isolate Largemouth bass virus; Grizzle et al., 2002), and the Australian EHNV (Langdon et al., 1986), which under formal ICTV classification also comprises the European catfish/sheatfish isolates (ECV/ESV; Ariel et al., 2010; Mavian et al., 2012a). The remaining three species are those predominantly associated with amphibian and reptile hosts, comprising the ancient North American ATV (Jancovich et al., 1997), the European CMTV (Balseiro et al., 2009), and FV3 (Granoff et al., 1965) which is distributed across North America, Europe, and Asia. These species, together with many named isolates awaiting formal classification, exist across a global distribution on all continents except Antarctica.

The seven species of *Ranavirus* contain an exceptionally broad host range. In the last extensive review of its kind, Duffus et al. (2015) concluded that ranaviruses have been found to infect at least 175 species across 52 families of poikilothermic vertebrates. However, this number undoubtably needs reassessing.

First, in the last two years alone, ranaviruses have been detected in several novel wild host species and/or hitherto unreported geographic regions. Examples include red-eared slider turtles (*Trachemys scripta elegans*) in Poland (Borzym et al., 2020), Dybowski's brown frogs (*Rana dybowskii*) in South Korea (Park et al., 2021), and several African species of *Hoplobatrachus* and *Ptychadena* amphibians in Chad, together with a species of *Pelomedus* turtle (Box et al., 2021). These accounts have both increased the number of host species ranaviruses are known to infect, whilst simultaneously extending our understanding of their known spatial distributions

– further increasing the likelihood that ranaviruses are associated with many more host species than currently recognised.

Secondly, it is becoming increasingly clear that the diversity of ranaviruses associated with fish and reptile hosts is far greater than currently understood. This may in part be due to missed detections, as species of these host classes do not always seem to suffer as severe pathologies as amphibians. As an example, the recently discovered ENARV species in marine Atlantic lumpfish were only detected through regulatory virus screening of visibly healthy fish stock (Stagg et al., 2020). However, the CMTV and FV3 amphibian-associated ranaviruses (AARV) are capable of infecting fish, both with evidence through experimental challenge (Brenes et al., 2014; Moody & Owens, 1994) and infections detected *in situ* (Ariel & Owens, 1997; Holopainen et al., 2016; Mao et al., 1999; Sriwanayos et al., 2020; Waltzek et al., 2014). It has been suggested that piscine and reptilian hosts of AARVs likely represent dead-end spillovers (Brenes et al., 2014; Price et al., 2017a), but several lines of evidence suggest a high prevalence of FV3-like viruses is maintained in non-amphibian host populations with subclinical infection. Examples include pumpkinseed fish (*Lepomis gibbosus*) in Portugal (Rosa et al., IN REVIEW), and multiple chelonian and squamate reptile species in Central America (Wynne, 2019), North America (Carstairs et al., 2020; McKenzie et al., 2019), Europe (Borzym et al., 2020), Australia (Maclaine et al., 2020), and Africa (Box et al., 2021), reiterating a potentially vast diversity of host associations awaiting detection.

As Gray & Chinchar (2015) noted in their editorial introduction to the book *Ranaviruses*: "We are just beginning to scratch the surface in understanding the complex interactions of these pathogens and their diverse hosts." Indeed, despite significant advances, there are still major outstanding questions surrounding ranaviruses that have likely been left unanswered largely due to the fact that they are not pathogens of mammals (Gray & Chinchar 2015). These knowledge gaps particularly relate to: i) the evolutionary rates of ranavirus, the extent of their diversity and the relationships they hold; ii) why their geographic distributions are patchy and the disease outbreaks they cause are sporadic; and iii) what genetic and environmental factors drive the differences observed in lineage pathogenicity and host range.

1.3 Thesis Aims

The overarching aim of this thesis is to quantitatively characterise *Ranavirus* diversity and evolutionary dynamics using cutting-edge tools. Quantifying *Ranavirus* biological diversity can broadly be broken down into genomic and phenotypic assessments. The latter requires data gathered through a variety of *in vitro* and *in vivo* frameworks, such as molecular assays conducted in cell culture or challenge experiments in host animals. Methods such as these are the only means to untangle host-pathogen interactions relating to immunological response of the host and associated pathological progressions. Virus challenge experiments in particular have been indispensable to our understanding of different ranavirus phenotypes, such as the restricted host range of the ATV lineage (Jancovich et al., 2001) or the fact that BIV was capable of infecting fish before it was realised *in situ* (Ariel & Owens, 1997; Moody & Owens, 1994). However, molecular techniques including electron microscopy, tissue histopathology and immunostaining, antigenic capture assays (e.g., ELISA; Ariel et al., 2010), restriction enzyme digestions (e.g., RFLP; Hyatt et al., 2000), and PCR and gene sequencing (e.g., MCP analysis; Mao et al., 1997) have provided the fundamental bulk to our understanding of ranavirus biological diversity.

Quantitative assessment of *Ranavirus* genomic diversity, on the other hand, has contributed less to our understanding of the genus. This is perhaps an unfair assertion to make, largely because genomic techniques exist in a state of comparative infancy, at least in terms of broad application, to many of the methodologies listed above. Nevertheless, whole-genome sequencing and its associated analytical frameworks have revolutionised all realms of evolutionary biology, and particularly to pathogen systems. The so-called Next Generation Sequencing revolution (Koboldt et al., 2013) has facilitated profound insights into evolutionary process not otherwise visible, such as deciphering cryptic pathogen diversity (Abbot et al., 2007; de Carvalho et al., 2020; Sepúlveda et al., 2017). Though, arguably the greatest asset to whole-genome methodologies is the unprecedented potential to uncover pathogen population demographic processes in space and time, from elucidating transmission chains (Biek et al., 2012; Kohl et al., 2014; Roach et al., 2015; van Dorp et al., 2019) to dating evolutionary events such as host shifts (Nadin-Davis et al., 2017; Weinert et al., 2012) and geographical introductions (Kamath et al., 2016; O'Hanlon et al., 2018). This latter framework is known as phylodynamics (Grenfell et al., 2004), and a core component to it is the accurate estimation of evolutionary rates through high-resolution phylogenetics. Large strides have been made in recent years in the application of such methods

to the ranavirus system (Claytor et al., 2017; Epstein & Storfer, 2016; Price, 2015; Stöhr et al., 2015; Vilaça et al., 2019). However, these studies have often lacked genomic resolution and have been based on a limited sample.

In this thesis, I focus on *Ranavirus* genomics as opposed to providing novel insights into the phenotypic diversity of the genus. I introduce the largest whole-genome dataset to date, which combines a novel collection of sequences and all those that were publicly available at the time of analysis. The dataset encompasses all currently described lineages of *Ranavirus* across their global distribution, which I scrutinise over three data chapters.

In Chapter 2, I initially delineate the pan-genome content of *Ranavirus*. I provide a substantial update to the functional annotation of ranavirus genes, and demarcate the core genome complement across all included isolates. I use this genomic backbone to reconstruct the global phylogeny of *Ranavirus*, which I compare to a reconstruction using an alignment-free methodology that makes use of all genomic information. In doing so, I provide the most detailed characterisation of *Ranavirus* genetic diversity to date.

Chapter 3 describes features of *Ranavirus* genome evolution, particularly relating to genetic recombination. Using the taxonomic groups identified in Chapter 2, I delineate lineage-specific core genome complements with a focus on the amphibian-associated ranaviruses (AARV). With these sets, I employ individual and concatenated gene assessments, including linkage disequilibrium decay profiles, phylogenetic incongruency recombination metrics and *dN*/*dS* ratio selection analyses. I identify regions of the genome that frequently recombine, and make rudimentary comparisons (in lieu of focused genomic sampling) of rates between lineages, whilst identifying genes that have undergone positive selection at key moments in the evolutionary history of the genus.

Chapter 4 focuses on the FV3-like ranaviruses, which is the most intensely sampled lineage in the dataset. Due to the high sampling effort, the group contains significant temporal signal in their rate of mutation accumulation. This permitted the employment of coalescent theory to perform a phylodynamic reconstruction of the FV3-like ranaviruses. Through this analysis, I provide an insight into the origins of the lineage, and a quantification of the frequency and timing of FV3 invasions into the UK.

Finally, in Chapter 5, I provide a general discussion on the major findings from the collective body of this work. I discuss the current inadequacies of *Ranavirus* taxonomy, and how recombination and genetic mosaicism has played a significant role in generating diversity amongst the genus. I also emphasise the pivotal role human-mediated translocation through animal trade has played both in facilitating geographic expansion of ranaviruses, but also in providing opportunity for the unnatural union of distinct genetic lineages, and the future implications of this.

2

Reconstructing the Global Genomic Diversity of Amphibian-Like Ranaviruses

2.1 Abstract

Quantifying viral genetic diversity can be a challenging but important task. Vast effective population sizes and fast evolutionary rates allow viruses to acquire large amounts of genetic diversity, often rendering the task of elucidating virial relationships with conventional molecular techniques difficult. Accurate viral classification is important for investigations into evolutionary dynamics, diagnostic typing, and drawing balanced policy considerations. *Ranavirus* is a viral genus of ecologically and commercially important wildlife pathogens with poorly resolved global diversity. In this chapter, I collated a large whole-genome dataset of 170 novel and publicly available isolates sampled across five continents to reconstruct the global phylogeny of *Ranavirus*. By delineating the pan-genome of the genus, I identified 217 orthologous open reading frames, with 49 core genes present in all isolates. I provide a substantial update in functional annotation, from ~30% to ~70% of the core genome (~50% of the pan-genome). Using the genetic backbone of the genus, I constructed a high-resolution Maximum Likelihood phylogeny, which I compared to alignment-free pairwise genetic distance estimates from complete genome assembles. Genetic diversity incompatible with current ICTV taxonomic classifications was uncovered, largely relating to biases introduced when performing phylogenetic reconstruction using restricted portions of genomic information. These results demonstrate revisions to ranavirus taxonomy are warranted, and I provide suggestions towards an updated framework of lineage classification. Finally, the structure of the global phylogeny suggests that a large extent of ranavirus genetic diversity awaits detection, highlighting a need for greater genomic surveillance efforts.

2.2 Introduction

Ranavirus is a genus of large double-stranded DNA viruses that parasitise an extraordinarily broad host range across three taxonomic classes of poikilothermic vertebrates (Chinchar, 2002). Due to the often-severe conservation and economic consequences incurred by ranavirosis disease outbreaks, ranaviruses have garnered a considerable and sustained research effort over the last five decades (Campbell et al., 2020; Wirth et al., 2021). This considerable research attention has yielded significant insights into their biology, ranging from epidemiological to molecular characterisations (Gray & Chinchar, 2015). Despite those efforts, the global genetic diversity of *Ranavirus* remains poorly characterised.

The genus belongs to the family *Iridoviridae* (phylum *Nucleocytoviricota*), which is divided into the two subfamilies of the *Betairidovirinae* that infect arthropods (including insects and crustaceans), and the *Alphairidovirinae* which infect ectothermic vertebrates (Chinchar et al., 2017b). The latter contains three genera that are all associated with boney fish hosts (*Lymphocystivirus*, *Megalocytivirus*, and *Ranavirus*), except for *Ranavirus*, which additionally are notable pathogens of amphibians and reptiles. Of the iridovird genera, *Ranavirus* is the best characterised largely because all member species currently recognised by the International Committee on the Taxonomy of Viruses (ICTV) are notifiable pathogens to the World Organisation of Animal health (OIE). This is due to: i) their emergent infectious disease status (Daszak et al., 1999); ii) the hosts they infect that are heavily implicated in global animal trade (Gratwicke et al., 2010; Picco & Collins, 2008; Schloegel et al., 2010); and iii) the often dire consequences of disease spread through amphibian communities and animal stocks in commercial aquaculture (Ariel et al., 2010; Langdon et al., 1986; Price et al., 2014; Zhang et al., 2001).

There are currently seven formally recognised species of *Ranavirus* classified by the ICTV (Chinchar et al., 2017a; Walker et al., 2020). However, the taxonomy of the genus has undergone many revisions over the years. Some isolates have proven difficult to classify taxonomically, having been described as one species, then reassigned to others (Chinchar et al., 2009; International Committee on Taxonomy of Viruses, 2012). These include the European catfish virus (ECV)-like isolates (Ariel et al., 2010; Mavian et al., 2012a), which are now characterised as belonging to the *Epizootic hematopoietic necrosis virus* (EHNV) species, as well as Bohle iridovirus (BIV)-like isolates (de Voe et al., 2004; Hick et al., 2016; Marschang

et al., 2005) which are now assigned to the *Frog virus 3* (FV3) species (Walker et al., 2020).

There are also named isolates recognised as tentative species by the ICTV that await formal classification. Currently these comprise three fish-associated viruses, including the closely related European marine Ranavirus maximus (RMax) and Cod iridovirus (CoIV; Ariel et al., 2016), and the distinctive Short-fined eel virus (SERV). The latter is a highly divergent isolate with a complex history involving importation of short-finned eels (*Anguilla australis*) to Italy, apparently from New Zealand (Subramaniam et al., 2016). Furthermore, there are several idiosyncratic viruses in terms of their isolation source (geography and host) and virulence profiles that are currently tentatively assigned to the FV3 species, which may need reconsideration. These include, amongst others, the aforementioned Australian and captive-sourced BIV-like isolates, the predominantly Asian Tiger frog virus (TFV; He et al., 2002; Sriwanayos et al., 2020; Box et al., 2021) isolates, and the closely related *Rana grylio* virus (RGV; Lei et al., 2012) and Soft-shelled turtle iridovirus (STIV; Huang et al., 2009) isolates, both associated with captive farmed animals in China.

Ranavirus systematics and taxonomy may seem like a purely academic pursuit. However, there are several non-mutually exclusive factors that make the process of accurate and balanced classifications of ranavirus genetic diversity an important practical task. First, despite the broad host range and generalist nature of different amphibian-like *Ranavirus* (ALRV) lineages, there is considerable variation in genotype-by-genotype interactions of host and virus (Duffus et al., 2014). Although different intrinsic virulence traits between ALRV lineages exist (Claytor et al., 2017; Price et al., 2014), disease outcome does not always proceed in a predictable manner depending on host-virus genotypes, which is further confounded by the differential effect host life stage has on successful infection for some ranaviruses (Hoverman et al., 2010; Miller et al., 2011). Teasing apart these interactions first requires accurate classification of the genotypes involved.

Second, given that ranavirus hosts are animals implicated in global trade worth billions of US dollars each year (Auliya et al., 2016; Gratwicke et al., 2010), it is crucial that ranaviruses detected in trade networks are accurately typed. An improved taxonomic framework will allow to better assess the risk to local herpetofauna for a detected ranavirus lineage, but also to properly determine the geographic and population of the source of invasions. There is an established and

growing body of evidence of the significant impact trade has had on the global spread of ranaviruses (Gratwicke et al., 2010; Hyatt et al., 2000; Picco & Collins, 2008; Une et al., 2009; Wombwell, 2014). Moreover, there are also several documented incidents of virulent recombinant ranaviruses that likely emerged as a result of amphibian culture and trade (Claytor et al., 2017; Ferreira et al., 2021; Price, 2015). Detailed classification of *Ranavirus* lineage diversity is required to identify donors implicated in recombination events and determine the source(s) of invasions. Taken together, an improved taxonomic classification framework will facilitate assessments of virulence and susceptibility risks to draw balanced mitigative policy for animal trade.

Host isolation source is the main route by which *Ranavirus* isolates are named. This scheme has resulted in an inflated number of proposed strains within the literature, that often contain little to no follow up evidence of a demonstrably divergent phenotype to type isolates. Nevertheless, species demarcation in *Ranavirus* taxonomy has primarily been based on genomic features, mainly comprising genome size and composition (gene presence/absence and GC content), genomic collinearity, nucleotide identity, restriction fragment length polymorphism (RFLP) profiles, and phylogenetic analysis of 26 genes core to the *Iridoviridae* (Chinchar et al., 2017b).

Genome sizes of *Ranavirus* range from approximately 105 – 120 kilobase pairs (Kbp) and are the smallest of the *Iridoviridae,* with genomes in other genera often exceeding 200 Kbp (Chinchar et al., 2017a). Indeed, the largest *Ranavirus* genomes belong to the basal fish-associated isolates, supporting them as distinct taxonomic entities given the larger genomes of the ancestral iridovirid state. Next, structural variation of genomic organisation has possibly played the most significant role in ranavirus taxonomic characterisations. All iridovirid genomes are linear with genes on both DNA strands, although gene order is circularly permuted and inversions are common (Mavian et al., 2012b; Price, 2015). However, ranaviruses are considered to frequently recombine (Chinchar & Granoff, 1986; Claytor et al., 2017; Vilaça et al., 2019), and coding sequence control appears to occur on an individual basis rather than in a linked manner (Mesnard et al., 1988). As such, the genomic arrangement amongst *Ranavirus* varies markedly, but is thought to be generally conserved between species based on dot-plot analyses of reference whole-genome sequences (Jancovich et al., 2015b). Many exceptions to coherent genomic arrangements within species however exist (Candido et al., 2019), and

isolates are often additionally typed based on whole or partial genomic pairwise nucleotide similarity (Ariel et al., 2010; Holopainen et al., 2009).

Chinchar et al. (2017b) highlight how these approaches can be problematic for accurate lineage classification, and that the number of species within iridovirid genera, particularly *Ranavirus*, may need revision. For instance, they postulate that if nucleotide identity of $> 90\%$ is held as the key taxonomic feature, then all currently recognised amphibian-like ranaviruses may be considered strains of the same species. Further, if genome collinearity is considered more important, then the SGIV-like fish-associated ranaviruses may be considered a separate genus, given a complete lack of concordance in genetic arrangement with any other ranavirus lineage (Chinchar et al., 2017b; Yu et al., 2020). Whilst considerations of genome collinearity and nucleotide similarity are adequate for typing the most divergent of species, their lack of resolution creates blind spots to finer-scale evolutionary processes that arguably only phylogenetic analysis of genomic data can illuminate.

Contemporary phylogenetic methodologies offer the most powerful means of reconstructing evolutionary relationships between taxa (Gibbs, 2013). The sophistication of distance-based phylogenetic methods applied to nucleotide sequence data using complex substitution models, such as those employing rate class variation, is unrivaled for modeling evolutionary divergences (Yang & Rannala, 2012). Until relatively recently, the main limiting factors to the application of these methods was the time and cost required for the initial generation of sequence data, and the computational capacity for analysis. However, these factors barely represent hurdles in the current day, thanks in large part to the Next Generation Sequencing revolution (Koboldt et al., 2013), which has led to a significant reduction in the cost of whole-genome sequencing.

The majority of studies that have performed any phylogenetic analysis of ranaviruses have used the major capsid protein (MCP) as the sole marker, due to its highly conserved nature and well established PCR and sequencing protocols (Mao et al., 1997). However, the MCP constitutes only approximately 1.3% of the average ranavirus genome. As such, a greater amount of genetic information taken from whole genome sequences has been increasingly included in ranavirus phylogenetics. For instance, both Stöhr et al. (2015) and Epstein & Storfer (2016) selected 17 core genes found in all ranavirus lineages. More detailed phylogenetic characterisations in ranavirus taxonomy have focused on the 26 core genes that

are found in all iridoviruses, which the ICTV now considers the gold standard for ranavirus lineage assessment (Eaton et al., 2007; Lei et al., 2012; Stagg et al., 2020; Chinchar et al., 2017a). Finally, Price (2015) performed the most detailed phylogenetic analysis of *Ranavirus* to date by taking both Maximum Likelihood and Bayesian approaches to analyse 51 core genes forming a genetic backbone to 17 isolates deemed characteristic reference isolates across the phylogeny of amphibian-like ranaviruses, which excludes the divergent SGIV-like lineage. Incorporating as much genomic information in phylogenetic analyses (with appropriate controls such as for recombination) is the standard that should be strived for; both to maximumly resolve the detailed evolutionary histories of *Ranavirus*, and to reduce any biases that may be introduced by selecting subsets of genes with idiosyncratic evolutionary pasts.

Today, arguably the greatest limitation to phylogenetic applications lies in their analytical foundation. Phylogenetic methods rely on the alignment of sequences to calculate pairwise genetic distances; if genetic sequence data cannot be aligned, molecular phylogenic analysis cannot be applied. This is a particular sticking point for phylogenetic analysis of highly diverse data, which commonly characterises viral genomics. This is the reason why many *Ranavirus* phylogenetic studies only consider a fraction of the genome, such as the 26 core iridovirid genes (Eaton et al., 2007; Chinchar et al., 2017a), as these are the portions of the genome that can be aligned. However, alternative alignment-free methodologies exist, which can offer valuable insights where phylogenetic methods cannot.

K-mer based algorithms offer a comparatively simple means of estimating pairwise genetic distances without having to produce multiple sequence alignments. This is achieved by decomposing sequences into constituent sets of *k*-mers of a given size, then assessing the shared abundance of the unique complement between sequence sets in a pairwise manner. Sequences of any length or complexity can be compared in this way, which in combination with conventional methods has contributed to significant advancements in prokaryotic taxonomy (Parks et al., 2018, 2020). One such study applied *k*-mer based distances to devise a unified typing system of bacterial plasmids, which are genetic entities likely only bested by viruses in terms of poorly resolved evolutionary relationships (Acman et al., 2020). Yet, to date, I am unaware of any such methodologies having been applied to ranavirus genetic diversity characterisations.

In this chapter, I sought to resolve the global genomic diversity of *Ranavirus* by leveraging whole-genome information. Specifically, my aims were to: i) conduct a pan-genome analysis to decipher the complete complement and core set of functionally orthologous ranavirus genes, and how they are partitioned among the lineages; ii) update the functional annotation of the ranavirus (pan-)genome; iii) reconstruct the global phylogeny of ranavirus using conventional methods with as much resolution as possible; and iv) consider the whole-genome complement in alignment-free characterisations to elucidate insights yielded by the accessory genome in resolving the relationships between *Ranavirus* lineages. To achieve these aims, I collated and analysed the largest set of complete genome sequences to date, by combining a novel dataset with all publicly available genome assemblies.

2.3 Methods

2.3.1 Data Acquisition

Novel sample acquisition. I acquired a set of novel *Ranavirus* whole-genome sequences (WGS) from Dr. Stephen Price, which were generated from previous projects (NERC grants: NE/M000338/1, NE/M000591/1, and NE/M00080X/1). The protocols implemented for generating the novel genomes are briefly outlined in the following paragraphs.

Library preparation and sequencing. Ranavirus PCR positive tissue samples or cultured isolates were acquired from collaborating researchers (Table S1). DNA was extracted using the DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions, and were tested for *Ranavirus* positivity using qPCR following the protocol of Leung et al. (2017). DNA concentrations were then measured by Qubit assay, and sheared using a Covaris instrument (Covaris, Inc. Woburn, MA, USA). Ranaviral DNA was enriched via bead-capture technique to separate viral from host DNA, using the Agillet SureSelect XT2 (Agillet Technologies, Inc. Santa Clara, CA, USA) target enrichment protocol. The instructions were followed according to the manufacturer, including all recommended quality control steps, to produce indexed libraries. Finally, ranavirus DNA libraries were pooled and sequenced using an Illumina HiSeq platform (Illumina Inc. San Diego, CA, USA).

Genomic sequence assembly. Short reads were demultiplexed using *bcl2fastq* software, provided by Illumina. The *BBMap* software suit

(https://sourceforge.net/projects/bbmap/) was then used to filter low quality reads, trim adaptors, and normalise read coverage to 100X. The *SPAdes* v3.12.0 (Prjibelski et al., 2020) software package was then used to assemble the isolate sequences, which were collapsed into closed assemblies when genomes were assembled into more than one contig.

Public sample acquisition. A portion of the whole genome dataset consisted of publicly available ranavirus sequencies acquired from the National Centre for Biotechnology Information (NCBI) Short Read Archive (SRA), which were assembled in the same manner to the novel sequences listed above (Table S2). In addition, I downloaded all public ranavirus WGS assemblies (available up to August 2021) from the NCBI GenBank database (Table S2). Duplicate assemblies generated from bioinformatic processing and merging of publicly available datasets were identified and discarded, favouring SRA assemblies generated by Dr. Price over those acquired from GenBank. Finally, I gave each isolate WGS a unique identifier to include in large-scale data visualisations (such as phylogenetic trees), which consisted of the two-to-three digit international country code from which the sample was isolated and two numeric digits representing the decade and year of sampling, if present, followed by an identification string inherited from the original sample name. If individual isolates are referred to in text, I state only the isolate sample name without the country-date code.

Metadata. As part of the handover from Dr. Stephen Price, comprehensive metadata on all novel ranavirus WGSs was also provided. Information included the source of the samples, such as location, host species, wild or captive host, sampled from tissue or culture, contributing researchers, and library preparation and sequencing methods used. Relevant information used in all future analyses is provided in Table S2. I additionally added metadata for the publicly acquired genomes, with all available information parsed from the GenBank files associated with the sequences using a custom Python script (https://github.com/bioinfo-chris/PhD.git; /scripts/gb_extract.py). I then cleaned the metadata by correcting obvious erroneous/conflicting entries, filling missing entries where possible by consulting the literature, and in rare cases contacting relevant researchers involved in the generation of the data. I then expanded metadata by adding sequence information such as length, GC content, as well as more detailed host taxonomy information and latitude-longitude coordinates (where missing) from geocoded location information. The latter process was carried out using *ggmap* (Kahle & Wickham, 2013), a package implemented in the interpreted

programming language *R* v4.0 (R Core Team, 2020). Finally, I calculated the Shannon H-Index of host diversity for each ranavirus lineage using the *R* package *vegan* (Oksanen et al., 2020).

2.3.2 Pan-Genome Analysis

Prokka-Roary pipeline. In order to characterise key genomic features specific to *Ranavirus* and to construct the largest-possible sequence alignment for phylogenetic reconstructions, I started by performing a genus-wide pan-genome analysis to delineate the *Ranavirus* core and accessory genome. First, I used the genome annotation tool *Prokka* v1.14.6 (Seemann, 2014) to identify coding sequence open reading frames (ORFs) in all ranavirus WGSs. *Prokka* acts as a wrapper for several other software packages to first predict genomic features using codon information such as coding sequences (CDS) and structural and non-coding RNAs, then annotates features through homology searches against various databases. I specified the *--kingdom Viruses* option to direct *Prokka* to use its curated virus protein database for CDS annotations, if present.

Next, I used the *Roary* v3.11.12 (Page et al., 2015) pipeline to cluster all identified *Ranavirus* CDS ORFs at a minimum amino-acid homology threshold of 80%. I selected this threshold as the International Committee on the Taxonomy of Viruses (ICTV) defines species within the genus *Ranavirus* as those with $> 95\%$ sequence homology within the 26 core iridovid genes (Chinchar et al., 2017a), and therefore, 80% homology stands as a liberal threshold for orthologous gene sequences between species or lineages. Furthermore, 80% homology has previously been used to demarcate *Ranavirus* orthologs on the basis that homologous protein function likely deteriorates below this threshold (Epstein & Storfer, 2016; Price, 2015). The *Roary* pipeline clusters CDS ORFs into ortholog clusters by taking annotated assemblies in GFF3 format, produced by *Prokka*, as input. From these files, the CDS regions are extracted and converted to protein sequences, which are initially clustered using CD-HIT (Fu et al., 2012). After which, a pairwise protein *BLAST* v2.9.0 (Camacho et al., 2009) search of all amino acid sequences is performed at the define homology threshold, and the results are used to cluster the sequences using Markov Cluster (MCL) algorithm (Enright et al., 2002). The CD-HIT and MCL results are then merged to create the pan-genome complement of orthologous ORF clusters.

Pan-genome analysis and core ORF extraction. One of the key files produced by the *Roary* analysis is the pan-genome presence/absence table. From this, I used *R* to

determine shared and private genes between phylogenetically inferred *Ranavirus* lineages (see below). I then considered core genes as orthologous ORF clusters present in all isolates. To extract the core ORFs from each isolate assembly, I used the ORF cluster reference sequence provided by the *Roary* output as a query sequence in a nucleotide *BLAST* search against a database of the ranavirus WGS dataset. As above, I set a minimum threshold of 80% for both nucleotide identity and high-scoring pair (HSP) alignment coverage, then extracted the resulting highest scoring aligned portions within each assembly as its respective core ORF set. Finally, in the event that isolates did not contain core genes within the 80% threshold, I performed *ad hoc* nucleotide *BLAST* searches at 70, 50, and 30% homology of all core ORFs against the problematic assemblies to elucidate the threshold of homology required to include them.

2.3.3 Orthologous Gene Annotation

Next, I used the reference sequence from each orthologous ORF cluster to perform homology searches to annotate the pan-genome, with functions where possible. I employed two approaches. I first performed a nucleotide *BLAST* search at 90% identity and HSP coverage against a database of all ranavirus CDS sequences published on NCBI, specifying a return of ten target sequences. Annotations were selected based on hits with greatest percentage homology, except in the case of hypothetical protein annotations, where I manually checked lower scoring matches for the event they contained functional information. I then translated the ortholog reference sequences into their amino acid complement (first frame based on presence of a start codon) using the *transseq* command of the *EMBOSS* v6.6.0 (Madeira et al., 2019) software suit to perform a protein family and domain scan against Pfam. Pfam is a database of all currently classified protein families and domains, and contains a seed alignment of representative sequences for each entry, with which profile hidden Markov models (HMM) are constructed from and used to search against (Mistry et al., 2021). I carried out the search using the *hmmscan* function of the command line software *HMMER* v3.3.2 (Mistry et al., 2013), suppling the ortholog amino acid sequences and the current release of the Pfam HMM database at the time of analysis (http://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/Pfam-A.hmm.gz; 19-Mar-2021 release, accessed Aug-2021). I selected top scoring hits based on the lowest E-values (chance of random match), and assigned annotations to ortholog ORF clusters from the consensus between the *BLAST* and Pfam search results.

2.3.4 Multiple Sequence Alignments

Multiple sequence alignments. I wrote and used a custom bash script (https://github.com/bioinfo-chris/PhD.git; /scripts/core_genome_pipeline_ALRV_n170.sh) to concatenate each core ORF in a common orientation for each isolate. Then, I aligned the concatenated core genomes using *MAFFT* v7.453 (Katoh & Standley, 2013) with the default settings. The multiple sequence alignment (MSA) was then trimmed of any gap positions found in 20% or more of isolates using the command line tool *trimAL* v1.4 (Capella-Gutiérrez et al., 2009). The resulting genus-wide alignment of all ranavirus concatenated core genomes was then subjected to the homoplastic-site pruning and Maximum Likelihood phylogenetic inference analyses described below. From the resulting genus wide *Ranavirus* phylogenetic tree, I identified monophyletic clades and characterised distinctive remaining paraphylies for candidate taxonomic groupings.

Homoplasy filtering. Models of phylogenetic inference used by common tree-building tools are limited to assuming a single evolutionary history of the sequences being analysed. Due to this, loci in MSAs that contain signal of recombination will lead to inaccurate phylogenetic reconstruction (Pond et al., 2006a). As genetic recombination of distinct ranavirus lineages has been known to occur (Claytor et al., 2017; Epstein & Storfer, 2016; Price, 2015; Vilaça et al., 2019), I screened and removed recombinant sites from all alignments used in downstream phylogenetic analyses. I achieved this by employing a perhaps overly stringent protocol based on detection and removal of homoplasies – genomic loci that cause phylogenetic conflicts in tree topology. Homoplasies can arise independently by means of convergent evolution, whereby distinct lineages derive the same molecular solution to a common selective pressure, or they can be loci derived via homologous recombination of genetic segments introduced across lineages (Smith & Smith, 1998).

The protocol followed that described by van Dorp et al. (2020). First, a Maximum Parsimony tree was produced using the rapid tree inference tool, *MPBoot* v1.1.0 (Hoang et al., 2018), using 1,000 bootstraps. The resulting tree and input alignment were then used to identify homoplastic sites with the tool *homoplasyFinder* v0.0.0.9 (Crispell et al., 2019), which is a package implemented in *R*. For each site in the alignment, *homoplasyFinder* provides a consistency index which captures the minimum number of state changes required on the Maximum Parsimony tree that explain the character state at the tips (Fitch, 1971). As recommended by the

authors of the tool, sites with a consistency index of ≤ 0.5 (highly inconsistent) were considered homoplastic and removed from MSAs, and thus any potential recombinant loci.

Finally, after filtering homoplasies, variant sites were called using the command line software *SNP-sites* v2.5.1 (Page et al., 2016) to obtain counts of single nucleotide polymorphisms (SNPs) for each MSA.

2.3.5 Phylogenetic Inference

To reconstruct the genomic diversity of *Ranavirus*, I used a Maximum Likelihood (ML) phylogenetic reconstruction approach. I utilised the command-line software *RAxML* v8.2.12 (Stamatakis, 2014) to infer the phylogenetic relationships between the homoplasy-filtered concatenated core genome alignment. I specified the general time-reversable (GTR) model of nucleotide substitution with a Gamma distribution of transition-transversion rate variation (specifying the flag *-m GTRCAT*, which is memory and time efficient approximation of the Gamma rate category model). I further specified 100 ML iterations, selecting the best tree. A separate run was also executed alongside with 1,000 bootstrap iterations, which I used to annotate the best ML tree for node support.

A quirk of using a time-reversable ML model is that the root of the phylogenetic tree can be placed anywhere in process of calculating branch lengths (Cho, 2012). As such, whilst *RAxML* produces rooted trees, it is an arbitrary, evolutionary-uninformed root. Because of this, I included an outgroup isolate selected *a priori*, with which the final phylogeny was re-rooted to upon drawing the tree. I used an isolate informed from the literature as the outgroup, which was the highly diverged *Short-finned eel ranavirus* (SERV; Ariel et al., 2010; Subramaniam et al., 2016; Chinchar et al., 2017b; Price et al., 2017a). Finally, I drew the final phylogenic tree in *R* using the package *ggtree* (Yu et al., 2017).

2.3.6 Alignment-Free Genetic Distances

Given that phylogenetic approaches rely on variants called from sequence alignments, highly diverse genomic datasets require discarding sequence information that cannot be aligned. Indeed, the core ORFs of *Ranavirus* comprised only approximately 50% of the genome (see Results 2.4.2), and as such, I employed alignment-free genomic characterisations of the genus in addition to the phylogenetic analyses described above. To compute pairwise genetic distances, I used the command line software *Mash* v2.1.1 (Ondov et al., 2016) which implements a *k*-mer based approach.

Mash decomposes sequences into a set of their constituent *k*-mers of a given base-pair (bp) length. The MinHash technique (Broder, 1997) is utilised to produce a compressed 'sketch' of the set, whereby a hashing function is used to obtain a 32 or 64-bit hash for each *k*-mer, depending on its size. Sketches are then compared to rapidly compute distances based on shared hashes. Two key distance metrics are drawn. First, the exact Jaccard Index is calculated as the number of shared hashes over the total number of hashes between two sketches. The so-called Mash Distance is also computed, which is metric that estimates a mutation rate between sequences directly from sketches, with apparent comparative performance to alignment-based approaches (Ondov et al., 2016).

Using the *sketch* function of *Mash*, I supplied the *Ranavirus* WGS dataset containing all isolate assemblies as input, specifying $k = 12$ bp with a maximum sketch size of ten-thousand non-redundant hashes. I then compared the resulting sketch against itself to compute pairwise genetic distances using the *dist* function. Next, to visualise the raw pairwise genetic distances, I performed the following using *R*. First, I converted Jaccard Index into Jaccard Distances (1 – Jaccard Index), and created symmetric n-by-n matrices of both the pairwise Jaccard and Mash Distances. I then used the package *ComplexHeatmap* (Gu et al., 2016) to gradient colour the Jaccard Distance matrix, with isolates arranged in the heatmap by hierarchical clustering. Finally, I constructed a Neighbour-Joining (NJ) tree using the Mash Distance matrix of pairwise mutation rates. I compared the NJ tree to the corresponding core genome ML phylogenetic tree in a node-optimised cophylogeny using the package *phytools* (Revell, 2012) to assess topology concordance. I did this to visualise how the alignment-free approach using all WGS information reconstructed *Ranavirus* genetic diversity compared to the alignment-based model that utilised only core genomic material.

2.3.7 Genetic Ordinations

Principal coordinate analysis (PCoA). To characterise the variance between and within *Ranavirus* clade memberships based on whole and core genome distances in two-dimensional space, I employed a PCoA analysis of metric multidimensional scaling. I used the symmetric matrix of Jaccard Distances as input to the *R* function *cmdscale* to compute the principal coordinate eigen values, and ultimately the pairwise dissimilarity distance coordinates in Euclidian space. I then plotted the ordination using the *R* package *ggplot2* (Wickham, 2016). Groups were coloured according to phylogenetically inferred clades, including 95%

confidence ellipses around clade groups assuming a multivariate normal distribution.

Principal component analysis (PCA). Next, I assessed how *Ranavirus* clade groupings compared between those yielded by the *k*-mer-based WGS distances (above), and information obtained from the corresponding alignment-based core genotypes. To do this, I performed multivariate PCA with the *base R* function *prcomp*, supplying a multi-allelic SNP array derived from the ranavirus core genome alignment as input. The SNP array was created by first calling variants using *SNP-sites* specifying the *-v* option for vcf format output, then editing the file to contain only isolate and variant information in n-by-p format. I then plotted the resulting coordinates of the first two principal components explaining the greatest amount of variation using the *R* package *ggplot2*, as before.

Discriminant analysis of principal components (DAPC). Lastly, I sought to characterise the genetic features driving the patterns obtained from the WGS and core genome multivariate scaling analyses described above. To do so, I utilised Discriminant Analysis (DA) which maximizes variance between groups whilst neglecting within-group variance. Specifically, I employed DAPC (Jombart et al., 2010), implemented as the *dapc* function of the *R* package *adegenet* (Jombart, 2008), which is a method to overcome the requirements of DA that otherwise commonly restrict its use with genetic data. For instance, DA necessitates that the number of variables does not exceed observations (Lachenbruch & Goldstein, 1979), which is often violated by large n-by-p SNP arrays. DAPC overcomes this, and other limitations, by reducing multivariate dimensionality and applying the discriminant functions to principal components.

I first applied DAPC to elucidate gene contributions to *Ranavirus* clade groupings to assess genomic features that may drive patterns observed amongst WGS genetic distances. Unfortunately, the *k*-mer-derived WGS distances could not be supplied directly as input for DAPC, given the lack of feature information (*k*-mers) lost in the process pairwise distance calculation. As an alternative means of capturing whole-genome contributions to observed structures, I used the pan-genome ORF presence/absence table as input. Secondly, I applied DAPC directly to the core genome SNP array to assess the contributions of SNPs (and the genes in which they occur), that may be driving the observed group variation between *Ranavirus* clades. For both DAPC analyses, I specified the five clades identified *a priori* as groups (phylogenetically-inferred EHNV-like, ATV-like, CMTV-like,

FV3-like, and TFV-like), the post-hoc inferred optimal number of PCs to include (using the *optim.a.score* function), and four discriminant axes to retain (n groups – 1).

2.4 Results

2.4.1 Isolate Details

Whole-genome sequences. I collated a total of 179 complete *Ranavirus* whole genome sequences from novel and public sources. After quality control for whole genome completeness and removal of duplicates, 173 unique isolates remained. Of these, 55 were novel and 118 were of public origin. Of the latter, 16 isolates were acquired from the SRA and *de novo* assembled with the novel sequences. The remaining 102 were assemblies downloaded from NCBI GenBank. A list of accessions (where applicable) and associated metadata are supplied in Table S2.

I derived clade classifications across the *Ranavirus* genus from the results of the alignment-based and alignment-free analyses reported below. However, for the sake of reporting clade-specific isolate data in the proceeding sections, I briefly outline them here first. Overall, I delineated five main *Ranavirus* clades, two of which contain paraphylies (see below). I classified the clade names using the current exemplar isolates for the *Ranavirus* species currently accepted by the ICTV (https://talk.ictvonline.org/ictv-reports/ictv_online_report/dsdna-viruses/w/iridoviridae/616/genus-ranavirus last accessed December 2021; Chinchar et al., 2017a), or by the dominant species within the given clade if paraphyletic. Altogether, the clade names used for the remainder of this thesis are the *Epizootic haematopoietic necrosis virus* (EHNV)-like, the *Ambystoma tigrinum virus* (ATV)-like, the *Common midwife toad virus* (CMTV)-like, the *Tiger frog virus* (TFV)-like, and the *Frog virus 3* (FV3)-like. See below for the detailed rationale and specificities to these classifications. The SGIV-like group comprising the fish-associated ranaviruses of two *Singapore grouper iridovirus* (SGIV)-like isolates and the single *Santee-Cooper ranavirus* (Largemouth bass virus; LMBV) did not fall into any of the five clades, and were excluded from all analyses, as I deemed them too divergent to be included with aim of the study in mind (see below). For simplicity, I henceforth refer to this group as the basal fish ranaviruses.

All isolate assemblies were sampled across 24 countries on five continents (Fig. 2.1), ranging between the years 1966 and 2017. The viruses were isolated from a total of 56 host species (43 genera) across 15 orders of the three vertebrate classes

Amphibia (34 spp.), Actinopterygii (16 spp.), and Reptilia (10 spp.). The host breadth of each clade of isolates varied considerably (Fig. 2.2). The ATV-like (22 isolates) contained the most restricted host range with a Shannon H-Index (SHI) of diversity of 0.923, isolated from three species of ambystomid salamanders in North America. Next, the paraphyletic EHNV-like (25 isolates) contained a SHI of 1.696, and were isolated from eight distinct species of fish across seven orders from Australia, Europe and the North Atlantic. Then the amphibian-associated CMTV-, FV3-, and TFV-like each contained the greatest taxonomic breadth of hosts, isolated from across all three host orders. The CMTV-like (46 isolates) contained the largest SHI of 2.448, and were isolated from 18 host species across

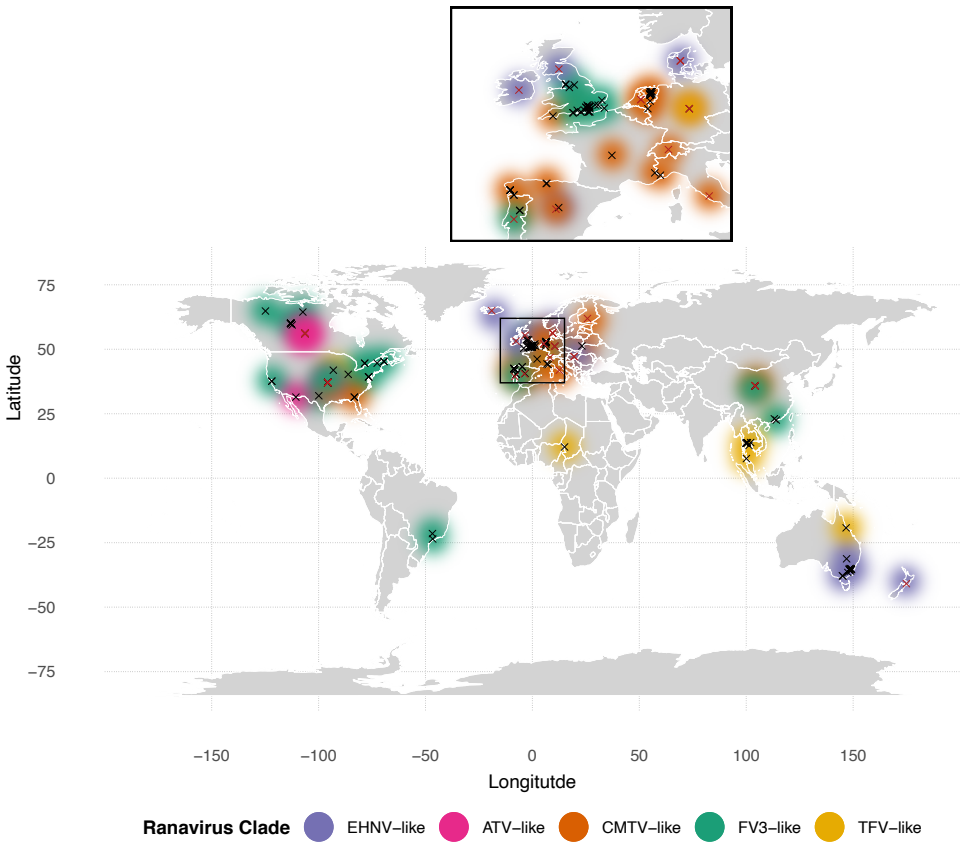


Figure 2.1. *Global* Ranavirus *isolate distribution map. Black crosses are isolate assemblies with specific known sampling locations, whereas red crosses are country geographic centroids for samples where the exact location is unknown. Coloured shading reflects the lineage of the ranavirus isolate.*

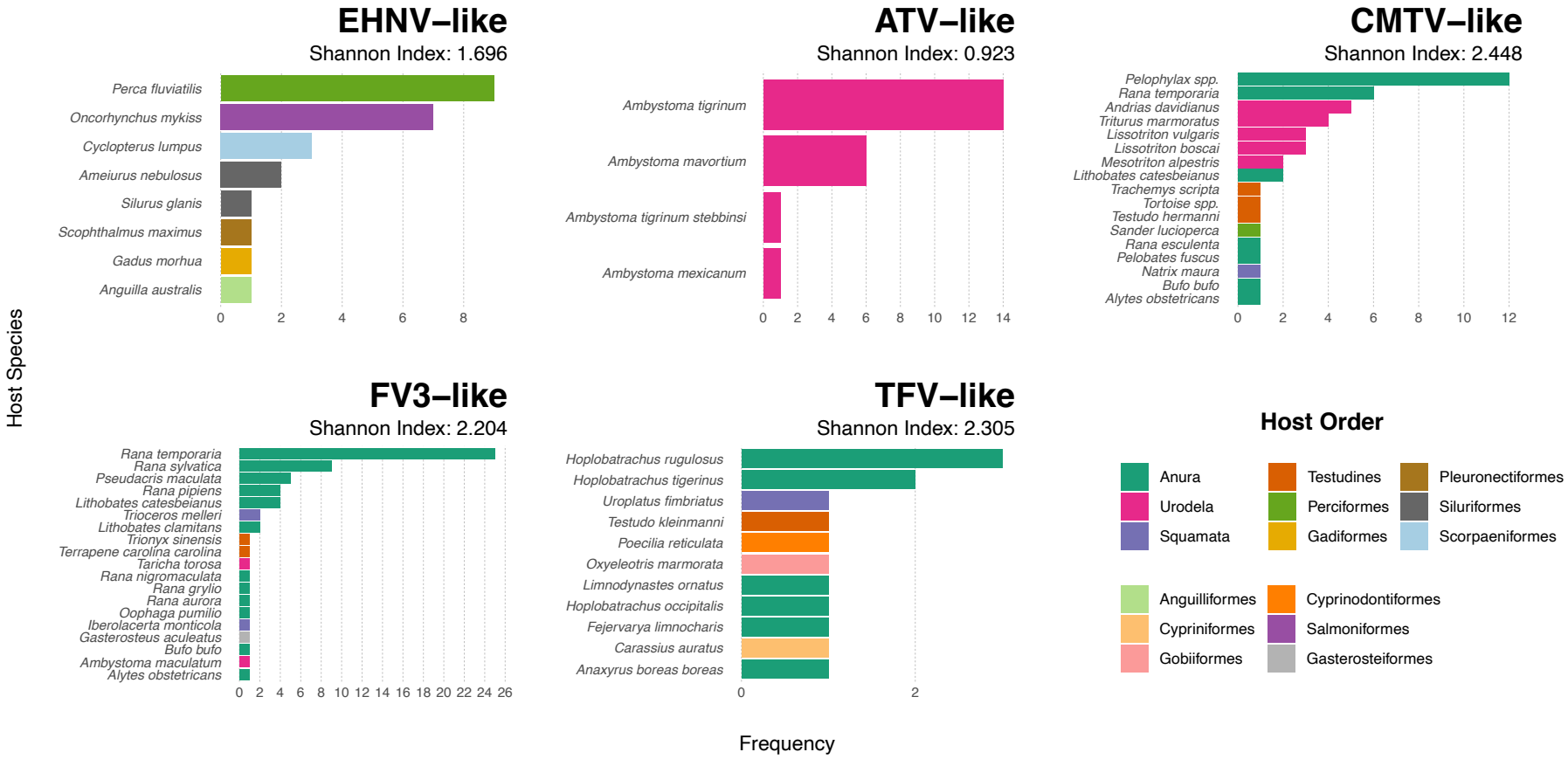


Figure 2.2. *Hosts from which ranavirus assemblies were isolated. Host species are given on the Y-axis of each panel, with the frequency of isolates on sampled from hosts on the X-axis. Bars for each host species are coloured according to their taxonomic order, which span the three vertebrate classes Actinopterygii, Amphibia, and Reptilia. Panels are divided from taxonomic grouping of ranavirus isolates according the phylogenetic characterisations outlined later in the chapter.*

Europe, Asia, and the USA (one captive isolate). Next, the FV3-like (63 isolates) contained second greatest SHI of 2.204, sampled from 20 host species (the largest number of unique hosts, but at lower abundance) across North and South America, Europe, and Asia. Finally, the paraphyletic TFV-like (14 isolates) contained an intermediate SHI of 2.305, sampled from 12 host species from Asia, Africa, Australia, Europe, and the USA.

2.4.2 Pan- and Core Genome, and Core Alignment

Pan- and accessory genome. Using the *Prokka-Roary* pipeline, I delineated 49 core orthologous ORF clusters at a minimum of 80% amino acid homology in a total 170 isolates. The three SGIV-like isolates (AY666015.1, AY521625.1, and MK681856.1) did not contain any ORFs that clustered at the homology cutoff (Fig. S1). Further, from the nucleotide *BLAST* search of all core ORFs against the SGIV-like genomes at 70, 50, and 30% minimum nucleotide homology, only three genes for LMBV were recovered at ~75% identity. The remainder of the lineage contained no nucleotide homology to any core ORF at all thresholds used. The degree of divergence of the basal fish ranaviruses was not surprising as it has been previously quantified (Epstein & Storfer, 2016; Price, 2015), and as discussed by Chinchar et al. (2017b), it is possible these viruses represent a basal sister genus to *Ranavirus*. Moreover, the two SGIV-like isolates contained significantly longer-than-average genomes compared to the remaining isolates of *Ranavirus* (median: 139.8 Kbp versus 106.1 Kbp), whereas the LMBV isolate was uncharacteristically small (99.2 Kbp). Taken together, I deemed these three basal fish-associated ranaviruses sufficiently divergent to the remainder of the ALRVs to excluded them from all future analyses, particularly for the reduction in amino acid homology in orthologous ORF clustering that would be required for their inclusion. As such, the pan-genome of the remaining 170 ALRV isolates consisted of 217 ortholog clusters.

The pan-genome was variably partitioned amongst the five ALRV clades, with each containing an average 98.6 shared and 34 private ORFs (Fig. 2.3 **A**). The most basal clade of EHNV-like contained a total of 154 ORFs and the greatest number of 59 private genes, with a median 107 ORFs per isolate. The ATV-like contained the smallest ORF sets, with 14 private genes of 109 total, and each isolate containing a median of 91 genes. Next, the CMTV- and FV3-like both contained a median of 93 ORFs per isolate, and also contained a similar complement of private and shared ORFs, with 104 and 102 shared and 35 and 36 private genes, respectively.

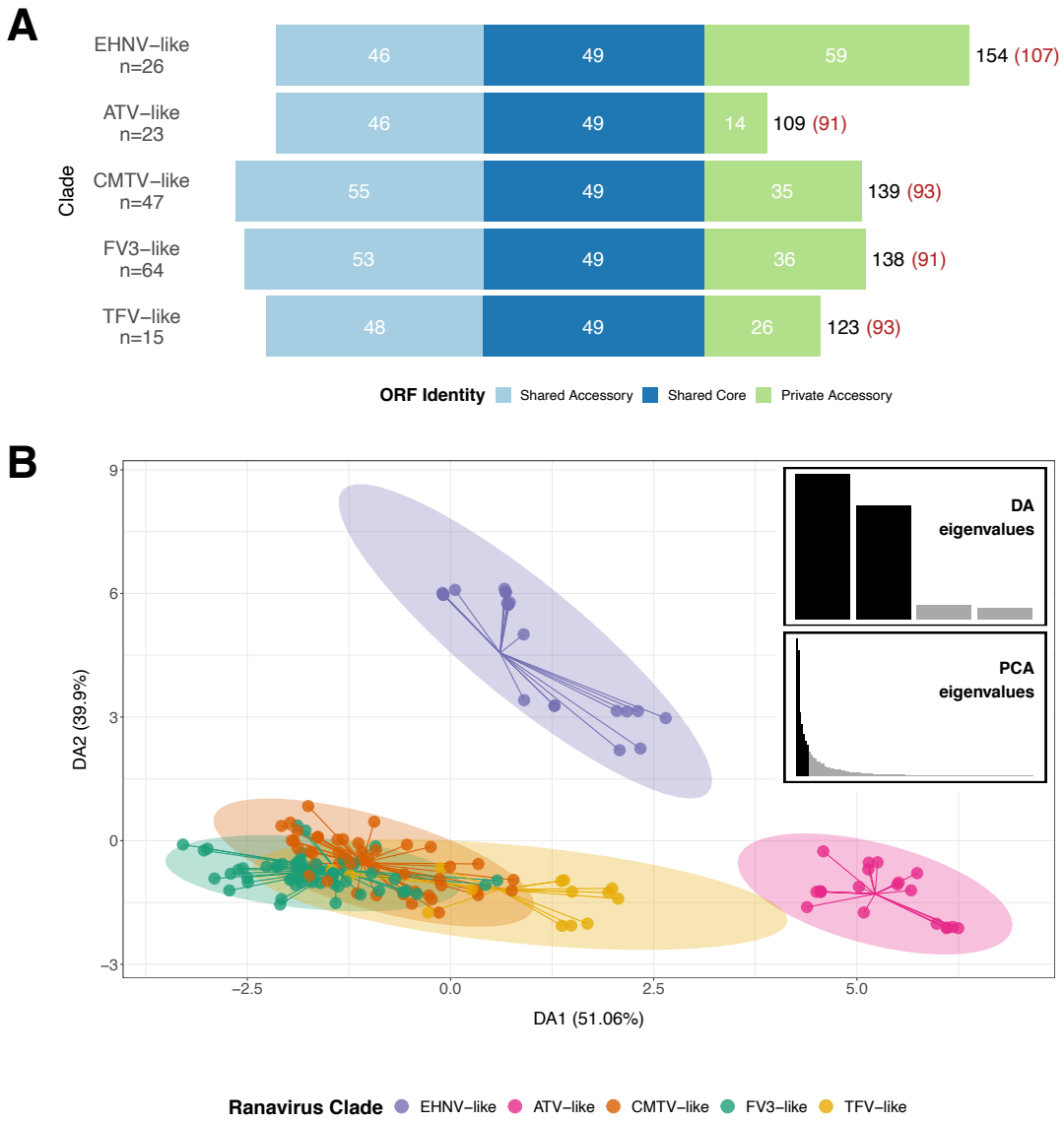


***Figure 2.3**. Pan-genome partitions between* Ranavirus *clades. Panel **A** shows the private, core, and shared accessory genome complement for the isolates of each clade, based on gene presence/absence of the total 217 orthologous ORFs clusters. The total number of unique ORFs is displayed outside the bars in black, and the median number of predicted ORFs per isolate of the clade parenthesised in red. Panel **B** illustrates variance between clades in terms of gene presence/absence, as determined by DAPC analysis performed using four discriminate functions on the optimum seven retained PCs, grouped by the five ALRV clades.*

Finally, the TFV contained a median of 91 ORFs per isolate, with 26 private genes. To visualise the variance of gene complements between clades, I performed a DAPC on clade isolates gene presence-absence (Fig. 2.3 **B**). This showed that both the ATV- and EHNV-like contained the most distinct ORF complement.

Meanwhile, there was almost complete overlap in variance between the CMTV- and FV3-like gene composition, and although the TFV-like contained a large amount of overlap as well, their degree of divergence from the latter two clades suggested a more unique complement of shared ORFs.

Functional annotation. By performing *BLASTn* searches of current *Ranavirus* gene annotations and HMM scans of Pfam protein domains, I was able to recover a total of 120 gene annotations, representing 55.3% of pan-genome (Table S3). Where annotations were present, most also contained functional information; only 15 annotated genes across the pan-genome contained a still unknown function via domain scans. Of the *Ranavirus* core genome, 34 (69.4%) genes were annotated, which are shown in Table 2.1.

Core genome alignment. The ALRV core genome alignment consisted of the concatenated 49 core genes (Table 2.1). After being trimmed of gaps, the MSA contained 45,169 bp positions. The *homoplasyFinder* protocol detected 2,261 homoplastic loci with a consistency index of ≤ 0.5. After removal, the final ALRV core alignment was a total of 42,908 bp in length (39.54% of the average ALRV genome) and contained 6,158 variant sites (SNPs).

2.4.3 Phylogenetic Reconstructions and Clade Classifications

Maximum Likelihood Phylogeny. The genus-wide ML phylogenetic reconstruction of the 170 ALRV ranaviruses contained excellent internal node support, including complete bootstrap support for the root (Fig. 2.4). Three clear monophyletic groupings were apparent, delineating the basal ATV-likes, intermediate CMTV-likes, and derived FV3-likes, each with total support except for positioning of the CMTVs which contained 87% confidence. Viruses with native piscine hosts were not resolved monophyletically, but were instead represented by three subclades. Overall, I named this basal paraphyletic clade the EHNV-like due to the dominance of that lineage, as well as a grouping supported by the alignment-free characterisations reported below. Nevertheless, the three monophyletic subclades comprised: i) the European sheatfish virus (ESV) and European catfish virus (ECV) isolates most basal to the root of the phylogeny; ii) the intermediate *Epizootic haematopoietic necrosis virus* isolates from Australia; and iii) the most divergent and derived marine European North Atlantic ranaviruses containing the *Lumpfish ranavirus*, *Ranavirus maximus*, and *Cod iridovirus* isolates (latter two are unclassified species, according to the ICTV).

***Table 2.1**. Amphibian-like* Ranavirus *core genome annotations. Of the 217 orthologous Roary clusters detected at 80% amino acid homology, 49 were found in all 170 ALRV isolates – the ALRV core genome. The first column is the simple annotation assigned by the* Roary *pipeline. Next is the annotation according to ORFs of the FV3 type genome (AY548484.1). The last column shows the consensus annotation, giving the protein information derived from* BLASTn *homology searches of existing ranavirus annotations, followed by functional domain information ascertained from translated core gene Pfam HMM scans, divided by a semi-colon. If both methods failed to return a match, ORF clusters are annotated as hypothetical proteins as the consensus annotation, labeled in grey.*

Roary ORF Cluster	RefSeq FV3 Annotation	Consensus Protein and/or Functional Annotation
BALF5	ORF 60R	DNA polymerase family B
I4L	ORF 38R	Ribonucleotide reductase, barrel domain
MCP	ORF 90R	Large eukaryotic DNA virus major capsid protein
group_40	ORF 18L	Serine/threonine-protein kinase
group_45	ORF 59L	Hypothetical protein
group_46	ORF 39R	Hydrolase of the metallo-beta-lactamase superfamily
group_47	ORF 4R	Hypothetical protein
group_58	ORF 23R	Hypothetical protein
group_59	ORF 20R	Hypothetical protein; Family of unknown function (DUF5850)
group_60	ORF 14R	Putative surface protein
group_62	ORF 56R	Hypothetical protein
group_68	ORF 34R	Human parainfluenza virus 1L-like protein
group_69	ORF 16R	Vertebrate interleukin-3 regulated transcription factor
group_70	ORF 96R	C-mannosyltransferase dpy-19; Family of unknown function (DUF5875)
group_72	ORF 81R	Transcription elongation factor S-II (TFIIS)
group_74	ORF 97R	Myeloid cell leukemia protein; Apoptosis regulator proteins, Bcl-2 family
group_75	ORF 3R	Hypothetical protein; Domain of unknown function (DUF1729)
group_76	ORF 75L	Lipopolysaccharide-induced TNF-alpha factor (LITAF)-like protein; LITAF-like zinc ribbon domain
group_78	ORF 58R	Mitochondrial resolvase Ydc2 / RNA splicing MRS1
group_85	ORF 85R	Deoxynucleoside kinase/thymidine kinase
group_89	ORF 57R	Putative phosphotransferase; Protein kinase domain
group_90	ORF 47L	Hypothetical protein
group_91	ORF 10R	Hypothetical protein
group_92	ORF 31R	Hypothetical protein
group_94	ORF 15R	AAA-ATPase; Poxvirus A32 protein
group_95	ORF 3R	Hypothetical protein
group_96	ORF 77L	Hypothetical protein
group_97	ORF 69R	Transmembrane domain protein; Domain of unknown function (DUF4444)
group_101	ORF 28R	RNA polymerase Rpb5, C-terminal domain
group_102	ORF 94L	Hypothetical protein; Protein of unknown function (DUF2726)
group_104	ORF 56R	Hypothetical protein
group_106	ORF 27R	Tyrosine kinase; Glycosyl transferase family 90
group_111	ORF 12L	Family of unknown function (DUF5832)
group_119	ORF 37R	NLI interacting factor-like phosphatase
group_120	ORF 33R	Tryptophan-associated transmembrane protein
group_121	ORF 11R	WXG100 protein secretion system (Wss), protein YukC
group_122	ORF 84R	Proliferating cell nuclear antigen
group_125	ORF 95R	Putative DNA repair protein RAD2; XPG I-region
group_126	ORF 88R	Thiol oxidoreductase; Erv1 / Alr family
group_127	ORF 80L	Ribonuclease III; Ribonuclease III domain
group_131	ORF 21L	Helicase family protein; Family of unknown function (DUF5767)
group_132	ORF 1R	Replicating factor; Poxvirus Late Transcription Factor VLTF3 like
group_133	ORF 83R	Cytosine DNA methyltransferase; C-5 cytosine-specific DNA methylase
group_136	ORF 9L	NTPase/helicase; Type III restriction enzyme, res subunit
group_137	ORF 25R	p31K protein; Protein of unknown function (DUF2738)
group_143	ORF 17L	Methyl-accepting chemotaxis sensory transducer; ATP synthase D chain, mitochondrial (ATP5H)
group_144	ORF 53R	Myristylated membrane protein; Lipid membrane protein of large eukaryotic DNA viruses
group_145	ORF 41R	Hypothetical protein; Family of unknown function (DUF5757)
group_217	ORF 22R	D5 family NTPase/ATPase; D5 N terminal like

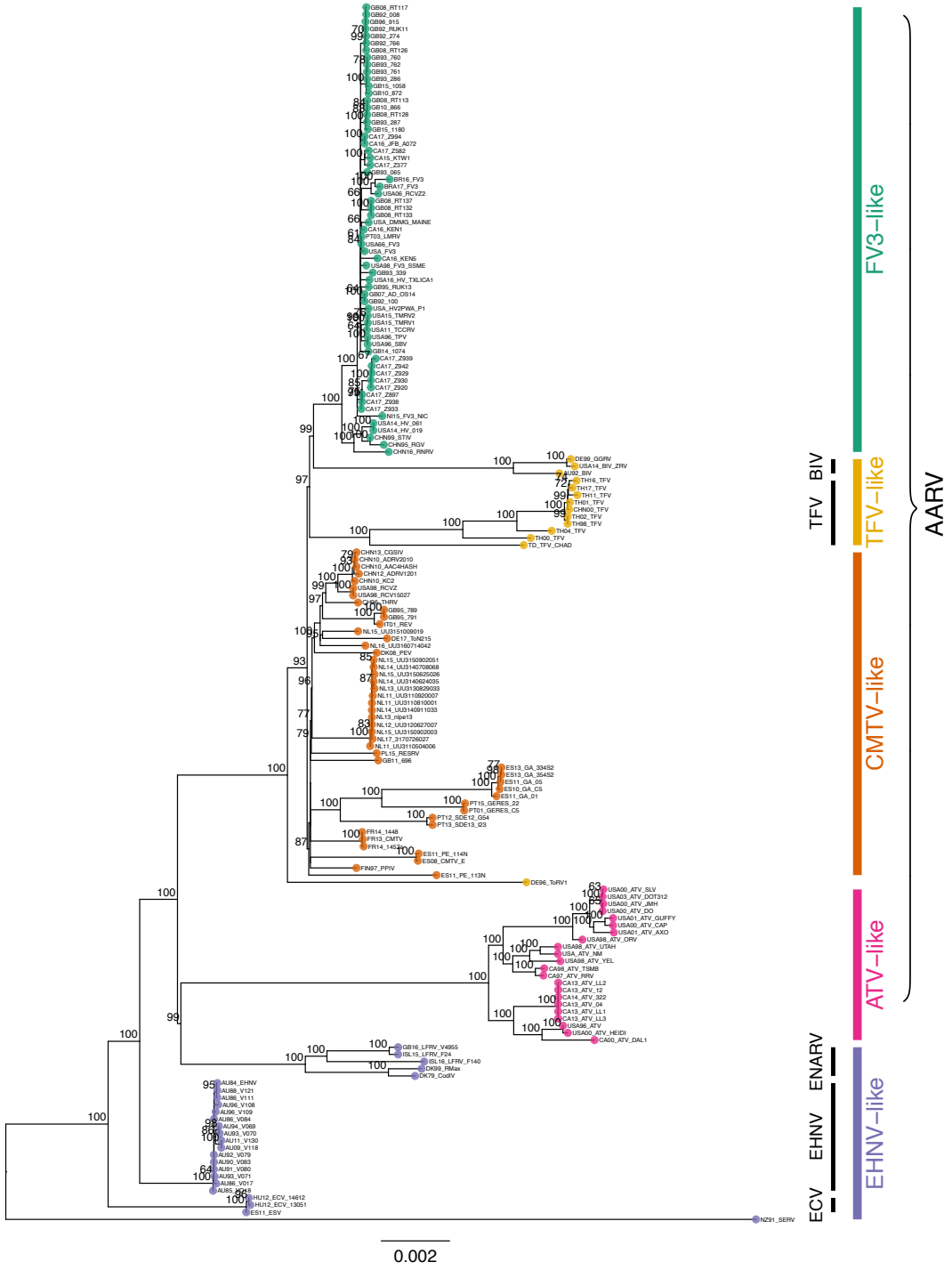


Figure 2.4. *Maximum Likelihood* Ranavirus *core phylogeny. Global phylogeny of 170 amphibian-like Ranavirus isolates, based on 49 core genes in each assembly at 80% amino acid homology. Clade bars are coloured, with lineages of paraphyletic clades in black. Amphibian-associated ranaviruses are indicated by the curly brace. The concatenated alignment was homoplasy filtered to control for recombination, leaving a total of 6,158 SNPs. The phylogeny was constructed using* RAxML *specifying the GTR+Γ (GTRCAT) substitution model and 1,000 bootstraps for node support. Branch lengths are scaled by substitutions per site. The phylogeny is rooted to the SERV isolate (KX353311.2). SGIV-like fish-associated ranaviruses were omitted due to core genome divergence beyond the homology cutoff.*

The remaining isolates were not resolved monophyletically, and instead comprised three paraphylies (with ≥ 97% support). I grouped these into a paraphyletic clade which I named the TFV-like, due the dominant contribution of that lineage and alignment-free groupings characterised below. This clade comprised two clear lineages of: i) the basal *Tiger frog virus* isolates, and ii) the derived *Bohle iridovirus* (BIV) isolates, which could be considered monophyletic with the FV3 clade, except that they contained highly divergent genetic distances. Whilst both these sub-lineages fell intermediate between the CMTV- and FV3-like, a single paraphyletic isolate to all clades, *Tortoise ranavirus-1* (ToRV1), grouped alone and basal to the derived AARV clades (CMTV-, TFV-, and FV3-like). This isolate could represent a unique lineage, although it was sampled from a zoo and has been reported as a mosaic virus based on its polyphyletic genomic makeup (Price 2015). Despite its singular basal position, I included ToRV1 with the paraphyletic TFV-like clade to resolve all lineages. All TFV-like sub lineages were situated individually on very long branches, highlighting an extreme divergence only comparable to the ATV-like. This unique genomic diversity could be indicative of undersampled diversity with an ancient geographic association, or novel diversity generated through recombination, or both. It is important to also note that all TFV-like ranaviruses were isolated from captive settings, except for an African TFV isolate from Chad, which was sampled from a wild amphibian (MW727505.1; Box et al., 2021).

Alignment-free k-mer distances. Raw pairwise genetic distances between the 170 ALRV isolates are shown in Figure 2.5. The heatmap of genetic dissimilarity, based on Jaccard Distances derived from shared 12-mers, strongly supported the clade groupings outlined above. Isolates were arranged on the heatmap by hierarchical clustering, which revealed large blocks of similarity within the monophyletic ATV-, CMTV-, and FV3-like clades. Of these, the FV3-like contained the least amount of diversity, which could be indicative of recent population expansion. The CMTV-like contained the most within-clade diversity, suggesting different demographic processes likely relating to geography and host breadth. Then, the fine-scale structure of the paraphyletic EHNV-like supported the three clear sub-lineages, and as above, the structure for the TFV-like supported two sub-lineages (the TFV and BIV groups) with two substantially divergent isolates (ToRV1 and TFV Chad). The heatmap also illustrated that *Ranavirus* genetic diversity did not seem to significantly diverge within clades if a member was isolated from a

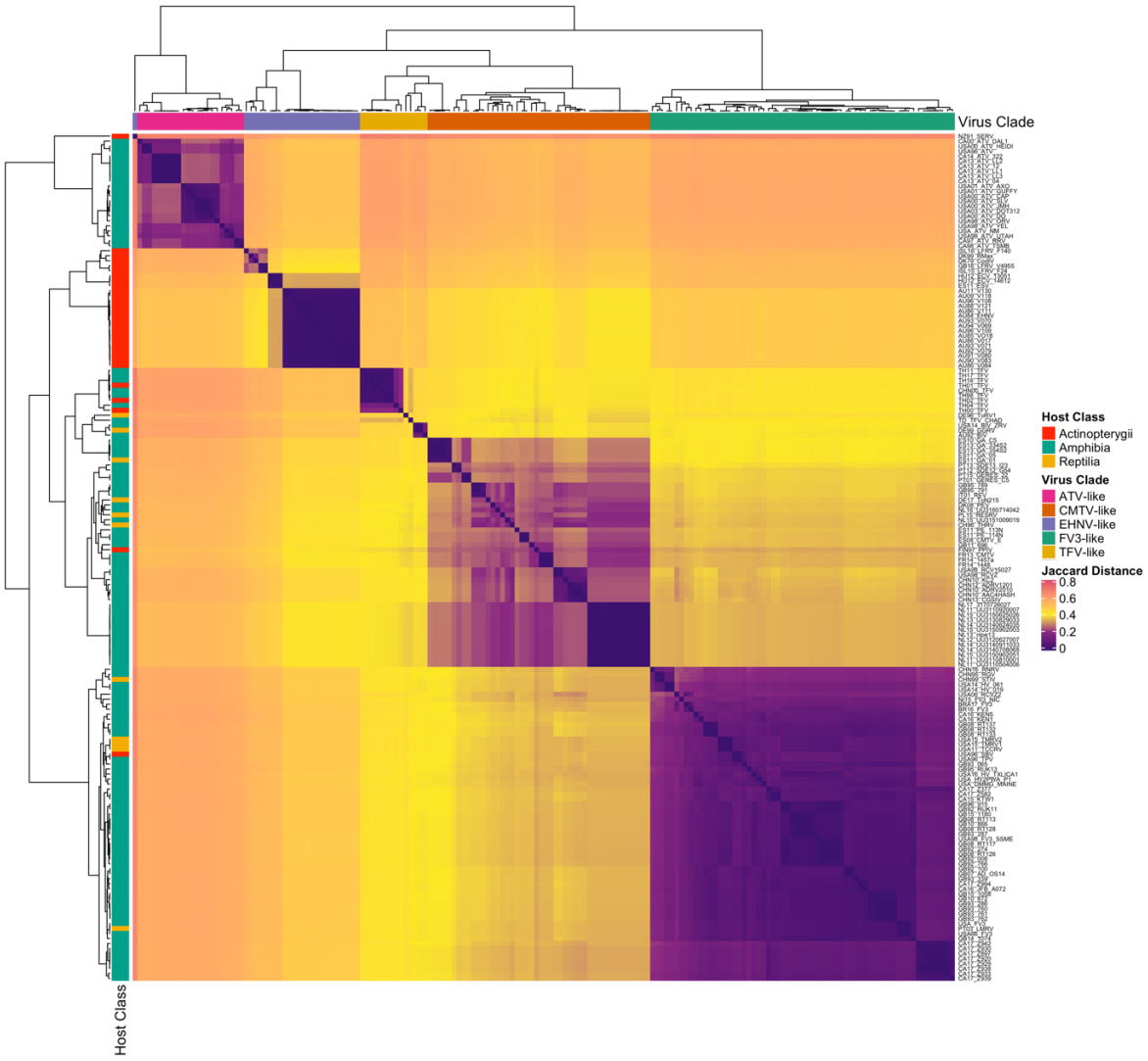


***Figure* 2.5**. *Heatmap of pairwise genetic distances between* Ranavirus *whole-genome assemblies. The 170 ALRV assemblies were decomposed into their constituent set of 12-mers using* Mash*, from which pairwise Jaccard Distances were derived; the heatmap is coloured purple for less distance (greater genetic similarity) to red for greater distances. Rows and columns are symmetric, and are ordered according to hierarchical clustering of pairwise distance. Top annotations are of ranavirus clade, whilst left annotations denote host taxonomic class. The FV3-, CMTV-, and ATV-like all comprise clear groupings, whilst the EHNV- (top blue annotation) and TFV-like (top yellow annotation) have clearly identifiable substructures reflecting their paraphyletic grouping.*

different taxonomic host class. One exception is the Pike perch iridovirus (PPIV; KX574341.1) isolate, the only CMTV-like of piscine origin, which interestingly contained a marked increase in genetic similarity to isolates within the FV3- and TFV-like clades.

After deriving Mash Distances from decomposed 12-mer sets of the ALRV assemblies, I produced a NJ-tree that exhibited two major topographical rearrangements from the alignment-based core phylogeny (Fig. 2.6). Both involved the EHNV- and TFV-like clades, and although they remained paraphyletic, their groupings were resolved more concordantly. For the EHNV-like clade, the ESV/ECV and EHNV isolates switched positions to cluster as derived subclades to the *European North Atlantic ranaviruses*, and as a sister clade to the ATV-like. Next, the TFV-likes grouped together (including ToRV1) basal to the CMTV- and FV3-like. The remaining AARV clades were again resolved monophyletically in the same relative topology observed by the by the core ML phylogenic reconstruction, with slight rearrangements within each clade.

Rearrangements within the FV3-like showed an interesting pattern of isolates from Asian and captive origin as most basal, to North American, then European isolates as predominately derived, whereas a greater degree of geographic paraphyly between continents was present in the ML core phylogeny. Similarly, within the CMTV-like clade, the Adrias davidanus ranavirus (ADRV) isolates sampled from captive-bred Chinese giant salamanders also fell as the most basal CMTV-like lineage, which conflicted with their derived position in the ML core reconstruction.

Genetic ordinations. Principal coordinate analysis of Jaccard Distances, derived from the shared 12-mers of the WGS assemblies and the core genome alignment, revealed clade groupings that were much more consistent based on whole-genome information. In principal coordinate space of WGS Jaccard Distances (Fig. 2.7 **A**), the ATV-, CMTV-, and FV3-like clustered as distinct individual groups, whereas the EHNV- and TFV-like clades contained slightly more within-group variation. Both patterns reflect their respective mono- and paraphyletic tree structures – the three sub-lineages of the EHNV-like could clearly be discerned. Importantly, however, the TFV-like clade grouping remained distinct from other clades, and clustered most closely to the CMTV-like, whereas the FV3-likes clustered distantly by a PCoA score of approximately 0.1 on the second axis.

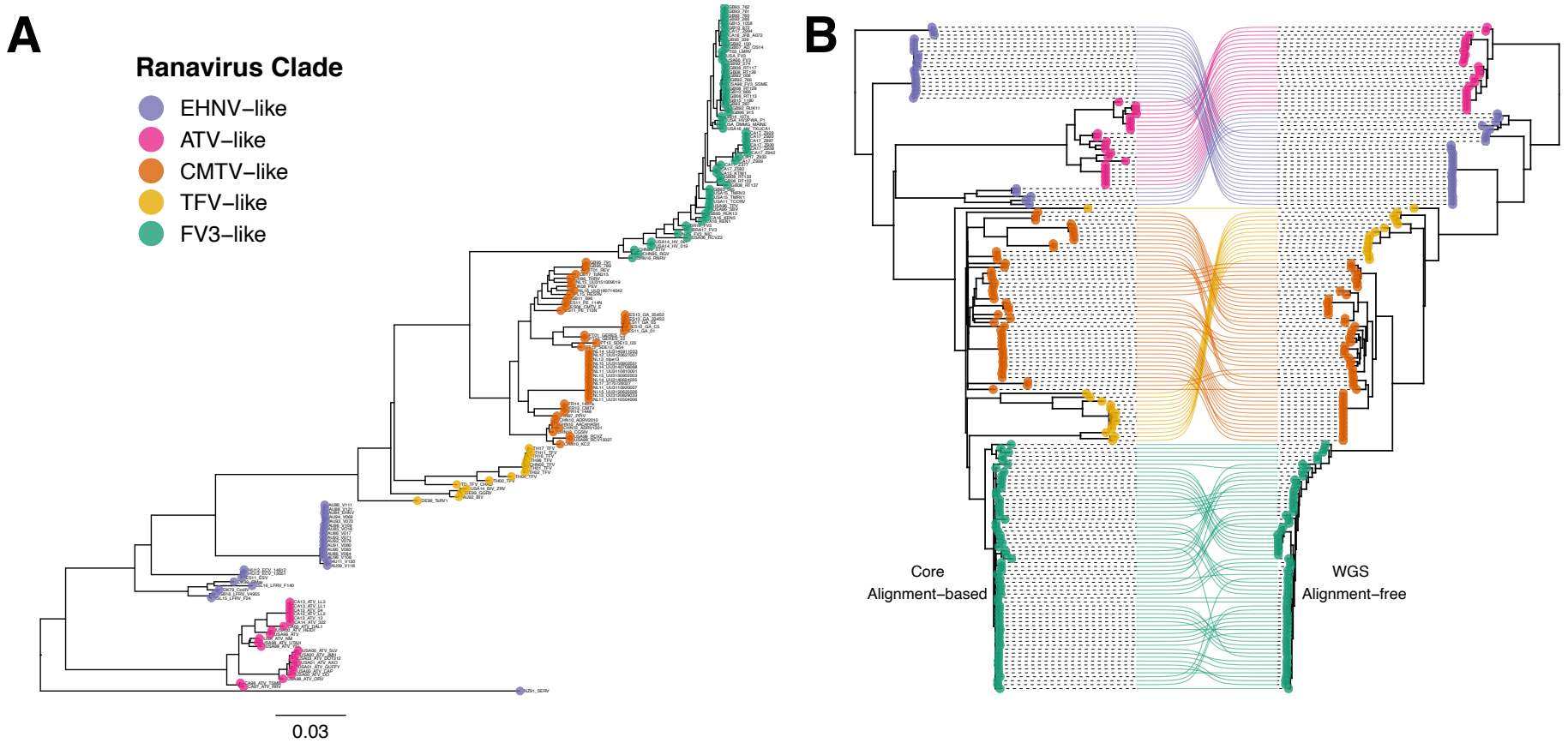


Figure 2.6. *Ranavirus whole-genome alignment-free tree. Panel **A** shows a Neighbor-Joining tree based on pairwise Mash Distances generated using matched 12-mers. Panel **B** is a cophylogeny displaying the differences in tree topology between the alignment-based ML phylogeny and the alignment-free Mash tree of panel **A**. To aid in visual comparison, both trees have had the SERV root isolate truncated together with node rotation to optimise congruency. Note the major positional rearrangement of the TFV-like (yellow) from a CMTV/FV3-like intermediate position on the ML phylogeny, to a basal position outside both these clades. Another major rearrangement involves the EHNV and ECV/ESV isolates towards a closer, yet still paraphyletic, grouping with the North Atlantic marine ranaviruses. Further minor rearrangements occur, but the stay contained within their respective clades.*

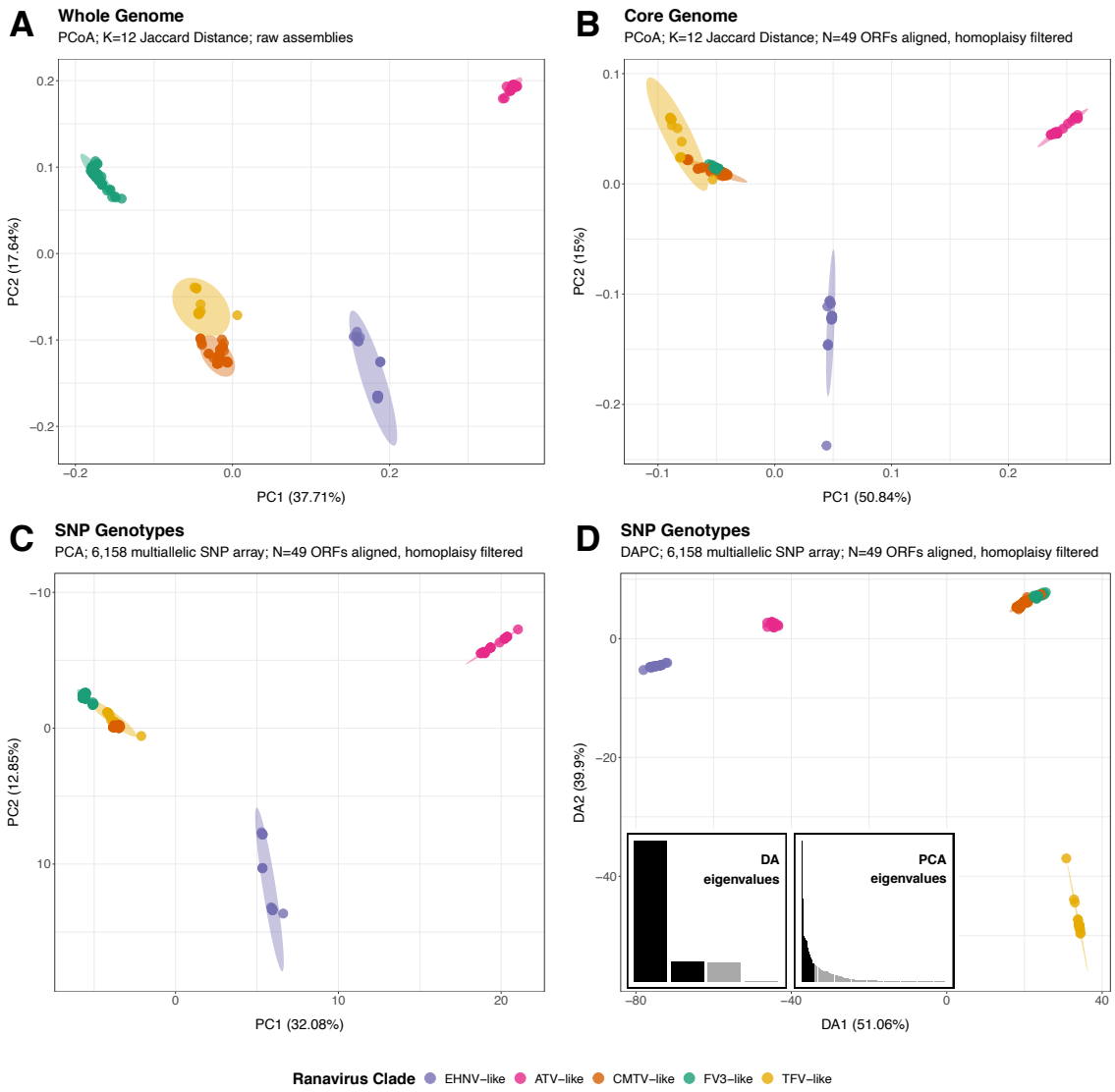


***Figure 2.7**. Genetic distance ordinations of amphibian-like ranaviruses. Panels **A** and **B** show Principal Coordinate analyses of Jaccard Distances derived from shared 12-mer sets of decomposed whole genome assembles and the 49 ORF core genome homoplasy-filtered alignment, respectively, the latter of which was used to build the ML core phylogeny (Fig. 2.4). Note that the distances yielded with only core genome information (39.54% of the average whole genome) in panel **B** fails to resolve the genetic diversity between the FV3-, CMTV- and TFV-like ranaviruses, when compared to considering the whole genome. Panel **C** shows the Principal Component analysis ordination of the multiallelic SNP array derived from the same core genome aliment used to produce panel **B**. Note the further reduction in differentiation of the CMTV/FV3/TFV-like clades. Panel **D** then illustrates a DAPC performed on the same SNP array using four discriminant functions on 13 retained PCs, which emphasises between-clade variance, and adequately resolves the unique diversity of the TFV-like ranaviruses.*

This pattern was not seen on the PCoA of core genome Jaccard Distances (Fig. 2.7 **B**), which demonstrated almost no between-group variance of the CMTV-, TFV-, and FV3-like clades, which clustered on top of each other even when considering the 3rd PC axis (not shown). The latter pattern is consistent with the consideration of the TFV and BIV isolates being 'FV3-like', as they are currently classified, and as is reflected by their position on the core ML phylogeny. However, this narrative becomes incompatible when the whole-genome is considered in assessing clade structures.

The distinctiveness of the TFV-like clade was further reflected by DAPC analysis of SNP genotype data. Panel **C** of Figure 2.7 shows a PCA of the SNP array, whilst panel **D** shows the subsequent discriminant analysis of the principal components. DAPC maximises between-group variance, and the optimum number of PCs to retain for the analysis was suggested to be 13 based on the *ad hoc* optimum *a*-score analysis. The resulting DAPC ordination showed a large degree of differentiation of the TFV-, EHNV-, and ATV-like genotypes, whilst the variance between the CMTV- and FV3-like was negligible. This pattern was in line with DAPC performed on clade gene presence-absence, except for the extreme variance observed for the TFV-like.

2.5 Discussion

Despite almost five decades of research (Granoff et al., 1965), the genetic diversity of *Ranavirus* remains poorly defined compared to related viruses studied over the same timeframe. Early characterisations of novel isolates employed what were then cutting-edge molecular techniques, including RFLP typing (Mao et al., 1997) and classifying structural variants upon the generation of the first genomic sequences (Jancovich et al., 2003). Unfortunately, much of the current considerations of ranavirus classifications still rely on these early findings and methodologies. What is more, where contemporary phylogenetic approaches have been applied, they have largely been based on a highly limited sample both in terms of the number of isolates characterised and the amount of genomic information incorporated (Epstein & Storfer, 2016; Stöhr et al., 2015).

The goal of this chapter was to quantify and describe the global genetic diversity of *Ranavirus*, making use of the largest collective genomic sampling effort available. Key to achieving this was the demarcation of the pan-genome content of the genomic assemblies. After elucidating the 49 genes present in all isolates –

the ALRV core genome – I applied conventional alignment-based Maximum Likelihood techniques to resolve the global phylogeny of *Ranavirus*, using the largest-possible amount of genomic information. I then characterised the systematics of *Ranavirus* to draw taxonomically informative classifications, based on the observed phylogenetic lineages in combination with alignment-free characterisations of the whole genome assemblies. In doing so, I uncovered a hitherto poorly defined genetic diversity of distinct ranavirus lineages, which underscore much of the confusion present in *Ranavirus* taxonomy.

2.5.1 *Ranavirus* Systematics and Taxonomy

It is important to note that in resolving the global genetic diversity of *Ranavirus*, I did not consider the highly divergent lineage of the SGIV-like basal fish-associated viruses that are currently considered members of the genus (Chinchar et al., 2017a). My decision to exclude the lineage was based primarily on the fact that no CDS sequences of its isolates clustered with other ranavirus orthologs at 80% amino acid homology. Further, *ad hoc BLASTn* searches failed to detect any nucleotide homology at as low as 30%, except for LMBV (MK681856.1), which contained three conserved genes with ~75% homology to the ALRV reference gene sequences. However, amino acid homology of this lineage to other ranaviruses does exist (Qin et al., 2003), and therefore my results suggest a very high degree of synonymous codon substitution. As such, the amino acid homology clustering threshold would had have to have been significantly lowered to include these isolates in the analyses of this thesis. Epstein & Storfer (2016) used an amino acid homology cutoff of 60% to include the SGIV-like, which delineated a *Ranavirus* core genome comprising 17 ORFs, whilst another study found SGIV-like nucleotide homology to ALRV genes to be < 35% (Stöhr et al., 2015). This suggests that a significantly lower threshold would be required to include even the 26 core *Iridoviridae* genes (Eaton et al., 2007). What is more, these viruses contain no collinearity in genomic arrangement with any other ranavirus isolates (Song et al., 2004; Tsai et al., 2005), which Chinchar et al. (2017b) suggest could be used as grounds to classify the lineage as a sister genus to *Ranavirus*. Altogether, these were the reasons for excluding the SGIV-like, particularly with the mind to maximally resolve the genetic diversity of the remaining amphibian-like ranaviruses that comprise the majority of the genus sampled to date.

In total, I reconstructed the genetic diversity of 170 amphibian-like *Ranavirus* isolates, which I classified into five taxonomically informative groups. Three of which are the well-established monophyletic lineages of the ATV-, CMTV-, and

FV3-like (Epstein & Storfer, 2016; Jancovich et al., 2015a; Price, 2015), whereas the remaining two classifications were the paraphyletic groups of the EHNV- and TFV-like. Both of the latter contained a substructural diversity of distinct lineages, each with different host associations and geographic ranges. Despite this, I justified the paraphyletic groupings based on whole genome characterisations, such as how the respective assemblies clustered together in genetic ordinations relative to all other lineages, including in whole-genome principal coordinate space and core SNP DAPC (Fig. 2.7 **D**). Nevertheless, it is clear the paraphyletic groups I define may comprise more than one ranavirus species, whereas the monophyletic clades likely constitute an evolving lineage of one distinct species each.

My drawing of two paraphyletic groups serves as an *incertae sedis* descriptor of the associated lineages within, of which each would benefit from additional sampling effort and characterisation in order to be fully resolved. For instance, my conclusion that the paraphylies contain multiple species diverges from the current ICTV species classifications of *Ranavirus*. The three lineages that I define as the EHNV-like comprise only two current formal species of the *Epizootic hematopoietic necrosis virus* and *European North Atlantic ranavirus* (Langdon et al., 1986; Stagg et al., 2020), yet the ESV and ECV isolates that form the additional monophyletic group are not officially recognised as a distinct taxon from *Epizootic hematopoietic necrosis virus* (Chinchar et al., 2017a).

The non-species status is also true for the constituent lineages of the paraphyletic group I define as TFV-like. The two lineages that are comprised of the respective TFV and BIV isolates (the latter including German gecko ranavirus KP266742.1 and Zoo ranavirus MK227779.1) are currently classified as members of the *Frog virus 3* species, despite their unique genetic diversity. For instance, these lineages were placed on highly divergent long branches on the core phylogeny (Fig. 2.4), and the average Jaccard Distance of each TFV-like whole genome assembly was closer to all CMTV-like assemblies than to all FV3-likes (0.403 verses 0.413, respectively). I further observed the same pattern for the isolate ToRV1 (KP266743.1), which is also currently classified as FV3, despite conflicting reports of its placement within the CMTV clade (Ferreira et al., 2021; Stöhr et al., 2015; Yu et al., 2020). This confusion likely arises from the fact this virus exhibits hallmarks of a mosaic genomic makeup from both CMTV- and FV3-like origins (Price, 2015), so its phylogenetic placement may alter depending on what genes are considered for analysis. Overall, however, the whole genome complement of ToRV1 had an

average Jaccard Distance of 0.392 to the CMTV-likes, compared to 0.429 to the FV3-likes, suggesting that it perhaps contains a greater proportion of CMTV-like genomic segments.

As the TFV and BIV isolates grouped with ToRV1 in whole-genome characterisations, it is possible the TFV-like comprise a group of highly mosaic viruses. Indeed, the majority of TFV-like isolates originate from captive settings; only TFV Chad (MW727505.1) was sampled from wild-caught animals (Box et al., 2021). It is possible that divergent ranaviruses co-circulating in captive settings and trade networks could increase the chance of otherwise unlikely coinfections of single hosts, and therefore the provide an environment for recombinant strains to emergence (Bowden et al., 2004; Claytor et al., 2017; Pisoni et al., 2007; Price, 2015). Further investigations into the degree of polyphyly amongst these lineages is required to determine the validity of this hypothesis. It is also not likely the TFV-like are completely mosaic, based on the fact they have a substantial private accessory genome complement. As such, an alternative hypothesis is that the TFV-like are endemic viruses native to Southeast Asia and Oceania that have naturally accrued genetic diversity in isolation of the FV3-like siter lineage. Of course, hypothesis of recombination-generated diversity versus natural diversity acquired in isolation by distance need not compete to be mutually exclusive. It is eminently possible both these processes have played a role in generating the unique complement of TFV-like genetic diversity, though further investigation is needed to determine their relative contributions.

Recent steps by the ICTV have been made towards a more quantitative approach for the classification of viral species (Gibbs, 2013; Peterson, 2014). Of chief importance within the revised framework is the monophyletic grouping of taxa for a virus species to be defined. Indeed, it is due to this that the TFV and BIV isolates are classified as belonging to the FV3 species. Despite their highly divergent distances, these isolates can be considered monophyletic with FV3-like in core genome phylogenies (Fig. 2.4; Price 2015; Stöhr et al., 2015; Epstein & Storfer 2016; Chinchar et al., 2017a). However, the monophyly is broken when considering the whole genome, as observed from the basal repositioning of the TFV-like in the Neighbour-Joining tree of Mash Distances, which simultaneously retained an intact monophyly of the FV3-like proper and the CMTV-like (Fig. 2.6). This is highly suggestive of a bias that is introduced when the *Ranavirus* phylogeny is built using only core genomic information, which, at most, represents less than half of the average ranavirus genome.

The extent of the core genome bias is demonstrated in Figure 2.7. Panels **B** and **C** show ordinations of the 49 core ORF alignment ascertained via PCoA of 12-mer Jaccard Distances and PCA of the core SNP array, respectively. Both reconstructed the core diversity very similarly, as would be expected, which importantly show an almost complete overlap in the genetic distances between the CMTV-, TFV-, and FV3-like groups. These clades however fall as distinct, separated groups when distances are drawn using all genomic information (Fig. 2.7 **A**). This highlights the source of confusion introduced in definitions of these lineages, which stems from the lack of genetic diversity in higher taxonomic raking core genes. It seems likely these core genes contain idiosyncratic evolutionary histories that do not well characterise the trajectories taken by the derived amphibian-associated lineages.

I observed definitive monophylies for the ATV-, CMTV-, and the remaining FV3-like ranaviruses (excluding the TFV and BIV linages). These lineages represent well characterised species of *Ranavirus,* each containing very distinctive traits and population demographics (Chinchar et al., 2009; Jancovich et al., 2003; Jancovich et al., 2015a; Price et al., 2014; Price et al., 2017a). The ATV-like are most basal, and entirely restricted to North America. Despite successful infection following experimental challenge of anuran amphibians, they appear to be restricted to the single genus of *Ambystoma* salamander hosts in the wild (Jancovich et al., 2001). Previous studies have suggested an ancient association of the ATV-like with their hosts based on their phylogenetic concordance and apparent coevolution (Storfer et al., 2007). The results I present support this, as the ATV-like contain a unique and ancient genetic diversity in terms of their gene content and the long branch of private polymorphisms on the core phylogeny.

Similarly, the CMTV-like are restricted predominantly to mainland Europe in the wild. However, in complete opposition to the ATV-like, that have a staggeringly broad host range, often infecting entire diverse amphibian communities during outbreaks (Price et al., 2014). Despite CMTV being one of the most recently discovered species, they form the basal lineage to the derived superclade of amphibian-associated ranaviruses (AARVs), together with the TFV- and FV3-like, suggesting they contain an ancient association to Europe. I base this on the partitioning of the genetic diversity within the lineage; on both the alignment-based and alignment-free trees, isolates grouped in strong, geographically defined subclades situated on long branches that each lack structure relating to host type or sampling date (Fig. 2.4; Fig. 2.6).

Despite containing an almost equally broad host range, the FV3-like stand in stark contrast to the CMTV-like in terms of genetic diversity. They contain the shallowest diversity of all lineages (Fig. 2.5), yet the largest geographic range across North America, Europe, and Asia. This suggests very different demographic process during their evolutionary history, indicative of recent range expansion which is consistent with their invasion into the UK (Hyatt et al., 2000; Price et al., 2016). What is more, the most basal isolates to the FV3-lineage comprise a subclade of two novel North American isolates together with the Chinese RGV and STIV ranaviruses. The latter are often characterised as distinctive ranaviruses potentially deserving of species status, based on their geography and STIV's host association to a chelonian (Huang et al., 2009; Jancovich et al., 2015a; Zhang et al., 2001). However, by comparison, the average WGS Jaccard Distances *within* ATV- (0.168) and the CMTV-like (0.224) lineages is similar, or far exceeds, that of the average distances RGV and STIV have to sister members of their own lineage (0.179 and 0.174, respectively), suggesting they are firmly of the FV3 species.

2.5.2 *Ranavirus* Pan-Genome

Amongst the 170 ALRV isolates, I delineated a total of 217 unique orthologous gene clusters. This marks a substantial increase on previous assessments, such as the last effort in which Price (2015) determined 130 orthologs from 17 genomes across the ALRV phylogeny. Other studies that have conducted whole genome analyses of *Ranavirus* predicted ORFs and transferred existing annotations for each genome based on *BLAST* homology searches, but no clustering steps were reported beyond identifying core orthologs (Epstein & Storfer, 2016; Stöhr et al., 2015). Of the pan-genome I characterise, I found 49 genes to be core to all isolates, which is slightly fewer than have been previously identified for the ALRV. Epstein & Storfer (2016) and Price (2015) report 52 and 51 ALRV core ORFs at 80% amino acid homology, respectively. However, these studies considered fewer isolates, and so it is likely that a small fraction of the many more genomes included here lacked two-to-three genes, which could be due to sequnecing/assembly errors or natural gene truncations which are known to frequntly occur (Lung et al., 2021; Morrison et al., 2014).

Together with the large increase in the number of described ranavirus genes (pan-genome), I also report a vast improvement on the number of functional annotations. To date, annotation has been provided for approximately 30% of

described *Ranavirus* genes (Price 2015; Epstein & Storfer 2016; Chinchar et al., 2017b), which I update to 69.4% of core genes, and 55.3% of the pan-genome (Table S3). Functional annotations of particular note include increased detection of US22 family proteins. I found eight across the pan-genome > 20% divergent from each other in amino acid substitutions, which therefore do not represent paralogs of distinct types. The US22 family are increasingly thought to be important host-acquired virulence factors involved in immune evasion (Carstairs et al., 2020; Chen et al., 2013; Zhang et al., 2011). Although no US22 protein was included in the core genome, each clade contained at least one (TFV-like), and up to five (EHNV- and FV3-like), US22 protein families. The EHNV-like interestingly have two distinct US22 proteins (*Roary* groups 175 and 196) in high abundance (present in 20 and 16 isolates, respectively) that are not found in any AARVs, and thus are likely specific to their fish hosts. The FV3-like ranaviruses also contained two US22 proteins (groups 202 and 208) private to their lineage, but at lower abundance.

Of all the ALRV clades, the paraphyletic EHNV-like contained the largest and most distinct complement of genes. The median number of predicted ORFs for each EHNV-like isolate was 107 (range 98–109), which differs from previous estimates, such as Jancovich et al. (2015b) who report ESV alone to contain 136 predicted ORFs. Nevertheless, the large excess of ORFs private to the EHNV-like is likely a reflection of adaptations to or genes acquired from the genetically diverse hosts that the paraphyletic group infects. For instance, even within the piscine taxonomic class of the Actinopterygii, the genetic distance between the orders of Anguilliformes, Salmoniformes and Perciformes that the EHNV-like infect is substantial, traversing basal to derived positions on a phylogeny that comprises nearly half of all living vertebrate species (Near et al., 2012).

DAPC analysis also described a very distinct set of genes for the ATV-like, but this was likely a product of the fact they contained the smallest complement of 109 unique ORFs, and therefore the greatest degree of gene absences. Furthermore, despite my result of an average of 91 ORFs per ATV assembly, 109 unique genes within the ATV-like is well within the range of predicted ORFs that has been reported within a single ATV genome previously (Epstein & Storfer, 2016; Jancovich et al., 2015b). Nevertheless, if total gene content reflects adaptation to host genotype diversity, then the small complement of ATV genes is in line with their highly restricted host range (SHI of host species diversity: 0.923).

Finally, the CMTV-, FV3- and TFV-like all contained a similar complement of total and private genes. Again, this is in line with the fact that these clades had a similar degree of host species diversity (SHI range: 2.204 – 2.448), which also spans three vertebrate classes for each clade (Fig. 2.2). However, these clades also contained large overlap in the variance of their gene presence/absences (Fig. 2.3 **A**), which could be due to their recent, close evolutionary relationships, or an indication of a high rate of recombinogenic gene conversion events between these lineages.

2.5.3 Limitations and Future Directions

A key limitation to dataset of whole genome sequence presented here is reflected in the degree of structured sampling. Despite comprising 170 isolates across a global distribution, the majority of samples originated from approximately five studies of large concerted sampling efforts (EHNV-like: Hick et al., 2017; ATV-like: Epstein & Storfer 2016; CMTV-like: Saucedo et al., 2018; FV3-like: Vilaça et al., 2019; TFV-like: Sriwanayos et al., 2020), in addition to the 55 novel CMTV-like and FV3-like genomes contributed by the present study. The remainder of assemblies represent a sporadic collection of genomes sampled opportunistically, many of which were sequenced from culture with an unknown number of passages (e.g., Tan et al., 2004; Price 2013). Furthermore, for the majority of available genomic assemblies, the authors have not provided the short-read sequence data, meaning inferences drawn from collections of such genomes are vulnerable to biases introduced from the multiple and varied assembly protocols used to produce them. Both these factors serve as potential confounders to describing the natural genetic diversity of *Ranavirus*. However, these limiting factors could be circumvented by establishing future initiatives of targeted genomic surveillance, together with requirements for raw sequence and meta data to be made transparent and available.

From the results I present in this chapter, it is apparent that the ranaviruses contain large foothold of ancient natural endemism in disparate regions of the globe. However, ranavirus emergence in geographic regions considered historically free from their presence is often purported to have occurred through human introduction, such as CMTV in Iberia (Price et al., 2014), and the FV3 in Asia (Zhang et al., 2001; Huang et al., 2009; Kwon et al., 2017). On the one hand it is clear ranaviruses have been anthropogenically introduced beyond their natural range (Hyatt et al., 2000; Price et al., 2017b; Majji et al., 2006). However, I present further lines of evidence of deep evolutionary diversities associated with distinct geographic regions, particularly amongst the CMTV-like in Europe and the TFV-

like in Southeast Asia and Oceania (despite captive association), which oppose recent invasion.

In addition, an increasing number of studies have demonstrated the importance of changing environmental conditions in influencing disease outcomes, such as temperature (Price et al., 2019) and host-microbiome interactions (Campbell et al., 2019). These findings open a door to the possibility that instances of ranavirus-associated disease outbreaks and emergence may often occur as a result of perturbations to endemic pathogen dynamics, rather than simply novel virus introductions (Rosa et al., IN REVIEW). Future research would benefit from undertaking surveys of ranavirus distributions with designs explicitly to test hypotheses surrounding natural endemicity (Campbell et al., 2018). Ideally, focus should be directed at transect or randomised sampling approaches to ranavirus genomic surveys in natural host populations of poorly characterised regions, such as Asia. Shifting importance away from the current reliance of sampling in response to outbreaks and animals exhibiting gross signs of disease, and predominantly those associated with aquaculture, is needed to move towards a more complete understanding of ranavirus genetic diversity and phylogeography.

Finally, it is well understood that ranaviruses recombine, both from *in vitro* and *in situ* lines of evidence (Chinchar & Granoff, 1986; Claytor et al., 2017; Vilaça et al., 2019). What is not clear are the rates at which ranaviruses recombine, either within, and particularly, between lineages. Understanding the extent to which the latter occurs is important to address whether recombination plays a significant role in generating ranavirus genomic diversity. As previously mentioned, the unique diversity of ToRV1 appears to be due to a polyphyletic evolutionary history (Price, 2015). Other examples of between-lineage gene conversion exist, such as with the isolate RCVZ2 (MF187209.1), which additionally has been shown to contain a more virulent phenotype (Claytor et al., 2017). Further research on the rates of ranavirus recombination *per se* is needed, which can now be conducted given the wealth of genomic information available.

2.5.4 Conclusions

By improving the genomic resolution with which to characterise *Ranavirus* genetic diversity, I have revealed intricacies to the evolution of the genus not previously quantified. Specifically, my phylogenetic characterisations suggest that *Ranavirus* contains several miss- and uncharacterised species given the emphasis the ICTV

now gives to the monophyletic grouping of taxa, together with distinct host and geographic associations (Peterson, 2014). In particular, the EHNV-like paraphyletic clade appears to contain an additional species (ESV/ECV) to the two currently recognised (EHNV and ENARV). The TFV-like, however, likely contains two species (TFV and BIV) that are currently assigned as FV3-like (Chinchar et al., 2017a). An additional third isolate, ToRV1, groups with the TFV-like. However, whether it represents a true species cannot currently be determined given its seemingly mosaic origins, together with the undefined role that recombination may play in generating ranavirus diversity. Further, whether the SGIV-like represent a sister genus to ranaviruses was not the aim of this study. However, the two-fold reduction in genomic material with reasonable homology for inclusion in phylogenetic reconstruction (Stöhr et al., 2015) suggests the lineage should not be considered in genus-wide characterisations of *Ranavirus*. The great confusion shrouding the field of ranavirus taxonomy stems solely from the ill-defined framework for classification of the genus, with considerations derived from largely obsolete molecular methodologies, and vague, artificially determined thresholds. Quantitative phylogenetics employed with whole genome approaches are now widely available and applicable, and as such, the ICTV should revise its recommendations regarding ranavirus systematics based on current genomic standards.

3

High Rates of Ancestral Recombination Provide a Major Process of Genomic Diversification in *Ranavirus*

3.1 Abstract

Genetic recombination is an important feature of DNA virus evolution. It is a process where novel combinations of existing nucleotide variants can be introduced into a single genome, which can rapidly generate genetic diversity despite slow mutation rates. Recombination provides a means for viruses to explore phenotypic space, and has been implicated in the emergence of virulent lineages, evasion of host immunity, and the expansion of host range. *Ranavirus* is a genus of important DNA viral pathogens of poikilothermic vertebrates, which are known to recombine. However, the recombinogenic rates of ranaviruses have not been explicitly investigated, nor have the implications that the process may have both for the evolutionary history of the genus and its contribution to risks posed to their hosts. In this chapter, I demonstrate high rates of ancestral recombination amongst all *Ranavirus* lineages. Specifically, each clade of amphibian-like ranaviruses contained isolates with evidence of polyphyletic gene histories originating across distinct lineages. The most extreme degree of polyphyly was observed amongst ranaviruses associated with captive hosts, with many isolates containing nearly complete phylogenetic gene assignments to clades other than their own. High rates of within-lineage recombination were also detected, with patterns of linkage disequilibrium decay along the length of genes, which were congruent with locations of recombination breakpoints detected through phylogenetic incongruency methods. High rates of intragenic recombination were detected in genes with ATP-hydrolysing functions,

suggesting that recombination may be associated with purifying selection. Altogether, these results suggest recombination is an integral facet of ranavirus evolution, standing as a dynamic process involved in both generating genetic diversity and possibly maintaining protein stability. This raises implications for ranavirus taxonomic classification considerations, as well as the risk of novel diversity emerging through coinfection of different lineages in host animal trade networks.

3.2 Introduction

Genetic recombination is a pervasive feature of DNA virus evolution. Whereas the evolutionary fate of RNA viruses is likely mostly decided through their high mutation rates, DNA viruses trundle by comparison (Sanjuán et al., 2010), and therefore may often rely on acquiring genetic diversity by means faster than their rate of mutation alone. Homologous recombination is the main process by which this is achieved (He et al., 2010; Pisoni et al., 2007). Mutation is the ultimate source of genetic diversity, in which genetic variants are introduced into the population through nucleotide substitution. Recombination is a process that subsequently permits the movement of variants across genomes, introducing new combinations of existing mutations into one molecule (Simon-Loriere & Holmes, 2011). High rates of recombination can therefore provide the necessary raw material for the genetic diversity required for rapid viral adaptation to proceed in the face of the relatively slow mutation rates of large DNA viruses (Sanjuán & Domingo-Calap, 2016).

Viral recombination takes place when two or more virus genomes infect a single host cell and exchange genomic segments. The likelihood of co-infection depends on several biological and epidemiological factors of both virus and host, such as the geographical distribution of hosts, prevalence and co-circulation of viral strains, tissue tropism, and viral loads during infection (Pérez-Losada et al., 2015). Homologous recombination specifically occurs at the same locus in the parent genomes, and is facilitated by high relatedness and homology between the two coinfecting viral strains for a viable recombinant to be produced (Galli et al., 2010). Non-homologous recombination of segments from different loci can also occur, however, this usually results in aberrant, non-functional genomic structures (Galli & Bukh, 2014). Nevertheless, genetic recombination allows viruses to explore adaptive and phenotypic space associated with the emergence of new and often more virulent lineages (Combelas et al., 2011; He et al., 2010), evasion of host

immunity (Elde et al., 2012), and the expansion of host range (Croizier et al., 1994; Kim et al., 2000; Kondo & Maeda, 1991).

There are several mechanisms of viral recombination, which differ for RNA and DNA viruses. Two main mechanisms exist for RNA viruses, which involve template switching during genome replication (Aaziz & Tepfer, 1999), and reassortment of segmented genomes in virion packaging (McDonald & Patton, 2011). Most of the knowledge surrounding recombination in DNA viruses, however, comes from Herpesvirus-1 (HSV-1) due to its extensive use as a model organism (Pérez-Losada et al., 2015). In the HSV-1 system, homologous recombination is intimately associated with genome replication and DNA repair pathways (Weller & Sawitzke, 2014), and further appears to be an important process for the maintenance of viral genomic stability in response to DNA damage (Wilkinson & Weller, 2004). However, other model DNA viruses, including lambdoid bacteriophages and adenoviruses, share the strategy of using the same pathways and enzymes during DNA replication, repair, and recombination. It is therefore thought that the interconnectedness of these processes stands as an ancestral mechanism for recombination in double-stranded DNA (dsDNA) viruses (Kamita et al., 2003; Robinson et al., 2011; Weller & Sawitzke, 2014). Random cleavage of DNA during replication is common (Kass & Jasin, 2010), and is an integral process to the replication of the HSV-1 genome which proceeds via means of a continuous rolling circle concatemer (Skaliter & Lehman, 1994). The DNA double-stranded breaks produced during HSV-1 genome replication therefore initiate cellular repair pathways via homologous recombination (Weller & Sawitzke, 2014). Given their seemingly conserved nature of replication, gene conversion of dsDNA virus genomic segments is likely to ubiquitously occur if a coinfecting strain is present in the host cell, due to the opportunity provided by an alternative homologous template for use in routine replicative repair.

Ranavirus is a genus of dsDNA virus with poorly defined rates of genetic recombination. They belong to the virus phylum of the *Nucleocytoviricota* (also known as the nucleocytoplasmic large DNA viruses; NCLDV), which contains many members known to frequently recombine, including poxviruses (Filée, 2013; Iyer et al., 2006; Koonin et al., 2020). Like their NCLDV relatives, ranaviruses also contain genomic signatures which suggest recombination has played an important role in their evolution and life history. For example, they contain linear dsDNA genomes with circularly permuted gene arrangements, and coding sequence transcription does not proceed in a linked manner, suggesting

individual genetic units prone to frequently recombine (Chinchar, 2002; Mesnard et al., 1988). Furthermore, ranavirus recombination has been demonstrated *in vitro* through the use of temperature sensitive mutants (Chinchar & Granoff, 1986), as well as through more recent *in situ* evidence from both wild (Vilaça et al., 2019) and captive host animals (Candido et al., 2019; Claytor et al., 2017; Ferreira et al., 2021). Lastly, recombination has been detected *in silico* amongst genomes of disparate origin for the main purpose of removing its signal in phylogenetic reconstruction (Price, 2015; Stöhr et al., 2015).

Quantifying the rates of recombination amongst *Ranavirus* has important practical implications. First, there is the general constraint recombination bares on analytical frameworks that rely on evolutionary rate estimations, including selection analyses and phylogenetics, as alluded to above. For instance, modern phylogenetic methodologies can only adequately reconstruct genetic diversity derived from a single evolutionary history. If recombination has occurred between genetic lineages, phylogenetic reconstruction of the genotypes involved will result in inaccurate tree topology (Pond et al., 2006a). This can be particularly problematic for viral systems that are highly mosaic in nature, such as HIV-1 which contains rates of recombination rivaling that of mutation (Zhuang et al., 2002). Therefore, quantifying the rate of recombination is important for phylogenetic applications, such as systematics and taxonomy, so it can appropriately be controlled for by either removing its signal in sequence alignments, or analysing only recombinant-free genomic partitions (Pond et al., 2006b).

Second are the specific implications genetic recombination has to the evolution of *Ranavirus* itself. Ranaviruses are notifiable pathogens to the World Organisation for Animal Health (OIE, 2019) due to the impact they have on the natural biodiversity of their amphibian hosts (Cunningham et al., 1996; Daszak et al., 1999; Jancovich et al., 1997; Price et al., 2014), and severity of economic loss associated with disease outbreaks of farmed fish (Ariel et al., 2010; Langdon et al., 1986). As such, there is grave concern for the potential emergence of more virulent recombinant ranavirus strains that could further exacerbate these plights. In fact, it is a concern that has been realised; one documented instance of ranaviral recombination occurred between two distinct lineages belonging to the *Common midwife toad virus* (CMTV)- and *Frog virus 3* (FV3)-like species, whose recombinant daughter strain exhibited increased virulence compared to a prior outbreak of the presumed parent (Claytor et al., 2017; Mazzoni et al., 2009). Farming and culture

practices that cohouse animals in high densities, which may have disparate trade origins, increases the likelihood of coinfections from two or more viral genotypes and lineages (Bowden et al., 2004; Pisoni et al., 2007). However, assessing the risk this may pose to the emergence of new and virulent ranavirus strains – for instance, in drawing mitigative aquaculture and trade policy – is dependent on an understanding of the propensity of ranavirus to recombine, and the relatedness between lineages that are able to do so successfully.

Additionally, a large knowledge gap exists on the extent the role of recombination has played in the genome evolution of *Ranavirus*. For instance, it is not understood whether recombination has had a pervasive effect of segregating derived adaptive sites throughout populations within or between respective lineages. High rates of this process could have far-reaching implications for the modes of ranavirus adaptation, such as facilitating host range expansion (Kondo & Maeda, 1991). For example, it is well understood that there is an exceptional difference in host breadth between ranavirus lineages, from *Ambystoma tigrinum virus* (ATV) that naturally infects only one host genus (Jancovich et al., 2003), to others such as the CMTV lineage that is capable of not only infecting entire amphibian communities, but also three vertebrate classes (Price et al., 2014; Tapiovaara et al., 1998; von Essen et al., 2020). The genetic determinants and influencers for these differences in host range remain elusive. Moreover, whilst there is tentative evidence for a high degree of cross-lineage polyphyletic origins for the isolate ToRV1 (Price, 2015), there has been no explicit examination of the rate at which lineages diverge based on mutation alone or through ancestral recombination events of divergent strains.

In this chapter, I sought to examine the rates of *Ranavirus* recombination, in order to consider the extent to which the evolutionary history of the genus may have been shaped by its influence. Specifically, my aims were to: i) assess the recombination landscape between lineages of the amphibian-like ranaviruses (ALRV); ii) determine the rate and position of genomic loci involved in recombination events between isolates within the derived amphibian-associated ranavirus (AARV) lineages; and iii) characterise the genes that contain evidence of having undergone episodic diversifying selection across the phylogeny of the genus, as well as those specific to AARV clades. A central goal was to elucidate congruencies between recombinogenic and adaptive sites, and whether these processes follow key evolutionary events such as instances associated with host switching, whilst characterising the functional significance of implicated genes.

To achieve these aims, I applied both phylogenetic and non-phylogenetic recombination and selection analyses to the dataset of 170 ALRV isolates produced in Chapter 2.

3.3 Methods

3.3.1 Sequence Data, Clade Classification, Alignments and Pi

My acquisition of the 170 *Ranavirus* assemblies used in the following analyses is outlined in Chapter 2 (Methods 2.3.1), along with the methodological rationale by which each isolate was assigned membership to the *Ranavirus* clade classifications of the EHNV-like, ATV-like, CMTV-like, FV3-like, and TFV-like. Given that these five distinct taxonomic classifications contain both mono- and paraphyletic groupings (i.e., the EHNV- and TFV-like), it is often more appropriate to refer to them as either a true lineage or a paraphyletic clade, respectively. I therefore subsequently use the term 'clade' interchangeably, but 'lineage' to refer specifically to those that are monophyletic.

Given that amphibians are considered to be the derived host state of *Ranavirus* (Price et al., 2017a) and emerging pathogens in these hosts, I sought to assess the genomic evolution that may characterise ancestral host shifts and recent demographic change. As such, I conducted recombination and selection analyses at two resolutions, both at the level of the whole genus, and for the specific clades of amphibian-associated ranaviruses (AARV) consisting of the ATV-like, CMTV-like, FV3-like, and the TFV-like. To perform the focused clade-specific analyses, I utilised the isolate open reading frame (ORF) coding sequence (CDS) datasets previously generated by *Prokka* (Seemann 2014) to conduct a pan-genome analysis for each of the four AARV clades using *Roary* (Page et al., 2015). I performed the analyses as outlined previously (Chapter 2, Methods 2.3.2), using a homology threshold of 80%, to isolate the core genes specific to each clade with which to conduct the following genome evolution analyses on.

All concatenated and individual gene (genus-wide and clade-specific) multiple sequence alignments (MSA) were constructed in the same manner as in Chapter 2 (Methods 2.3.4), except for the step of removing recombination signal by homoplastic site filtering (see below). Briefly, all alignments were built using *MAFFT* (Katoh & Standley, 2013) and trimmed of gaps in ≥ 20% of isolates with *trimAL* (Capella-Gutiérrez et al., 2009). For each AARV clade, I used the metric pi to summarise the nucleotide diversity across concatenated core genome

alignments both comprising clade specific ORF sets, as well those common across the genus (ALRV core). To do so, I used the *R* package *PopGenome* v2.7.5 (Pfeifer et al., 2014) to estimate pi over a sliding window of 500 bp with a 125 bp jump. I then graphically summarised respective genomic diversities with custom genome schematics produced using *ggplot2* (Wickham, 2016), with which I also plotted recombination results generated below to characterise patterns of clade-specific AARV genome evolution.

3.3.2 Recombination Between *Ranavirus* Clades

SplitsTree. To assess the potential recombination landscape across the *Ranavirus* genus, I began by characterising phylogenetic incongruencies between isolates of all lineages and clades. Such conflicts, termed homoplasies, are recurrent mutations that arise independently at disparate positions on a phylogeny as a result of either convergent evolution or homologous recombination (Maynard Smith & Smith, 1998). I utilised the software *SplitsTree4* v4.16.0 (Huson & Bryant, 2006) to construct a genus-wide phylogenetic network highlighting where phylogenetic conflicts amongst *Ranavirus* exist. I provided *SplitsTree4* with the gap-trimmed MSA of all 170 isolates, and specified the estimation of pairwise genetic distances based on rates yielded using the HKY85 substitution model with a transition/transversion split ratio of 2.0 and equal base frequencies. After, I used the neighbor-net method of phylogenetic network reconstruction, in which 'split' edges are drawn between branches containing tree topology conflicts. Splits in the network therefore represent potential recombinogenic sites within the alignment, and between which sequences they occur.

Isolate gene histories. Another method of inferring ancestral recombination is by means of assessing genetic mosaicism. Specifically, the degree of detectable polyphyletic structure – multiple phylogenetic origins – of the viral genome. Various software exist that employ different methods to detect virus mosaicism, usually by scanning for statistically significant non-random patterns of homologous genomic regions in parent and donor sequences (Lam et al., 2018; Martin et al., 2015). However, due to computational and analytical limitations relating to operating system requirements of such tools and the structure of the genus-wide *Ranavirus* dataset, I developed and employed my own descriptive framework for assessing genetic mosaicism of each *Ranavirus* isolate.

I began by constructing individual gene alignments of each ALRV core ORF with all isolates using the method outlined above. I then built Maximum Likelihood

(ML) phylogenies from the resulting alignments using *RAxML* v8.2.12 (Stamatakis, 2014), specifying the *-m GTRCAT* model, as outlined in Chapter 2 (Methods 2.3.5). Within the ML trees of each gene, I assessed the phylogenetic placement for each isolate, and specifically, which *Ranavirus* lineage or clade each belonged to. The ideal method to carry this out would follow that of the clade classification rationale outlined in Chapter 2, by assessing the membership of isolates to monophyletic groupings. However, such a method would be unfeasible due to the manual tree inspection required. Moreover, the degree of paraphyletic topologies can vary greatly depending on the genetic diversity amongst individual genes, making it problematic to define clade memberships. As such, I opted to define clade classifications for the origin each isolate's genes using an automated process based on pairwise cophenetic distances. Specifically, I computed a pairwise cophenetic distance matrix for each gene tree using the *R* package *ape* (Paradis & Schliep, 2019). I then estimated the mean average distance of each isolate to the member isolates of each of the five *Ranavirus* clades – defined from the *Ranavirus* concatenated core-gene phylogeny – and assigned gene origins based on the clade with the closest average distance to the isolate.

Next, I performed two statistical checks and corrections to discard potentially erroneous assignments. First, I assessed whether individual gene tree topologies could be trusted to meaningfully assess clade placements for gene assignments. To do so, I performed a Mantel test comparing two cophenetic distance matrices of the concatenated core gene tree and the respective individual gene trees, rearranged to match dimensions, and applied the Bonferroni correction for the number of gene trees analysed (the core gene set). Gene trees that did not contain a significant correlation to the reconstructed phylogenetic distances of the core phylogeny were deemed not to contain accurate phylogenetic signal, and were discarded. Second, I performed a test to assess the phylogenetic distinctiveness of the relationship the isolate's genes had to their clade assignments. I achieved this by applying a Mann-Whitney U test to assess whether the closest average clade distance significantly differed from the second-closest average clade distance, and again applying the Bonferroni correction. Non-significantly distinct assignments were not considered to have likely originated definitively from one clade over another, and were therefore discarded.

Finally, I visualised the individual gene histories of each isolate through genome schematics produced using *ggplot2*. Based on the CDS ORF data of each isolate output by *Prokka* (Chapter 2, Methods 2.3.2), I plotted the position and

strandedness of ORFs, which I coloured by the phylogenetic clade assignments described above. Core ORFs that failed phylogenetic assignment checks were labeled unassigned and coloured grey, along with accessory genes that could not be assessed.

3.3.3 Recombination Within *Ranavirus* Clades

Linkage disequilibrium decay. After assessing the recombination landscape between ranavirus clades, I next sought to characterise and compare the rates of recombination within each AARV clade. However, a larger comparative sampling window of a given group could bias the likelihood of observing recombination events when making comparisons between clades. As such, for the following clade-specific recombination analyses, I rarified the sample set for each clade to contain a sampling window spanning the same period of time. To make initial characterisations, I began by using a non-phylogenetic method of characterising and comparing linkage disequilibrium (LD) decay over physical genomic distance.

LD describes the correlation of individual allele frequencies and the observed haplotype frequencies on which the alleles are found, between pairs of variant sites (single nucleotide polymorphisms; SNPs). A rate of genetic recombination causes breakpoints along the genome, which decorrelates SNPs linked on a haplotype, bringing the allele frequencies back toward linkage equilibrium. The effect of which becomes more pronounced with greater physical distance between loci, as SNPs physically closer together have less chance of incurring a breakpoint between them. Characterising LD and its decay along the whole genome therefore stands as a robust means of detecting a recombining population, with the strength of decay acting as a function of the recombinogenic rate.

The coefficient of linkage disequilibrium, known as D, is calculated as follows:

$$D_{(AB)} = P_{(AB)} - P_{(A)}P_{(B)}$$

Where $P_{(AB)}$ is the haplotype frequency of biallelic variant sites A and B, and $P_{(A)}P_{(B)}$ is the product of the independent frequencies of the two polymorphisms. Sites are considered to be in LD (and therefore contain alleles more often linked on a haplotype) if D deviates from zero for whatever reason (Hill & Robertson, 1968). However, based on the specific combination of allele frequencies, D can be negative or positive, reflecting whether alleles are found less or more often, respectively, on the same haplotype than would be expected by chance. This

property of D therefore complicates the means of assessing linear decay in SNP correlation by genomic distance. As such, normalisations of D are often implemented. One such statistic is the quantity r^2 (not to be confused with regression R-squared), which takes into account the four individual allele frequencies of the two sites, and is defined as:

$$r_{AB}^2 = \frac{D_{AB}^2}{p_A(1-p_A)p_B(1-p_B)}$$

Where $p_{A|B}$ is the major allele frequencies of sites A or B. Like the Pearson correlation coefficient, the r^2 measure of LD is always a positive value between zero and one, representing complete SNP independence and linkage, respectively.

I initially characterised LD decay across the entire concatenated core genome alignments of the AARV clades, which I wrote a custom *bash* script to carry out (https://github.com/bioinfo-chris/PhD.git; /scripts/LD/LD_pipeline.sh). The steps of the pipeline were as follows. First, the respective core ORFs specific to each clade were positioned, orientated and aligned in relation to a reference whole-genome sequence (WGS). Each reference WGS was compiled into a *BLAST* database, which was used to conduct a nucleotide search with the core ORFs belonging to the reference used as queries. The start and end positions of each core ORF in relation their respective WGS assembly revealed the gene orientation and genomic position. Using these coordinates, all isolate core ORFs were reordered and reoriented according to their position and direction in the reference WGS, using the *revseq* command of the *EMBOSS* v6.6.0.0 (Rice et al., 2000) software suit. The reordered core ORFs were then profile-aligned to the reference WGS using *MUSCLE* v3.8.1551 (Edgar, 2004), so that the alignment contained gaps where non-core genetic material was otherwise located. After which, the reference WGS was removed from the alignment, and variants were called using the software *SNP-sites* v2.5.1 (Page et al., 2016) specifying the *-v* option for variant calling format (vcf) file output. The resulting vcf file contained raw multiallelic variants called with whole-genome coordinates in relation to the reference sequence, in haploid format. Due to downstream mathematical and computational limitations, all variants were then converted to biallelic homozygote diploid, retaining the reference and dominant alternate allele for multiallelic variants. Finally, the command line software *Tomahawk* v0.7.0 (https://mklarqvist.github.io/tomahawk/) was called to calculate pairwise LD statistics between all SNPs, based on the whole genome coordinates of the reference used.

Ranaviruses contain circularly permuted linear genomes, and genetic rearrangements frequently occur (Price, 2015). Assemblies of all clades indeed contained many rearrangements, which may have resulted either by natural replication processes or by the often-spurious nature of whole-genome assembly. As such, the choice of the reference sequence in which to orient and align ORFs, then call variants from, affects the LD results obtained. In order to characterise this effect and ultimately select LD results representative of the respective clade populations, I reran the entire LD pipeline described using each isolate assembly in each clade in turn as the reference. I then performed linear regressions between the r^2 measure of LD and pairwise genomic distance in *R* to assess decay for each reference-specific set of variants. However, due to the pairwise nature of the r^2 and distance metrics, statistical independence of the data was violated for the use and interpretation of linear regression. Therefore, I implemented permutation tests as a circumvention to assess the significance of the model outputs. To do this, the pairwise genomic distances between each variant pair were randomly shuffled before refitting the model, which was repeated one-thousand times to delineate a null distribution of regression coefficients and their corresponding *R*-squared values. These distributions were then used to compare any significant decay results using observed distances. Finally, I selected the set of results produced using the reference sequence that yielded the median regression *R*-squared amongst all reference assemblies used for each clade, which most likely reflected the most common genome arrangement of the given population.

Next, I reran the LD pipeline on alignments of all individual core ORFs specific to each of the four AARV clades (i.e., not the genus-wide core genome). I carried this out to characterise regions of the genome specific to each clade where intragenic recombination occurs, specifically to identify genes that may be non-constrained to permit recombination, and might indicate adaptive recombinogenic processes (see 3.3.4 Episodic Selection Analyses, below). Given that genetic distances were only being considered within individual ORFs, I used a truncated version of the *bash* pipeline outlined above, which omitted the ORF referenced-based reordering and reorienting alignment processes. All permutation *P*-values were then corrected for multiple testing using the Benjamini-Hochberg procedure, given the number of core ORFs specific to each clade analysed.

HyPhy GARD analyses. In addition to characterising LD decay both across the genome and within individual core genes, I also sought to employ a phylogenetic

approach to characterising recombination within the AARV clades. To do so, I employed the *GARD* module (Pond et al., 2006b) of the *HyPhy* v2.5.32 (Pond et al., 2005) software suite of phylogenetic likelihood-based analyses. *GARD* (genetic algorithm for recombination detection) utilises a phylogenetic incongruency-based methodology to detect the most likely recombination breakpoints at variant sites within an alignment. Exact breakpoints may be situated between SNPs, although these are impossible to detect through alignment-based approaches. To begin with, *GARD* reconstructs a phylogenetic tree based on the whole alignment using a Maximum Likelihood approach, and evaluates the goodness-of-fit for the model using the small-sample Akaike's Information Criterion (AICc). It then fits a single breakpoint model by taking a brute-force approach to consider every variant site as a potential recombination breakpoint. The same process is used to build a phylogenetic tree for each fragment of the alignment either side of each potential breakpoint, and uses phylogenetic incongruency of the two trees and the goodness-of-fit to identify the most likely site of the single breakpoint. After which, *GARD* fits successive multiple breakpoint models in the same alignment-partitioning manner. However, for these models, an aggressive population-based hill-climber genetic algorithm heuristic is used to search breakpoint location space, as a brute-force approach that considers more than one breakpoint site is computationally insurmountable. If recombination is detected in each preceding model, an increasing number of multiple breakpoints models are run until the AICc goodness-of-fit deprecates, at which point the algorithm terminates.

I supplied *GARD* with both AARV clade-specific individual core gene alignments, as well as concatenated core gene alignments common to all clades (ALRV), specifying the GTR substitution model for phylogenetic reconstruction. As these concatenated core alignments were comprised of the genes common to all clades, the frequency and distribution of breakpoints between AARV clades could be compared. In the event that no intragenic recombination was detected in individual gene alignments, *GARD* analysis of the concatenated core genomes may detect breakpoints at SNPs close to gene boundaries where recombination occurred in intergenic non-coding regions, or non-core genes. To quantitatively assess the proportion of instances where breakpoints occurred close to gene boundaries in each concatenated alignment, I calculated the mean base-pair distance plus two standard deviations of SNPs closest to gene boundary positions, per clade. I deemed breakpoints outside this range to be sufficiently within coding regions to classify intragenic breakpoints within concatenated alignments.

Finally, to assess the significance of the probability of observing the co-occurrence rates of *GARD* breakpoints detected within regions LD decay, and vice versa, I wrote a script to perform permutation tests with 1,000 iterations. Specifically, for each clade analysis, I randomly assigned positions along the length of the alignment for the respective number of breakpoints observed for each, to count the proportion that fell within the observed regions of LD decay. I then conversely carried out the same process, but with randomly assigned regions of LD decay using the observed breakpoints. In doing so, I derived *P*-values from the iterations where proportions of random co-occurrence exceeded those observed. The script and data can be found at (https://github.com/bioinfo-chris/PhD.git; /scripts/ld-bp_permTest.R)

Homoplasy distributions. I then characterised the frequency and distribution of homoplastic loci within each AARV clade. As described above, homoplasies occur as a result of recombination or convergent evolution, and therefore stand as either an indicator of recombinogenic loci or adaptive sites under common selective pressure. As such, the distribution of homoplasies can be used to support concordant recombination and selection analysis results, particularly when they occur at high frequencies. The protocol I used to identify homoplastic sites followed that described in Chapter 2 (Methods 2.3.4). Briefly, for each concatenated alignment of core genes common between clades, I constructed a Maximum Parsimony tree using *MPBoot* v1.1.0 (Hoang et al., 2018). The trees together with the alignments used to produce them, were provided as input to *homoplasyFinder* v0.0.0.9 (Crispell et al., 2019) to calculate a consistency index for each variant site. Sites with a consistency index ≤ 0.5 (highly inconsistent) were considered homoplasies.

3.3.4 Episodic Selection Analyses

HyPhy BUSTED analyses. I sought to characterise whether core genes across the genus contained evidence of having evolved under positive diversifying selection, firstly to identify whether recombinogenic regions were adaptive. Methods based on the non-synonymous to synonymous codon substitution ratio – *dN*/*dS*, denoted as ω – have been successful in the past in detecting episodic selection amongst specific small groups of *Ranavirus* isolates (Epstein & Storfer, 2016; Price, 2015). However, given the greater taxonomic breadth and genomic resolution of the *Ranavirus* isolates characterised here, I aimed to update these findings by first assessing positive selection with a gene-wide branch-site model approach, implemented by the *BUSTED* module of *Hyphy* v2.5.32 (Murrell et al., 2015). The

advantage of using *BUSTED* (branch-site unrestricted statistical test for episodic selection) lies within its statistical power to detect episodic selection through its use of a stochastic unrestricted branch-site random effects likelihood (BS-REL) model. The core idea of the BS-REL model is to issue three rate categories of ω to each branch of a tree (purifying selection $0 \leq \omega < 1$; neutral selection $\omega = 1$; positive selection $\omega > 1$), and allow each codon site to evolve at any of these classes. Therefore, a tree with *B* branches will be modeled with 3^B configurations of ω for each codon site (Smith et al., 2015). This allows detection of selection on a subset of foreground branches and sites within the gene, which other methods that average ω over codon sites and/or branches may otherwise fail to detect. *BUSTED* however does so in a stochastic site non-specific manner, and therefore the model output will only provide evidence for a gene having undergone diversifying selection at a minimum of at least one site on a foreground branch, but not where exactly these occur.

To perform selection analyses (using both *BUSTED* and *aBSREL* below), I began by first producing individual *Ranavirus* core ORF codon alignments. As opposed to standard nucleotide alignments, codon alignments must contain total sequence lengths in multiples of three, and are aligned to preserve the reading frame of amino acid translation. I used the alignment tool *PRANK* v.170427 (Löytynoja & Goldman, 2005) specifying the *-codon* option to perform codon alignments of all *Ranavirus* core genes. In the event sequences were not in multiples of three and codon-aligning failed, I used the back translation protocol of *PRANK* by specifying *-translate* flag. This performs an amino acid alignment after translation of the sequences, whilst simultaneously aligning the nucleotide sequences in the corresponding amino acid positions. I then manually inspected alignments for misaligned isolate sequences containing frameshifts, which I removed for downstream analyses. After which, I applied the BS-REL model considering all branches foreground by supplying *BUSTED* with the individual *Ranavirus* core gene codon alignments, together with its corresponding ML tree, as outlined above (constructed using the regular nucleotide alignment; see *Isolate gene histories*).

HyPhy aBSREL analyses. Finally, together with identifying the genes with evidence of having undergone positive diversifying selection, I sought to further pinpoint specific taxa or ancestral states of *Ranavirus* in which gene adaptation had occurred. A key aim in doing so was to identify genes that had been positively selected for in response to host shifts, given that amphibians are considered to be the derived host state of *Ranavirus* (Price et al., 2017a). To do so, I utilised the *aBSREL* module (Smith et al., 2015) in *HyPhy*, which employs a so-called adaptive

branch-site random effects likelihood model, a variant of the BS-REL. The key difference to the standard BS-REL is that the *aBSREL* model is adaptive to the structure of the dataset, calculating a variable number of ω rate categories to be inferred for each branch, given the complexity of the data. The optimal number of rate categories across branches can then be discerned though AICc goodness-of-fit, permitting the identification of which branches experienced significant diversifying selection of the given gene through model comparison. As before, I supplied *aBSREL* with each of the *Ranavirus* core ORF codon alignments and corresponding ML gene trees, however, I ran five different analyses. First, I performed an exploratory analysis across the whole genus, which considered all branches of the tree as foreground in the ω rate assessments. The power to detect selection is impacted by this approach, given the Holm-Bonferroni correction applied to account for the number of tests performed for each branch. As such, I in turn carried the same analysis for each ALRV clade using trees annotated to identify the member isolates. In doing so, each clade-specific analysis only considered the respective terminal branches foreground. This approach increased power for isolate-specific selection detection within each clade, given the lower number of foreground branches tested.

3.4 Results

3.4.1 AARV Clade-Specific Core Genomes

To delineate the pan-genome content for the four amphibian-associated *Ranavirus* taxonomic subsets, I used the *Roary* pipeline together with the *Prokka* CDS output yielded in Chapter 2. A total of 77 ATV-like, 76 CMTV-like, 69 FV3-like, and 68 TFV-like core orthologous ORFs were found to cluster all isolates of the respective clades at 80% amino acid homology.

3.4.2 Between Clade Recombination

Genus-wide phylogenetic incongruencies. The *SplitsTree* phylogenetic network of the ALRV core genome alignment (n = 49 ORFs) reconstructed the ALRV genetic diversity in accordance with the previously outlined clade characterisations (Fig. 3.1). However, the TFV-like presented a key exception, whereby the group were found to be especially paraphyletic, and did not diverge from a single ancestral branch unlike all other clades, including the paraphyletic EHNV-like. Instead, the TFV-like contained a web of basal split edges furcating from other major clade branches, predominantly from the FV3-like, indicating extensive phylogenetic conflict within the alignment, and therefore potential ancestral recombination between the implicated clades. The isolate ToRV1 was the most extreme, existing

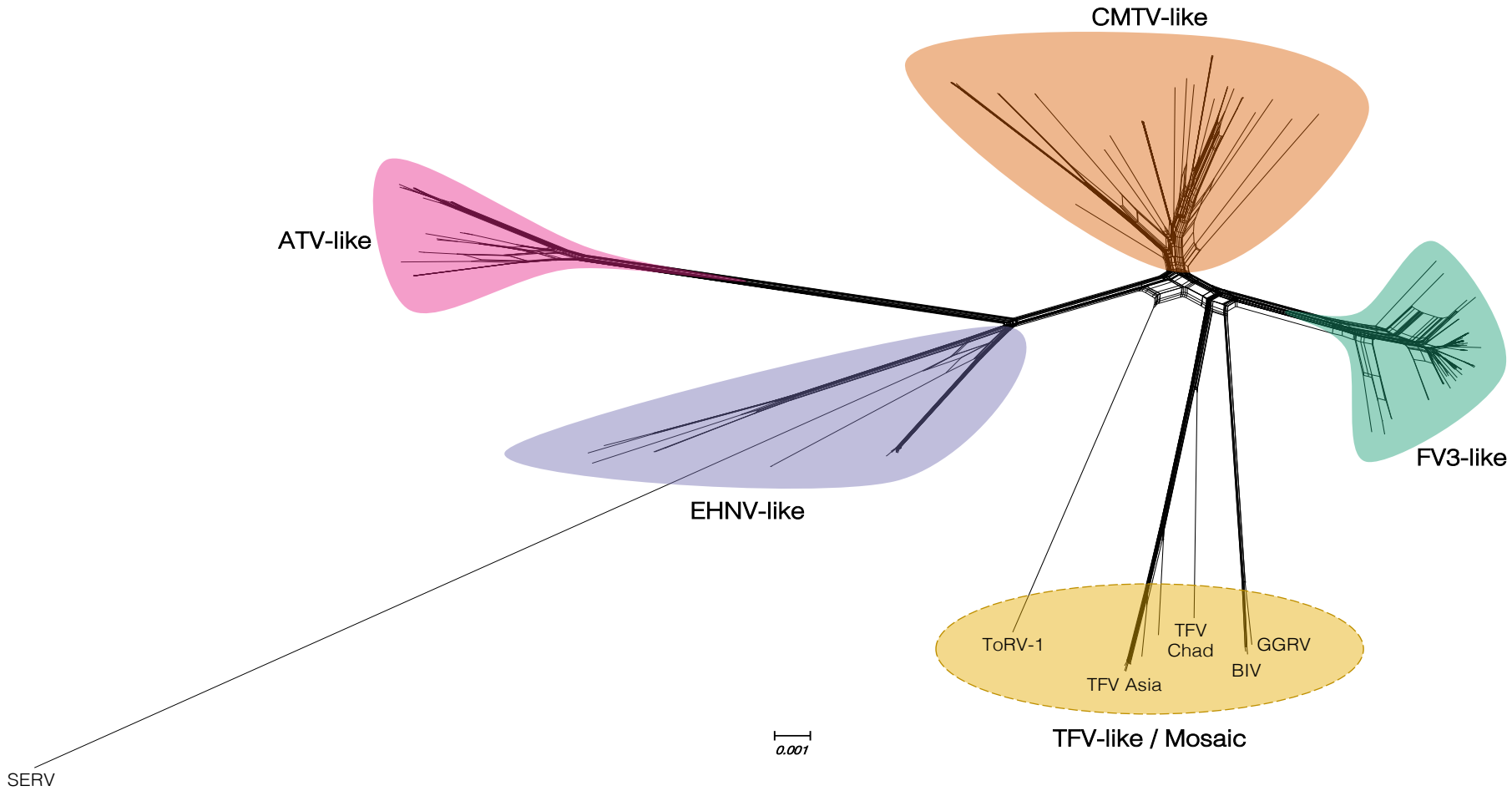


Figure 3.1. *Phylogenetic network of the amphibian-like ranaviruses. The network was constructed with* SplitsTree, *using a concatenated sequence alignment of 49 core ALRV genes across 170 isolates. Pairwise distances were calculated using the HKY85 substitution model, scaled by substitutions/site^{-1}. Edges between branches indicate sites in the alignment where phylogenetic conflicts exist. Note the high degree of splits implicating the TFV-like, and particularly the isolate ToRV1.*

alone on a long branch extending entirely from split edges between the EHNV-, CMTV-, and FV3-like. This may indicate mosaic origins for the TFV-like group, particularly for ToRV1. Intriguingly, amongst the EHNV-like, the divergent SERV isolate – used as a root for the genus – along with the ECV/ESV isolates, also exhibited a similar pattern of extensive basal splits between the two major EHNV and *European North Atlantic ranavirus* (ENARV) branches, possibly indicating further recombinogenic origins. All remaining ALRV clades contained many splits between branches of their member isolates, with the greatest degree of a derived web-like clade structure of phylogenetic conflicts present in the CMTV- and FV3-likes.

Isolate gene histories. In assigning the individual phylogenetic gene histories of each ALRV isolate, each clade contained a similar proportion of unassigned and/or non-core ORFs. Together the clades contained a mean 71.6% unassigned or non-core ORFs (range 69.2 – 74.2%) after controlling for genes with discordant genetic distance reconstructions to the concatenated core genome phylogeny, as well as eliminating significantly indistinct gene assignments. Of the isolate genes that achieved a phylogenetic assignment, the majority were congruent with the isolate's given classification based on its position on the genus core phylogeny, suggesting common evolutionary histories in most instances. However, all clades contained isolates with at least one gene assignment outside of their respective clade (Fig. 3.2 ; Fig. S2).

Thirty-four (of 63; 54%) FV3-like isolates, the most of any group, contained between one and six genes with evolutionary histories originating outside the clade (Fig. S2). The majority of their polyphyletic genes originated from the CMTV-like, with the largest mosaic composition belonging to RCVZ2 (MF187209.1). It had six assigned ORFs (22% of assigned ORFs) originating from the CMTV-like, which is in agreement with previous findings of this isolate's cross-lineage mosaic composition (Claytor et al., 2017). Interestingly, five of the most basal FV3-like isolates (two from the USA and three from Chinese aquaculture facilities) had genes originating from the EHNV-like. Next, the CMTV-like contained 19 isolates (of 46; 41%) with at least one ORF originating from either the FV3-like or EHNV-like clades, and in three instances, a gene with the closest relationship to the TFV-like. The EHNV-like contained five isolates with either one or two genes that grouped with the ATV-like clade. Next, ATV-like all contained single genetic origins within their clade, except for the isolate ATV_NM (KR075880.1) that had a single gene with FV3-like origins, which is plausible based on the shared geographic ranges of these lineages.

The TFV-like clade contained the greatest degree of polyphyly, in agreement with the possible mosaic origins identified by the *SplitsTree* phylogenetic network. Further supporting its unusual phylogenetic placement, the isolate ToRV1 had 26 ORFs (96%) assigned outside the clade. This may well indicate its misclassification to the TFV-like group, however, the genes were almost equally distributed between the CMTV- (13 ORFs) and FV3-like (9 ORFs) clades, with an additional four ORFs falling with the EHNV-like, and only one with the TFV-like (Fig. 3.2). A similar pattern was observed for the BIV isolates, which each had approximately 15 ORFs assigned outside the TFV-like clade. As an example, BIV (NC_038507.1) had ten ORFs clustering with the FV3-like, four with the CMTV-like and two with the EHNV-like, whilst a higher proportion of nine ORFs were assigned to the TFV-like (Fig. 3.2). Despite its close, but basal, relatedness to the TFV lineage proper, the TFV from Chad (MW727505.1) also demonstrated a high degree of polyphyly with six genes each falling closest to the FV3- and CMTV-like. The remainder of the TFV isolates contained less-to-no polyphyly, which could indicate that this subclade represents a natural population of *Ranavirus* endemic to Asia, whilst the BIVs, TFV Chad, and ToRV1 are polyphyletic mosaics.

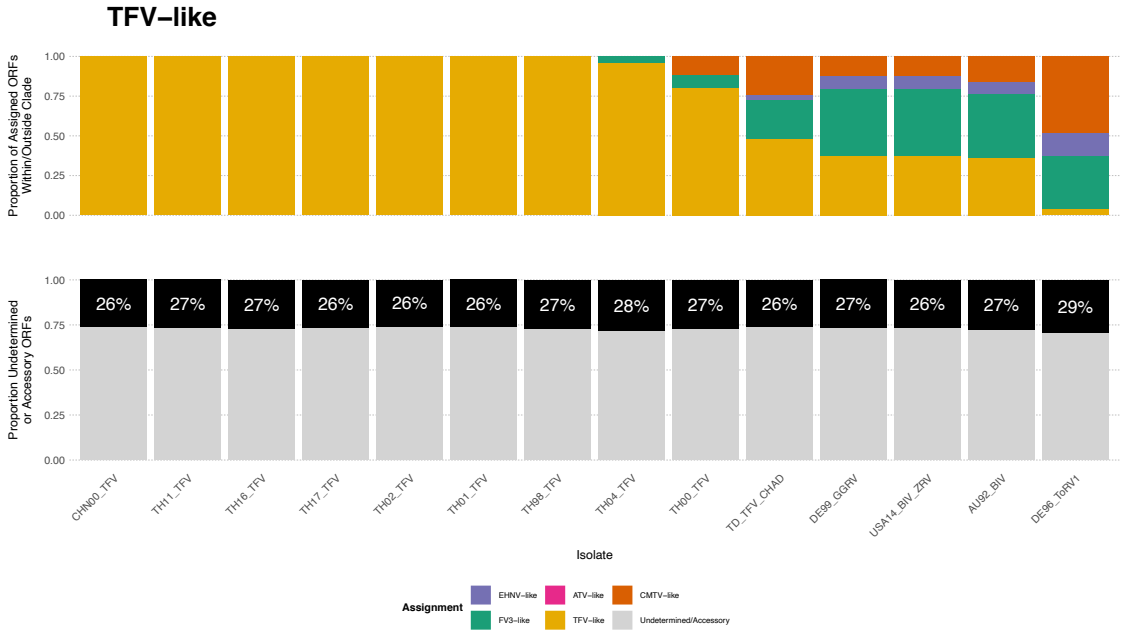


***Figure 3.2**. Gene histories of the paraphyletic TFV-like clade. Each column represents an isolate, where the top panel illustrates the proportion of assigned genes originating from particular clades, according to the ALRV core phylogeny. The bottom panel shows the proportion of genes that achieved an unambiguous phylogenetic assignment (black) amongst all identified CDSs (median 93 per isolate). The assignment process is based on closest average cophenetic distance of each clade to the given isolate within individual gene trees.*

Figure 3.3 shows genome schematics of exemplar isolates from all clades with the greatest proportion of polyphyletic genes, and from which clades they originated. Altogether, the individual gene histories amongst *Ranavirus* suggest a high degree of often spatially clustered mosaicism amongst isolates of distinct lineages. Intriguingly, there was a tendency for isolates from captivity to contain the highest proportion of polyphyletic genes. However, the association was non-significant when considering all isolates together, as well as partitioned by clade (Fig. S3). Nevertheless, the extraordinarily high degree of polyphyly amongst the TFV-like suggest this clade might be formed of isolates with mosaic origins, which may be reflective of their near-complete association to captive and farmed host animals. Indeed, FV3-like isolates from captive settings (e.g., RCVZ2 and FV3_NIC; MF360246.1) are also among the most polyphyletic of their clade (Fig. S2; Fig. S3). Given the degree of polyphyly, assigning TFV-like isolates such as ToRV1 to a single clade is problematic at best, and accounts for the variability in assignments of this group reported in the ranavirus literature (Ferreira et al., 2021; Stöhr et al., 2015; Yu et al., 2020).

3.4.3 Within Clade Recombination

Temporal rarefaction. The sampling dates of all isolates ranged between 1966 and 2017. However, the ATV-like clade (22 isolates) contained the shortest sampling span of 1996 – 2015. As such, the remaining three AARV clades were down-sampled to contain only isolates collected within the same 19-year timeframe to try and control for temporal biases when making within-clade recombination rate comparisons. This resulted in 47 CMTV-like, 46 FV3-like, and 10 TFV-like isolates being retained for clade-specific sequence alignments that were used to generate the results of the following sections.

LD decay by genomic and genetic distance. Evidence was present for highly significant LD decay along the genome for all AARV clade-specific core genome alignments. However, for the models that yielded the most significant patterns of LD decay (alignment orientations with minimum-possible permutation *P*-value = 0.001), the majority explained very little variance ($R^2 < 1\%$) in decay signal (Fig. 3.4 **A**). This pattern held true even when SNPs with minor allele frequencies (MAF) < 0.2 were excluded to reduce pairwise noise generated from the many low MAF variants. The exception was the FV3-like clade, which had increasingly more variance explained as rare SNPs were excluded; Figure 3.4 (**B)** shows the decay signal of the FV3-like alignment according to the reference (MG953518.1) that yielded the median R^2 fit of 18.6%, with a regression coefficient of $-1.1e^{-06}$.

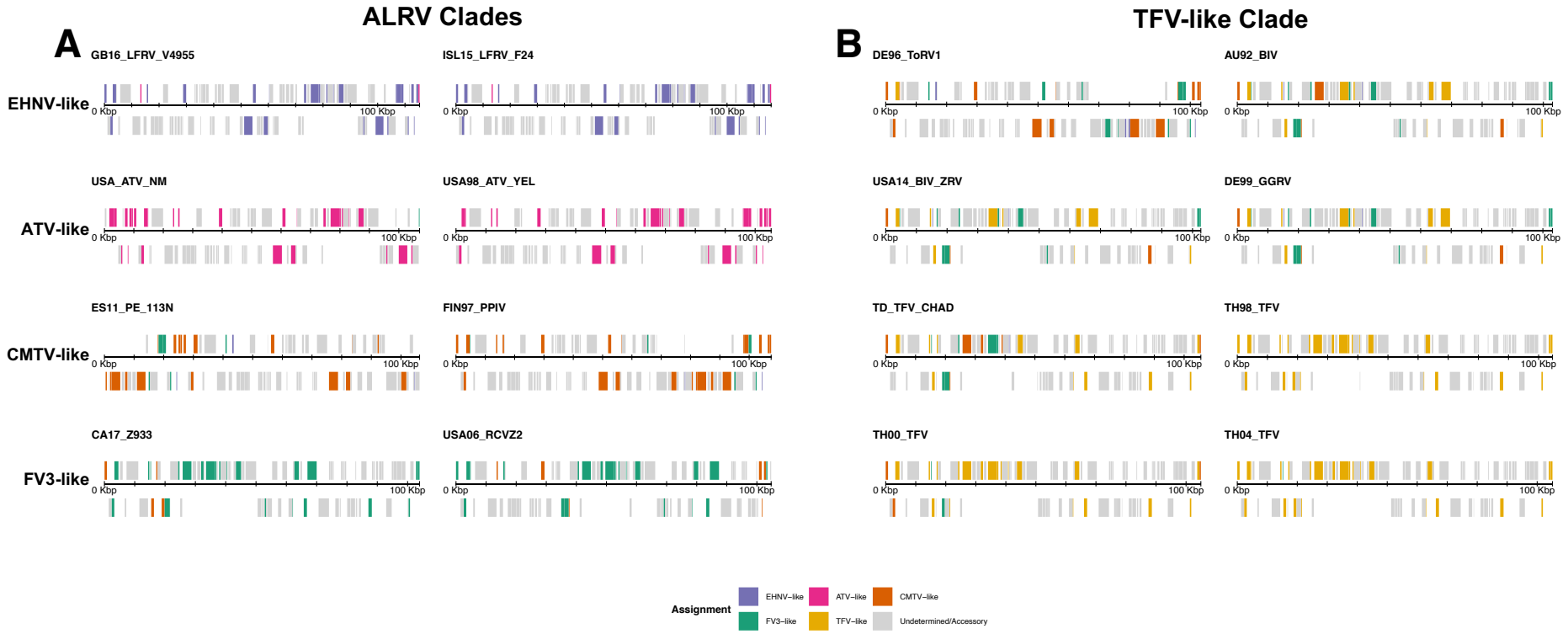


Figure 3.3. *Genome schematics of exemplar* Ranavirus *isolates and their gene histories. Panel* ***A*** *shows two genomes each from the EHNV-like, ATV-like, CMTV-like and FV3-like, top to bottom, that each contain high degrees of polyphyly amongst their clade (except the ATV-like, in which only one genome contained evidence of polyphyly). Panel* ***B*** *shows eight of the most polyphyletic genomes from the paraphyletic TFV-like clade. Notably, the isolates ToRV1, BIV and TFV Chad contain the most extreme degree of polyphyletic genetic composition. Each gene block represents a predicted ORF sequence and its strandedness (above scale bar is 5′ → 3′), where grey denotes either non-core genes (therefore could not be aligned) and core genes that failed unambiguous phylogenetic assignment.*

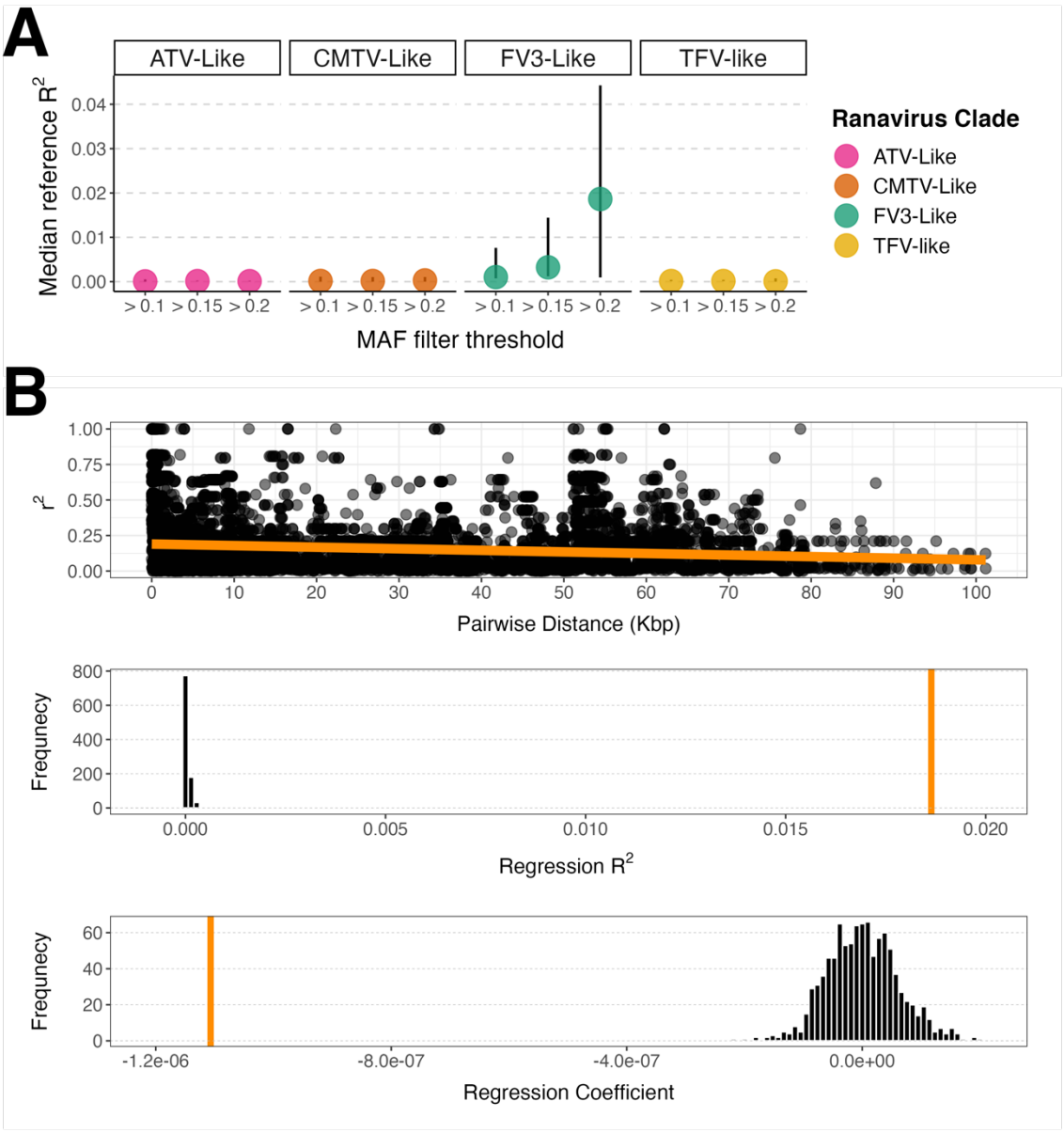


Figure 3.4. *Linkage Disequilibrium decay by genomic distance in amphibian-associated ranaviruses. Panel* ***A*** *shows the strength of LD decay by regression* R^2 *along core genome alignments (n = 49 ORFs, aligned in WGS positions) for each AARV clade. Each alignment was rarified to contain isolates sampled between 1996 – 2015 in an effort to control for temporal bias (number of observed recombination events). A distribution of (permutation test) significant LD decay* R^2 *values are displayed for each clade, showing the range and median, yielded from analyses reperformed for each reference sequence WGS arrangement used. Each set of analyses were performed by filtering out rare variants, only retaining SNPs with minor allele frequencies (MAF) >10%, >15% and >20% to reduce noise. Panel* ***B*** *shows the exemplar reference alignment for the FV3-like clade that yielded the median regression* R^2 *value for the MAF >20% analysis. The top panel shows LD decay by pairwise genomic distance, whilst the bottom panels show the permutation test results for the regression* R^2 *and the regression coefficient, where orange bars are the observed value and back bars are the permuted distribution (n = 1,000 iterations).*

The weak fit of most models appeared to be due to the large amount of genetic diversity in all clade-specific alignments, except the FV3-like. Such diversity, even with the applied MAF filtering thresholds, resulted in hundreds of thousands of pairwise SNP comparisons across all positions over the relatively short genomic distance of the ranavirus genome (~105 Kbp), rendering linear regression as likely inappropriate to assess decay in this manner, at least in this system.

Nevertheless, significant and abundant evidence was present that the isolates within each AARV clade were sampled from frequently recombining populations. First, the individual gene alignments of core ORFs specific to each clade showed many instances of intragenic LD decay after multiple testing corrections were applied (Table 3.1; Fig. 3.5). The ATV-like contained 15 (of 77) clade-specific core genes with intragenic LD decay, the highest frequency of all clades, whereas the CMTV-like contained the fewest with 7 (of 76). There were also multiple, but fewer, instances of individual gene alignments from each clade with evidence of recombination breakpoints detected by *GARD* analysis (Table 3.1), although the genes identified were not always congruent with those showing LD decay. The FV3-like core ORF set was an exception, where *GARD* did not find any intragenic breakpoints. Next, after concatenating and aligning the core genes common to each clade (ALRV core; n = 49), an average of 10 breakpoints (range 4 – 14) were detected by *GARD* for each clade (Table 3.1), with large ΔAICc scores between the null model with no breakpoints (ΔAICc range: 63 – 3,966).

The incongruency between the breakpoints detected in individual and concatenated gene alignments suggested that, in most instances, recombination breakpoints occurred in regions outside the core genome, or perhaps in non-coding intergenic regions. This is not surprising given the ALRV core genome only constitutes ~40% of the average *Ranavirus* whole genome. That said, only 20% of *GARD* breakpoints were positioned within the respective mean base-pair distance of SNPs closest to gene boundaries, per clade alignment (Fig. 3.6). This indicated that 80% of all breakpoints detected were positioned sufficiently distant from gene boundaries to suggest true intragenic recombination, despite a comparative lack of detection in individual genetic units. Furthermore, there was no significant difference in breakpoints preferentially falling close to gene boundaries for different clades, suggesting similar levels of intra-genic recombination detection (Kruskal-Wallis $\chi_3^2 = 5.94, P = 0.11$). Finally, 35% of observed *GARD* breakpoints were detected in regions of LD decay (mean permuted proportion = 0.204; P= 0.009), whereas 47.8% of LD decay regions

Table 3.1. *Recombination analyses within lineages of amphibian-associated ranaviruses. Two sets of recombination results are displayed; those yielded from individual core gene alignments specific to each AARV clade, and those yielded from concatenated alignments of the 49 core genes common to the ALRV. The number of isolates retained after temporal rarefaction for each AARV clade is in parentheses in the first column, next displaying the total number core ORFs specific each clade. The total number of ORFs with presence of LD decay also shows in parentheses the number that correspond to the ALRV core, with mean regression* R^2 *based on all genes with detected LD decay. The individual genes with GARD breakpoints column also displays the number that are ALRV core in parentheses. Lastly, the number of homoplasies and GARD breakpoints detected for each clade-specific concatenated core alignment is given in the final two columns. The ΔAICc is given for the best-fitting multi-breakpoint GARD model compared to the null model with no-breakpoints.*

	Clade-Specific Core ORFs				**Concatenated Core Genome (ALRV n=49)**	
AARV Clade	**Number of Core ORFs**	**ORFs with LD Decay**	**Mean LD Decay Regression R^2**	**ORFs with GARD Breakpoints**	**Homoplasies**	**GARD Breakpoints**
ATV-like (n=22)	77	15 (6)	2.25%	4 (2)	165	14 (ΔAICc 754)
CMTV-like (n=47)	76	7 (4)	0.42%	2 (1)	392	13 (ΔAICc 1783)
FV3-like (n=46)	69	9 (6)	13.46%	0 (0)	191	9 (ΔAICc 3966)
TFV-like (n=10)	68	9 (7)	2.43%	2 (1)	129	4 (ΔAICc 63)

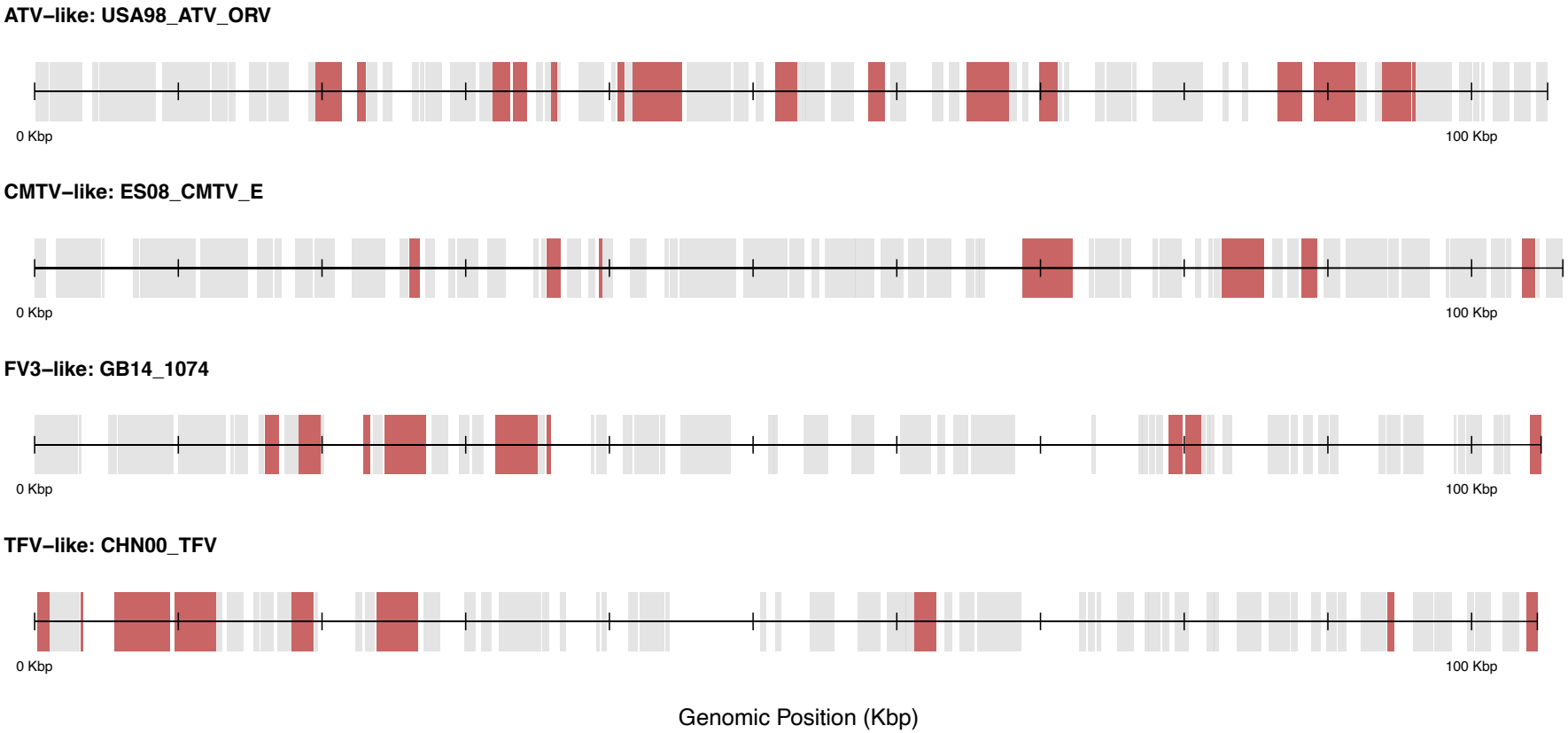


Figure 3.5. *Genome schematics of recombinant regions in each AARV clade. Shown is the core-gene complement specific for each AARV clade, comprising 77 ATV-like, 76 CMTV-like, 69 FV3-like, and 68 TFV-like core orthologous ORFs. Schematics are based on alignments of ORFs arranged and oriented to the whole-genome positions of a representative reference sequence (isolate name in schematic title); grey bars indicate core ORF positions and white space indicates non-core genomic material. Red shaded genes are those that contain evidence of LD decay, with significance ascertained from regression permutation tests corrected for multiple testing.*

contained breakpoints (mean permuted proportion = 0.154; $P < 0.001$). The statistically significant association between the location of breakpoints and regions displaying signs of LD decay suggests that the two methods capture a common signal of underlying genetic recombination (Fig. 3.6).

Four genes that contained congruent evidence of intragenic recombination were found to be common between clades, even though the recombination analyses were carried out independently on the individual clade-specific alignments. Specifically, these were a phosphotransferase (group 89), tyrosine kinase (group 106), an NTP/ATPase Type III restriction enzyme (group 136), and the D5 family NTP/ATPase (group 217). It may be interesting to note that the affected core genes – as well as additional ones including deoxynucleoside kinase/thymidine kinase (group 85) specific to the ATV-like – appear to be involved in cellular metabolic processes primarily catalysing ATP hydrolysis.

3.4.4 Episodic Diversifying Selection

In order to assess the selection landscape across *Ranavirus*, I employed two *dN*/*dS* ratio-based methods on the 49 individual ALRV core genes. After producing codon alignments for each gene, the selection analyses failed for two – an AAA-ATPase (group 94) and a helicase-like protein (group 131) – as these genes appeared truncated by a high frequency of stop codons throughout the length of the reading frame. Both the *BUSTED* and *aBSREL* methods successfully completed for all remaining genes, and together 12 were found to have significant ($P < 0.05$) evidence of positive diversifying selection, following Holm-Bonferroni correction for the number of branches tested (Table 3.2). Specifically, the site non-specific *BUSTED* method detected positive selection in five genes across all branches, whereas the *aBSREL* methods (see below) detected selection on 13 branches of ten genes. There were congruent results between the methods on three genes, which were a phosphotransferase protein kinase (group 89), the myristylated membrane protein (group 144) and the D5 family NTP/ATPase (group 217). Interestingly, there was also a degree of congruency between genes that contained evidence of selection and genes displaying strong evidence for recombination. Specifically, these were the phosphotransferase protein kinase (group 89), the D5 family NTP/ATPase (group 217), an RNA polymerase Rpb5 (group 101), and the NTP/ATPase Type III restriction enzyme (group 136).

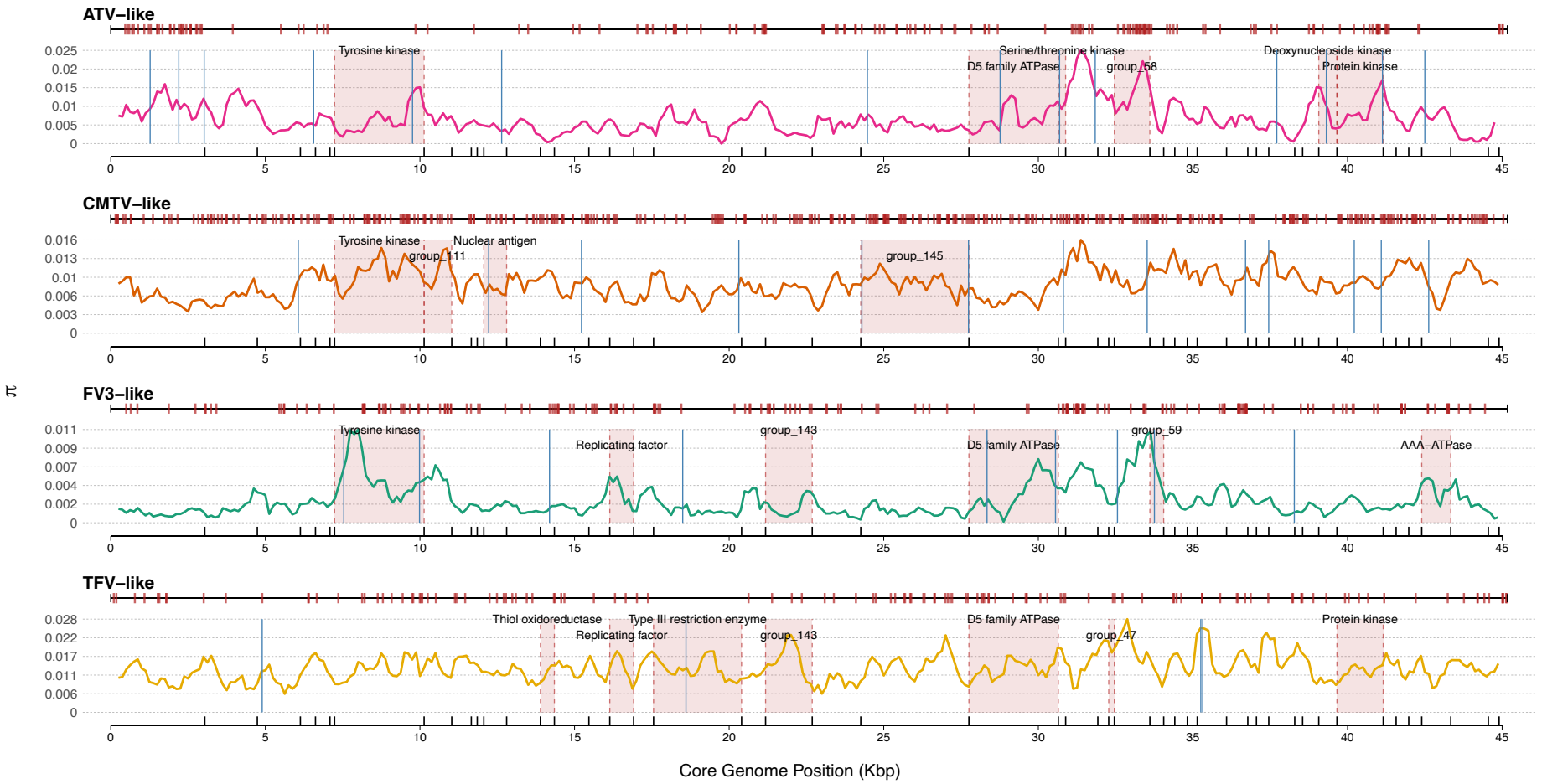


Figure 3.6. *Concatenated core genome schematics of amphibian-associated ranavirus clades. Each genome schematic shows the ALRV core genes ($n = 49$), concatenated and aligned in a common orientation and order for each clade alignment, with gene boundaries displayed as ticks at the bottom of schematics. Alignments were rarified to contain isolates sampled between 1996 – 2015. Nucleotide diversity (pi) was calculated over 500 bp windows. Displayed above each schematic is the distribution of homoplasies.* GARD *breakpoints are displayed as blue lines. Regions of LD*

*[**Figure 3.6 continued**] decay were mapped on to the schematics from decay profiles of individual core gene alignments specific to each clade. Functional annotation (with* Roary *group for hypothetical proteins) of the ALRV core gene is displayed for clade-specific genes with LD decay.*

***Table 3.2**. Episodic diversifying selection results amongst core genes of the ALRV (n = 49). The results of two selection analyses are given, implementing variations of the branch-site random effects likelihood model (BS-REL). First are genes with significant evidence of diversifying selection under the site non-specific* BUSTED *model. Next are the results ascertained from the* aBSREL *model, which identifies the branch, and therefore implicated taxa, that contain evidence of having undergone positive selection. Rows of results overlap where evidence was congruent between models for a given gene. Finally, the functional annotation of genes is given in the final column, where known.*

BUSTED		aBSREL					
Roary ORF	***P*-value**	**Roary ORF**	**Model**	**Branch**	**Taxa**	***P*-value**	**Functional Annotation**
BALF5	0.0114						DNA polymerase family B
group_126	$< 1e^{-5}$						Thiol oxidoreductase; Erv1 / Alr family
group_89	$< 1e^{-5}$	group_89	All branches	CA17_Z377	FV3-like terminal	0.00402	Putative phosphotransferase; Protein kinase domain
			CMTV foreground	CHN12_ADRV1201	CMTV-like terminal	0.01855	
group_144	0.0395	group_144	All branches	Node 221	TFV-like MRCA	0.00173	Myristylated membrane protein; Lipid membrane protein of large eukaryotic DNA viruses
group_217	$< 1e^{-5}$	group_217	All branches	CA16_KEN5	FV3-like terminal	$< 1e^{-5}$	D5 family NTPase/ATPase; D5 N terminal like
		I4L	All branches	Node 2	ALRV MRCA	0.00002	Ribonucleotide reductase, barrel domain
		group_68	TFV foreground	TH04_TFV	TFV-like terminal	0.02754	Human parainfluenza virus 1L-like protein
		group_74	All branches	Node 170	CMTV-like subclade	0.02799	Myeloid cell leukemia protein; Apoptosis regulator proteins, Bcl-2 family
				Node 184	CMTV-like subclade	0.02798	
		group_95	CMTV foreground	CHN12_ADRV1201	CMTV-like terminal	0.04443	Hypothetical protein
			FV3 foreground	CA16_KEN5	FV3-like terminal	0.00002	
		group_101	All branches	Node 5	CMTV/FV3/TFV MRCA	0.00958	RNA polymerase Rpb5, C-terminal domain
		group_121	All branches	Node 81	CMTV-like subclade	0.01189	WXG100 protein secretion system (Wss), protein YukC
		group_136	All branches	NZ91_SERV	EHNV-like terminal	0.00002	NTPase/helicase; Type III restriction enzyme, res subunit

The all-branches *aBSREL* model detected a number of genes having undergone positive selection at key sites specific to amphibian-associated ranavirus evolution (Fig. 3.7). The earliest of these was the ribonuclease reductase barrel domain (group I4L), which contained strong evidence of positive selection ($P < 0.0001$) on the ancestral branch to all ALRV ranaviruses, only excluding the *European North Atlantic ranavirus* subclade of the EHNV-like. Next, another RNA polymerase Rpb5 (group 101) showed evidence of diversifying selection ($P = 0.0096$) on the ancestral branch to the CMTV-, FV3-, and TFV-like superclade. Two genes further implicated ancestral taxa of specific amphibian-associated clades. First, the myristylated membrane protein (group 144; also detected by *BUSTED*) was positively selected for in the ancestor to the TFV-like clade ($P = 0.0017$), which for this gene formed a distinct monophyletic clade only excluding ToRV1. Second, the myeloid cell leukaemia protein (group 74), a Bcl-2 family apoptosis regulator, had positive selection detected on two CMTV-like branches (both P values $= 0.028$). These implicated one ancestor to the subclade of the North European CMTV isolates, and an ancestor to the trade-linked Chinese giant salamander ranavirus (ADRV) and USA RCVZ (MF187210.1) isolates. Finally, there were also genes that had positive selection detected on terminal branches, suggesting perhaps recent selective pressures faced by the implicated isolates (Table 3.2; Fig. 3.7). These viruses involved the wild-caught Canadian FV3-like (MK959605.1; MK959608.1), a captive Chinese ADRV CMTV-like isolate (KC865735.1), and one captive Thai TFV-like isolate (MT512497.1).

3.5 Discussion

Recombination is an important and conserved feature of dsDNA virus evolution (Kamita et al., 2003; Robinson et al., 2011; Weller & Sawitzke, 2014). In this chapter, I have shown that its role in the evolution of *Ranavirus* stands as no exception. Through conventional phylogenetic and non-phylogenetic means, I have demonstrated significant and extensive evidence that, amongst all clades of *Ranavirus,* the genomic samples featured in this study were drawn from frequently recombining populations. This was true for populations both within clades – which could be expected given shared spatial distributions – but also between distinct phylogenetic lineages, suggesting possible ancestral genetic exchanges. In addition to the many recombinant genomic regions identified in alignments of all AARV all clades, each group (except the ATV-like) also contained several isolates with polyphyletic gene histories, which indicates that

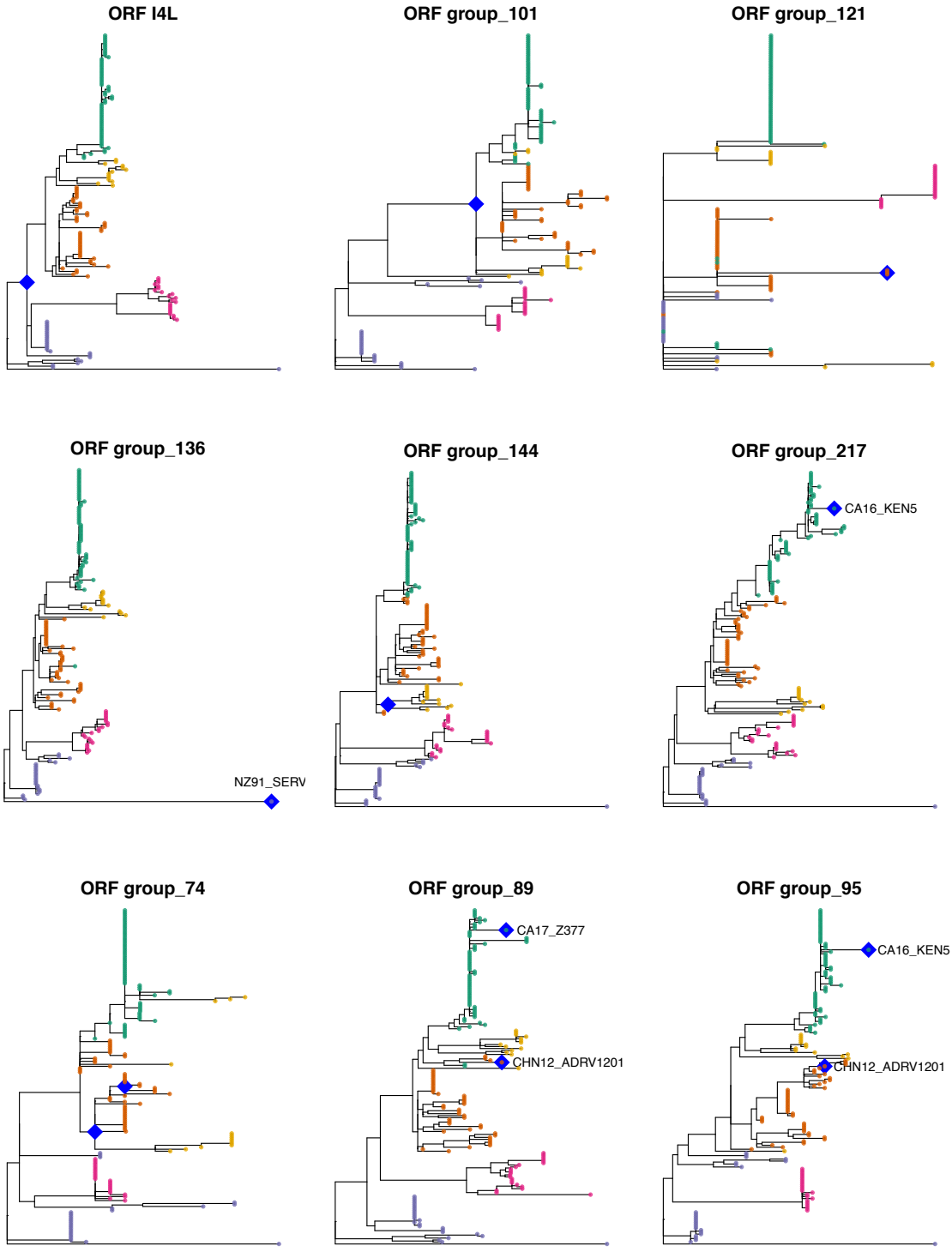


Figure 3.7. *ALRV core genes with evidence of episodic diversifying selection. Individual core gene trees with 170 isolates across all clades, with evidence of positive selection ascertained from the* aBSREL *model implemented in* HyPhy. *Branches under selection are displayed with blue diamonds, and where the branch is terminal, the implicated isolate name is given. Note selection at the key evolutionary event on the gene tree of ORF I4L, which implicates the ancestral taxa of the amphibian-associated ranaviruses, likely indicating selection following the first host-shift away from fish. A similar pattern is seen for ORF group 101.*

recombination is a pervasive process that has contributed to the generation of *Ranavirus* genetic diversity. Overall, my results suggest a far greater extent of ancestral recombination and polyphyly amongst ranaviruses than had been previously identified (Candido et al., 2019; Chinchar & Granoff, 1986; Claytor et al., 2017; Price, 2015; Vilaça et al., 2019).

3.5.1 Patterns of Recombination Amongst *Ranavirus*

To date, there has been no explicit investigation into the rates of ranaviral recombination, and the exact mechanism by which ranaviruses may recombine is not known. However, recombination in dsDNA viruses appears to be a conserved process tightly linked to DNA replication and repair amongst diverse groups (Kamita et al., 2003; Robinson et al., 2011; Weller & Sawitzke, 2014). Specifically, segment exchange occurs via homologous repair pathways of DNA double-stranded breaks (DSB) created during replication (Weller & Sawitzke, 2014). Indeed, ranavirus DNA replication involves the production of larger-than-unit-length genomic concatemers requiring cleavage, inducing DSBs (Goorha & Murti, 1982). As such, although awaiting experimental validation, it is not unreasonable to assume genetic recombination of distinct ranavirus genotypes occurs by way of the same conserved pathways as other model dsDNA viruses. An outstanding question, however, is the degree of homology that is required for homologous recombination to proceed, particularly concerning genetic exchange between divergent ranaviruses such as the EHNV-like and the derived AARV lineages. It has been shown that a surprisingly little amount of sequence homology can permit recombination in vaccina virus (*Poxviridae*) and SV40 (*Polyomaviridae*) dsDNA virus models, although a sharp drop off in rates was also observed for segments approximately $> 20\%$ divergent (Rubnitz & Subramani, 1984; Yao & Evans, 2001). Though, structure of the molecule also plays an important role, where linear segments tend to recombine more readily than circular DNA (Yao & Evans, 2001). As both these DNA structures are produced during ranavirus genome replication (Goorha & Murti, 1982) and sequence homology amongst core genes of the ALRV generally exceeds 90%, it is eminently plausible that all lineages of *Ranavirus* are able to recombine, given the opportunity. The greatest rate limiting factors are therefore likely epidemiological, related to host tropism ability of ranavirus lineages, host ecology and host-virus spatial distributions.

Currently, there are two documented instances of between-lineage recombination observed for ranavirus *in situ*. Both Claytor et al. (2017) and Vilaça et al. (2019) report North American FV3-like isolates with genomic segments derived from

CMTV-like ranaviruses. The studies used the same approach to detect recombination events amongst their samples and a selected group of reference genomes by considering the convergence of results yielded from six or more detection algorithms implemented in RDP4 software suit (Martin et al., 2015). Using the phylogenetic approach I developed to assess isolate placement within core gene trees, I recapitulate these findings. Specifically, the FV3-like ranaviruses with greatest degree of polyphyly with CMTV-like genetic diversity were indeed RCVZ2 (Claytor et al., 2017) and the several Canadian FV3 isolates reported by Vilaça et al. (2019). Interestingly, I also discovered a similarly high degree of polyphyly for the FV3 NIC isolate (MF360246.1; Saucedo et al., 2017), which was sampled from captive poison-dart frogs (*Oophaga pumilio*) imported to the Netherlands from Nicaragua.

The ultimate factor that dictates whether recombination can proceed to generate virus genetic diversity is coinfection of a single host with two or more distinct viral genotypes. Divergent *Ranavirus* lineages do overlap in both geographic range and host type in the wild, such as FV3- and CMTV-likes in Europe (Rosa et al., 2017). However, ranaviruses circulating in trade networks that infect captive hosts cohoused at high densities greatly increases the chance of divergent lineages co-occurring, as with the precedence set by other systems (Bowden et al., 2004; Pisoni et al., 2007). Indeed, in my gene history analysis, all FV3-like ranaviruses associated with captivity contained polyphylies, including the Chinese RNRV (MG791866.1), RGV (JQ654586.1) and STIV (EU627010.1) isolates, and those mentioned above with the highest rates of CMTV-like introgression. However, the remaining North American polyphyletic FV3-like genomes were associated with wild hosts, yet they also seemingly contained genes derived CMTV-like origins. Given that all CMTV-like ranaviruses are naturally restricted to a Eurasian geographic range, it appears likely anthropogenic-mediated translocation of ranaviruses has enabled genetic recombination between these two distinct linages.

The risk posed by trade and aquaculture of captive herpetofauna and fish to generating recombinant strains appears particularly relevant for the TFV-like ranaviruses. This polyphyletic clade is almost entirely associated with captivity, except for a TFV isolate from Chad (MW727505.1), which was sampled from a wild crowned bullfrog (*Hoplobatrachus occipitalis*; Box et al., 2021). Notwithstanding, this predominantly captivity-associated clade concordantly exhibited the greatest degree of gene polyphylies among its member isolates. This was particularly apparent for the BIV sub-lineage, whilst the assigned genes of

ToRV1 suggested its genomic makeup was almost completely polyphyletic, equally comprised of FV3-like and CMTV-like origins with an intriguingly high frequency of EHNV-like genes. The exact circumstances that could have permitted coinfection of these diverse lineages remains unclear. Whatever the case, it does however seem apparent that human practices involving animal trade have increased the likelihood of lineages to co-occur, providing opportunity for coinfection and recombination. But further than merely contributing to the generation of diversity within existing ranavirus lineages, trade and aquaculture practices may have possibly facilitated the emergence of the TFV-like clade altogether, borne through repeated recombination events of distinct ranavirus lineages.

It must be noted that there are potential pitfalls to my *Ranavirus* lineage mosaicism assessments. Despite the two statistical checks implemented to discard erroneous gene assignments, it is possible that in some instances the degree of isolate polyphyly could be overestimated. For example, a gene containing a correlation between its reconstructed pairwise distances to those of the concatenated core phylogeny could yet contain an aberrant topology, which may yield significantly distinct mean cophenetic distances between clades for a given isolate's placement. In such a case, a polyphyletic assignment would pass the checks, but depending on the level of diversity within the gene and the resulting tree topology, average cophenetic distances between member isolates of each clade may be misleading. Circumventions to the problem might include manual tree inspections, or inclusion of a further check to assess overly aberrant tree topologies. However, tree comparison metrics such as Robinson-Foulds (Robinson & Foulds, 1981) tend to be extremely sensitive to phylogenies with high diversity involving large numbers of taxa, and therefore are overpowered for purposes of assessing subtle changes in clade memberships (Kuhner & Yamato, 2015). Nevertheless, my results recapitulate previous findings of polyphyly amongst FV3-like ranaviruses, supporting the validity of the approach. However, caution is warranted. Performing tests using simulated data of mosaic genomes with differing levels of genetic diversity in the generated gene units would be an appropriate next step to assess the analytical limits of my approach.

3.5.2 Influence of Recombination on the Genome Evolution of Ranaviruses

Whether recombinogenic sites occur within the genome at inter- or intra-genic loci carries particular significance. Inter-genic recombination occurring between genes in non-coding regions can, on the one hand, be less disruptive as there is a lower

risk of producing deleterious genomic structures compared to introducing variants within protein coding regions (Voigt et al., 2002). On the other hand, intra-genic recombination acts as a major route for proteins to diversify, and more often introduces alleles to genes whilst maintaining stable protein structures, relative to the acquisition of random mutation (Drummond et al., 2005). As such, a balance exists between structural protein constraints and diversifying selective pressures that dictates whether inter- or intra-genic recombination is favoured. Based on within *Ranavirus*-clade alignments of both individual and concatenated core genes, it is apparent that recombination occurs in both inter- and intra-genic regions. Although I found significant evidence for LD decay across each concatenated core alignment, there were many stronger instances of decay within individual genetic units. What is more, recombination breakpoints detected in the concatenated core genomes occurred at the highest rates within genic regions away from gene boundaries (irrespective of clade), and with a non-random association with regions of genic LD decay. Together, this suggests that ranavirus recombination preferentially occurs within coding regions, a trait which may have been selected for as a means to increase protein diversification.

In my characterisations of *Ranavirus* recombination rates, I observed lineage-specific differences. For instance, the ATV-like clade contained the greatest total number of genes with evidence of LD decay, together with greatest number of breakpoints detected in the concatenated core genome. The ATV-like also contained the second lowest number of genomes put forward for recombination analysis, which could be taken as evidence for a highly recombining population given the comparatively limited sample. However, despite my effort to control for temporal confounders by rarefying genomes to a common sampling window, there are other impeding factors that could lead to observed differential rates between clades. For instance, simulations have shown that regions of high genetic diversity decreases the analytical power of detection, masking sites of ancestral recombination (Wiuf et al., 2001). The CMTV-like clade contained a particularly high count of homoplasies, approximately two-fold greater compared to the other clades, together with a very low average regression R^2 strength (0.42%) of LD decay detected along individual genes, which could indicate a potentially high degree of interfering diversity impeding recombination detection. Other factors relating to epidemiology could also have influenced the underlying recombination rates, such as host range. For instance, it could be hypothesised that narrower host ranges could increase coinfection rates of single host animals. Indeed, the ATV-like did contain the most restricted host breadth, infecting only

one genus. However, Shannon's H-Index of host diversity of the included isolates (not shown) was not a significant predictor for the number of breakpoint or regions LD decay observed.

Intriguingly, the genes with strongest evidence for recombination – which also occurred at high frequencies across clades – all appeared to contain metabolic functions involving the hydrolysis of ATP. A subset of these genes were helicases linked to DNA replication, such as the D5 NTP/ATPase (Evans et al., 1995). This particular gene has previously been flagged as a target of recombination in ranaviruses (Price, 2015), which is not necessarily surprising considering the same genomes were included in present analyses. However, the D5 NTP/ATPase gene together with another helicase NTP/ATPase Type III restriction enzyme were detected to recombine in at least three clades – concluded by independent analyses – supporting the pervasive nature of this family of proteins to recombine. Why might this be? I consider two hypotheses. First, these proteins involve immediate-early expressed genes that are crucial to initial stages to DNA replication, such as genome uncoating upon cell infection (Kilcher et al., 2014). Helicase protein domains associated with DNA-binding and ATP hydrolysis are exceptionally conserved (Hutin et al., 2016), and therefore may be highly constrained to incur mutations that alter protein structure in any way. In such a case, rather than as a means to generate genetic diversity, recombination could be used as a strategy to purge deleterious mutation accumulation and promote protein structural integrity, which is known to occur in other organisms (Kass & Jasin, 2010; Sijmons et al., 2015; Wilkinson & Weller, 2004). Second, and contrary to the latter purifying selection hypothesis, these helicase and metabolic proteins may be under selective pressures to diversify and adapt to the many various host factors they interact with. For example, to optimise to the various host proteins recruited to form the complexes required to carry out viral protein function (Wilkinson & Weller, 2004). However, this hypothesis seems less likely, as all clades were sampled from multiple host genera except the ATV-like, which incidentally contained the highest rates of recombination, including amongst said metabolic and helicase proteins.

One potential line of evidence for the diversification of the helicases and metabolic proteins was given through the branch-site selection models I employed. Many of the same proteins discussed above, including the D5 NTP/ATPase, NTP/ATPase Type III restriction enzyme, and a protein kinase were found to have undergone significant diversifying selection. The former two proteins were supported by

both the *BUSTED* and *aBSREL* models, which could be taken as strong evidence that recombination is acting as an adaptive diversifying process in these genes, perhaps in response to host shifts. However, there is a large confounder to this interpretation, which stems from recombinogenic sites known to be overrepresented among false-positives detected by *dN*/*dS* ratio characterisations (Arenas & Posada 2010; Pond et al., 2006a). The principal reason being that recombination can inflate non-synonymous codon substitutions in the absence of adaptive process, which phylogenetic ω rate methodologies are particularly sensitive to (Anisimova et al., 2003).

The problem is further compounded by the fact that true adaptive sites segregating throughout a non-clonal population precludes genetic recombination, creating a chicken-egg conundrum. The only way to navigate the problem is to control for recombination prior to characterising ω rates (Pérez-Losada et al., 2015). One common way to do so is to employ the *GARD* algorithm (Pond et al., 2006b) to detect recombination within individual genes and proceed with ω ratio analyses using the gene alignments partitioned at detected breakpoints (Pérez-Losada et al., 2015; Pond et al., 2006b). That said, I employed *GARD* in this way, but failed to detect any recombination breakpoints within the individual gene alignments containing all isolates across the *Ranavirus* phylogeny, which were those submitted to episodic diversifying selection analysis. As such, I could not employ this technique for recombination control prior to performing selection analyses, despite detecting breakpoints in gene alignments amongst subsets of taxa. Somewhat perversely, the power to detect recombination breakpoints is inversely related to the frequency that they occur, rendering genomic regions with high rates of ancestral recombination difficult to detect (Wiuf et al., 2001). As such, the fact that I failed to detect recombination in genus-wide gene alignments cannot be taken as evidence to legitimise the true-positive detection of diversifying selection of the helicases and metabolic proteins.

Nonetheless, several other genes had significant evidence of diversifying selection that did not co-occur with high detected rates of recombination. Three genes had undergone positive selection on branches of the ALRV phylogeny that coincided with key evolutionary events of *Ranavirus*. First, the ribonuclease reductase barrel domain, an RNA polymerase, implicated the most recent common ancestor (MRCA) to nearly all ALRV taxa, only excluding the *European North Atlantic ranavirus* subclade of the EHNV-like. This is suggestive of the earliest identified adaptation of ancestral ranaviruses key in permitting host tropism of amphibian

and reptile taxa, considering this point on the ranavirus phylogeny represents the first host shift away from the ancestral piscine state. What is more, a second RNA polymerase, the C-terminal domain of the Rpb5 protein, contained significant evidence of having undergone positive selection on the branch of the MRCA to the CMTV-, TFV-, and FV3-like ranaviruses. This implicates the ancestral viral taxa that are thought have switched into amphibian hosts for the second time in *Ranavirus* evolution (Price et al., 2017a), and therefore provides further evidence of key adaptation of ranaviral RNA polymerases required for tropism of herpetofauna. RNA polymerases play a central role in eukaryotic transcription, which involves their recruitment to nuclear promotors and various other transcriptional machineries (Todone et al., 2000). The initial stages of ranavirus mRNA synthesis, and indeed DNA replication, also takes place in the nucleus of the host cell (Jancovich et al., 2015b). It is then not surprising that RNA polymerases might face intense pressure to structurally optimise and conform to the various translational factors of non-native hosts, particularly those of a different taxonomic class whilst switching from fish to amphibians. Finally, a third key protein was positively selected for in MRCA to the TFV-like clade, which was the myristylated membrane protein. This class of DNA virus lipid membrane proteins have been identified as targets for potent neutralizing monoclonal antibodies (Wolffe et al., 1995), and therefore likely represents an adaptation to immune evasion. Intriguingly, on the gene tree of this protein (group 144; Fig. 3.7), both subclades of the BIV and TFV lineages formed a monophyletic clade derived from the CMTV-like. This may therefore represent adaptation to immune evasion in an ancestral CMTV-like lineage.

3.5.3 Conclusions and Future Directions

Recombination has clearly contributed significantly to the evolution of *Ranavirus*. Here, I have presented lines of evidence that suggest that genetic recombination may be an ancestral and inherent process enabling ranaviruses to diversify and adapt, but may also contribute to important roles such as maintenance of genomic stability. The pervasive nature of ranaviruses to recombine raises immediate concerns relating to the human-assisted movement of host animals implicated in trade – particularly those associated with aquaculture – given the increased virulence observed in recombinant strains (Claytor et al., 2017). High rates of recombination across the phylogeny of ranaviruses also raises implications for their taxonomy, as mosaic lineages do not only seem plausible, but likely. In particular, the captivity-associated TFV-like clade appear highly polyphyletic, rendering their phylogenetic reconstruction on a single tree problematic. Future

formal classification of the genus implementing phylogenetic methodology should take this fact into consideration.

The differential rates of recombination between lineages that I identified were largely equivocal. Future work is needed to firstly determine whether the differences are true and not an artefact of sampling biases, before undertaking an explicit examination of whether differences are due to genomic or epidemiological factors. Ideally, rates of *Ranavirus* recombination should be quantified in units of substitutions per site per year (or genome replication) to thoroughly characterise the rate which genetic diversity is generated beyond mutation alone. For this to be achieved, a much greater resolution of genomic sampling for each lineage through time is required to ensure the presence of a robust temporal signal of genetic diversification (Drummond et al., 2003). Lab-based experimental designs should then be developed to determine genomic factors contributing to any true differential rates, such as the effect of genetic distance. Finally, experimental work is needed to determine the functional significance of highly recombining sites and resulting phenotypes, particularly as a quantification towards *in situ* risk assessments of recombinant strain emergence.

4

Phylogenetic Dating Points to Global Spread of *Frog Virus 3* Over a Century Before Disease Detection

4.1 Abstract

Globalisation is a key contributor to the spread of infectious diseases. Phylogenetic analysis of pathogen genomes offers opportunities to reconstruct their origin and spread in space and time. Such efforts in recent years have yielded vital insights on host-pathogen associations at different evolutionary scales. In this chapter, I investigate the historical origins and translocations of *Ranavirus*, amphibian viral pathogens of major conservation concern, with a particular focus on characterising the introductions of the FV3 lineage into the United Kingdom. To do so, I applied a phylodynamic approach to a dataset of 58 FV3-like ranavirus whole-genome sequences sampled globally to reconstruct their genetic diversity through time. I uncovered extensive human-mediated translocations of FV3-like ranaviruses between North America and Europe potentially dating back as early as the 19th Century, preceding the first detection of amphibian ranaviral disease in the UK by up to nearly two centuries. Furthermore, analysis of population structure revealed inferred ancestry components shared by viruses implicated in outbreaks amongst frog-farming facilities on separate continents, suggesting that intercontinental trade-linked networks have contributed to the geographic distribution of ranaviruses today. This chapter highlights the risk that animal trade poses to disease emergence in wildlife pathogens, both since the onset of globalisation and in current times.

4.2 Introduction

One of the core facets of modern globalisation is the stark increase in international trade that began accelerating in the 19th Century (Vanham, 2019). Despite the undeniable benefits reaped through the sharing of ideas, technologies and cultural beliefs, the drastic change in the movement of people and goods quickly became key contributors to the spread of infectious diseases. It is often argued that emerging infectious diseases (EIDs) are increasing in incidence (although, see: Rosenberg et al., 2013), which is thought to predominantly be driven by pathogens with a propensity to switch hosts, such as zoonoses shared between humans and animals (Jones et al., 2008).

Pathogen spillover events from one host species population to another are ultimately caused by substantial changes to host-pathogen dynamics. Whilst there are important evolutionary components to such changes, environmentally-driven perturbations to population demographics, health, and spatial distributions most often cause pathogen spillover from a disease reservoir population (Daszak et al., 2000, 2003). Whether or not the occurrence of EIDs linked to spillover events is elevated beyond the natural baseline is debatable. However, it is abundantly clear that the consequences of anthropogenic activity such as climate change, non-native species invasions, and habitat loss and destruction are important proximate drivers to perturbations of host-pathogen dynamics and disease spillover (Allen et al., 2017; Borremans et al., 2019; Cohen et al., 2019; Faust et al., 2018).

Animal trade is another important proximate cause of emergent diseases, as it can create varied opportunities for pathogens to expand in range and spillover into new and naïve host populations. Anthropogenic-mediated disease spread of this kind has been termed 'pathogen pollution' (Cunningham et al., 2003; Daszak et al., 2003), given its sometimes comparable impact to ecosystems as other forms of environmental pollution. Phylogenetic distance between host species is an important determinant of successful sharing of pathogens between animals (Shaw et al., 2020; Woolhouse & Gowtage-Sequeria, 2005). Given the taxonomic diversity of animals traded globally for exotic pets and commercial aqua- and agricultural practices, pathogen pollution through animal trade poses a grave risk to local wildlife health in regions with large animal trade economies. Numerous examples exist of animal trade-linked epizootic outbreaks in native fauna, including: rinderpest transmitted from cattle to African Bovidae (Kock et al., 1999); crayfish plague introduced throughout Europe from imported invasive North American

signal crayfish (Holdich et al., 2009); and chytridiomycosis caused by the panzootic lineage of chytrid fungus that has resulted in expatriations and extinctions of numerous species of amphibians across the globe (O'Hanlon et al., 2018).

Ranavirus, a genus of double-stranded DNA viruses of the family *Iridoviridae*, are also wildlife pathogens of amphibians that have been introduced to multiple regions beyond their considered native distributions. There are three main lineages of amphibian-associated *Ranavirus* (AARV) found globally; it is thought the *Frog virus 3* (FV3)-like and *Ambystoma tigrinum virus* (ATV)-like lineages originated in North America, whilst the *Common midwife toad virus* (CMTV)-like lineage is endemic to Europe (Price et al., 2017a). Species of all three lineages have demonstrated range expansions that most likely occurred through animal trade, with ATV-likes spreading within North America via anglers using diseased salamanders sold as bait (Epstein & Storfer, 2016), and both CMTV- and FV3-likes emerging in Asian farmed amphibians with transcontinental trade links (Chen et al., 1999; Meng et al., 2014; Mu et al., 2018; Zhang et al., 2001), with possible spillover in wild amphibian populations (Xu et al., 2010). Moreover, the FV3-like lineage is particularly notable for its expansion into the historically naïve common frog (*Rana temporaria*) population of the United Kingdom, as it is one of the few cited examples of persistent ranavirosis-driven population declines without recovery (Teacher et al., 2010).

The FV3 lineage is thought to have emerged in the UK in the late 1980s, following the detection of the first ranavirus-associated mortality of native common frogs (Cunningham et al., 1996). Invasion into the UK was later inferred based on the isolates exhibiting a high degree of molecular sequence homology to the North American FV3 lineage that emerged earlier in the 1960s, together with the complete lack of ranavirus detection on any of the smaller British Isles and the Republic of Ireland (Hyatt et al., 2000; Price et al., 2016). In response to the unusual mortality events in the UK, a long-term citizen science initiative named the Frog Mortality Project (FMP) was established with the aim to record the health of UK amphibian populations (Cunningham et al., 1996). The extensive epidemiological data collected through the FMP over nearly three decades revealed two major localities of Lancashire and Greater London with increased ranavirosis prevalence, both of which have been suggested as the points of introduction into the UK (Price et al., 2016).

Molecular characterisations of ranaviruses have refined our understanding of ranavirus association to the UK. Based on the relationship of early ranaviruses isolated from diseased amphibian carcasses in the UK, a longstanding belief was established that the FV3 lineage was the only one in circulation following its introduction (Duffus et al., 2015; Hyatt et al., 2000). Recently, however, examination of archived FMP samples identified two phylogenetically distinct CMTV-like virus-associated incidents separated by 300 Km and 16 years, suggesting repeated introductions despite no detection of CMTV in the UK since (Price et al., 2017b). Moreover, Duffus et al. (2017) performed the most comprehensive phylogeographic survey to date using sequences of two partial loci of 24 UK FV3-like isolates sampled across the country. They found three monophyletic clusters partially correlating with geographic locations, which were taken by the authors as support for at least three independent introductions of FV3-like ranaviruses into the UK.

Whole genome sequence (WGS) methods employing high-resolution phylogenetic reconstruction arguably offers the most valuable approach towards a more complete understanding of the timing and frequency of ranavirus introductions into the UK. First, WGS resolution permits access to any relevant signal a set of genomes may impart for evolutionary reconstruction. For example, *Mycobacterium bovis*, the aetiological agent of bovine tuberculosis (bTB), is a clonally reproducing bacteria that does not recombine and has slow evolutionary rates. However, whole-genome sequencing of 31 isolates over a small spatiotemporal scale permitted genotyping to 39 polymorphic sites, which proved sufficient to adequately reconstruct transmission between two host species and identify spatial patterns of the genotypes causing bTB outbreaks on farms in Northern Ireland (Biek et al., 2012). Continued and more intense genomic sampling later revealed a slow heterogenous spatial dispersal of the different genotypes, averaging 2 Km per year (Trewby et al., 2016).

Second, sufficient longitudinal sampling of WGS data can capture evolutionary change in real time, permitting time-calibration of phylogenetic trees. Together with epidemiological and spatial components, time-calibration of pathogen genealogies stands as the core pillar to what is known as phylodynamic reconstruction, which often stands as the only means of elucidating wildlife disease emergence and spread (Drummond et al., 2003; Grenfell et al., 2004). For instance, such reconstructions can be used to retrace disease outbreaks and identify reservoir populations in multihost systems, whilst also obtaining direct

estimates of dates pathogen introduction and spillover events occur. As an example, in the face of much contention given an aggressive eradication campaign, phylodynamic analysis conclusively identified the sources of recurrent outbreaks of the bacterial pathogen *Brucella abortis* amongst cattle livestock in the Greater Yellowstone Ecosystem (Kamath et al., 2016). A large-scale genomic sample illustrated that *B. abortis* was being maintained in wild bison and elk, but that isolates sampled from the elk population were responsible for the highest rates of host switching into commercial livestock. Results such as these demonstrate how important insights yielded from phylodynamic analysis can be necessary to form effective wildlife disease control and mitigation strategies.

In this chapter, I build on previous efforts to characterise the historical origins and movements of the FV3 lineage using whole-genome phylogenetic approaches, with a particular focus on *Ranavirus* presence in the UK. Specifically, I sought to: i) reconstruct the global phylogeography of the FV3 lineage using phylodynamic methodology; ii) determine the sources, frequency, and timing of ranavirus invasions into UK; and iii) characterise the population structure of all available FV3-like isolates using genetic marker clustering to uncover ancestry components shared between populations that may be linked by trade networks. To achieve these aims, I acquired 34 new complete genomes of UK ranaviruses and combined these with all publicly available FV3-like genomes sampled globally to produce a dataset of 58 complete genomes spanning from the 1960s to the present.

4.3 Methods

4.3.1 Sample Acquisition

As outlined in Chapter 2, all publicly available complete FV3-like ranavirus whole genome assemblies were downloaded from the NCBI Nucleotide database. I additionally included Bohle iridovirus (BIV; NC_038507.1) as an outgroup isolate for phylogenetic inference, based on it being sufficiently and equally diverged from all FV3-like isolates. Novel samples were acquired from Dr. Stephen Price, who obtained and sequenced isolates from archived amphibian tissues of either liver or toe/tail clips from carcasses collected by the FMP, and later GWH, spanning the years 1992 to 2017. Sample processing, library preparation, and genome sequencing are outlined in Chapter 2 (Methods 2.3.1). After merging the novel and public datasets, I performed quality control steps which included excluding isolates lacking reported sampling dates, which are necessary for phylogenetic time calibration analysis (see below).

4.3.2 Core Genome Sequence Alignments

Core genome delineation. I employed a pan-genome analysis approach to delineate and align all core genes within the FV3-like *Ranavirus* sample set, which is outlined in Chapter 3 (Methods 3.3.1). Briefly, I first used the genome annotation tool *Prokka* v1.14.6 (Seemann, 2014) to identify coding sequence (CDS) open reading frames (ORFs) in all FV3-like WGS assemblies. Next, I applied the *Roary* pipeline v3.11.12 (Croucher et al., 2015) to cluster all CDSs at a minimum amino-acid homology threshold of 80%. Core genes were considered as orthologous ORF clusters that were present in all FV3-like isolates (including the outgroup).

Multiple sequence alignments. Depending on the analysis, all core ORFs or subsets for each FV3-like ranavirus isolate were concatenated into contiguous sequences. Multiple sequence alignments (MSA) of concatenated genes were produced using *MAFFT* v7.453 (Katoh & Standley, 2013) with the default settings. All MSAs were then trimmed for any gap positions found in 20% or more of isolates using the command line tool *trimAL* v1.4 (Capella-Gutiérrez et al., 2009).

4.3.3 Phylogenetic Inference

Control for recombination. I screened for and removed putative recombinant sites from all alignments used in downstream phylogenetic analyses by employing a stringent protocol based on detection and removal of homoplasies, as in the preceding chapters. First, I produced a Maximum Parsimony tree using the rapid tree inference tool *MPBoot* v1.1.0 (Hoang et al., 2018) with 1,000 bootstraps. The resulting tree and input alignment were then input to *homoplasyFinder* v0.0.0.9 (Crispell et al., 2019). For each site in the alignment, *homoplasyFinder* provides a consistency index which captures the minimum number of state changes required on the Maximum Parsimony tree to explain the character state at the tips (Fitch, 1971). As recommended by the authors, sites with a consistency index of ≤ 0.5 were considered homoplastic and removed from MSAs.

Maximum Likelihood inference. To delineate the relationships between isolates of the homoplasy-filtered concatenated alignment containing all FV3-like core genes, I used a Maximum Likelihood (ML) phylogenetic reconstruction approach implemented in *RAxML* v8.2.12 (Stamatakis, 2014). The general time-reversable model of nucleotide substitution was specified with a Gamma distribution of transition-transversion rate variation. I ran 100 ML iterations, and selected the best tree. A separate bootstrapping run was executed alongside with 1,000 iterations, which I used to annotate the best ML tree for node support.

4.3.4 Phylogenetic Time-Calibration

Temporal signal assessment. An essential prerequisite to calibrating phylogenetic trees by units of time is the presence of temporal signal, or clock-like evolution (Drummond et al., 2003). I employed a regression-based approach (Rambaut et al., 2016) to assess evidence of temporal signal within the MSA using the *roottotip* function of the *R* package *BactDating* (Didelot et al., 2018). The FV3-like ML tree scaled by absolute substitutions was used as the input, and tip-dates were converted to decimal-date format. I determined the presence of significant temporal signal using a date randomisation test following 10,000 permutations. Amongst all assemblies, only two (novel) isolates lacked a sample date required for phylogenetic time calibration, and therefore had to be omitted from temporal analyses.

I identified globally weak temporal signal within the concatenated alignment of all core FV3-like ORFs. As such, I employed a gene-by-gene approach to maximise alignment temporality by constraining the inclusion of genes in concatenated alignments to those with elevated degrees of temporal signal. Specifically, I developed a protocol that assessed the temporal signal in discrete alignments of individual core ORFs. I then selected all core ORFs exhibiting the strongest temporal signal, determined by the lowest permutation *P*-values (relaxed to < 0.06) and adequate regression R^2 ($> 1\%$) to concatenate, align, homoplasy-filter and use to build an ML tree, as previously described. The final tree, built using the concatenated subset of ORFs, was then subsequently reassessed for improved temporal signal, and the resultant MSA was carried forwards for phylogenetic time calibration.

Bayesian phylogenetic inference. To calibrate the FV3-like ranavirus phylogeny by time, I implemented the *BEAST2* v2.6.3 (Bouckaert et al., 2019) framework. Bayesian Evolutionary Analysis by Sampling Trees version 2 (*BEAST2*) is a platform for conducting Bayesian phylogenetic inference by Markov chain Monte Carlo (MCMC) analysis. Bayes' Theorem, at its core, is a way of modelling the conditional probability of an event given a prior understanding relevant to that event. Ultimately, the posterior probability is derived, which represents a prediction of the event based on updating the prior given the data. This paves a way to comprehensively explore parameter space and uncertainty around a model. It is given as follows:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

In understanding each component in terms of a *BEAST2* model, it is perhaps easiest to start with the prior, $P(A)$. This is the probability of the model parameters, or the prior belief in the hypotheses behind specifying those parameters. These take the shape of the chosen demographic model, substitution model, clock rate, amongst others, and requires some prior understanding of the biology of the system being analysed. Next is the likelihood, $P(B|A)$, which is calculated based on the probability of the data given the parameters defined by the model. The posterior, $P(A|B)$, is calculated using the updated probability for the model parameters in light of the data. Finally, the marginal likelihood, $P(B)$, represents the probability of the data *under* the model, and therefore any combination of parameters therein. Its calculation is almost insurmountable to achieve, and is therefore only briefly approximated. However, the marginal likelihood has a key use in model comparison, and so several methods have been developed for its more detailed approximation for these purposes (see bleow; Russel et al., 2019).

Initially, I ran the *bModelTest* (Bouckaert & Drummond, 2017), which employs a reversable jump MCMC algorithm, allowing the Markov chain to jump between states representing the different available substitution models. This method can be used as a model-averaging technique when appropriate parameters are not known, whilst simultaneously estimating the trees. However, model-averaging runs can be used to assess the proportion of time the MCMC spent moving between states. As such, the state in which the chain spent the greatest proportion of time represents posterior support for that given model, whilst also providing initial estimates for unknown parameters, including the clock rate and the proportion of sites invariant to rate heterogeneity.

After, I ran four different models each using both the restricted alignment containing strong temporal signal as well as the MSA of all FV3 core genes. I used two variant demographic models: the coalescent constant and the coalescent exponential. In each, I implemented a strict clock with a log-normal distribution of rates imposed over all lineages, and a relaxed log-normal clock for the event that lineages evolved at different rates. Common to each model was the transversion model (TVM; 123451) of nucleotide substitution (allowing variable transversion, but not transition, rate frequencies; estimated using the *bModelTest*),

as well as four Gamma count categories of site heterogeneity, an estimated proportion of invariant sites, and 50 million MCMC generations with trees and model parameters sampled every 50 thousand steps. After the completion of all runs, I examined the log files using *Tracer* v1.7.1 (Rambaut et al., 2018) to assess model convergences, requiring Effective Sample Size (ESS) values > 200. I then performed model selection using Bayes Factors (BF), calculated by comparing marginal likelihood values derived from each model in a pairwise manner. These values were estimated using the nested sampling method, with a sub-chain length of 20 million steps and 38 particles for each model (Russel et al., 2019). Finally, I used *TreeAnnotator* to select and annotate the tree with greatest posterior support from the model favoured by the obtained Bayes Factors.

Lastly, all phylogenies described thus far were drawn in *R* using the packages *ape* (Paradis & Schliep, 2019), *phytools* (Revell, 2012) and *ggtree* (Yu et al., 2017). A *post hoc* analysis was also run using *ape*, in which I used the coalescence events of the time-calibrated phylogeny of the FV3-like sub-clades to estimate changes in effective population size through time. This was done using a generalised skyline plot, which is an extension of the classic skyline plot first implemented by Pybus et al. (2000) using maximum likelihood methodology.

4.3.5 Population Structure

To assess the degree of population structure in FV3-like ranaviruses and to investigate the influence of trade networks through shared genomic signatures, I applied the model-based clustering algorithm *ADMIXTURE* v1.3.0 (Alexander & Lange, 2011). As *ADMIXTURE* estimates individual ancestry proportions using multi-locus genotype data in approximate linkage equilibrium, I used a concatenated alignment of all core FV3-like ORFs without homoplasy pruning to consider the maximum possible number of genomic markers. Though, to filter out tightly linked single nucleotide polymorphisms (SNPs), I developed a protocol following Richard et al. (2020) to calculate pairwise linkage disequilibrium (LD) statistics between all variant sites (the pipeline is outlined in Chapter 3, Methods 3.3.3). Briefly, all FV3-like core ORFs were positioned, orientated, and aligned in relation to a reference WGS with a genomic arrangement representative of the population so that variants were called with whole-genome coordinates using the command line tool *SNP-sites* v2.5.1 (Page et al., 2016). I then used *Tomahawk* v0.7.0 (https://mklarqvist.github.io/tomahawk/) to calculate pairwise LD statistics between all SNPs, and excluded highly linked sites with an LD $r^2 \geq 0.5$. The remaining SNPs, pruned of background LD, were then used to infer the optimal

number of FV3-like ancestral components, based on the optimal number of clusters $'K'$ identified by the lowest cross-validation error score.

4.4 Results

4.4.1 Genomic Dataset and Phylogenetic Reconstructions

The curated whole-genome dataset contained a total of 58 complete FV3-like ranavirus assemblies comprising 34 novel whole-genomes and 24 publicly available assemblies acquired from NCBI GenBank. Together, the isolates were sampled across seven countries spanning four continents, collected between the years 1966 to 2017 (Table S2).

Following identification of the ORF coding sequences within each assembly using *Prokka*, I employed the *Roary* pipeline to delineate the pan- and core genome content. I identified a total of 69 core ORFs present in all included FV3-like assemblies when using an 80% protein homology threshold. After being trimmed of gap positions, the concatenated alignment of all core ORFs was a total length of 59,404 bp, covering 56.1% of the RefSeq FV3 genome (NC_005946.1).

To identify sites putatively deriving from recombination which could potentially lead to inaccurate phylogenetic reconstruction, I applied homoplasy filtering which identified 553 sites within the complete FV3-like core genome alignment. After removal of these sites, the remaining core genome MSA contained 1,813 SNPs. An ML phylogenetic tree built on this alignment revealed two well supported major clades of FV3-like ranaviruses: a basal clade made up of Asian and North American lineages, and another, predominantly New World clade, containing mixed European (UK) and American lineages (Fig. 4.1). The backbone of the latter, larger clade contained poor support for the topology of various lineages within, owing to low genetic diversity amongst the sequences, despite distinct geographic origins and a wide sampling window ($1966 - 2017$).

4.4.2 Phylodynamic Analysis

As noted in the Methods, the concatenated alignment using the complete set of core genes contained weak temporal signal ($R^2 = 0.06$). As such, I employed an approach to filter the genomic data included in the alignment with the view to maximise temporal signal. To do so, I assessed individual gene temporal signal in

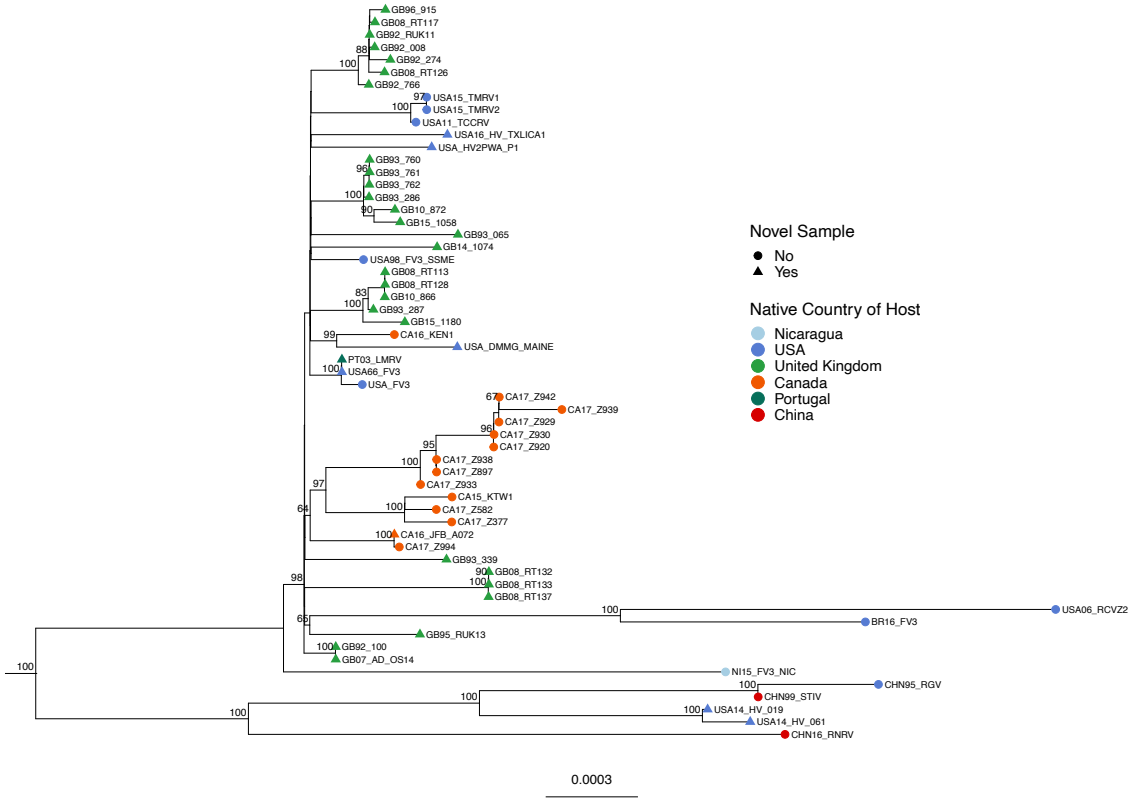


Figure 4.1. *Maximum Likelihood phylogeny of 58 FV3-like ranavirus isolates sampled globally. The tree was built using a concatenated MSA of core genes with RAxML, specifying the GTR + **Γ** substitution model. The MSA contains 69 core genes and 1,813 SNPs after removing homoplasic sites to control for recombination. Node support values are derived from 1,000 bootstrap replicates. The tree is scaled by substitutions/site^{-1} and has had the BIV root isolate truncated for visual clarity.*

the core ORF set. I identified 15 core FV3-like ORFs containing strong signal based on low root-to-tip regression permutation *P*-values and adequate regression R^2 (Table S4). After I concatenated, aligned and homoplasy-filtered the subset of 15 ORFs, the MSA comprised 19,093 sites (18% of the FV3 reference genome) varying over 619 SNPs. A resulting ML tree constructed from the MSA exhibited a highly significant root-to-tip regression, suggesting strong temporal signal within the alignment ($R^2 = 0.12$, permutation $P = 0.00303$; Fig. S4).

I proceeded with time calibration of both the 69 ORF (all core) and 15 ORF (strong temporal signal) MSAs using the Bayesian tip-dating framework implemented in *BEAST2*, testing a range of demographic and clock models (see Methods 4.3.4). The model fit of converged runs for each alignment were assessed by BF support, derived from approximated marginal likelihood values computed by nested

sampling runs. However, given the inappropriateness of using BF support to select the best-fitting models *between* alignments (using different datasets), I additionally assessed the congruency in the modelled uncertainty between significant events in the phylogenies reconstructed by each alignment. To do so, I compared the degree of overlap in the HPD intervals around the date estimates of important nodes on the respective trees (Fig. S5). I observed a high degree of overlap in all nodes, suggesting congruent phylogenetic signal was present in both alignments. Thus, given the stronger temporal signal, I chose to move forward to infer the timing of evolutionary events based on the 15 ORF MSA.

Of the four variant models tested using the 15 ORF MSA, I found that a strict clock model was always favoured over a relaxed clock, and that the coalescent constant demographic model was overwhelmingly favoured over the exponential with BF support of 12.102 (coalescent constant Marginal Likelihood: -32364.987 SD 2.47 versus coalescent exponential Marginal Likelihood: -32377.089 SD 2.59). The coalescent constant model with a strict clock yielded a clock rate estimate of $5.845e^{-06}$ (95% HPD $3.299e^{-06} - 8.214e^{-06}$) substitutions per site per year and 0.291 (95% HPD $7.734e^{-08} - 0.561$) proportion of estimated invariant sites across the 15 ORFs (~18% of the average FV3-like genome).

The topology of the FV3-like ranavirus time calibrated phylogeny was largely in agreement with its ML counterpart, despite that the inference necessarily used less genomic data (34.14% of SNPs). All sub-clades clustered consistently, including the basal USA and Chinese lineages (Fig. 4.2). Using this approach, I infer the time to most recent common ancestor (tMRCA) of all FV3-like isolates included in the analysis to the mid 16th Century (1550.34; 95% HPD $1333.56 - 1743.50$). This date marks the divergence of the most basal USA and Asian lineages from other North American and later European lineages. Note that only one European isolate included in the analysis was sourced from outside the UK, sampled in Portugal. Nevertheless, the parent node to all lineages of European isolates coalesced at the date 1802.23 (95% HPD $1698.71 - 1889.21$), indicating a time period consistent with the earliest *possible* presence, and therefore introduction, of FV3 ranaviruses into the UK.

However, the time-calibrated phylogeny resolved the sub-clade with all UK isolates to also contain ranaviruses sampled from the Americas, rendering the coalescence of all UK samples an unlikely date of first introduction to the British

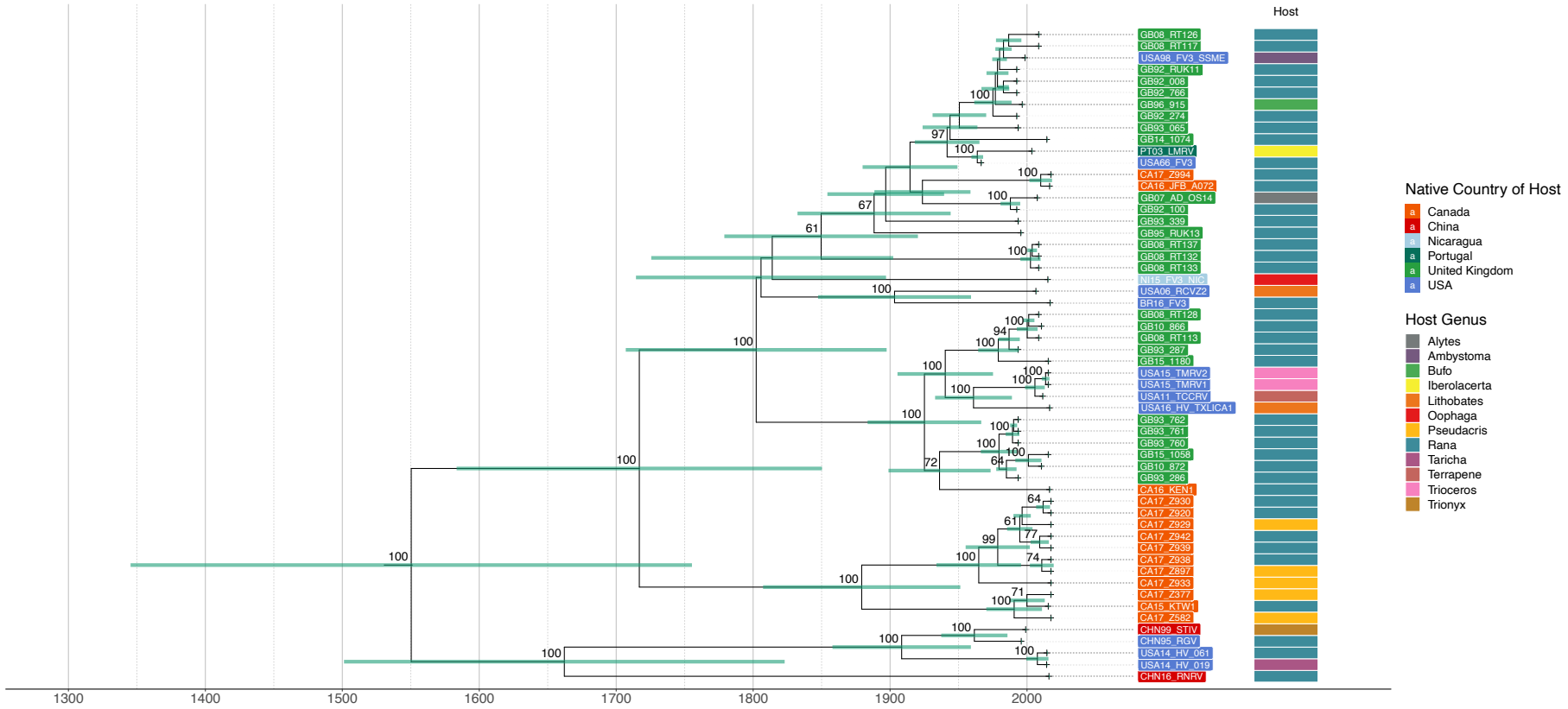


Figure 4.2. *FV3–like ranavirus time–calibrated phylogeny. The tree is scaled by substitutions/site^{-1}year^{-1} and was built using a concatenated multiple sequence alignment of 15 core genes extracted from 56 isolates with known sample dates. The MSA contained 619 SNPs after homoplastic site filtering. A coalescent constant model was implemented in The BEAST2 framework, imposing a strict lognormal clock and the TVM substitution model. Node labels are posterior probabilities with only supports >60% shown, and bars represent node height 95% HPD intervals. Isolate labels are coloured according to the native country of each host, and the heatmap shows the genus of the host.*

Isles. Specifically, the phylogeographic conflicts consisted of six coalescence events: three with USA isolates, two from Canada, and one from Brazil (the latter however contained $< 60\%$ posterior support with two USA isolates, making it difficult to infer the direct ancestor). Two of the conflicts formed comparatively recent (tMRCA ~1980) geographically monophyletic UK sub-clades with a basal non-UK ancestor, which I inferred as likely independent introduction events. However, the largest and earliest sub-clade (tMRCA ~1830, top clade of tree; Fig. 4.2) of UK isolates was geographically paraphyletic, including many isolates sampled from North America with basal UK ancestors. Whether the three implicated topographical conflicts represent subsequent recurrent translocation events into the UK or back into North America is challenging to discern from the phylogeny alone. What is more, the paraphyletic structure of this early clade contained many UK isolates at the end of long, isolated branches, indicating that the clade may in fact be formed of multiple recurrent introductions into the UK. As such, although the sub-clade likely contained multiple introductions, I inferred only the lineage with a basal non-UK ancestor as a likely UK importation, which could have occurred as early as 1828.84 (95% HPD $1729.92 - 1906.65$).

Taken together, given the current dataset, I infer a minimum of three independent introductions into the UK from North America, with three further translocation events which could either represent subsequent recurrent UK introductions, or possible back introductions into North America (Fig. 4.3 **A**). However, the structure of the time calibrated phylogeny, together with considerations of the effective population size of the basal North American source population, likely render this number of UK invasions as a substantial underestimate.

4.4.3 Phylogeography of UK Ranaviruses

The North American and UK FV3-like ranavirus isolates exhibited very different phylogeographic patterns. I observed that all geographically monophyletic UK sub-clades within the FV3-like time-calibrated phylogeny also contained a heterogenous distribution of isolates from discrete regions, often spanning the length of the UK, whilst the North American isolates generally exhibited greater regional phylogeographic concordance (Fig. S6). Due to this, I could not use the phylogeography of the UK FV3-likes as an additional indicator for the number of potential introductions into the UK. The pattern does, however, suggest widespread translocations of genetic lineages of FV3s across the UK, as identified previously (Price et al., 2016).

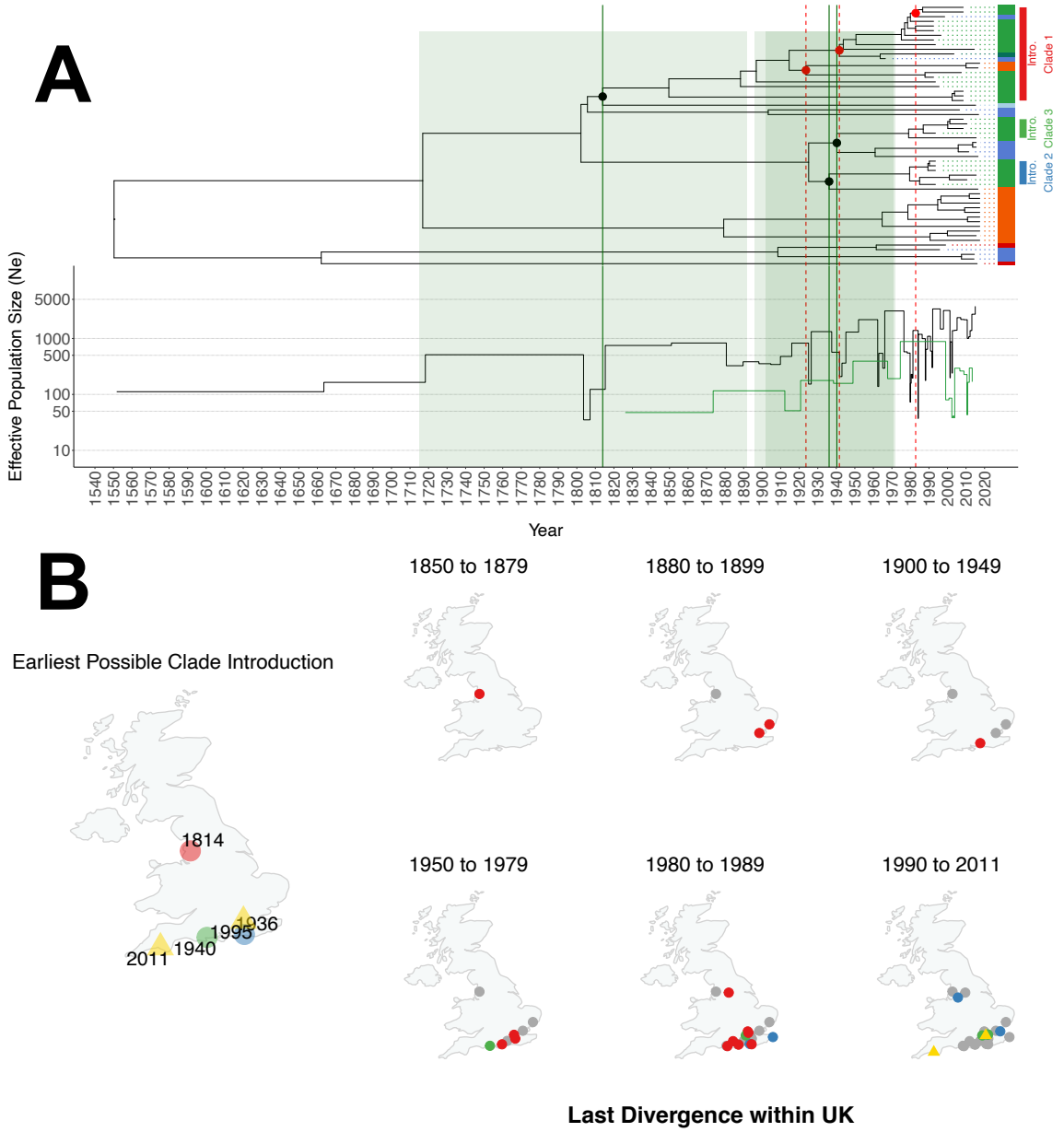


Figure 4.3. *Phylogenetically inferred dates of FV3-like ranavirus introductions and phylogeographic emergences in the UK. Panel* ***A*** *shows the tMRCA between UK and non-UK isolates, used to infer the date and frequency of ranavirus translocations between countries (colour coded heatmap according to Fig. 4.2: China* red, *USA* blue, *Canada* orange, *UK* green, *Nicaragua* light blue, *Portugal* dark green*). Changes in effective population size (N_e) were estimated for the entire sampled population (black N_e line) and the UK-only population (green N_e line), using coalescence events within the whole phylogeny and between only UK isolates only, respectively. Vertical dark green lines and shaded bars represent confident UK introduction date point estimates (black node points) and 95% HPD intervals, whilst dashed lines (red node points) are ambiguous translocation events (see text). Panel* ***B*** *shows the basal sequence location and earliest-possible date of each inferred introduction clade, and the conservative estimate for the location date of each sampled isolate of each introduction clade based on the last divergence within the UK (terminal branch parent node within UK sub-clades). Yellow triangles denote sample locations of British CMTV-like isolates.*

Based on the associated locations and last divergence dates between lineages of isolates within the UK, I infer that the first UK introductions occurred in the north of the UK in Lancashire (Fig. 4.3 **B**), close to the city of Manchester (latitude and longitude: 53°45′13.64″N, 2°21′31.07″W). However, overlapping HPD intervals may cast uncertainty over this location. Nevertheless, the most basal isolates (RT132, RT133, and RT137) stemming from this event are triplicate samples collected from the same outbreak. As such, the parent node of this small subclade provides the earliest-possible estimate that the ancestor to these isolates was present within the UK at 1849.65 (95% HPD 1776.39 − 1917.77). Assuming the clade of taxa associated with first introduction were descended from it, the proceeding divergences occurred over the following 50 years and involved isolates sampled in Southeast England in the counties of Norfolk and Essex (approximately 300 Km away), indicating likely human-mediated translocations.

Two subsequent introductions both occurred in the South of England at 1936.12 and 1940.30 (95% HPD 1895.97 − 1970.56 and 1902.90 − 1971.86, respectively). Similarly to the earliest invasion, I observed isolates derived from both later introductions across all localities from the South Coast of England, Greater London and Lancashire over a 50-year time span (Fig. 4.3 **B**), which points to extensive human-mediated translocations given the isolate relatedness and distances travelled.

Finally, despite not being considered in the phylodynamic analysis, the two CMTV lineages were concurrently sampled in the South of England; in Greater London in 1995 (51°18′43.2″N, 0°18′43.2″W) and in Plymouth in 2011 (50°23′2.4″N, 4°3′57.6″W).

4.4.4 Population Structure of the FV3 Lineage

To formally test for population structure, I applied the allele-frequency based clustering algorithm *ADMIXTURE*. After filtering tightly linked sites ($r^2 \geq 0.5$), a total of 728 SNPs were preserved to carry forward for SNP clustering. Further, the cross-validation error scoring suggested that $K = 4$ was the optimal number of presumed ancestral populations for the FV3-like isolates included in the analysis (Fig. S7). I further observed that the majority of the FV3-like ranavirus isolates were suggested to belong to a single inferred ancestral population, regardless of geographic origin of sampling (Fig. 4.4). The greatest amount of ancestral diversity was found in North America, despite a marginally larger European (almost

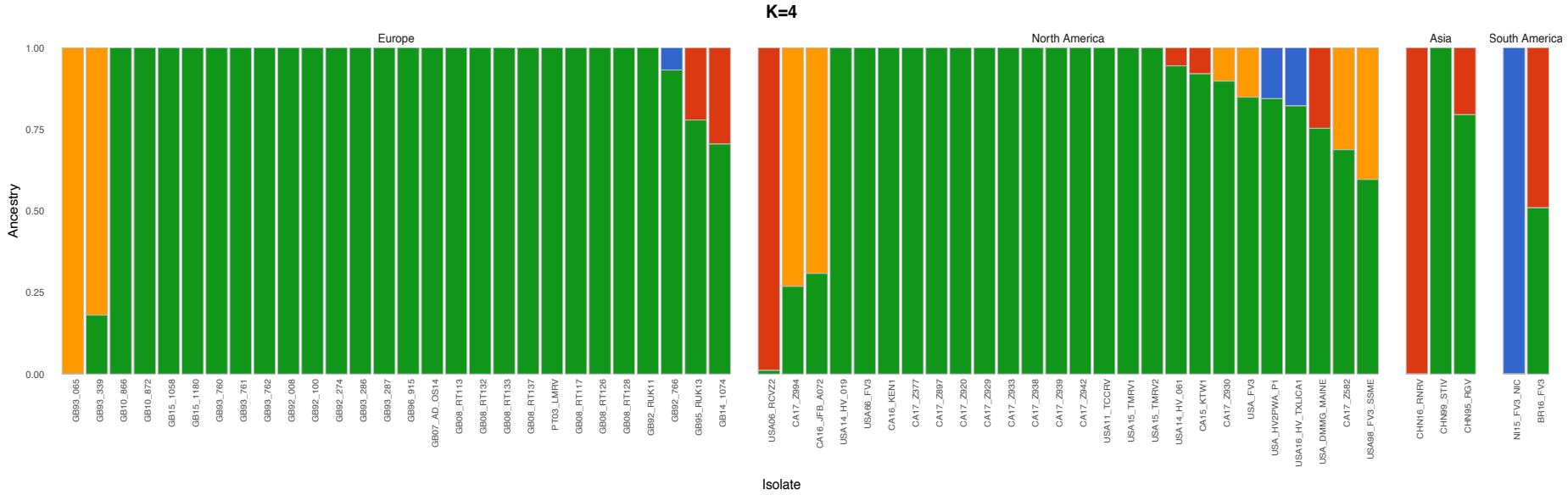


Figure 4.4. *Ancestral population structure of 58 FV3-like ranavirus isolates. The core genome alignment of 69 ORFs was used to call SNPs. Sites with an LD* r^2 *correlation* >0.5 *were excluded, retaining 728 SNPs in approximate linkage equilibrium. SNPs were clustered using the* ADMIXTURE *algorithm to estimate the proportion of ancestry for each isolate. Optimal value of* $\mathrm{K} = 4$ *ancestral populations was determined by cross-validation error analysis (Fig. S7).*

exclusively UK) sample size. One population was ascribed to the UK, as it comprised 100% of the genotype of novel UK isolate GB93_065. This population also contributed 82% of ancestral component to another UK isolate (GB93_339), as well as to multiple Canadian isolates and one USA isolate, to a lesser proportion. I interpret this pattern as consistent with UK virus introgression into the North American FV3-like ranavirus population. Moreover, I suggest this could further offer support to a direction of movement for the ambiguous translocation events outlined by the time calibrated phylogeny, especially as the North American sequences USA66_FV3 (AY548484.1), USA98_FV3_SSME (KJ175144.1), CA16_JFB_A072 (novel isolate), CA17_Z994 (MK959621.1) involved in these evets contained the ancestry component enriched in UK isolates. The genotype of the Chinese isolate RNRV (MG791866.1) was also ascribed entirely to one ancestral population group, which comprised 98.9% of the ancestry of the isolate RCVZ2 (MF187209.1), along with 49% ancestry of the Brazilian isolate from 2016 (MH351268.1). Interestingly, these isolates were sampled from captive ranaculture facilities across three continents all breeding American bullfrogs (*Rana catesbeianus*), which I take as evidence indicating a network of FV3-like ranaviruses translocated through ranaculture trade.

4.5 Discussion

Since their emergence in wild amphibian populations, ranaviruses and their associated disease in amphibians have been a global conservation concern. Whilst large strides have been made in the characterisation of ranavirus biology, factors implicated in their emergence and spread have remained elusive, having often relied on speculations warranting further lines of evidence. In this chapter, I have taken advantage of the largest collective genomic sampling effort of ranaviruses to date, applying phylogenetic and population genetic techniques to provide quantitative insights into the origins and historical translocations of the FV3 lineage.

4.5.1 Origins and Translocations of the FV3 Lineage

Early molecular characterisations of FV3-like isolates suggested that the lineage originated in North America, based on the timing of emergence and sequence homology with later described European isolates (Hyatt et al., 2000). However, it has more recently been hypothesised that the FV3 lineage could be ancestral to Asia, given the most basal and ancient positioning of Chinese isolates in FV3-like phylogenetic reconstructions (Price, 2013; Vilaça et al., 2019). Indeed, in my

phylogenetic reconstructions of globally distributed FV3-like ranaviruses, both the ML and Bayesian approaches recapitulated the basal positioning of isolates from China, along with two sampled in California (Fig. 4.1; Fig. 4.2). It is from this phylogenetic perspective that there is an emerging debate over "out of America" or "out of Asia" hypotheses for the FV3 lineage. However, such inferences are currently complicated by the fact that Asian FV3-like ranaviruses have been poorly sampled compared to their North American and European relatives. Furthermore, despite ranavirus detection in many countries throughout Asia, all WGS data has been generated from China alone. It is also important to consider that the genomic data from Asian ranaviruses has been sampled exclusively from captive amphibian and reptile aquaculture facilities (Chen et al., 1999; Mu et al., 2018; Zhang et al., 2001).

Amphibian aquaculture facilities around the world are notorious for farming North American species, particularly the American bullfrog, *Rana catesbeianus*. Indeed, the term 'ranaculture' was coined early in the 20th Century to describe farming large American ranid species for food (Schorsch, 1933). The practice, along with its species, became an American export and was initialised in Asia as early as 1918 in Japan (Une et al., 2009), but proliferated throughout Asia predominantly from the 1980s on. The American bullfrog and related species have been farmed across Asia, with the largest importers/exporters being China, Indonesia, Taiwan and Vietnam (Gratwicke et al., 2010; Schloegel et al., 2009).

The North American pig frog, *Rana grylio,* was one such species first imported into China for ranaculture in the 1980s. The so-called *Rana grylio* virus (RGV) was isolated during disease outbreaks that emerged in these pig frog farms in 1995, which marked the first detection of amphibian-associated ranaviruses in Asia (Zhang et al., 2001). Lei et al. (2012) sequenced the RGV genome, and upon discovering its placement in the FV3 lineage, the authors proposed it was an imported virus. They further speculated that a host shift subsequently occurred into farmed softshell turtles (*Trionyx sinensis*), based on the close relatedness of RGV to the Soft-shelled turtle iridovirus (STIV) isolate. Mu et al. (2018) propose a similar exogenous route of emergence for the *Rana nigromaculata* ranavirus (RNRV) isolate in cultured Chinese black-spotted frogs. Whilst there have been few reports of ranavirus incidence in wild Asian amphibian populations, FV3-like ranaviruses have been detected and associated with mortality of *Rana dybowskii* in China and South Korea (Park et al., 2021; Xu et al., 2010) and *Rana huanrenensis*, as well as in South Korea (Kwon et al., 2017). The dearth of reports is likely in part

due to poor surveillance effort of disease in wild Asian amphibians, but also a probable lack of natural widespread endemism of amphibian ranaviruses to Asia. Moreover, due to the importation of farmed amphibians native to America (Une et al., 2009; Zhang et al., 2001), it is not unreasonable to assume ranavirus has spilled over into wild Asian amphibian populations from aquaculture facilities, just as spillover has been increasingly documented between farmed species.

Despite the fact that FV3-like ranaviruses emerged in Asia following the widespread establishment of ranaculture practice, further lines of evidence are required to rule out the "out of Asia" hypothesis. For instance, ranaviruses emerged in the USA within a similar timeframe (Fijan et al., 1991; Granoff et al., 1965), and the country arguably holds the greatest risk of ranavirus invasion through animal trade. This is because the USA is one of the largest single countries importing amphibians, both of farmed frogs for food (along with France and Belgium) and wild species for the pet trade (Carpenter et al., 2014; Gratwicke et al., 2010; Picco & Collins, 2008; Wombwell, 2014). However, in this study I offer two main lines of evidence that point towards the likely origin of the FV3 lineage in North America.

The first is given by the time to most recent common ancestor (tMRCA) for all isolates included in the FV3-like time-calibrated phylogeny. The tMRCA was inferred to be around the year 1550 (95% HPD 1333.56 – 1743.50), a timeframe that is more consistent with the onset of global trade with North America than Asia, which would have extended several hundred years prior to this period. Secondly, my population structure analysis of the FV3 lineage suggested North American ranaviruses contained the greatest amount of diversity of all sampled continents; all four of the predicted FV3 ancestral populations were partitioned amongst the genotypes of the isolates from the USA, which was not recapitulated by any other geographic region, though this could be biased by proportions of sample contributions. Nevertheless, this pattern is not consistent with a bottleneck in genetic diversity that would be expected to follow importation events into the USA. What is more, the entire genotype of the Chinese RNRV isolate was assigned to one ancestral population, which subsequently contributed ancestry only to other isolates from captively farmed animals. These included the only Brazilian isolate (MH351268.1) and the USA isolate RCVZ2 (MF187209.1) – both from outbreaks in captive farmed American bullfrogs (Candido et al., 2019; Claytor et al., 2017) – suggesting a possible trade-linked population network involving hosts native to North America.

Irrespective of the continental origin of the FV3 lineage, the results I present contribute the most compelling evidence to date that the lineage was introduced into Europe and the UK. Both FV3-like phylogenetic reconstruction approaches demonstrated extensive paraphyly of isolates sampled in both North America and the UK. As I discuss above, the most basal ranaviruses originate from both Asia and North America, including the Canadian isolates reported by Vilaça et al. (2019). All UK ranaviruses sampled to date, however, are more phylogenetically derived. From their collective coalescence on the time-calibrated phylogeny, I consider the earliest possible point estimate for their presence within the British Isles at the turn of the 19th Century (1802.23; 95% HPD 1698.70 – 1889.21). The phylogeographic conflicts posed by the paraphyletic grouping with North American isolates after this time supports translocation events of FV3-like ranaviruses between the two continents, including one event with Portugal. Based on the most recent common ancestor between transcontinental isolates, the most parsimonious scenario given the dataset is that six translocations in total occurred between North America and the UK (with an additional seventh involving Portugal). This either suggests six introductions into the UK, or a smaller number of introduction events into the UK with a proportion of back introductions into North America. Whatever the case, this number represents a mere sample of reconstructed translocations, which points towards a pattern of UK invasions that began sometime in the 1800s and accelerated during the mid-20th Century.

For three of the reconstructed translocation events, I was able to infer the direction of introduction from North America to the UK based on a phylogenetically basal set of North American isolates (Fig. 4.3). However, the direction of movement between continents for the remaining three cases was challenging to infer with confidence, due to the basal positioning of UK isolates to the events. A line of evidence exists in inferring that at least one of the directionally ambiguous events constituted a re-introduction of UK ranaviruses back into North America. Within the FV3-like time-calibrated phylogeny, I detected two intercontinental translocations involving Canadian ranaviruses reported by Vilaça et al. (2019). Two of these isolates – KEN1 (MK959604.1) and Z994 (MK959621.1) – were reported in their study to have undergone a recombination event with an unknown donor. The authors concluded that this unresolved event likely involved an unsampled Canadian ranavirus strain or lineage. Intriguingly, Z994 was implicated in one of the three phylogeographic conflicts with an ambiguous direction of translocation that I report. Rather than inferring recombination with an unknown source, as reported by Vilaça et al. (2019), the possible re-

introduction source from the UK presents an alternative hypothesis for the unexplained genetic diversity amongst the highlighted Canadian isolates. This hypothesis is further supported by the population structure analysis I undertook, in that an approximately homogenous ancestral population was assigned to UK diversity, which contributed a proportion of ancestry to two isolates from Maine in the USA, and four from Canada. Intriguingly, the largest shared ancestry component enriched in UK isolates was with isolate Z994 (Fig. 4.4), suggesting the unresolved recombination event could have involved UK ranavirus genetic diversity.

4.5.2 Emergence and Spread of UK Ranaviruses

In Figure 4.3 (**B**), I visualise the timing of the dispersal of all UK ranavirus lineages using the dates of coalescence events from FV3-like time-calibrated phylogeny. In addition, I included the location and sample dates of the CMTV-like isolates identified by Price et al. (2017), as these were not phylogenetically reconstructed, but likely represent independent introduction events given distinct phylogenetic and geographic origins sampled 16 years apart. I inferred the most probable location and earliest-possible date of each of the three phylogenetically resolved FV3 introductions by using the tMRCA date of the clade borne from the invasion, and the geographic coordinates of the most basal isolate therein. However, the earliest possible point estimate of each introduction represents the ancestral divergence between UK and non-UK ranaviruses. This date it does not necessarily imply that the ancestor of UK samples was indeed found at the given location during that time, as the translocation could have taken place at any time along the distal branch. I therefore reason that the most confident earliest date estimate of an isolate's ancestor being present where it was sampled is given by the last divergence between each isolate and its sister taxa *within* the UK (parent node of the terminal branch), as it is less parsimonious to assume this divergence occurred outside the UK. As such, I used the date of the last divergence of each isolate (parent node) within each of the three UK introduction clades to visualise the timing of their phylogeographic relationships, and therefore the most plausible estimate for the timing of emergence for each isolate.

With this approach, I found striking exceptions to what has been previously considered regarding the timing of FV3-like introductions to the UK; chiefly, that ranavirus presence in the British Isles likely exists on the scale of centuries, rather than decades. Using a regression-based estimate, Price et al. (2016) note that the genetic diversity amongst seven isolates (and a restricted number of loci) of UK

ranaviruses could not be explained by the 30 years of evolution since their detected emergence. Indeed, the reconstructions I present suggest that the first FV3-like clade arrived as early as 1814 (95% HPD 1715.04 – 1891.77). The overlapping 95% HPD intervals cast uncertainty around which isolates within the clade were truly most basal, but the results based on current sampling suggest that the first introduction occurred in the north of the UK in Lancashire, followed by sister isolates appearing in Norfolk and Essex approximately 300 Km farther south within 50 years. Similarly, the proceeding two mid-century introductions occurring in the South each exhibited a geographic distribution of related isolates across the South Coast of England and Greater London over a 20 – 40 year period. In one case, an isolate belonging to the clade borne from a southern introduction in 1936 (95% HPD 1895.97 – 1970.56) had an ancestor associated with the North, again in Lancashire, only two decades later in 2001 (95% HPD 1990.85 – 2009.99). Despite the uncertainty surrounding the exact point of each introduction (true basal isolate of the clade), it seems implausible that the natural migration of host reservoirs could explain all the dispersal observed across the UK, given the often-short timeframe between divergences of related viruses and large geographic range between locations of the sampled descendants. This pattern is highly suggestive of extensive human-mediated translocation of ranaviruses following introduction into the UK.

Human-mediated range expansion strongly supports the conclusion drawn by Price et al. (2016), who used the FMP ranavirosis reports to reconstruct the patterns of emergence of UK ranaviruses. The authors used a spatiotemporal model with two components modelling the contributions of local transmission characteristic of pond-to-pond frog dispersal, and infections uncharacteristic to the local system such as importation events. After controlling for reporting effort, the authors found the largest proportion of disease reports were indicative of local transmission in line with host dispersal, but a significant proportion of events were explained by non-local processes. Environmental factors such as climate were ruled out as correlates to non-local expansions. Instead, the covariate that best predicted new outbreak events was human population density. The pattern strongly suggests that the country-wide range expansion of UK ranaviruses is most likely due to human activity, which is consistent with the phylogeographic timing of events I report.

The timing of the early 19th Century and cluster of mid-20th Century UK introductions is concurrent with known patterns of transatlantic movement of

people and goods. Victorian-era British naturalists were famed for their collections of exotic species, many of which were imported live to stock arboretums, zoos and private collections. Exotic collections could well explain the earliest-possible UK introduction of FV3-like ranaviruses in 1814, for instance through the first American bullfrogs introduced to garden ponds. The isolates implicated in the first introduction I infer were most closely related to genetic diversity from the USA based on SNP clustering, but interestingly the clade established from this event was also implicated in the potential back introduction into Canada. Given that Canada was a British Colony from 1763 to 1867, a large amount of transatlantic trade of goods, such as fur and timber, occurred both during and after this period. With the establishment of colonies to accommodate the harvesting of traded goods, the practice of stocking freshwater bodies with non-native European fishes for food and leisure was widespread, such as the species well documented in the Great Lakes Basin (Mandrak & Cudmore, 2010). Whilst there is scant documentation of FV3-like ranavirus infecting fish (however, see: Rosa et al., IN REVIEW; Waltzek et al., 2014), early fish stocking practices carried out between the UK and North America would have provided ample opportunity for the translocation of infectious materials between the two continents during this early period.

The cluster of mid-to-late 20th Century translocations I inferred likely resulted from a very different form of trade. As discussed above, amphibians (particularly from the genus *Rana*) began to be routinely traded globally during the 20th Century for ranaculture, but also as research animals and pets. In the UK, there are four border inspection ports licensed to handle amphibians (Wombwell, 2014). These are the airports of Edinburgh, Manchester, Heathrow, and Gatwick. Intriguingly, Heathrow and Gatwick are both situated in the South of England around Greater London, and Manchester Airport is in Lancashire – the main areas of ranavirosis prevalence in the UK. What is more, Manchester Airport was established in 1938, followed by Heathrow in 1946 and Gatwick in 1958. Heathrow handles the vast majority of amphibian shipments into the country (reportedly around 85%; Wombwell 2014), where in turn approximately 85% of animals are bound for domestic trade and 25% re-exported (Peel et al., 2012; Wombwell et al., 2016). Due to strict biosecurity measures, it is unlikely animals are released or escape directly from these inspection ports. However, adherence to such measures is likely more variable amongst the many distribution centres and businesses the ports supply. Whatever the specific route of entry, feral populations of non-native amphibians are currently extant throughout the UK, particularly in the South (Langton et al.,

2011). American bullfrogs were a heavily imported species (Banks, 2000) that currently maintain a southern foothold in the UK, but have since been eradicated from a once more widespread distribution (Ficetola et al., 2007, 2008).

Given that the estimates I infer of *Ranavirus* introduction into the UK long predate the earliest reports of disease, the question is posed as to why the pathogen was only detected towards the end of the 20th Century. In previous work, I and my colleagues have shown that ranavirus disease severity is strongly correlated with temperature (Price et al., 2019). Specifically, that FV3-like ranaviruses linearly replicate to much high titres *in vitro* up to ~30°C. This pattern extended to the *in vivo* model, where UK common frogs experimentally challenged with the same isolates suffered significantly worse and faster disease outcomes at 20°C versus 10°C. Lastly, using *in situ* epidemiological data from the FMP, we found increases in UK ranavirus disease occurrence strongly correlated with the rising temperatures throughout the 1990s, which was further supported by decreases in reports as average temperatures began to fall over the following decade (Price et al., 2019). Altogether, these findings suggest that ranaviruses could have persisted unnoticed amongst UK common frog populations through subclinical infections, which is supported with recent evidence of widespread ranavirus detection in UK *R. temporaria* populations with no signs or history of disease (Campbell et al., 2018). Rising temperatures throughout the late 20th Century then stands as the most plausible factor for the widespread symptomatic disease that prompted ranavirus detection.

4.5.3 Concluding Remarks

The role human activity has played in the global emergence of *Ranavirus* is clear. In the UK, I have shown that an increase in the frequency of introductions coincided with the radical change in the movement of people and traded goods as air travel became commonplace. However, my conclusions are constrained by a sparsely sampled dataset, and a large question remains as to whether ranavirus presence in the UK was seeded by a burst of introductions between the 19th to 20th centuries, or if invasions are ongoing in the present day. Continued and more extensive genomic surveillance is required to better elucidate the specific sources and key routes of ranavirus translocations, particularly for the aim of developing global mitigation strategies. For instance, as ranaculture has played a key role in ranavirus translocation and emergence, focused surveillance in and around aquaculture facilities would offer a powerful means for assessing where biosecurity measures may have failed and better enforcement is needed.

The risk to global health from emergent infectious diseases of both humans and wildlife arising through anthropogenic processes has long been recognised, and has become poignantly salient in recent times. Regular calls to update policy internationally have been and continue to be made (Auliya et al., 2016; Fisher & Murray, 2021; Hanisch et al., 2012; Horton et al., 2014; Marano et al., 2007; Patz et al., 2004). Response to such pleas is reflected in academic initiatives such as One Health, Conservation Medicine, EcoHealth and Planetary Health, which have recently been termed the integrative environment-health sciences (IEHS) in the latest call to more urgent and effective global action (Fisher & Murray, 2021). The evidence I have presented here demonstrates how the animal trade industry has contributed to the global emergence of FV3-like ranaviruses. Pathogen pollution of this sort, resulting from animal trade, is precisely the kind consequential risk that the IEHS seek to reduce (Marano et al., 2007). This study contributes yet another example to an ever-growing list, which begs the question of what exactly is required before definitive, concerted steps towards effective preventative action of emergent infectious diseases may be taken.

5

General Discussion

Over the last five decades, ranaviruses have generated much intrigue and concern, as they contain one of the broadest host breadths of any genus of eukaryotic virus, whilst simultaneously posing a significant risk to amphibian biodiversity and aquaculture commerce. For this reason, these viruses have attracted researchers that intersect veterinary, ecological and evolutionary disciplines. Their continual global emergences and strain discoveries have promoted many challenging questions, which have been primarily addressed through epidemiological and molecular characterisations. In this thesis, I have extended these characterisations by providing the largest single synthesis of *Ranavirus* genomic diversity to gain insight into their evolutionary histories and dynamics.

Whole genome sequences contain a wealth of information. However, this information has not been exploited in ranavirus biology to its current potential. This is perhaps most apparent when considering the genomic relationships between the 170 isolates included in this thesis. In Chapter 2, I reveal stark differences in the genetic distances yielded from the core genome (approximately 50% of the genome) compared to all available genomic information. A clear subdivision of lineages emerges, which calls into question the species membership of many isolates. Particularly affected are those of the paraphyletic TFV-like clade, in which the TFV and BIV sub-lineages likely represent distinct species from the FV3-like. So too do the ECV, ENARV, and EHNV sub-lineages of the EHNV-like clade.

Chapters 3 and 4 reveal striking findings on *Ranavirus* evolution and historical dynamics, which again were only achievable through leveraging whole genome information. For instance, the significant temporal signal in the mutation rate of FV3-like lineage could only be uncovered through intense genomic sampling. Through phylodynamic analysis of this signal, I estimate invasion of the lineage

into UK on a time scale that predates our previous understanding by up to a century. Furthermore, I reveal a striking degree of polyphyletic gene history contributing to the genomic makeup of single isolates within each ranavirus lineage, and reveal that extensive recombination occurs in regions throughout the length of the genome. This demonstrates a hitherto underappreciated role of recombination in generating genomic diversity in *Ranavirus*.

5.1 *Ranavirus* Diversity: More than Meets the Eye

Like viral taxonomy as a whole, current *Ranavirus* taxonomy contains many inadequacies (Chinchar et al., 2017b; Gibbs 2013). It is almost certain that the genus contains a vastly underappreciated diversity, but its delimitation has been hampered by imprecise molecular methods and a focus on the 26 genes that are core to the *Iridoviridae* family, together with artificially selected distance threshold cutoffs (Eaton et al., 2007; Chinchar et al., 2017a). Whereas early viral taxonomy held factors of molecular structure, phenotype and epidemiology key in systematics classifications, these features have been downplayed in recent years in favour of phylogenetic monophyly (Gibbs, 2013; Koonin et al., 2020). This is in large part because metagenomic analyses are discovering a vast diversity of viruses that cannot be cultured, and so these classically important defining factors cannot be quantified (Simmonds et al., 2017). Thus, the emphasis on monophyly begets a responsibility for as-accurate-as-possible phylogenetic analysis, which requires capturing the greatest possible amount of phylogenetic signal in the system under consideration (Chung et al., 2018).

In my attempt to perform phylogenetic analysis of ranavirus with the highest-possible resolution, I took an approach to identify orthologous coding sequences to align and carry forward. Recently, two related studies by Waltzek and colleagues (Box et al., 2021; Sriwanayos et al., 2020) both performed a similar analysis, but implemented a different methodological approach to identify so-called localised colinear blocks of homologous regions in approximately 45 whole genomes across the ALRV genus. Only one of the papers (Box et al., 2021) reported that the method achieved a sequence alignment of ~145 Kbp in size, including gaps. However, neither study reported the homology thresholds used and the proportion of genomes retained. It is not possible that their alignment gaps result from true indels considering the average genome length is ~105 Kbp, with the largest included approximately 125 Kbp in size (SERV; KX353311.2). As such, accurate phylogenetic distances reflective of the rates of evolutionary process cannot be obtained from such an alignment. Further, both studies only report

cladograms, which unlike phylograms, have meaningless branch lengths that only show relative relationships between sequences. This is analogous to asking for directions, and only being told which way but not how far. Knowing the distance in addition to the bearing is important, as it influences the considerations needed to accurately arrive at the destination.

Distance in addition to the demarcation of monophyly becomes particularly relevant for the TFV and BIV lineages of the paraphyletic TFV-like group. All its comprising isolates are currently classified as FV3-like based on their similar genome arrangement and their monophyletic grouping of core genome information (Chinchar et al., 2017a; Sriwanayos et al., 2020). However, drawing the hierarchical line for monophyletic inclusion can be unclear. For example, all of the derived AARV (excluding the ATV-like) form a higher order monophyletic grouping, but the CMTV-like are distinguished as a distinct species to the FV3-like based primarily on their genomic arrangement in addition to various epidemiological factors. Aside from genome arrangement, the TFV and BIV sub-lineages contain similar features that distinguish them from the FV3-like to same degree as the CMTV-like, including geography and host types. However, their distinctiveness becomes particularly apparent when phylogenetic distance is considered. Based on core-phylogeny branch lengths, the average distances of the FV3-likes to the CMTV-likes are in fact closer than to both the TFV and BIV lineages, despite the monophyletic grouping of the latter with the FV3s. This is indicative of a recent ancestral relationship of the core genome content between the FV3- and TFV-like, whilst revealing very distinct evolutionary processes at play for both groups, which cladograms are unable to reveal.

It seems clear that the derived AARV of the CMTV-, TFV-, and FV3-like lineages are descended from a common ancestor, but it is unhelpful to delimitate species with arbitrary distance thresholds, such as $> 90\%$ which is sometimes used (Chinchar et al., 2017b). Cutoffs such as these leave much to speculation, opening the door to confusion. Instead, monophyletic groupings should be considered together with flexible genetic distance thresholds based on significant step changes between clades. Under such a framework, each of the sub-lineages I identify within the paraphyletic TFV- and EHNV-like clades would certainly stand as distinct species of *Ranavirus*, which become bolstered when their host associations and geographic distributions are factored in.

In Chapter 3, I reveal pervasive recombination amongst the entire genus of the ALRV ranaviruses. This calls into question the appropriateness to analyse the phylogenetic relationship of highly mosaic isolates on a single tree, or at the very least, without rigorous correction measures. On the other extreme, my analysis of raw genomic distances through a *k*-mer based approach without any such correction, reveals intriguing relationships. Chiefly, that the highly mosaic TFV-like cluster more closely to the CMTV-like in principal coordinate space, despite the apparent monophyly of the core genome with the FV3-like. This could suggest that a large portion of the accessory genome is contributed by the CMTV-like, in addition to the polyphyletic core (~50% of the genome). Whatever the case, high rates of ancestral recombination and the pattern of polyphyly demonstrates how unique ranavirus species diversity can be generated through the union of existing lineages. Patterns of mosaicism carry implications for the permitting circumstances – such as the involvement of aquaculture facilities – which may go unnoticed if typing is carried out using restricted number of phylogenetic markers, as recent studies have sought to do (Ballard et al., 2020; Yu et al., 2020).

5.2 The Human Factor

There is increasing evidence that the expansion of ranavirus spatial distributions at both local and global scales has been facilitated by human activity. Through my phylodynamic reconstruction in Chapter 4, I estimate dates of ancestral emergences of FV3 isolates in the UK. The emergences are highly suggestive of human-facilitated movement, owing to short intervening timeframes of ancestral divergence and disparate sampling locations of the decedents across the country. This is consistent with a body of epidemiological evidence linking a positive association of ranavirus prevalence with human population density (Price et al., 2016; St-Amour et al., 2008). Proximate mechanisms for local translocation have been identified in the movement contaminated materials associated with leisure activities such as angling (Epstein & Storfer, 2016) and even watersports (Casais et al., 2019).

The greatest contributor to long-range spread of ranaviruses is through animal trade. My key finding in Chapter 4 is the estimate that animal trade has been mediating invasions of FV3 ranaviruses for approximately a century between North America and UK. Furthermore, there is a particular association of ranaviruses reported in captive settings across the globe. On the one hand, this may seem obvious given the detectability of disease outbreaks compared to those in the wild. However, ranaviruses detected in these settings may also be the first

point of local emergence in otherwise disease-free areas, given the high rates of ranavirus prevalence in animals associated with the pet trade (Wombwell, 2014). I report further genomic evidence to support this pattern, where a population structure of FV3 ranaviruses exists amongst those sampled from aquaculture facilities in North America and Asia.

It has been experimentally demonstrated that ranaviruses isolated from captivity can be more virulent than wildtype relatives (Brunner et al., 2015; Hoverman et al., 2011; Storfer et al., 2007). Several hypotheses have been put forward to explain the evolution of increased virulence in captive isolates, which primarily relate to host population demography. These include a reduction of the cost of virulence resulting in host death amongst the high population densities in captive settings, relating to a reduced need for rapid transmission for the epidemic to be sustained (Brunner et al., 2015). However, an alternative hypothesis may be given through the evidence of polyphyly amongst captive-associated isolates and the high rates of ranaviral recombination. It is possible that increased virulence could arise as a product of novel combinations of virulence-associated alleles from distinct lineages within a single virus (Combelas et al., 2011; He et al., 2010). If increased virulence is strongly associated with recombination in ranaviruses, mitigating co-occurrence of lineages circulating in traded animals may take on new significance for policy decision makers.

Recombination in distinct ranavirus species appears to take place in the wild (Vilaça et al., 2019). However, the probabilistic weighting of such events occurring in natural populations compared to those in captive settings, such as in aquaculture or zoo collections, is unclear. On the one hand, it is possible that higher comparative rates of coinfection may be yielded in the wild. The effective population sizes of wild susceptible host assemblages in overlapping enzootics are likely to be far greater than the populations affected by co-circulating infections linked amongst networks of captive rearing facilities. However, rates of recombination between different ranavirus lineages appears more frequent in ranaviruses from captive settings in the genomic sample, which I base firstly on the observation of high degrees of polyphyly associated with isolates sampled from captivity. Secondly, the greatest contributors to polyphyletic isolates were the FV3- and CMTV-like clades, which are not considered to geographically co-occur at high prevalence in the wild. CMTV has been found outside its natural distribution in captivity in North America (Majji et al., 2006) and in two archived samples in the UK (Price et al., 2017b), but to date, it has not been attributed to

contemporary dies offs in natural settings in either region. Conversely, FV3 is predominantly associated with North America and the UK, although there is some evidence for it circulating on mainland Europe (Rosa et al., 2017). Furthermore, whilst there is large overlap in the host range of both lineages, it may be difficult to quantify the extent that overlapping enzootics occur in a single species and habitat at any one time. Considering these factors in explaining the polyphyletic patterns observed amongst the CMTV-, FV3-, and TFV-like, it seems likely that lineages co-circulating in captivity and trade have a greater opportunity to cause coinfections. At the very least, the evidence presented suggests it is logical to assume that animal trade practices have accelerated the generation of ranavirus genetic diversity through recombinogenic means.

5.3 Prospects

From a certain perspective, it could be seen that we have entered a third era of ranavirus research; one of genomic exploration. The first ranavirus whole genome sequence was deposited in the NCBI GenBank database in 2003 (Jancovich et al., 2003). At the time of writing, the database now holds 129 ranavirus whole genome sequences (containing an additional five genomes since I last performed the search four months ago in August 2021). In Chapter 2, I used these genomes together with a large novel set to delineate the pan-genome content of the ALRV, and discovered many private genes to each lineage. Each clade-specific core genome ranges from 68 – 77 genes amongst an average 95 predicted ORFs per isolate. What are these private accessory genes? They could be host acquired (Filée, 2009); certainly ranaviruses recombine frequently as demonstrated in Chapter 3, so it is eminently possible that illegitimate recombination events with host genomic material could account for the gene diversity amongst closely related isolates. The many divergent US22 family proteins I describe within the pan-genome go some way to support this view. Alternatively, the accessory genes could be truncated relics no longer required for the particular host group being exploited (Chen et al., 2011). Whatever the case, with every new whole genome, a rich history awaits discovery.

In moving forwards, ranavirus research will benefit most greatly from continued genomic surveillance, particularly in an effort to move away from the reliance on sampling severe outbreaks and animals with gross signs of disease. Such efforts are needed to capture the full extent of *Ranavirus* diversity, which is clearly far greater than currently understood. In this vein, surveillance is particularly needed amongst geographic regions lacking reports of wild prevalence, and amongst

under-reported piscine and reptilian host taxa. A particularly fruitful product of genomic surveillance would stem from the sequencing ranavirus genomes isolated from wild animals with sub-clinical disease, as these data would facilitate research efforts seeking to elucidate the genetic determinates of ranavirus virulence (e.g., Morrison et al., 2014). Finally, sustained genomic surveillance would permit longitudinal analyses to be conducted. The lack of temporal signal in the substitution rates of the currently available genomic sample – except the FV3 lineage – is the key limiting factor preventing phylodynamic reconstruction of ranaviruses, which continued genomic surveillance would remedy. Phylodynamic approaches provide the only means into which insights can be gained into the ancient past lives of ranaviruses, including their origins.

Beyond surveillance, experimental designs aimed at elucidating ranavirus phenotypes and host pathology in tandem with genomic quantification remain greatly needed. In Chapter 2, I provided a substantial update to the functional annotation of not only the core genome (from ~30% to ~70%), but also the first effort to annotate the *Ranavirus* pan-genome. Despite this improvement, we do not understand how the majority of these genes influence host interactions in the context of disease progression and outcome. Gene knock-out and knock-in methods remain under-utilised since the establishment of transgenic ranavirus models, and their recognised potential (Robert & Jancovich, 2016). We live in an age where the importance and impact of viral discovery and quantification is painfully relevant. Though ranaviruses pose no direct threat to humans, they do to amphibians – the most widely threatened group of all animal taxa. Continued research and characterisation of *Ranavirus* therefore remains a justified and important endeavour, and particularly considering the current advancement of genomic technologies, we have never had so much to gain.

Appendices

Supplementary Figures & Captions

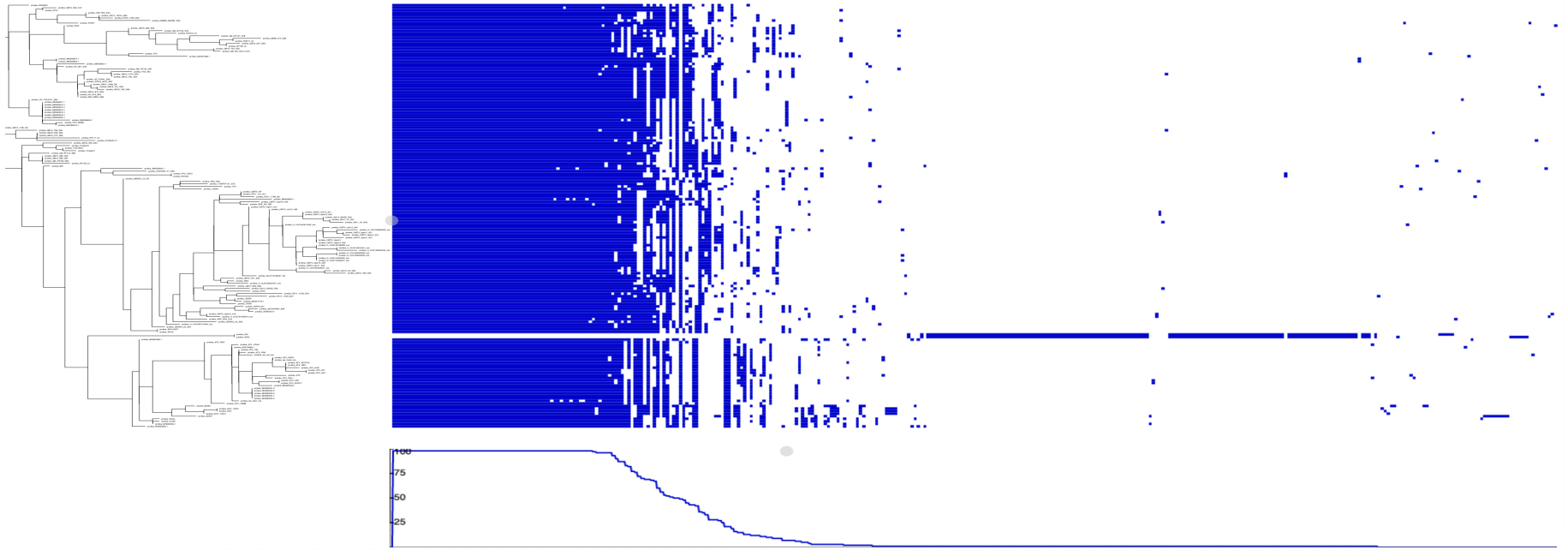


Figure S1. Ranavirus *pan-genome presence/absence of orthologous ORFs clustered at 80% amino acid homology. Shown on the left is a tree based on the shared accessory genome complement to roughly group isolates. The heatmap to the right shows individual ortholog clusters on the X-axis and their presence (blue) or absence in each isolate. The core genome is determined as clusters present in 100% of isolates, as seen in the plot at the bottom. Note the SGIV-like isolates that contain no orthologous clusters at the 80% homology cutoff (gaps in the heatmap). Removal of these isolates and their divergent ORFs retained 217 gene clusters amongst the remaining 170 amphibian-like ranaviruses.*

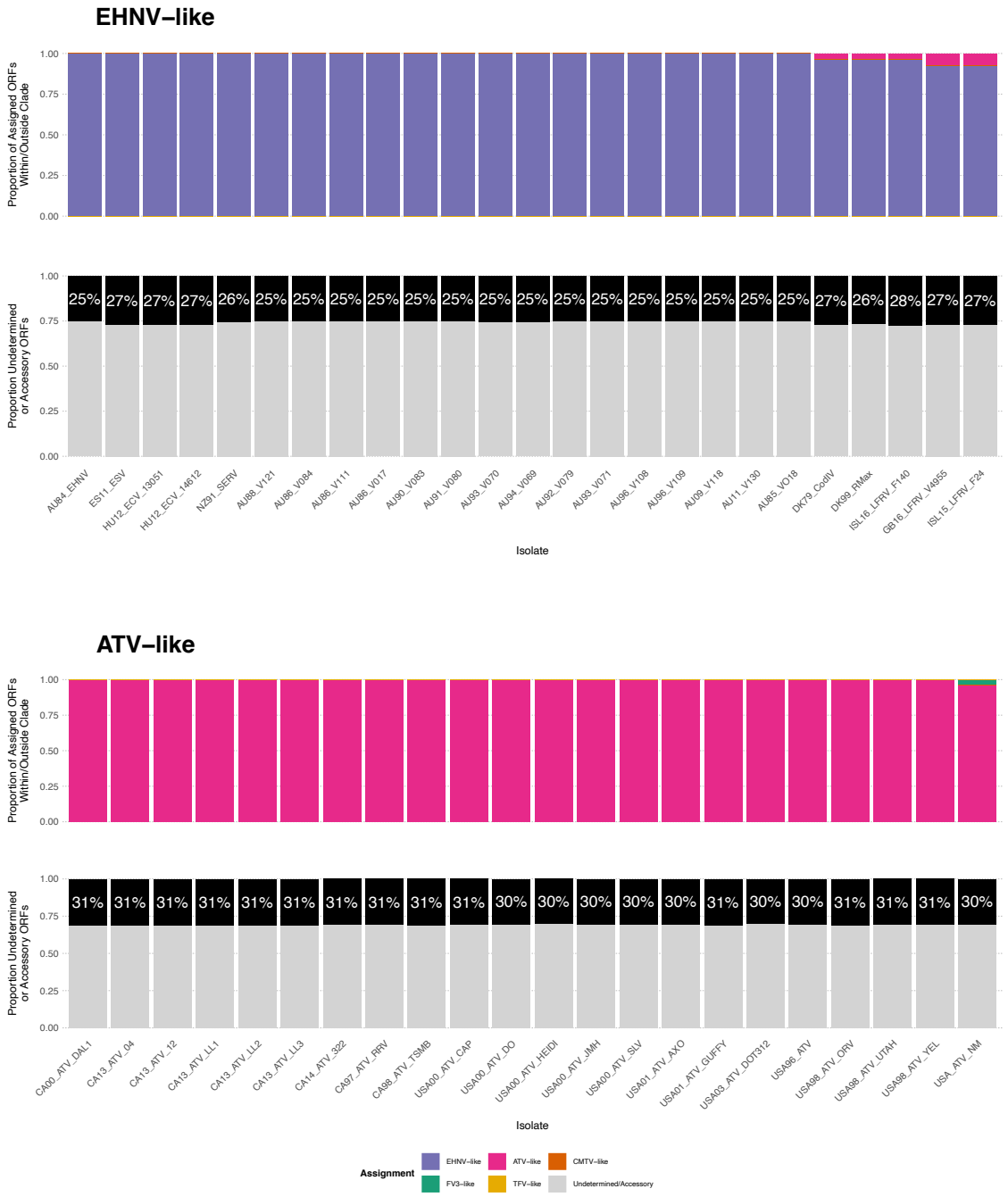


Figure S2. *Isolate gene histories of ALRV clades. Each column represents an isolate, where the top panel for each of the EHNV-like and ATV-like clades illustrates the proportion of assigned genes originating from particular clades, according to the ALRV core phylogeny. The bottom panel for each clade shows the proportion of genes that achieved an unambiguous phylogenetic assignment (black). The assignment process is based on closest average cophenetic distance of each clade to the given isolate within individual gene trees.*

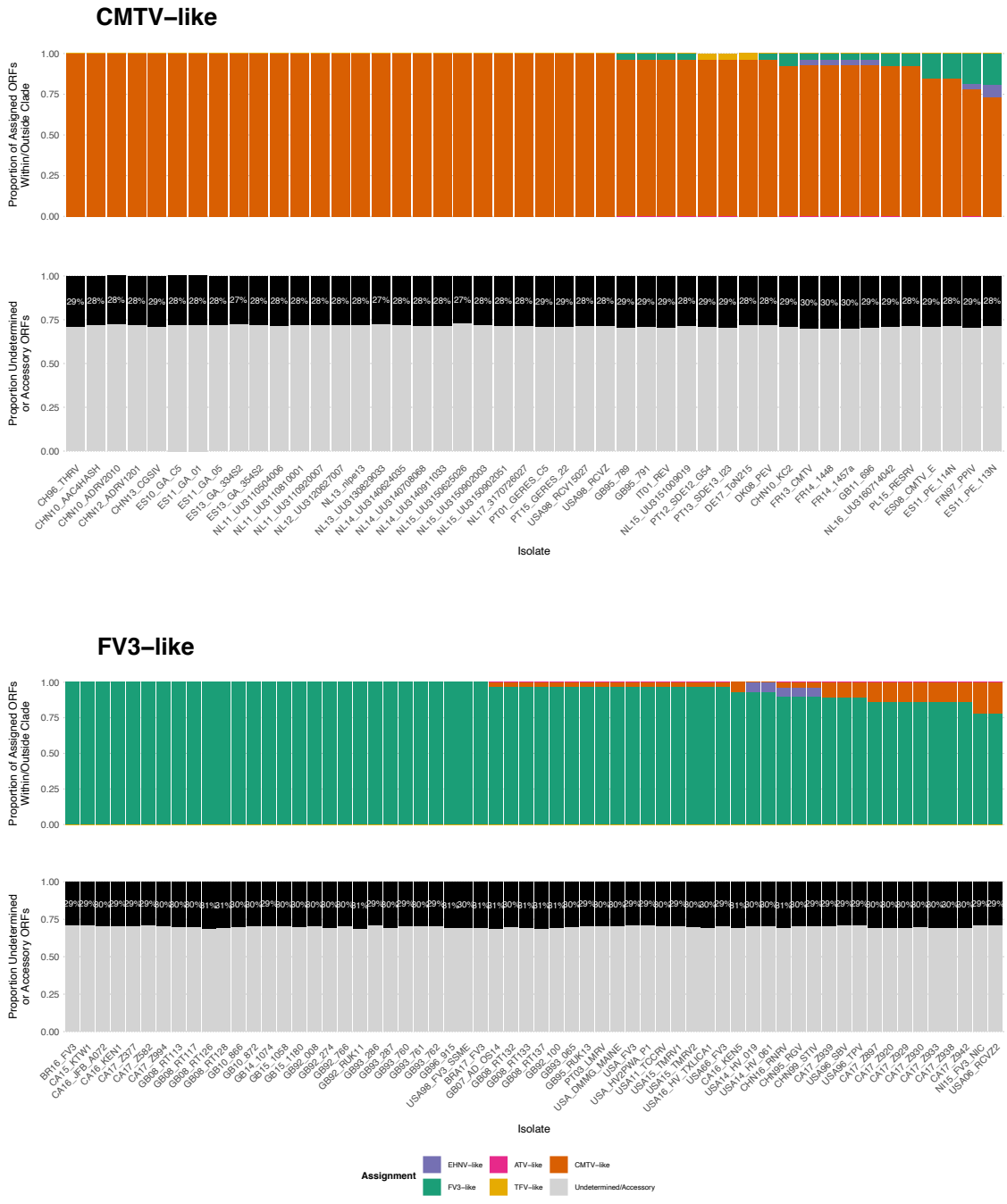


Figure S2 [continued]. *Isolate gene histories of ALRV clades. Each column represents an isolate, where the top panel for each of the CMTV-like and FV3-like clades illustrates the proportion of assigned genes originating from particular clades, according to the ALRV core phylogeny. The bottom panel for each clade shows the proportion of genes that achieved an unambiguous phylogenetic assignment (black). The assignment process is based on closest average cophenetic distance of each clade to the given isolate within individual gene trees.*

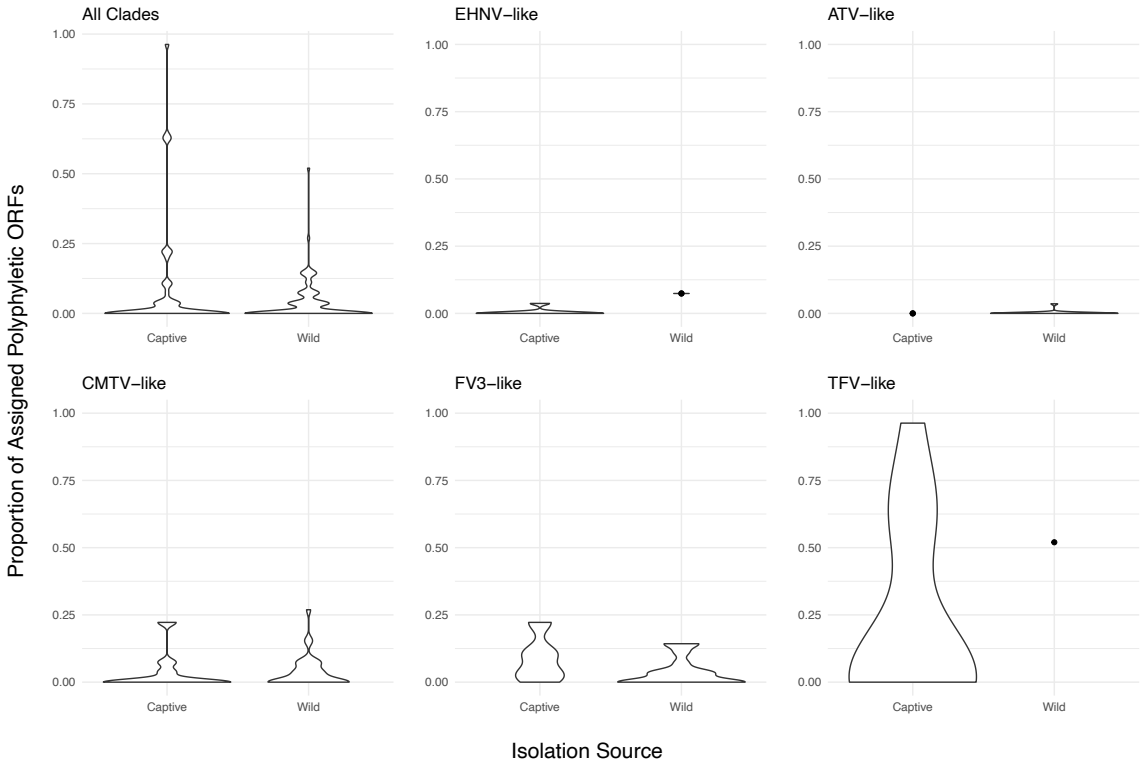


***Figure S3**. Proportion of polyphyletic genes associated with* Ranavirus *isolates from captive and wild sources. Violin distributions, with points for single observations, for all isolates together and each ALRV clade individually. Differences in the proportion of polyphyletic genes associated with captive and wild isolates were non-significant in every case. Nevertheless, the most extensively polyphyletic isolates tended to be sampled from captivity, driven primarily by those in the FV3- and TFV-like clades.*

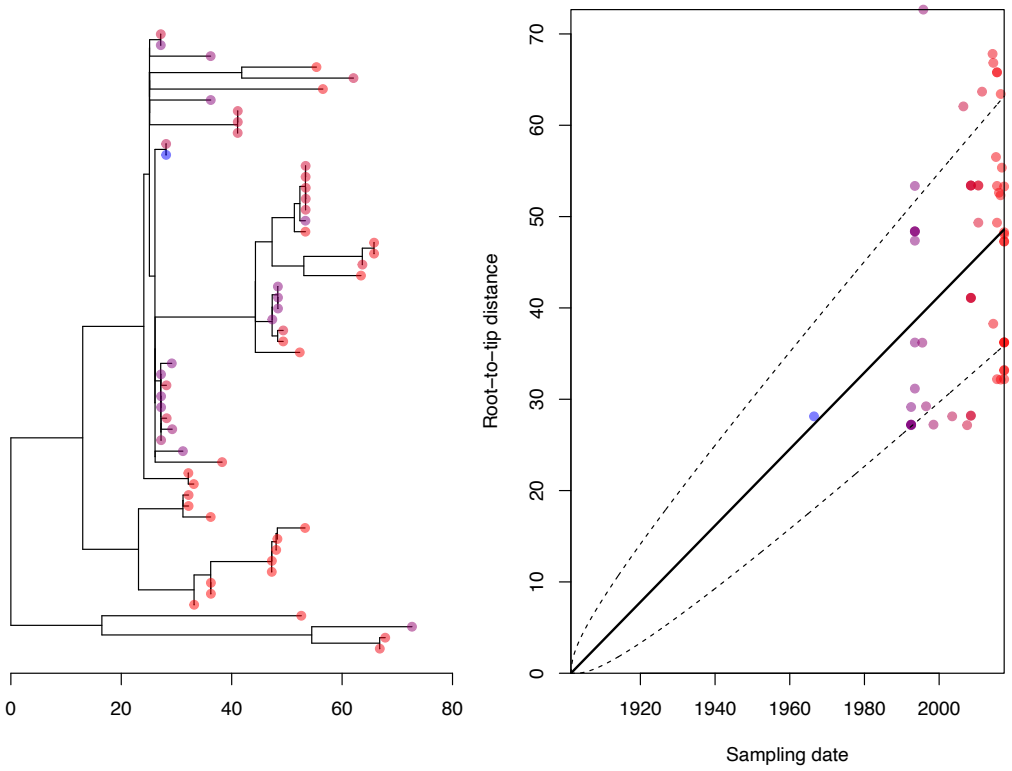
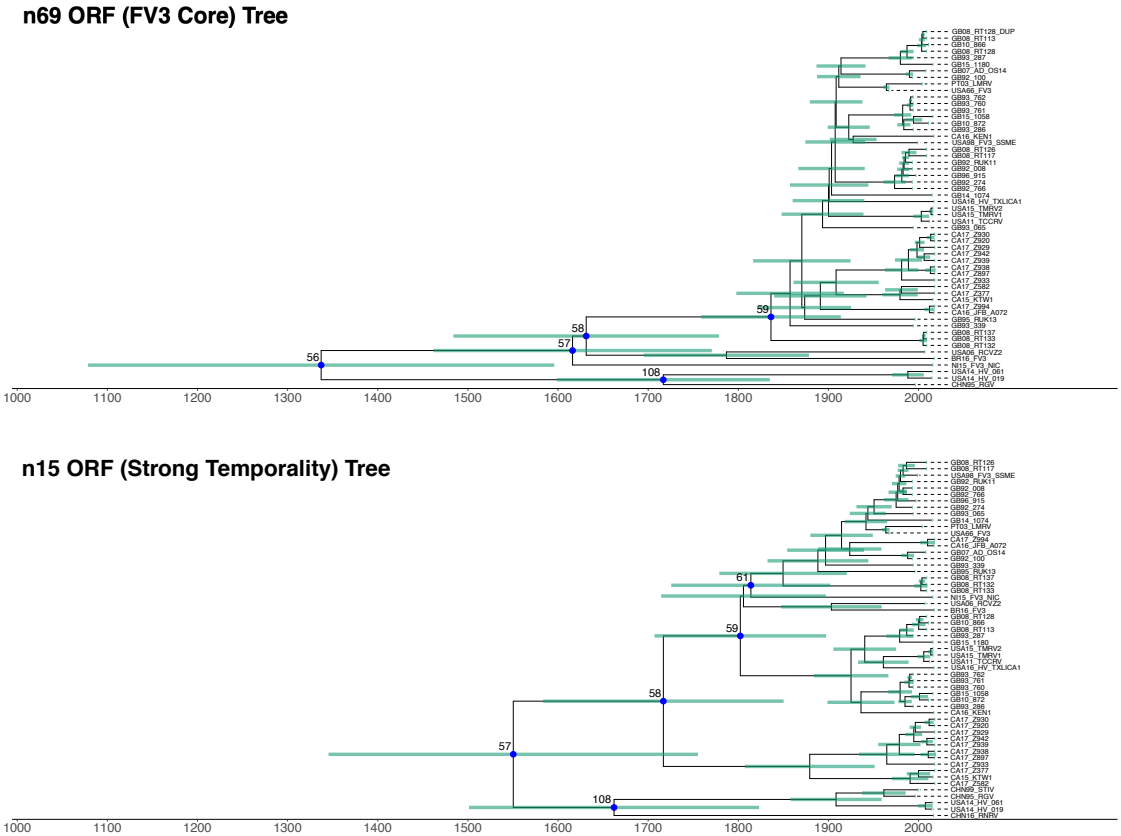


***Figure S4**. Root-to-tip regression of FV3-like isolates with known sample dates. The MSA was built using 15 concatenated core genes containing strong individual gene-tree temporal signal (Table S4), and contained 619 SNPs after controlling for recombination. The analysis was performed using* BactDating*, supplied with an ML tree built using* RAxML*, rescaled by absolute number of SNPs. The regression inferred rate of 0.0419 substitutions/year^{-1} and contained an $R^2 = 0.12$, with permutation significance of $P = 0.00303$ after 10,000 iterations.*



Node Event	n69 ORF Tree		n15 ORF Tree		HPD Overlap
	Number	Height HPD Interval	Number	Height HPD Interval	
Root	56	958.729 — 441.504	57	683.848 — 273.913	242.344
Basal Clade	108	425.643 — 189.231	108	524.978 — 203.202	335.747
New World Clade	57	561.447 — 252.489	58	442.749 — 175.932	190.260
UK Isolate Coalescence	59	265.736 — 110.735	59	318.699 — 128.197	207.964
Earliest Possible UK Invasion	58	532.789 — 237.898	61	287.486 — 110.759	49.588

Figure S5*. Comparison of overlapping HPD intervals between phylodynamic reconstructions of FV3-like ranaviruses. Two reconstructions were conducted, one using the alignment of the complete FV3-like core genome (n = 69 ORFs; top tree) and the other using the alignment of 15 ORFs with strong temporal signal (bottom tree). The combined table gives the 95% HPD interval (years before most recent sample date) of nodes representing significant evolutionary events, and the degree of HPD overlap in years between the two reconstructions.*

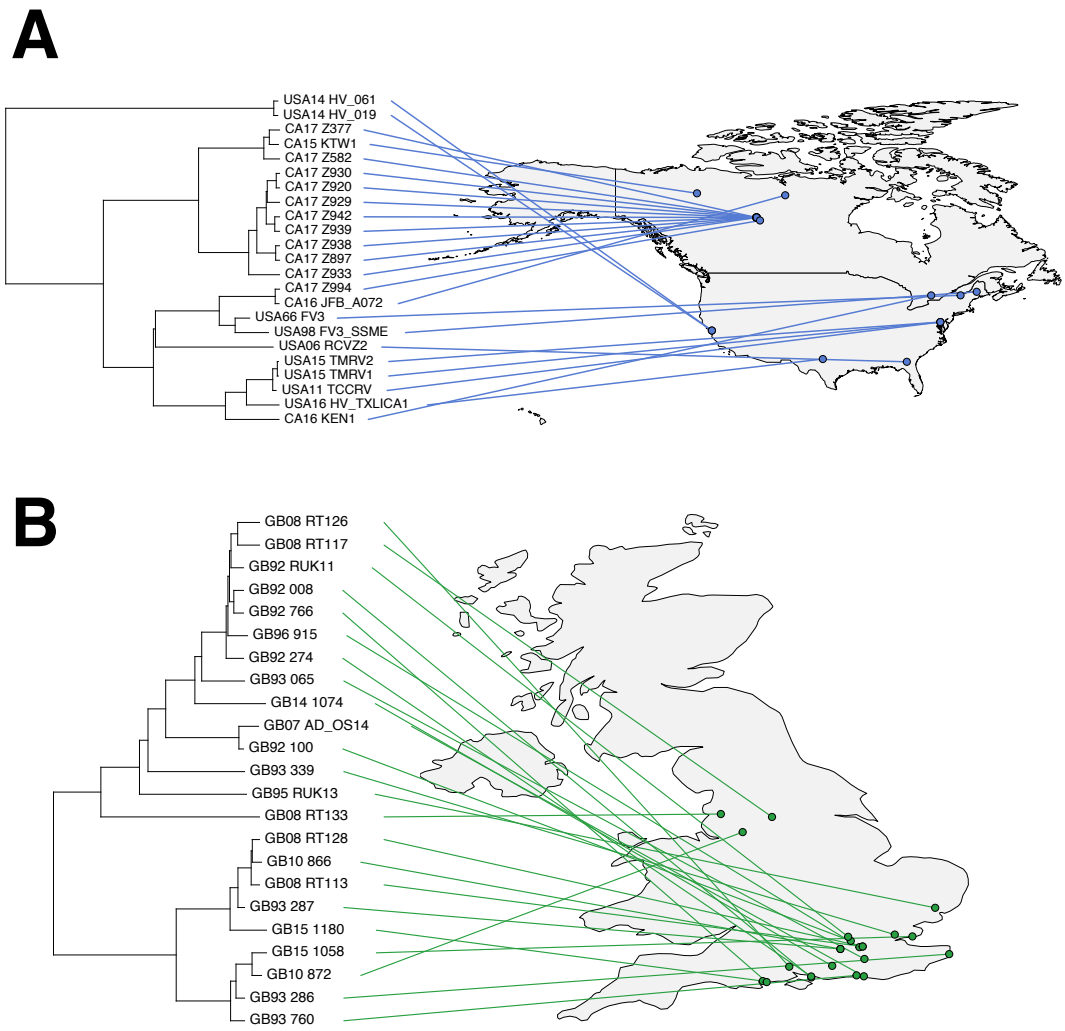


***Figure S6**. Phylogeographic mapping of FV3-like ranaviruses from the USA and the UK. Trees were derived from the time calibrated phylogeny (Fig. 4.2) by dropping tips and keeping only isolates from the USA (**A**) and the UK (**B**). Note that ranaviruses from the USA contain more phylogeographic structure compared to the UK relatives, where monophyletic clades contain viruses sampled from multiple regions.*

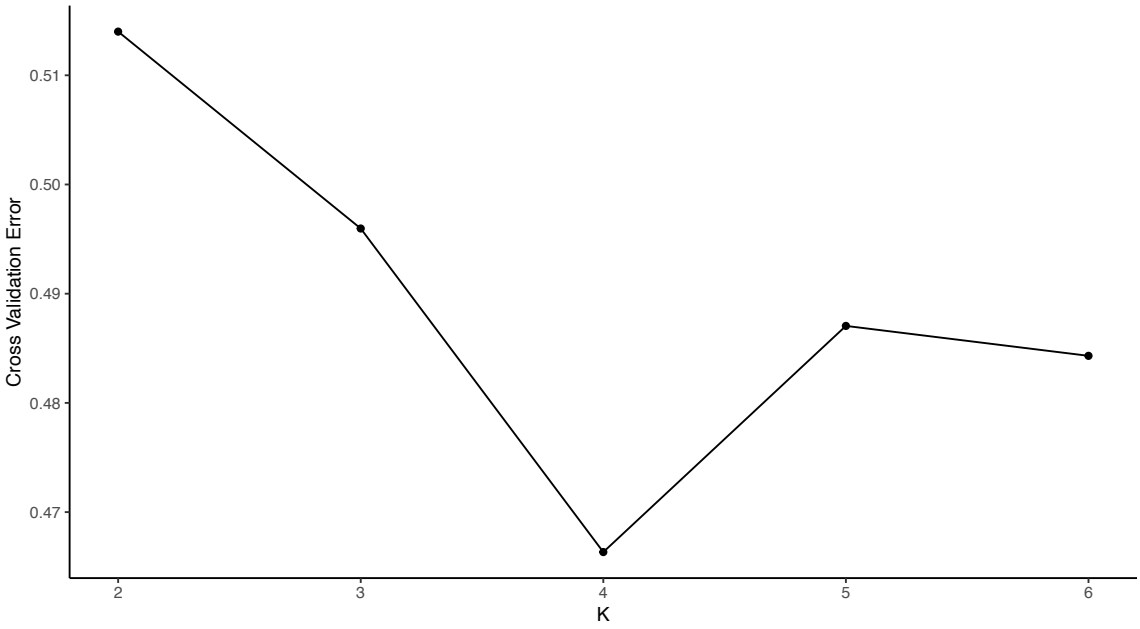


Figure S7. ADMIXTURE *cross-validation error scores giving support to the optimal value of* K *ancestral populations for FV3-like ranavirus isolates.*

Supplementary Tables & Captions

Table S1. *Contributing researchers who provided novel samples that were whole genome sequenced.*

Novel Sample ID	Contributing Researcher
CHN10_AAC4HASH	Andrew Cunningham
GB92_RUK11	Andrew Cunningham
GB95_RUK13	Andrew Cunningham
GB07_AD_OS14	Amanda Duffus
GB08_RT113	Amanda Duffus
GB08_RT117	Amanda Duffus
GB08_RT126	Amanda Duffus
GB08_RT128	Amanda Duffus
GB08_RT132	Amanda Duffus
GB08_RT133	Amanda Duffus
GB08_RT137	Amanda Duffus
GB11_696	Becki Lawson
GB95_789	Becki Lawson
GB95_791	Becki Lawson
GB10_866	Becki Lawson
GB10_872	Becki Lawson
GB14_1074	Becki Lawson
GB15_1058	Becki Lawson
GB15_1180	Becki Lawson
GB92_008	Becki Lawson
GB92_100	Becki Lawson
GB92_274	Becki Lawson
GB92_766	Becki Lawson
GB93_065	Becki Lawson
GB93_286	Becki Lawson
GB93_287	Becki Lawson
GB93_339	Becki Lawson
GB93_760	Becki Lawson
GB93_761	Becki Lawson
GB93_762	Becki Lawson
GB96_915	Becki Lawson
NL17_3170726027	Bernardo Saucedo
ES10_GA_C5	Cesar Ayres
ES11_GA_01	Cesar Ayres
ES11_GA_05	Cesar Ayres
ES13_GA_334S2	Cesar Ayres
ES13_GA_354S2	Cesar Ayres
FR13_CMTV	Claude Miaud
FR14_1448	Claude Miaud
FR14_1457a	Claude Miaud
PT01_GERES_C5	Gonçalo Rosa
PT12_SDE12_G54	Gonçalo Rosa
PT13_SDE13_I23	Gonçalo Rosa
PT15_GERES_22	Gonçalo Rosa
CA16_JFB_A072	Joefelix Benneteau
ES11_PE_113N	Jaime Bosch
ES11_PE_114N	Jaime Bosch
USA_HV2PWA_P1	Jason Hoverman
USA14_HV_019	Jason Hoverman
USA14_HV_061	Jason Hoverman
USA66_FV3	Jacques Robert
USA_DMMG_MAINE	Matt Gray
USA16_HV_TXLICA1	Matt Gray
DE17_ToN215	Rachael Marschang
PT03_LMRV	Rachael Marschang

Table S2. *Metadata for the 170 ranavirus whole genome sequences included in this thesis. Note that parenthesised countries in Isolation Country are the native country of the host. The metadata is also available in CSV format at https://github.com/bioinfo-chris/PhD.git.*

Sample ID	Novel	GenBank Accession	Secondary Accession or ID	NCBI SRA Name	Ranavirus Clade	Sequence Length	Sample Date	Continent	Isolation Country	Isolation Source	Longitude	Latitude	Host Species	Host Genus	Host Order	Host Class
CA00_ATV_DAL1	No	KR075884.1	ATV_DAL1	NA	ATV-like	105354	01/07/2000	N. America	Canada	Wild	-106.347	56.130	Ambystoma tigrinum	Ambystoma	Urodela	Amphibia
CA13_ATV_04	No	MK580531.2	ATV_2013_CAD_04	NA	ATV-like	106258	16/08/2013	N. America	Canada	Wild	-106.347	56.130	Ambystoma mavortium	Ambystoma	Urodela	Amphibia
CA13_ATV_12	No	MK580532.2	ATV_2013_CAD_12	NA	ATV-like	106652	16/08/2013	N. America	Canada	Wild	-106.347	56.130	Ambystoma mavortium	Ambystoma	Urodela	Amphibia
CA13_ATV_LL1	No	MK580533.2	ATV_2013_CAD_LL1	NA	ATV-like	106578	13/07/2013	N. America	Canada	Wild	-106.347	56.130	Ambystoma mavortium	Ambystoma	Urodela	Amphibia
CA13_ATV_LL2	No	MK580534.1	ATV_2013_CAD_LL2	NA	ATV-like	106730	13/07/2013	N. America	Canada	Wild	-106.347	56.130	Ambystoma mavortium	Ambystoma	Urodela	Amphibia
CA13_ATV_LL3	No	MK580535.2	ATV_2013_CAD_LL3	NA	ATV-like	106629	13/07/2013	N. America	Canada	Wild	-106.347	56.130	Ambystoma mavortium	Ambystoma	Urodela	Amphibia
CA14_ATV_322	No	MK580536.1	ATV_2014_CAD_322	NA	ATV-like	106915	12/08/2014	N. America	Canada	Wild	-106.347	56.130	Ambystoma mavortium	Ambystoma	Urodela	Amphibia
CA97_ATV_RRV	No	KR075879.1	ATV_RRV	NA	ATV-like	106971	01/07/1997	N. America	Canada	Wild	-106.347	56.130	Ambystoma tigrinum	Ambystoma	Urodela	Amphibia
CA98_ATV_TSMB	No	KR075875.1	ATV_TSMB	NA	ATV-like	106526	01/07/1998	N. America	Canada	Wild	-106.347	56.130	Ambystoma tigrinum	Ambystoma	Urodela	Amphibia
USA_ATV_NM	No	KR075880.1	ATV_NM	NA	ATV-like	107471	01/07/2004	N. America	USA	Wild	-95.713	37.090	Ambystoma tigrinum	Ambystoma	Urodela	Amphibia
USA00_ATV_CAP	No	KR075886.1	ATV_CAP	NA	ATV-like	106004	01/07/2000	N. America	USA	Wild	-95.713	37.090	Ambystoma tigrinum	Ambystoma	Urodela	Amphibia
USA00_ATV_DO	No	KR075885.1	ATV_DO	NA	ATV-like	105936	01/07/2000	N. America	USA	Wild	-95.713	37.090	Ambystoma tigrinum	Ambystoma	Urodela	Amphibia
USA00_ATV_HEIDI	No	KR075873.1	ATV_HEDI	NA	ATV-like	106230	01/07/2000	N. America	USA	Wild	-95.713	37.090	Ambystoma tigrinum	Ambystoma	Urodela	Amphibia
USA00_ATV_JMH	No	KR075881.1	ATV_JMH	NA	ATV-like	106380	01/07/2000	N. America	USA	Wild	-95.713	37.090	Ambystoma tigrinum	Ambystoma	Urodela	Amphibia
USA00_ATV_SLV	No	KR075878.1	ATV_SLV	NA	ATV-like	106722	01/07/2000	N. America	USA	Wild	-95.713	37.090	Ambystoma tigrinum	Ambystoma	Urodela	Amphibia
USA01_ATV_AXO	No	KR075872.1	ATV_AXO	NA	ATV-like	105504	01/07/2001	N. America	USA	Captive	-95.713	37.090	Ambystoma mexicanum	Ambystoma	Urodela	Amphibia
USA01_ATV_GUFFY	No	KR075882.1	ATV_GUFFY	NA	ATV-like	106437	01/07/2001	N. America	USA	Wild	-95.713	37.090	Ambystoma tigrinum	Ambystoma	Urodela	Amphibia
USA03_ATV_DOT312	No	KR075883.1	ATV_DOT312	NA	ATV-like	107829	01/07/2003	N. America	USA	Wild	-95.713	37.090	Ambystoma tigrinum	Ambystoma	Urodela	Amphibia

USA96_ATV	No	NC_005832.1	AY150217.1	NA	ATV-like	106332	01/07/1996	N. America	USA	Wild	-110.631	31.455	Ambystoma t. stebbinsi	Ambystoma	Urodela	Amphibia
USA98_ATV_ORV	No	KR075874.1	ATV_ORV	NA	ATV-like	106018	01/07/1998	N. America	USA	Captive	-95.713	37.090	Ambystoma tigrinum	Ambystoma	Urodela	Amphibia
USA98_ATV_UTAH	No	KR075877.1	ATV_UTAH	NA	ATV-like	106198	01/07/1998	N. America	USA	Wild	-95.713	37.090	Ambystoma tigrinum	Ambystoma	Urodela	Amphibia
USA98_ATV_YEL	No	KR075876.1	ATV_YEL	NA	ATV-like	105922	01/07/1998	N. America	USA	Wild	-95.713	37.090	Ambystoma tigrinum	Ambystoma	Urodela	Amphibia
CH96_THRV	No	KP266741.1	THRV	NA	CMTV-like	105811	01/07/1996	Europe	Switzerland	Captive	8.228	46.818	Testudo hermanni	Testudo	Testudines	Reptilia
CHN10_AAC4HASH	Yes	NA	NA	CHI10Ad_N126	CMTV-like	105945	01/07/2010	Asia	China	Captive	104.195	35.862	Andrias davidianus	Andrias	Urodela	Amphibia
CHN10_ADRV2010	No	KF033124.1	ADRV_2010	NA	CMTV-like	106719	01/07/2010	Asia	China	Captive	104.195	35.862	Andrias davidianus	Andrias	Urodela	Amphibia
CHN10_KC2	No	KC243313.1	CGSIV_2014	NA	CMTV-like	104373	05/11/2010	Asia	China	Captive	104.195	35.862	Andrias davidianus	Andrias	Urodela	Amphibia
CHN12_ADRV1201	No	KC865735.1	ADRV_1201	NA	CMTV-like	106734	30/04/2012	Asia	China	Captive	104.195	35.862	Andrias davidianus	Andrias	Urodela	Amphibia
CHN13_CGSIV	No	KF512820.1	CGSIV_2013	NA	CMTV-like	105375	29/07/2013	Asia	China	Captive	104.195	35.862	Andrias davidianus	Andrias	Urodela	Amphibia
DE17_ToN215	Yes	NA	ToN215	GER17Ts_N215	CMTV-like	105910	01/07/2017	Europe	Germany	Captive	10.452	51.166	Tortoise spp.	Tortoise	Testudines	Reptilia
DK08_PEV	No	MF538627.1	PEV	NA	CMTV-like	107469	01/07/2008	Europe	Denmark	Wild	9.502	56.264	Pelophylax spp.	Pelophylax	Anura	Amphibia
ES08_CMTV_E	No	JQ231222.1	CMTV_SP	NA	CMTV-like	106878	01/08/2008	Europe	Spain	Wild	-3.749	40.464	Mesotriton alpestris	Mesotriton	Urodela	Amphibia
ES10_GA_C5	Yes	NA	NA	SPA10Lb_N133	CMTV-like	106724	01/07/2010	Europe	Spain	Wild	-8.614	42.499	Lissotriton boscai	Lissotriton	Urodela	Amphibia
ES11_GA_01	Yes	NA	NA	SPA11Nm_N247	CMTV-like	106633	01/07/2011	Europe	Spain	Wild	-8.616	42.500	Natrix maura	Natrix	Squamata	Reptilia
ES11_GA_05	Yes	NA	NA	SPA11Tm_N146	CMTV-like	106754	01/07/2011	Europe	Spain	Wild	-8.614	42.499	Triturus marmoratus	Triturus	Urodela	Amphibia
ES11_PE_113N	Yes	NA	AAOV	SPA11Ao_N127	CMTV-like	106303	15/07/2011	Europe	Spain	Wild	-4.716	43.213	Alytes obstetricans	Alytes	Anura	Amphibia
ES11_PE_114N	Yes	NA	NA	SPA11Ma_N108	CMTV-like	106200	28/09/2011	Europe	Spain	Wild	-4.786	43.197	Mesotriton alpestris	Mesotriton	Urodela	Amphibia
ES13_GA_334S2	Yes	NA	NA	SPA13Tm_N140	CMTV-like	106966	10/05/2013	Europe	Spain	Wild	-8.616	42.500	Triturus marmoratus	Triturus	Urodela	Amphibia
ES13_GA_354S2	Yes	NA	NA	SPA13Tm_N150	CMTV-like	106756	10/05/2013	Europe	Spain	Wild	-8.616	42.500	Triturus marmoratus	Triturus	Urodela	Amphibia
FIN97_PPIV	No	KX574341.1	PPIV	NA	CMTV-like	108041	01/07/1997	Europe	Finland	Captive	25.748	61.924	Sander lucioperca	Sander	Perciformes	Actinopterygii

FR13_CMTV	Yes	NA	NA	FRA13Rt_N243	CMTV-like	105916	01/07/2013	Europe	France	Wild	2.210	46.200	Rana temporaria	Rana	Anura	Amphibia
FR14_1448	Yes	NA	NA	FRA14Rt_N122	CMTV-like	105955	24/07/2014	Europe	France	Wild	6.776	44.332	Rana temporaria	Rana	Anura	Amphibia
FR14_1457a	Yes	NA	NA	FRA14Rt_N295	CMTV-like	105878	19/09/2014	Europe	France	Wild	7.375	44.081	Rana temporaria	Rana	Anura	Amphibia
GB11_696	Yes	NA	NA	GBR11Rt_N139	CMTV-like	106509	01/07/2011	Europe	UK	Wild	-4.067	50.384	Rana temporaria	Rana	Anura	Amphibia
GB95_789	Yes	NA	NA	GBR95Rt_N125	CMTV-like	106508	01/07/1995	Europe	UK	Wild	-0.313	51.313	Rana temporaria	Rana	Anura	Amphibia
GB95_791	Yes	NA	NA	GBR95Rt_N252	CMTV-like	106172	01/07/1995	Europe	UK	Wild	-0.313	51.313	Rana temporaria	Rana	Anura	Amphibia
IT01_REV	No	MF538628.1	REV	NA	CMTV-like	107444	01/07/2001	Europe	Italy	Wild	12.567	41.872	Rana esculenta	Rana	Anura	Amphibia
NL11_UU3110504006	No	MF004271.1	UU3110504006	NET11Ps_sraNA	CMTV-like	107445	04/04/2011	Europe	Netherlands	Wild	6.310	52.650	Pelophylax spp.	Pelophylax	Anura	Amphibia
NL11_UU3110810001	No	MF033604.1	UU3110810001	NET11Ps_sraNA	CMTV-like	107765	10/08/2011	Europe	Netherlands	Wild	6.210	52.770	Pelophylax spp.	Pelophylax	Anura	Amphibia
NL11_UU3110920007	No	MF038789.1	UU3110920007	NET11Ps_sraNA	CMTV-like	107832	20/09/2011	Europe	Netherlands	Wild	6.470	52.750	Pelophylax spp.	Pelophylax	Anura	Amphibia
NL12_UU3120627007	No	MF062693.1	UU3120627007	NET12Pf_sraNA	CMTV-like	107681	27/06/2012	Europe	Netherlands	Wild	6.280	52.630	Pelobates fuscus	Pelobates	Anura	Amphibia
NL13_nlpe13	No	NC_039034.1	KP056312.1	NA	CMTV-like	107772	22/08/2013	Europe	Netherlands	Wild	5.291	52.133	Pelophylax spp.	Pelophylax	Anura	Amphibia
NL13_UU3130829033	No	MF062694.1	UU3130829033	NET13Lv_sraNA	CMTV-like	107662	29/08/2013	Europe	Netherlands	Wild	6.290	52.940	Lissotriton vulgaris	Lissotriton	Urodela	Amphibia
NL14_UU3140624035	No	MF062695.1	UU3140624035	NET14Ps_sraNA	CMTV-like	107714	24/06/2014	Europe	Netherlands	Wild	6.290	52.940	Pelophylax spp.	Pelophylax	Anura	Amphibia
NL14_UU3140708068	No	MF093732.1	UU3140708068	NET14Ps_sraNA	CMTV-like	107907	08/07/2014	Europe	Netherlands	Wild	6.380	52.970	Pelophylax spp.	Pelophylax	Anura	Amphibia
NL14_UU3140911033	No	MF102028.1	UU3140911033	NET14Ps_sraNA	CMTV-like	107304	11/09/2014	Europe	Netherlands	Wild	6.370	52.730	Pelophylax spp.	Pelophylax	Anura	Amphibia
NL15_UU3150625026	No	MF102029.1	UU3150625026	NET15Lv_sraNA	CMTV-like	107217	25/06/2015	Europe	Netherlands	Wild	6.280	52.630	Lissotriton vulgaris	Lissotriton	Urodela	Amphibia
NL15_UU3150902003	No	MF102030.1	UU3150902003	NET15Ps_sraNA	CMTV-like	107650	20/09/2015	Europe	Netherlands	Wild	6.500	52.980	Pelophylax spp.	Pelophylax	Anura	Amphibia
NL15_UU3150902051	No	MF004272.1	UU3150902051	NET15Ps_sraNA	CMTV-like	107423	02/09/2015	Europe	Netherlands	Wild	6.380	52.970	Lissotriton vulgaris	Lissotriton	Urodela	Amphibia
NL15_UU3151009019	No	MF125269.1	UU3151009019	NET15Ps_sraNA	CMTV-like	106658	09/10/2015	Europe	Netherlands	Wild	6.280	51.980	Pelophylax spp.	Pelophylax	Anura	Amphibia
NL16_UU3160714042	No	MF125270.1	UU3160714042	NET16Ps_sraNA	CMTV-like	106404	14/07/2016	Europe	Netherlands	Wild	6.040	51.160	Pelophylax spp.	Pelophylax	Anura	Amphibia

NL17_3170726027	Yes	NA	NA	NET17Pe_N204	CMTV-like	106876	01/07/2017	Europe	Netherlands	Wild	6.318	52.700	Pelophylax spp.	Pelophylax	Anura	Amphibia
PT01_GERES_C5	Yes	NA	NA	POR01Lb_N104	CMTV-like	106457	01/07/2001	Europe	Portugal	Wild	-8.221	41.974	Lissotriton boscai	Lissotriton	Urodela	Amphibia
PT12_SDE12_G54	Yes	NA	NA	POR12Bb_N118	CMTV-like	106097	02/09/2012	Europe	Portugal	Wild	-7.613	40.322	Bufo bufo	Bufo	Anura	Amphibia
PT13_SDE13_I23	Yes	NA	NA	POR13Lb_N287	CMTV-like	108127	18/08/2013	Europe	Portugal	Wild	-7.593	40.336	Lissotriton boscai	Lissotriton	Urodela	Amphibia
PT15_GERES_22	Yes	NA	NA	POR15Tm_N120	CMTV-like	106195	01/07/2015	Europe	Portugal	Wild	-8.221	41.974	Triturus marmoratus	Triturus	Urodela	Amphibia
USA98_RCV15027	No	KX397571.1	NA	NA	CMTV-like	106890	01/05/1998	N. America	USA	Captive	-95.713	37.090	Lithobates catesbeianus	Lithobates	Anura	Amphibia
USA98_RCVZ	No	MF187210.1	RCVZ	NA	CMTV-like	106890	01/07/1998	N. America	USA	Captive	-83.248	31.423	Lithobates catesbeianus	Lithobates	Anura	Amphibia
PL15_RESRV	No	MT452035.1	RESRV	NA	CMTV-like	106878	01/07/2015	Europe	Poland	Wild	23.141	51.289	Trachemys scripta	Trachemys	Testudines	Reptilia
AU84_EHNV	No	FJ433873.1	EHNV	NA	EHNV-like	127011	01/07/1984	Oceania	Australia	Captive	144.963	-37.814	Perca fluviatilis	Perca	Perciformes	Actinopterygii
DK79_CodIV	No	KX574342.1	CodIV	NA	EHNV-like	114865	01/07/1979	Europe	Denmark	Captive	9.502	56.264	Gadus morhua	Gadus	Gadiformes	Actinopterygii
DK99_RMax	No	KX574343.1	RMax	NA	EHNV-like	115510	01/07/1999	Europe	Denmark	Captive	9.502	56.264	Scophthalmus maximus	Scophthalmus	Pleuronectiformes	Actinopterygii
ES11_ESV	No	JQ724856.1	ESV	NA	EHNV-like	127732	01/07/2011	Europe	Spain	Captive	-3.453	40.639	Silurus glanis	Silurus	Siluriformes	Actinopterygii
GB16_LFRV_V4955	No	MH665360.1	LFRV_V4955	NA	EHNV-like	115744	01/04/2016	Europe	UK	Wild	-3.436	55.378	Cyclopterus lumpus	Cyclopterus	Scorpaeniformes	Actinopterygii
HU12_ECV_13051	No	KT989884.1	ECV_13051	NA	EHNV-like	127751	01/07/2012	Europe	Hungary	Captive	19.503	47.162	Ameiurus nebulosus	Ameiurus	Siluriformes	Actinopterygii
HU12_ECV_14612	No	KT989885.1	ECV_14612	NA	EHNV-like	127549	01/07/2012	Europe	Hungary	Captive	19.503	47.162	Ameiurus nebulosus	Ameiurus	Siluriformes	Actinopterygii
ISL15_LFRV_F24	No	MH665358.1	LFRV_F24-15	NA	EHNV-like	116726	01/04/2015	Europe	Iceland	Wild	-19.021	64.963	Cyclopterus lumpus	Cyclopterus	Scorpaeniformes	Actinopterygii
ISL16_LFRV_F140	No	MH665359.1	LFRV_F140-16	NA	EHNV-like	115947	01/07/2016	Europe	Ireland	Captive	-7.692	53.142	Cyclopterus lumpus	Cyclopterus	Scorpaeniformes	Actinopterygii
NZ91_SERV	No	KX353311.2	SERV	NA	EHNV-like	126965	01/01/1991	Oceania	New Zealand	Captive	174.886	-40.901	Anguilla australis	Anguilla	Anguilliformes	Actinopterygii
AU88_V121	No	MT510729.1	V121	NA	EHNV-like	125923	01/07/1988	Oceania	Australia	Captive	147.524	-36.558	Oncorhynchus mykiss	Oncorhynchus	Salmoniformes	Actinopterygii
AU86_V084	No	MT510730.1	V084	NA	EHNV-like	127487	01/07/1986	Oceania	Australia	Captive	148.774	-35.996	Perca fluviatilis	Perca	Perciformes	Actinopterygii
AU86_V111	No	MT510731.1	V111	NA	EHNV-like	125883	01/07/1986	Oceania	Australia	Captive	148.233	-35.321	Oncorhynchus mykiss	Oncorhynchus	Salmoniformes	Actinopterygii

AU86_V017	No	MT510732.1	V017	NA	EHNV-like	125588	01/07/1986	Oceania	Australia	Captive	148.272	-35.507	Oncorhynchus mykiss	Oncorhynchus	Salmoniformes	Actinopterygii
AU90_V083	No	MT510733.1	V083	NA	EHNV-like	125838	01/07/1990	Oceania	Australia	Captive	145.214	-37.985	Perca fluviatilis	Perca	Perciformes	Actinopterygii
AU91_V080	No	MT510734.1	V080	NA	EHNV-like	125627	01/07/1991	Oceania	Australia	Captive	148.272	-35.507	Perca fluviatilis	Perca	Perciformes	Actinopterygii
AU93_V070	No	MT510735.1	V070	NA	EHNV-like	125670	01/07/1993	Oceania	Australia	Captive	148.635	-34.999	Perca fluviatilis	Perca	Perciformes	Actinopterygii
AU94_V069	No	MT510736.1	V069	NA	EHNV-like	126027	01/07/1994	Oceania	Australia	Captive	148.233	-35.321	Perca fluviatilis	Perca	Perciformes	Actinopterygii
AU92_V079	No	MT510737.1	V079	NA	EHNV-like	126019	01/07/1992	Oceania	Australia	Captive	148.233	-35.321	Oncorhynchus mykiss	Oncorhynchus	Salmoniformes	Actinopterygii
AU93_V071	No	MT510738.1	V071	NA	EHNV-like	125803	01/07/1993	Oceania	Australia	Captive	149.114	-35.295	Oncorhynchus mykiss	Oncorhynchus	Salmoniformes	Actinopterygii
AU96_V108	No	MT510739.1	V108	NA	EHNV-like	125635	01/07/1996	Oceania	Australia	Captive	149.263	-35.418	Oncorhynchus mykiss	Oncorhynchus	Salmoniformes	Actinopterygii
AU96_V109	No	MT510740.1	V109	NA	EHNV-like	125882	01/07/1996	Oceania	Australia	Captive	148.438	-35.422	Oncorhynchus mykiss	Oncorhynchus	Salmoniformes	Actinopterygii
AU09_V118	No	MT510741.1	V118	NA	EHNV-like	125960	01/07/2009	Oceania	Australia	Captive	146.921	-31.253	Perca fluviatilis	Perca	Perciformes	Actinopterygii
AU11_V130	No	MT510742.1	V130	NA	EHNV-like	125860	01/07/2011	Oceania	Australia	Captive	149.068	-35.232	Perca fluviatilis	Perca	Perciformes	Actinopterygii
AU85_VO18	No	MT510743.1	VO18	NA	EHNV-like	125591	01/07/1985	Oceania	Australia	Captive	149.068	-35.232	Perca fluviatilis	Perca	Perciformes	Actinopterygii
BR12_Rana-Bra-01	No	MH351268.1	Rana-Bra-01	NA	FV3-like	105080	20/10/2016	S. America	Brazil (USA)	Captive	-46.640	-23.556	Lithobates catesbeianus	Rana	Anura	Amphibia
CA15_KTW1	No	MK959607.1	FV3_KTW1	NA	FV3-like	106586	01/07/2015	N. America	Canada	Wild	-124.846	64.826	Rana sylvatica	Rana	Anura	Amphibia
CA16_JFB_A072	Yes	NA	NA	CAN16Rs_N267	FV3-like	105540	01/07/2016	N. America	Canada	Wild	-107.355	64.438	Rana sylvatica	Rana	Anura	Amphibia
CA16_KEN1	No	MK959604.1	FV3_KEN1	NA	FV3-like	103148	09/06/2016	N. America	Canada	Wild	-78.427	44.580	Rana pipiens	Rana	Anura	Amphibia
CA16_KEN5	No	MK959605.1	FV3_KEN5	NA	FV3-like	102577	09/06/2016	N. America	Canada	Wild	-78.427	44.580	Rana pipiens	Rana	Anura	Amphibia
CA17_Z377	No	MK959608.1	FV3_Z377	NA	FV3-like	104680	01/07/2017	N. America	Canada	Wild	-113.127	60.034	Pseudacris maculata	Pseudacris	Anura	Amphibia
CA17_Z582	No	MK959609.1	FV3_Z582	NA	FV3-like	106549	01/07/2017	N. America	Canada	Wild	-113.127	60.034	Pseudacris maculata	Pseudacris	Anura	Amphibia
CA17_Z897	No	MK959611.1	FV3_Z897	NA	FV3-like	104643	01/07/2017	N. America	Canada	Wild	-113.026	60.028	Pseudacris maculata	Pseudacris	Anura	Amphibia
CA17_Z920	No	MK959613.1	FV3_Z920	NA	FV3-like	104603	01/07/2017	N. America	Canada	Wild	-113.026	60.028	Rana sylvatica	Rana	Anura	Amphibia

CA17_Z929	No	MK959614.1	FV3_Z929	NA	FV3-like	104566	01/07/2017	N. America	Canada	Wild	-113.026	60.028	Pseudacris maculata	Pseudacris	Anura	Amphibia
CA17_Z930	No	MK959615.1	FV3_Z930	NA	FV3-like	104245	01/07/2017	N. America	Canada	Wild	-113.026	60.028	Rana sylvatica	Rana	Anura	Amphibia
CA17_Z933	No	MK959616.1	FV3_Z933	NA	FV3-like	105037	01/07/2017	N. America	Canada	Wild	-113.026	60.028	Pseudacris maculata	Pseudacris	Anura	Amphibia
CA17_Z938	No	MK959618.1	FV3_Z938	NA	FV3-like	104549	01/07/2017	N. America	Canada	Wild	-113.026	60.028	Rana sylvatica	Rana	Anura	Amphibia
CA17_Z939	No	MK959619.1	FV3_Z939	NA	FV3-like	104416	01/07/2017	N. America	Canada	Wild	-113.026	60.028	Rana sylvatica	Rana	Anura	Amphibia
CA17_Z942	No	MK959620.1	FV3_Z942	NA	FV3-like	103673	01/07/2017	N. America	Canada	Wild	-113.026	60.028	Rana sylvatica	Rana	Anura	Amphibia
CA17_Z994	No	MK959621.1	FV3_Z994	NA	FV3-like	104788	01/07/2017	N. America	Canada	Wild	-112.355	59.438	Rana sylvatica	Rana	Anura	Amphibia
CHN16_RNRV	No	MG791866.1	RNRV	NA	FV3-like	104286	01/07/2016	Asia	China	Captive	104.195	35.862	Rana nigromaculata	Rana	Anura	Amphibia
CHN95_RGV	No	JQ654586.1	RGV	NA	FV3-like	105791	09/10/1995	Asia	China (USA)	Captive	104.195	35.862	Rana grylio	Rana	Anura	Amphibia
CHN99_STIV	No	EU627010.1	STIV	NA	FV3-like	105890	01/01/1999	Asia	China	Captive	114.058	22.543	Trionyx sinensis	Trionyx	Testudines	Reptilia
GB07_AD_OS14	Yes	NA	NA	GBR07Ao_N175	FV3-like	104493	01/07/2007	Europe	UK	Wild	-0.137	50.823	Alytes obstetricans	Alytes	Anura	Amphibia
GB08_RT113	Yes	NA	NA	GBR08Rt_N186	FV3-like	104752	01/07/2008	Europe	UK	Wild	-0.159	51.360	Rana temporaria	Rana	Anura	Amphibia
GB08_RT117	Yes	NA	NA	GBR08Rt_H106	FV3-like	104962	01/07/2008	Europe	UK	Wild	-1.792	53.693	Rana temporaria	Rana	Anura	Amphibia
GB08_RT126	Yes	NA	NA	GBR08Rt_H108	FV3-like	104534	01/07/2008	Europe	UK	Wild	-1.482	50.995	Rana temporaria	Rana	Anura	Amphibia
GB08_RT128	Yes	NA	NA	GBR08Rt_H109	FV3-like	105049	01/07/2008	Europe	UK	Wild	-0.159	51.360	Rana temporaria	Rana	Anura	Amphibia
GB08_RT132	Yes	NA	NA	GBR08Rt_N173	FV3-like	105180	01/07/2008	Europe	UK	Wild	-2.714	53.746	Rana temporaria	Rana	Anura	Amphibia
GB08_RT133	Yes	NA	NA	GBR08Rt_N187	FV3-like	105387	01/07/2008	Europe	UK	Wild	-2.714	53.746	Rana temporaria	Rana	Anura	Amphibia
GB08_RT137	Yes	NA	NA	GBR08Rt_N178	FV3-like	104103	01/07/2008	Europe	UK	Wild	-2.714	53.746	Rana temporaria	Rana	Anura	Amphibia
GB10_866	Yes	NA	NA	GBR10Rt_N254	FV3-like	104467	01/07/2010	Europe	UK	Wild	-0.560	51.315	Rana temporaria	Rana	Anura	Amphibia
GB10_872	Yes	NA	NA	GBR10Rt_N156	FV3-like	105534	01/07/2010	Europe	UK	Wild	-2.321	53.421	Rana temporaria	Rana	Anura	Amphibia
GB14_1074	Yes	NA	NA	GBR14Rt_N153	FV3-like	105556	01/07/2014	Europe	UK	Wild	-0.707	51.011	Rana temporaria	Rana	Anura	Amphibia

GB15_1058	Yes	NA	NA	GBR15Rt_N208	FV3-like	105498	01/07/2015	Europe	UK	Wild	0.738	51.541	Rana temporaria	Rana	Anura	Amphibia
GB15_1180	Yes	NA	NA	GBR15Rt_N209	FV3-like	104860	01/07/2015	Europe	UK	Wild	-1.959	50.737	Rana temporaria	Rana	Anura	Amphibia
GB92_008	Yes	NA	NA	GBR92Rt_N190	FV3-like	104828	01/07/1992	Europe	UK	Wild	-1.084	50.802	Rana temporaria	Rana	Anura	Amphibia
GB92_100	Yes	NA	NA	GBR92Rt_N132	FV3-like	104274	01/07/1992	Europe	UK	Wild	-0.373	51.461	Rana temporaria	Rana	Anura	Amphibia
GB92_274	Yes	NA	NA	GBR92Rt_N192	FV3-like	104429	01/07/1992	Europe	UK	Wild	-1.086	50.820	Rana temporaria	Rana	Anura	Amphibia
GB92_766	Yes	NA	NA	GBR92Rt_N230	FV3-like	104599	01/07/1992	Europe	UK	Wild	-1.888	50.719	Rana temporaria	Rana	Anura	Amphibia
GB92_RUK11	Yes	NA	RUK11	GBR92Rt_H110	FV3-like	103483	01/07/1992	Europe	UK	Wild	-0.421	51.535	Rana temporaria	Rana	Anura	Amphibia
GB93_065	Yes	NA	NA	GBR93Rt_N130	FV3-like	106796	01/07/1993	Europe	UK	Wild	-0.126	51.135	Rana temporaria	Rana	Anura	Amphibia
GB93_286	Yes	NA	NA	GBR93Rt_N135	FV3-like	105843	01/07/1993	Europe	UK	Wild	1.401	51.223	Rana temporaria	Rana	Anura	Amphibia
GB93_287	Yes	NA	NA	GBR93Rt_N148	FV3-like	105234	01/07/1993	Europe	UK	Wild	-0.560	51.315	Rana temporaria	Rana	Anura	Amphibia
GB93_339	Yes	NA	NA	GBR93Rt_N112	FV3-like	105009	01/07/1993	Europe	UK	Wild	0.420	51.574	Rana temporaria	Rana	Anura	Amphibia
GB93_760	Yes	NA	NA	GBR93Rt_N224	FV3-like	105675	01/07/1993	Europe	UK	Wild	-0.269	50.839	Rana temporaria	Rana	Anura	Amphibia
GB93_761	Yes	NA	NA	GBR93Rt_N225	FV3-like	105277	01/07/1993	Europe	UK	Wild	-0.269	50.839	Rana temporaria	Rana	Anura	Amphibia
GB93_762	Yes	NA	NA	GBR93Rt_N226	FV3-like	105550	01/07/1993	Europe	UK	Wild	-0.269	50.839	Rana temporaria	Rana	Anura	Amphibia
GB95_RUK13	Yes	KJ538546.1	RUK13	GBR95Rt_H111	FV3-like	106170	01/07/1995	Europe	UK	Wild	1.148	52.057	Rana temporaria	Rana	Anura	Amphibia
GB96_915	Yes	NA	NA	GBR96Bb_N193	FV3-like	103631	01/07/1996	Europe	UK	Wild	-0.217	51.348	Bufo bufo	Bufo	Anura	Amphibia
NI15_FV3_NIC	No	MF360246.1	FV3_NIC	NA	FV3-like	107035	24/03/2015	S. America	Netherlands	Captive	5.291	52.133	Oophaga pumilio	Oophaga	Anura	Amphibia
PT03_LMRV	Yes	NA	LMRV	POR03Im_N286	FV3-like	105329	01/07/2003	Europe	Portugal	Wild	-8.224	39.400	Iberolacerta monticola	Iberolacerta	Squamata	Reptilia
USA_DMMG_MAINE	Yes	NA	DMMG	USAUNRs_N159	FV3-like	106471	NA	N. America	USA	Wild	-69.445	45.254	Rana sylvatica	Rana	Anura	Amphibia
USA_FV3	No	AY548484.1	NA	NA	FV3-like	105903	NA	N. America	USA	Wild	-95.713	37.090	Rana pipiens	Rana	Anura	Amphibia
USA_HV2PWA_P1	Yes	NA	NA	USAUNLc_N265	FV3-like	106955	NA	N. America	USA	Wild	-86.135	40.267	Lithobates clamitans	Lithobates	Anura	Amphibia

USA06_RCVZ2	No	MF187209.1	RCVZ2	NA	FV3-like	104968	30/06/2006	N. America	USA	Captive	-83.248	31.423	Lithobates catesbeianus	Lithobates	Anura	Amphibia
USA11_TCCRV	No	MG953518.1	TCCRV	NA	FV3-like	104894	01/07/2011	N. America	USA	Captive	-76.600	39.300	Terrapene carolina carolina	Terrapene	Testudines	Reptilia
USA14_HV_019	Yes	NA	NA	USA14Tt_N263	FV3-like	105390	26/04/2014	N. America	USA	Wild	-121.896	37.617	Taricha torosa	Taricha	Urodela	Amphibia
USA14_HV_061	Yes	NA	NA	USA14Rc_N143	FV3-like	106784	07/07/2014	N. America	USA	Wild	-121.896	37.617	Lithobates catesbeianus	Rana	Anura	Amphibia
USA15_TMRV1	No	MG953519.1	TMRV1	NA	FV3-like	104803	01/07/2015	N. America	USA	Captive	-76.600	39.300	Trioceros melleri	Trioceros	Squamata	Reptilia
USA15_TMRV2	No	MG953520.1	TMRV2	NA	FV3-like	105030	01/07/2015	N. America	USA	Captive	-76.600	39.300	Trioceros melleri	Trioceros	Squamata	Reptilia
USA16_HV_TXLICA1	Yes	NA	NA	USA16Lc_N264	FV3-like	105232	01/07/2016	N. America	USA	Wild	-99.902	31.969	Lithobates clamitans	Lithobates	Anura	Amphibia
USA66_FV3	Yes	NA	FV3	USA66Rp_N162	FV3-like	105647	01/07/1966	N. America	USA	Wild	-72.600	44.600	Rana pipiens	Rana	Anura	Amphibia
USA98_FV3_SSME	No	KJ175144.1	SSME	NA	FV3-like	105070	01/07/1998	N. America	USA	Wild	-69.400	45.300	Ambystoma maculatum	Ambystoma	Urodela	Amphibia
BR17_Rana-Bra-17	No	MT578298.1	Rana-Bra-17	NA	FV3-like	104744	01/07/2017	S. America	Brazil (USA)	Captive	-46.747	-21.472	Lithobates catesbeianus	Lithobates	Anura	Amphibia
USA96_SBV	No	MZ514903.1	Stickleback virus	NA	FV3-like	105673	01/07/1996	N. America	USA	Wild	-95.713	37.090	Gasterosteus aculeatus	Gasterosteus	Gasterosteiformes	Actinopterygii
USA96_TPV	No	MZ514904.1	Tadpole virus 2	NA	FV3-like	105671	01/07/1996	N. America	USA	Wild	-95.713	37.090	Rana aurora	Rana	Anura	Amphibia
AU92_BIV	No	NC_038507.1	KX185156.1	NA	TFV-like	103531	01/01/1992	Oceania	Australia	Captive	146.719	-19.254	Limnodynastes ornatus	Limnodynastes	Anura	Amphibia
CHN00_TFV	No	AF389451.1	TFV	NA	TFV-like	105057	01/07/2000	Asia	China	Captive	113.143	23.029	Hoplobatrachus tigerinus	Hoplobatrachus	Anura	Amphibia
DE96_ToRV1	No	KP266743.1	ToRV1	NA	TFV-like	103876	01/07/1996	Europe	Germany	Captive	10.452	51.166	Testudo kleinmanni	Testudo	Testudines	Reptilia
DE99_GGRV	No	KP266742.1	GGRV	NA	TFV-like	103681	01/07/1999	Europe	Germany	Captive	10.452	51.166	Uroplatus fimbriatus	Uroplatus	Squamata	Reptilia
USA14_ZRV	No	MK227779.1	ZRV	NA	TFV-like	103266	01/07/2014	N. America	USA	Captive	-93.098	41.878	Anaxyrus boreas boreas	Anaxyrus	Anura	Amphibia
TH04_TFV	No	MT512497.1	TFV_D03-034	NA	TFV-like	105529	01/07/2004	Asia	Thailand (Cambodia)	Captive	102.066	13.822	Fejervarya limnocharis	Fejervarya	Anura	Amphibia
TH11_TFV	No	MT512498.1	TFV_D11-067	NA	TFV-like	105418	01/07/2011	Asia	Thailand	Captive	100.074	7.617	Hoplobatrachus rugulosus	Hoplobatrachus	Anura	Amphibia
TH16_TFV	No	MT512499.1	TFV_VD-16-006	NA	TFV-like	105206	01/07/2016	Asia	Thailand	Captive	99.813	13.528	Hoplobatrachus rugulosus	Hoplobatrachus	Anura	Amphibia
TH17_TFV	No	MT512500.1	TFV_VD-17-007	NA	TFV-like	105114	01/07/2017	Asia	Thailand	Captive	101.282	12.681	Hoplobatrachus rugulosus	Hoplobatrachus	Anura	Amphibia

TH02_TFV	No	MT512501.1	TF V_F0207	NA	TFV-like	106226	01/07/2002	Asia	Thailand	Captive	100.502	13.756	Carassius auratus	Carassius	Cypriniformes	Actinopterygii
TH00_TFV	No	MT512502.1	TFV_D2008	NA	TFV-like	105405	01/07/2000	Asia	Thailand	Captive	100.037	13.814	Oxyeleotris marmorata	Oxyeleotris	Gobiiformes	Actinopterygii
TH01_TFV	No	MT512503.1	TFV_F2112	NA	TFV-like	105249	01/07/2001	Asia	Thailand	Captive	100.274	13.550	Poecilia reticulata	Poecilia	Cyprinodontiformes	Actinopterygii
TH98_TFV	No	MT512504.1	TFV_AV9803	NA	TFV-like	105022	01/07/1998	Asia	Thailand	Captive	100.502	13.756	Hoplobatrachus tigerinus	Hoplobatrachus	Anura	Amphibia
TD_TFV_CHAD	No	MW727505.1	NA	NA	TFV-like	106120	NA	Africa	Chad	Wild	15.033	12.117	Hoplobatrachus occipitalis	Hoplobatrachus	Anura	Amphibia

***Table S3**. Protein functional annotation of the amphibian-like* Ranavirus *(ALRV) pan-genome. All identified coding sequences for each isolate (n = 170) were clustered at 80% amino acid homology using the* Roary *pipeline. 217 ORF clusters are listed with their reference sequence GC content, their closest* BLASTn *match to published ranavirus CDSs, and their best translated amino acid match to the Pfam database yielded from hidden Markov model (HMM) scans. The consensus annotation gives protein information followed functional domain information, which is divided by a semi-colon. Failed matches by either method are denoted by NA, and if both methods failed, ORF clusters were annotated as hypothetical protein as the consensus annotation.*

Roary ORF Cluster (80%)	GC Content	Consensus Protein and/or Functional Annotation	BLASTn Match	Pfam HMM Match
BALF5	0.59	DNA polymerase family B	DNA polymerase	DNA polymerase family B
DUT	0.63	dUTPase	dUTPase	dUTPase
F4L	0.56	Ribonucleotide reductase, small chain	Ribonucleotide reductase small subunit	Ribonucleotide reductase, small chain
I4L	0.61	Ribonucleotide reductase, barrel domain	Ribonucleoside diphosphate reductase alpha subunit	Ribonucleotide reductase, barrel domain
MCP	0.59	Large eukaryotic DNA virus major capsid protein	Major capsid protein	Large eukaryotic DNA virus major capsid protein
RPO1	0.62	RNA polymerase Rpb1, domain 2	DNA-dependent RNA polymerase II largest subunit	RNA polymerase Rpb1, domain 2
RPO2	0.61	RNA polymerase Rpb2, domain 6	DNA-dependent RNA polymerase b subunit	RNA polymerase Rpb2, domain 6
TD	0.56	Thymidylate synthase	Thymidylate synthase	Thymidylate synthase
TD_2	0.56	Thymidylate synthase	Thymidylate synthase	Thymidylate synthase
group_1	0.55	Neurofilament triplet H1-like protein	Neurofilament triplet H1-like protein	NA
group_2	0.56	Neurofilament triplet H1-like protein; Rho termination factor, N-terminal domain	Neurofilament triplet H1-like protein	Rho termination factor, N-terminal domain
group_3	0.55	Neurofilament triplet H1-like protein; Rho termination factor, N-terminal domain	Neurofilament triplet H1-like protein	Rho termination factor, N-terminal domain
group_4	0.5	Neurofilament triplet H1-like protein	Neurofilament triplet H1-like protein	NA
group_5	0.57	Hypothetical protein	Hypothetical protein	NA
group_6	0.57	Hypothetical protein	Hypothetical protein	NA
group_7	0.56	Hypothetical protein	Hypothetical protein	NA
group_8	0.53	Neurofilament triplet H1-like protein; Repeat of unknown function (DUF1388)	Neurofilament triplet H1-like protein	Repeat of unknown function (DUF1388)

group_9	0.51	SAP domain-containinig protein; SAP domain	Putative SAP domain-containinig protein	SAP domain
group_10	0.52	LCDV1 orf58-like protein; SAP domain	LCDV1 orf58-like protein	SAP domain
group_11	0.56	Hypothetical protein; Family of unknown function (DUF5892)	Hypothetical protein	Family of unknown function (DUF5892)
group_12	0.58	Hypothetical protein; Family of unknown function (DUF5852)	Hypothetical protein	Family of unknown function (DUF5852)
group_13	0.57	Hypothetical protein; Family of unknown function (DUF5852)	Hypothetical protein	Family of unknown function (DUF5852)
group_14	0.66	Hypothetical protein	Hypothetical protein	NA
group_15	0.58	CARD-like caspase; Caspase recruitment domain	CARD-like caspase	Caspase recruitment domain
group_16	0.61	Myristylated membrane protein; Pox virus entry-fusion-complex G9/A16	Myristylated membrane protein	Pox virus entry-fusion-complex G9/A16
group_17	0.56	Multi-glycosylated core protein 24 (MGC-24), sialomucin	Hypothetical protein	Multi-glycosylated core protein 24 (MGC-24), sialomucin
group_18	0.48	US22 family protein	US22 family protein	US22 like
group_19	0.57	Fibroblast growth factor	Hypothetical protein	Fibroblast growth factor
group_20	0.58	PaaX-like protein	Hypothetical protein	PaaX-like protein
group_21	0.56	Hypothetical protein	Hypothetical protein	NA
group_22	0.58	PaaX-like protein	Hypothetical protein	PaaX-like protein
group_23	0.58	PaaX-like protein	Hypothetical protein	PaaX-like protein
group_24	0.58	PaaX-like protein	Hypothetical protein	PaaX-like protein
group_25	0.56	Putative ATPase-dependent protease	Putative ATPase-dependent protease	NA
group_26	0.57	Hypothetical protein	Hypothetical protein	NA
group_27	0.57	Putative eIF2-alpha like protein	Putative eIF2-alpha like protein	NA
group_28	0.56	Starch binding domain	Hypothetical protein	Starch binding domain
group_29	0.56	Starch binding domain	Hypothetical protein	Starch binding domain
group_30	0.56	Hypothetical protein	Hypothetical protein	NA
group_31	0.56	Hypothetical protein	Hypothetical protein	NA
group_32	0.55	Hypothetical protein	Hypothetical protein	NA
group_33	0.56	Starch binding domain	NA	Starch binding domain

group_34	0.61	2-cysteine adaptor domain protein	Putative 2-cysteine adaptor domain protein	2-cysteine adaptor domain
group_37	0.56	Hypothetical protein	Hypothetical protein	NA
group_38	0.5	Hypothetical protein	Hypothetical protein	NA
group_39	0.65	Collagen triple helix repeat (20 copies)	Hypothetical protein	Collagen triple helix repeat (20 copies)
group_40	0.64	Serine/threonine-protein kinase	Serine/threonine-protein kinase	NA
group_41	0.59	Putative surface protein	surface protein	NA
group_42	0.48	Hypothetical protein	NA	NA
group_43	0.56	Hypothetical protein; Domain of unknown function (DUF4082)	Hypothetical protein	Domain of unknown function (DUF4082)
group_44	0.55	Hypothetical protein	Hypothetical protein	NA
group_45	0.58	Hypothetical protein	Hypothetical protein	NA
group_46	0.56	Hydrolase of the metallo-beta-lactamasesuperfamily	Hydrolase of the metallo-beta-lactamasesuperfamily	NA
group_47	0.54	Hypothetical protein	Hypothetical protein	NA
group_48	0.64	Hypothetical protein	Hypothetical protein	NA
group_49	0.48	Hypothetical protein; Family of unknown function (DUF5877)	Hypothetical protein	Family of unknown function (DUF5877)
group_50	0.66	Collagen triple helix repeat (20 copies)	Hypothetical protein	Collagen triple helix repeat (20 copies)
group_51	0.64	Collagen triple helix repeat (20 copies)	Hypothetical protein	Collagen triple helix repeat (20 copies)
group_52	0.69	Collagen triple helix repeat (20 copies)	Hypothetical protein	Collagen triple helix repeat (20 copies)
group_53	0.69	Collagen triple helix repeat (20 copies)	Hypothetical protein	Collagen triple helix repeat (20 copies)
group_54	0.66	Collagen triple helix repeat (20 copies)	Hypothetical protein	Collagen triple helix repeat (20 copies)
group_55	0.7	Collagen triple helix repeat (20 copies)	Hypothetical protein	Collagen triple helix repeat (20 copies)
group_56	0.57	Putative immediate-early protein; Protein of unknown function (DUF2888)	Putative immediate-early protein	Protein of unknown function (DUF2888)
group_57	0.49	Polymer-forming cytoskeletal	Hypothetical protein	Polymer-forming cytoskeletal
group_58	0.55	Hypothetical protein	Hypothetical protein	NA
group_59	0.59	Hypothetical protein; Family of unknown function (DUF5850)	Hypothetical protein	Family of unknown function (DUF5850)
group_60	0.56	Putative surface protein	surface protein	NA

group_61	0.54	IBR domain, a half RING-finger domain	Hypothetical protein	IBR domain, a half RING-finger domain
group_62	0.55	Hypothetical protein	Hypothetical protein	NA
group_63	0.56	3-beta hydroxysteroid dehydrogenase; 3-beta hydroxysteroid dehydrogenase/isomerase family	3beta-hydroxysteroid dehydrogenase	3-beta hydroxysteroid dehydrogenase/isomerase family
group_64	0.49	3-beta hydroxy-delta-5-C27 steroid oxidoreductase-like protein	3-beta-hydroxy-delta-5-C27 steroid oxidoreductase-like protein	NA
group_65	0.49	3-beta hydroxy-delta-5-C27 steroid oxidoreductase-like protein	3-beta-hydroxy-delta-5-C27 steroid oxidoreductase-like protein	NA
group_66	0.56	3-beta hydroxysteroid dehydrogenase/isomerase family	3-beta hydroxysteroid dehydrogenase	3-beta hydroxysteroid dehydrogenase/isomerase family
group_67	0.53	Hypothetical protein	Hypothetical protein	NA
group_68	0.56	Human parainfluenza virus 1L-like protein	Human parainfluenza virus 1L-like protein	NA
group_69	0.6	Integrase-like protein; Vertebrate interleukin-3 regulated transcription factor	Integrase-like protein	Vertebrate interleukin-3 regulated transcription factor
group_70	0.58	C-mannosyltransferase dpy-19; Family of unknown function (DUF5875)	C-mannosyltransferase dpy-19	Family of unknown function (DUF5875)
group_71	0.53	Hypothetical protein	Hypothetical protein	NA
group_72	0.62	Transcription elongation factor S-II (TFIIS)	Transcription elongation factor S-II	Transcription factor S-II (TFIIS)
group_73	0.6	Hypothetical protein	Hypothetical protein	NA
group_74	0.56	Myeloid cell leukemia protein; Apoptosis regulator proteins, Bcl-2 family	Myeloid cell leukemia protein	Apoptosis regulator proteins, Bcl-2 family
group_75	0.65	Hypothetical protein; Domain of unknown function (DUF1729)	Hypothetical protein	Domain of unknown function (DUF1729)
group_76	0.56	Lipopolysaccharide-induced TNF-alpha factor (LITAF)-like protein; LITAF-like zinc ribbon domain	Lipopolysaccharide-induced TNF-alpha factor (LITAF)-like protein	LITAF-like zinc ribbon domain
group_77	0.64	Collagen triple helix repeat (20 copies)	Hypothetical protein	Collagen triple helix repeat (20 copies)
group_78	0.59	Mitochondrial resolvase Ydc2 / RNA splicing MRS1	Hypothetical protein	Mitochondrial resolvase Ydc2 / RNA splicing MRS1
group_79	0.55	Hypothetical protein	Hypothetical protein	NA
group_80	0.59	Hypothetical protein	Hypothetical protein	NA
group_81	0.58	Hypothetical protein	Hypothetical protein	NA
group_82	0.58	Hypothetical protein	Hypothetical protein	NA
group_83	0.56	Immediate early protein ICP-46	Immediate early protein ICP-46	NA
group_84	0.56	Immediate early protein ICP-46	Immediate early protein ICP-46	NA

group_85	0.53	Deoxynucleoside kinase/thymidine kinase	Deoxynucleoside kinase/thymidine kinase	Deoxynucleoside kinase
group_86	0.65	Hypothetical protein	Hypothetical protein	NA
group_87	0.64	Putative DEXDc superfamily helicase-like protein; Type III restriction enzyme, res subunit	putative DEXDc superfamily helicase-like protein	Type III restriction enzyme, res subunit
group_88	0.64	Helicase-like protein; Type III restriction enzyme, res subunit	helicase-like protein	Type III restriction enzyme, res subunit
group_89	0.6	Putative phosphotransferase; Protein kinase domain	putative phosphotransferase	Protein kinase domain
group_90	0.58	Hypothetical protein	Hypothetical protein	NA
group_91	0.6	Hypothetical protein	Hypothetical protein	NA
group_92	0.55	Hypothetical protein	Hypothetical protein	NA
group_93	0.56	Hypothetical protein	Hypothetical protein	NA
group_94	0.56	AAA-ATPase; Poxvirus A32 protein	AAA-ATPase	Poxvirus A32 protein
group_95	0.59	Hypothetical protein	Hypothetical protein	NA
group_96	0.61	Hypothetical protein	Hypothetical protein	NA
group_97	0.65	Transmembrane domain protein; Domain of unknown function (DUF4444)	Transmembrane domain protein	Domain of unknown function (DUF4444)
group_99	0.61	DNA-dependent RNA polymerase b subunit; RNA polymerase Rpb2, domain 6	DNA-dependent RNA polymerase b subunit	RNA polymerase Rpb2, domain 6
group_100	0.62	DNA-dependent RNA polymerase b subunit; RNA polymerase Rpb2, domain 6	DNA-dependent RNA polymerase b subunit	RNA polymerase Rpb2, domain 6
group_101	0.57	RNA polymerase Rpb5, C-terminal domain	Hypothetical protein	RNA polymerase Rpb5, C-terminal domain
group_102	0.57	Hypothetical protein; Protein of unknown function (DUF2726)	Hypothetical protein	Protein of unknown function (DUF2726)
group_103	0.6	Hypothetical protein	Hypothetical protein	NA
group_104	0.58	Hypothetical protein	Hypothetical protein	NA
group_105	0.48	Hypothetical protein	Hypothetical protein	NA
group_106	0.64	Tyrosine kinase; Glycosyl transferase family 90	tyrosine kinase	Glycosyl transferase family 90
group_107	0.6	Hypothetical protein	Hypothetical protein	NA
group_109	0.58	Hypothetical protein; Family of unknown function (DUF5876)	Hypothetical protein	Family of unknown function (DUF5876)
group_110	0.53	Capsid maturation protease	Capsid maturation protease	NA
group_111	0.6	Hypothetical protein; Family of unknown function (DUF5832)	Hypothetical protein	Family of unknown function (DUF5832)

group_113	0.52	Hypothetical protein	Hypothetical protein	NA
group_114	0.47	Hemimethylated DNA-binding protein YccV like	NA	Hemimethylated DNA-binding protein YccV like
group_115	0.67	Hypothetical protein	Hypothetical protein	NA
group_116	0.64	Hypothetical protein	Hypothetical protein	NA
group_117	0.69	Hypothetical protein	Hypothetical protein	NA
group_118	0.52	Hypothetical protein	Hypothetical protein	NA
group_119	0.58	NLI interacting factor-like phosphatase	putative NIF/NLI interacting factor	NLI interacting factor-like phosphatase
group_120	0.56	Tryptophan-associated transmembrane protein	Hypothetical protein	Tryptophan-associated transmembrane protein (Trp_oprn_chp)
group_121	0.49	WXG100 protein secretion system (Wss), protein YukC	Hypothetical protein	WXG100 protein secretion system (Wss), protein YukC
group_122	0.58	Proliferating cell nuclear antigen	proliferating cell nuclear antigen	NA
group_123	0.54	Hypothetical protein; Family of unknown function (DUF5770)	Hypothetical protein	Family of unknown function (DUF5770)
group_124	0.61	Hypothetical protein	Hypothetical protein	NA
group_125	0.61	Putative DNA repair protein RAD2; XPG I-region	putative DNA repair protein RAD2	XPG I-region
group_126	0.52	Thiol oxidoreductase; Erv1 / Alr family	thiol oxidoreductase	Erv1 / Alr family
group_127	0.58	Ribonuclease III; Ribonuclease III domain	ribonuclease III	Ribonuclease III domain
group_128	0.58	Hypothetical protein; Domain of unknown function (DUF4754)	Hypothetical protein	Domain of unknown function (DUF4754)
group_129	0.53	Hypothetical protein	Hypothetical protein	NA
group_130	0.52	Hypothetical protein	Hypothetical protein	NA
group_131	0.54	Helicase family protein; Family of unknown function (DUF5767)	Helicase family protein	Family of unknown function (DUF5767)
group_132	0.53	Replicating factor; Poxvirus Late Transcription Factor VLTF3 like	Replicating factor	Poxvirus Late Transcription Factor VLTF3 like
group_133	0.58	Cytosine DNA methyltransferase; C-5 cytosine-specific DNA methylase	Cytosine DNA methyltransferase	C-5 cytosine-specific DNA methylase
group_134	0.56	Ribonucleotide reductase, small chain	ribonucleotide reductase beta subunit	Ribonucleotide reductase, small chain
group_136	0.62	NTPase/helicase; Type III restriction enzyme, res subunit	NTPase/helicase	Type III restriction enzyme, res subunit
group_137	0.52	p31K protein; Protein of unknown function (DUF2738)	p31K protein	Protein of unknown function (DUF2738)
group_138	0.42	US22 like	Hypothetical protein	US22 like

group_139	0.66	Collagen triple helix repeat (20 copies)	Hypothetical protein	Collagen triple helix repeat (20 copies)
group_140	0.54	Hypothetical protein	Hypothetical protein	NA
group_141	0.56	Hypothetical protein	Hypothetical protein	NA
group_142	0.56	Hypothetical protein	Hypothetical protein	NA
group_143	0.62	Methyl-accepting chemotaxis sensory transducer; ATP synthase D chain, mitochondrial (ATP5H)	Methyl-accepting chemotaxis sensory transducer	ATP synthase D chain, mitochondrial (ATP5H)
group_144	0.6	Myristylated membrane protein; Lipid membrane protein of large eukaryotic DNA viruses	Myristylated membrane protein	Lipid membrane protein of large eukaryotic DNA viruses
group_145	0.62	Hypothetical protein; Family of unknown function (DUF5757)	Hypothetical protein	Family of unknown function (DUF5757)
group_147	0.54	Hypothetical protein	Hypothetical protein	NA
group_148	0.54	Hypothetical protein	Hypothetical protein	NA
group_149	0.59	Hypothetical protein	Hypothetical protein	NA
group_150	0.53	Hypothetical protein	Hypothetical protein	NA
group_151	0.59	Hypothetical protein	Hypothetical protein	NA
group_152	0.6	Hypothetical protein	Hypothetical protein	NA
group_153	0.59	Hypothetical protein	Hypothetical protein	NA
group_154	0.59	Hypothetical protein	Hypothetical protein	NA
group_155	0.42	Hypothetical protein	Hypothetical protein	NA
group_156	0.52	Neurofilament triplet H1-like protein; Ferredoxin I 4Fe-4S cluster domain	Neurofilament triplet H1-like protein	Ferredoxin I 4Fe-4S cluster domain
group_157	0.55	Hypothetical protein	NA	NA
group_158	0.67	Collagen triple helix repeat (20 copies)	Hypothetical protein	Collagen triple helix repeat (20 copies)
group_160	0.39	US22 family protein; Formiminotransferase-cyclodeaminase	US22 family protein	Formiminotransferase-cyclodeaminase
group_161	0.63	Hypothetical protein	Hypothetical protein	NA
group_162	0.41	ZIP Zinc transporter	Hypothetical protein	ZIP Zinc transporter
group_164	0.52	Hypothetical protein	Hypothetical protein	NA
group_165	0.55	Hypothetical protein	Hypothetical protein	NA
group_166	0.51	US22 like	US22 family protein	US22 like

group_167	0.54	Hypothetical protein; Family of unknown function (DUF5852)	Hypothetical protein	Family of unknown function (DUF5852)
group_168	0.52	Hypothetical protein	Hypothetical protein	NA
group_169	0.54	Hypothetical protein	Hypothetical protein	NA
group_170	0.58	Hypothetical protein; N-terminal domain of unknown function (DUF4140)	Hypothetical protein	N-terminal domain of unknown function (DUF4140)
group_171	0.53	Dihydrofolate reductase	Dihydrofolate reductase	Dihydrofolate reductase
group_172	0.73	Collagen triple helix repeat (20 copies)	Hypothetical protein	Collagen triple helix repeat (20 copies)
group_173	0.58	Putative organic mercuric lyase merb2	Putative organic mercuric lyase merb2	NA
group_174	0.61	Hypothetical protein	Hypothetical protein	NA
group_175	0.55	US22 family protein; Family of unknown function (DUF6506)	US22 family protein	Family of unknown function (DUF6506)
group_176	0.41	Malarial early transcribed membrane protein (ETRAMP)	Hypothetical protein	Malarial early transcribed membrane protein (ETRAMP)
group_177	0.6	Enoyl-CoA hydratase/isomerase	NA	Enoyl-CoA hydratase/isomerase
group_178	0.57	Hypothetical protein	Hypothetical protein	NA
group_179	0.49	Hypothetical protein	Hypothetical protein	NA
group_180	0.61	Hypothetical protein	Hypothetical protein	NA
group_181	0.59	Hypothetical protein	Hypothetical protein	NA
group_182	0.46	Pigment-dispersing hormone (PDH)	Hypothetical protein	Pigment-dispersing hormone (PDH)
group_183	0.52	Hypothetical protein	Hypothetical protein	NA
group_184	0.63	Collagen triple helix repeat (20 copies)	Hypothetical protein	Collagen triple helix repeat (20 copies)
group_185	0.64	Collagen triple helix repeat (20 copies)	Hypothetical protein	Collagen triple helix repeat (20 copies)
group_186	0.32	Hypothetical protein	Hypothetical protein	NA
group_187	0.58	Hypothetical protein	Hypothetical protein	NA
group_188	0.59	Hypothetical protein	Hypothetical protein	NA
group_189	0.61	Hypothetical protein	NA	NA
group_190	0.42	Keratin, high sulfur B2 protein	NA	Keratin, high sulfur B2 protein
group_191	0.54	RGI orf47L-like protein; Tongue Cancer Chemotherapy Resistant Protein 1	RGI orf47L-like protein	Tongue Cancer Chemotherapy Resistant Protein 1

group_192	0.52	Neurofilament triplet H1-like protein; Malonate transporter MadL subunit	Neurofilament triplet H1-like protein	Malonate transporter MadL subunit
group_193	0.65	Putative myristylated membrane	Putative myristylated membrane	NA
group_194	0.53	Neurofilament triplet H1-like protein	Neurofilament triplet H1-like protein	NA
group_195	0.48	Hypothetical protein	NA	NA
group_196	0.58	US22 like	NA	US22 like
group_197	0.3	Hypothetical protein	NA	NA
group_198	0.44	CD20-like family	NA	CD20-like family
group_199	0.63	Putative surface protein; Domain of unknown function (DUF6457)	Surface protein	Domain of unknown function (DUF6457)
group_200	0.44	Hypothetical protein	NA	NA
group_201	0.46	Hypothetical protein	US22 family protein	NA
group_202	0.43	US22 like	Hypothetical protein	US22 like
group_203	0.65	Hypothetical protein	Hypothetical protein	NA
group_204	0.49	Hypothetical protein	Hypothetical protein	NA
group_205	0.45	Hypothetical protein	Hypothetical protein	NA
group_206	0.38	Histone H1-like nucleoprotein HC2	Hypothetical protein	Histone H1-like nucleoprotein HC2
group_207	0.64	Collagen triple helix repeat (20 copies)	Hypothetical protein	Collagen triple helix repeat (20 copies)
group_208	0.51	US22 like	Hypothetical protein	US22 like
group_209	0.25	Hypothetical protein	Hypothetical protein	NA
group_210	0.54	Virulence promoting factor	Hypothetical protein	Virulence promoting factor
group_211	0.64	Collagen triple helix repeat (20 copies)	NA	Collagen triple helix repeat (20 copies)
group_212	0.68	Collagen triple helix repeat (20 copies)	Hypothetical protein	Collagen triple helix repeat (20 copies)
group_213	0.27	Hypothetical protein	NA	NA
group_214	0.47	Hypothetical protein	Hypothetical protein	NA
group_215	0.51	Neurofilament triplet H1-like protein	Neurofilament triplet H1-like protein	NA
group_216	0.52	Neurofilament triplet H1-like protein	Neurofilament triplet H1-like protein	NA
group_217	0.6	D5 family NTPase/ATPase; D5 N terminal like	D5 family NTPase/ATPase	D5 N terminal like

Table S4. *FV3-like core genes with strong temporal signal based on* BactDating *permutation* P*-values and regression* R^2 *values yielded using individual gene trees. Functional annotations ascertained from* BLAST *searches are given where known.*

Roary Annotation	Functional Annotation	Permutation *P*-value	Regression R^2
I4L	Ribonucleotide reductase, barrel domain	0.031	5.3%
MCP	Major capsid protein	< 0.001	8.0%
RP01	RNA polymerase Rpb1, domain 2	0.0082	5.7%
group_16	Myristylated membrane protein	0.052	2.0%
group_18	US22 family protein	< 0.0001	25.8%
group_58	Hypothetical protein	< 0.001	13.3%
group_60	Putative surface protein	0.022	4.2%
group_69	Vertebrate interleukin-3 regulated transcription factor	0.016	4.5%
group_71	Hypothetical protein	0.056	3.0%
group_74	Myeloid cell leukemia protein; Apoptosis regulator proteins, Bcl-2 family	0.015	5.7%
group_76	Lipopolysaccharide-induced TNF-alpha factor (LIT AF)-like protein	0.048	2.9%
group_101	Hypothetical protein	0.012	5.9%
group_106	Lipopolysaccharide-induced TNF-alpha factor (LITAF)-like protein; zinc ribbon domain	0.054	2.4%
group_125	Putative DNA repair protein RAD2; XPG I-region	0.0089	6.6%
group_217	D5 family NTPase/ATPase; D5 N terminal like	0.0052	9.1%

References

Aaziz, R., & Tepfer, M. (1999). Recombination in RNA viruses and in virus-resistant transgenic plants. *Journal of General Virology*, *80*(6), 1339–1346. https://doi.org/10.1099/0022-1317-80-6-1339

Abbot, P., Aviles, A. E., Eller, L., & Durden, L. A. (2007). Mixed infections, cryptic diversity, and vector-borne pathogens: Evidence from Polygenis fleas and Bartonella species. *Applied and Environmental Microbiology*, *73*(19), 6045–6052. https://doi.org/10.1128/AEM.00228-07

Acman, M., van Dorp, L., Santini, J. M., & Balloux, F. (2020). Large-scale network analysis captures biological features of bacterial plasmids. *Nature Communications*, *11*(1), 1–11. https://doi.org/10.1038/s41467-020-16282-w

Ahne, W., Schlotfeldt, H. J., & Thomsen, I. (1989). Fish Viruses: Isolation of an Icosahedral Cytoplasmic Deoxyribovirus from Sheatfish (*Silurus glanis*). *Journal of Veterinary Medicine, Series B*, *36*(1–10), 333–336. https://doi.org/https://doi.org/10.1111/j.1439-0450.1989.tb00611.x

Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, *12*. https://doi.org/10.1186/1471-2105-12-246

Allen, T., Murray, K. A., Zambrana-Torrelio, C., Morse, S. S., Rondinini, C., Di Marco, M., Breit, N., Olival, K. J., & Daszak, P. (2017). Global hotspots and correlates of emerging zoonotic diseases. *Nature Communications*, *8*(1). https://doi.org/10.1038/s41467-017-00923-8

Anisimova, M., Nielsen, R., & Yang, Z. (2003). Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics*, *164*(3), 1229–1236. https://doi.org/10.1093/genetics/164.3.1229

Arenas, M., & Posada, D. (2010). Coalescent simulation of intracodon recombination. *Genetics*, *184*(2), 429–437. https://doi.org/10.1534/genetics.109.109736

Ariel, E., Holopainen, R., Olesen, N. J., & Tapiovaara, H. (2010). Comparative study of ranavirus isolates from cod (*Gadus morhua*) and turbot (*Psetta maxima*) with reference to other ranaviruses. *Archives of Virology*, *155*(8), 1261–1271. https://doi.org/10.1007/s00705-010-0715-z

Ariel, E., & Owens, L. (1997). Epizootic mortalities in tilapia *Oreochromis mossambicus*. *Diseases of Aquatic Organisms*, *29*, 1–6.

Ariel, E., Steckler, N. K., Subramaniam, K., Olesen, N. J., & Waltzek, T. B. (2016). Genomic sequencing of ranaviruses isolated from turbot (*Scophthalmus maximus*) and Atlantic Cod (*Gadus morhua*). *Genome Announcements*, *4*(6), 5–6. https://doi.org/10.1128/genomeA.01393-16

Auliya, M., García-Moreno, J., Schmidt, B. R., Schmeller, D. S., Hoogmoed, M. S., Fisher, M. C., Pasmans, F., Henle, K., Bickford, D., & Martel, A. (2016). The global amphibian trade flows through Europe: the need for enforcing and improving legislation. *Biodiversity and Conservation*, *25*(13), 2581–2595. https://doi.org/10.1007/s10531-016-1193-8

Ballard, D. R., Davis, A. J., Fuller, R. B., Garner, A. R., Mileham, A. D., Serna, J. D., Brue, D. E., Harding, C. M., Dodgen, C. D., Rosario, S. E., & Duffus, A. L. J. (2020). An examination of the Iridovirus core genes for reconstructing *Ranavirus* phylogenies. *Facets*, *5*(1), 523–533. https://doi.org/10.1139/facets-2020-0009

Balseiro, A., Dalton, K. P., Cerro, A., Marquez, I., Cunningham, A. A., Parra, F., Prieto, J. M., & Casais, R. (2009). Pathology, isolation and molecular characterisation of a ranavirus from the common midwife toad *Alytes obstetricans* on the Iberian Peninsula. *Disease of Aquatic Organisms*, *84*, 95–104. https://doi.org/10.3354/dao02032

Banks, B. (2000). British Bullfrogs? *British Wildlife*, *June 2000*.

Biek, R., O'Hare, A., Wright, D., Mallon, T., McCormick, C., Orton, R. J., McDowell, S., Trewby, H., Skuce, R. A., & Kao, R. R. (2012). Whole Genome Sequencing Reveals Local Transmission Patterns of *Mycobacterium bovis* in Sympatric Cattle and Badger Populations. *PLoS Pathogens*, *8*(11). https://doi.org/10.1371/journal.ppat.1003008

Bollinger, T. K., Mao, J., Schock, D., Brigham, R. M., & Chinchar, V. G. (1999). Pathology, Isolation, and Preliminary Molecular Characterization of a Novel Iridovirus From Tiger Salamanders in Saskatchewan. *Journal of Wildlife Diseases*, *35*(3), 413–429. https://doi.org/10.1007/s10100-009-0130-2

Borremans, B., Faust, C., Manlove, K. R., Sokolow, S. H., & Lloyd-Smith, J. O. (2019). Cross-species pathogen spillover across ecosystem boundaries: Mechanisms and theory. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1782). https://doi.org/10.1098/rstb.2018.0344

Borzym, E., Stachnik, M., Reichert, M., Rzeżutka, A., Agnieszka, J., B., W. T., & Kuttichantran, S. (2020). Genome Sequence of a *Ranavirus* Isolated from a Red-Eared Slider (*Trachemys scripta elegans*) in Poland. *Microbiology Resource Announcements*, *9*(47), 19–21.

Bouckaert, R. R., & Drummond, A. J. (2017). bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC Evolutionary Biology*, *17*(1), 1–11. https://doi.org/10.1186/s12862-017-0890-6

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., Matschiner, M., Mendes, F. K., Müller, N. F., Ogilvie, H. A., du Plessis, L., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., … Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, *15*(4). https://doi.org/10.1371/journal.pcbi.1006650

Bowden, R., Sakaoka, H., Donnelly, P., & Ward, R. (2004). High recombination rate in herpes simplex virus type 1 natural populations suggests significant co-infection. *Infection, Genetics and Evolution*, *4*(2), 115–123. https://doi.org/10.1016/j.meegid.2004.01.009

Box, E. K., Cleveland, C. A., Subramaniam, K., B, T., & Yabsley, M. J. (2021). Molecular Confirmation of Ranavirus Infection in Amphibians From Chad, Africa. *Frontiers in Veterinary Science*, *8*(September), 1–8. https://doi.org/10.3389/fvets.2021.733939

Braunwald, J., Tripier, F., & Kirn, A. (1979). Comparison of the properties of enveloped and naked frog virus 3 (FV 3) particles. *Journal of General Virology*, *45*(3), 673–682. https://doi.org/10.1099/0022-1317-45-3-673

Brenes, R., Gray, M. J., Waltzek, T. B., Wilkes, R. P., & Miller, D. L. (2014). Transmission of ranavirus between ectothermic vertebrate hosts. *PLoS ONE*, *9*(3), 1–6. https://doi.org/10.1371/journal.pone.0092476

Broder, A. Z. (1997). On the resemblance and containment of documents. *Proceedings of the International Conference on Compression and Complexity of Sequences*, 21–29. https://doi.org/10.1109/sequen.1997.666900

Brunner, J. L., Storfer, A., Gray, M. J., & Hoverman, J. T. (2015). Ranavirus Ecology and Evolution: from Epidemiology to Extinction. In M. J. Gray & V. G. Chinchar (Eds.), *Ranaviruses* (pp. 71–104). Springer International Publishing.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 1–9. https://doi.org/10.1186/1471-2105-10-421

Campbell, L. J., Garner, T. W. J., Hopkins, K., Griffiths, A. G. F., & Harrison, X. A. (2019). Outbreaks of an Emerging Viral Disease Covary With Differences in the Composition of the Skin Microbiome of a Wild United Kingdom Amphibian. *Frontiers in Microbiology*, *10*(June), 1–12. https://doi.org/10.3389/fmicb.2019.01245

Campbell, L. J., Hammond, S. A., Price, S. J., Sharma, M. D., Garner, T. W. J., Birol, I., Helbing, C. C., Wilfert, L., & Griffiths, A. G. F. (2018). A novel approach to wildlife transcriptomics provides evidence of disease-mediated differential expression and changes to the microbiome of amphibian populations. *Molecular Ecology*, *27*(6), 1413–1427. https://doi.org/10.1111/mec.14528

Campbell, L. J., Pawlik, A. H., & Harrison, X. A. (2020). Amphibian ranaviruses in Europe: Important directions for future research. *Facets*, *5*(1), 598–614. https://doi.org/10.1139/FACETS-2020-0007

Candido, M., Tavares, L. S., Alencar, A. L. F., Ferreira, C. M., Queiroz, S. R. de A., Fernandes, A. M., & Sousa, R. L. M. de. (2019). Genome analysis of *Ranavirus frog virus 3* isolated from American Bullfrog (*Lithobates catesbeianus*) in South America. *Scientific Reports*, *9*(1), 1–7. https://doi.org/10.1038/s41598-019-53626-z

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Carpenter, A. I., Andreone, F., Moore, R. D., & Griffiths, R. A. (2014). A review of the international trade in amphibians: The types, levels and dynamics of trade in CITES-listed species. *Oryx*, *48*(4), 565–574. https://doi.org/10.1017/S0030605312001627

Carstairs, S. J., Kyle, C. J., & Vilaça, S. T. (2020). High prevalence of subclinical frog virus 3 infection in freshwater turtles of Ontario, Canada. *Virology*, *543*(November 2019), 76–83. https://doi.org/10.1016/j.virol.2020.01.016

Casais, R., Larrinaga, A. R., Dalton, K. P., Domínguez Lapido, P., Márquez, I., Bécares, E., Carter, E. D., Gray, M. J., Miller, D. L., & Balseiro, A. (2019). Water sports could contribute to the translocation of ranaviruses. *Scientific*

Reports, *9*(1), 1–6. https://doi.org/10.1038/s41598-019-39674-5

Chen, G., Ward, B. M., Yu, K. H., Chinchar, V. G., & Robert, J. (2011). Improved Knockout Methodology Reveals That Frog Virus 3 Mutants Lacking either the *18K* Immediate-Early Gene or the Truncated *vIF-2* Gene Are Defective for Replication and Growth *In Vivo*. *Journal of Virology*, *85*(21), 11131–11138. https://doi.org/10.1128/JVI.05589-11

Chen, Z., Gui, J., Gao, X., Pei, C., Hong, Y., & Zhang, Q. (2013). Genome architecture changes and major gene variations of *Andrias davidianus* ranavirus (ADRV). *Veterinary Research*, *44*(1), 1–13. https://doi.org/10.1186/1297-9716-44-101

Chen, Z., Zheng, J., & Jiang, Y. (1999). Short communication A new iridovirus isolated from soft-shelled turtle. *Virus Research*, *63*, 147–151.

Chinchar, V. G. (2002). Ranaviruses (family *Iridoviridae*): emerging cold-blooded killers. *Archives of Virology*, *147*, 447–470.

Chinchar, V. G., & Granoff, A. (1986). Temperature-sensitive mutants of frog virus 3: biochemical and genetic characterization. *Journal of Virology*, *58*(1), 192–202. https://doi.org/10.1128/jvi.58.1.192-202.1986

Chinchar, V. G., Hick, P., Ince, I., Jancovich, J., Marschang, R., Qin, Q., Waltzek, T., Zhang, Q., William, T., & Whittington, R. (2017a). ICTV Virus taxonomy Profile: *Iridoviridae*. *Journal of General Virology*, *98*, 890–891.

Chinchar, V. G., Hyatt, A., Miyazaki, T., & Williams, T. (2009). Family *Iridoviridae*: Poor Viral Relations No Longer. *Current Topics in Microbiology and Immunology*, *328*, 123–170.

Chinchar, V. G., Waltzek, T. B., & Subramaniam, K. (2017b). Ranaviruses and other members of the family *Iridoviridae*: Their place in the virosphere. *Virology*, *511*(June), 259–271. https://doi.org/10.1016/j.virol.2017.06.007

Cho, A. (2012). Constructing Phylogenetic Trees Using Maximum Likelihood. In *Scripps Senior Theses* (Vol. 46).

Chua, F. H. C., Ng, M. L., Ng, K. L., Loo, J. J., & Wee, J. Y. (1994). Investigation of outbreaks of a novel disease, "Sleepy Grouper Disease", affecting the brown-spotted grouper, *Epinephelus tauvina* Forskal. *Journal of Fish Diseases*, *17*, 417–427.

Chung, M., Munro, J. B., Tettelin, H., & Dunning Hotopp, J. C. (2018). Using Core Genome Alignments To Assign Bacterial Species. *MSystems*, *3*(6), 1–21. https://doi.org/10.1128/msystems.00236-18

Claytor, S. C., Subramaniam, K., Landrau-Giovannetti, N., Chinchar, V. G., Gray, M. J., Miller, D. L., Mavian, C., Salemi, M., Wisely, S., & Waltzek, T. B. (2017). Ranavirus phylogenomics: Signatures of recombination and inversions among bullfrog ranaculture isolates. *Virology*, *511*(May), 330–343. https://doi.org/10.1016/j.virol.2017.07.028

Cohen, J. M., Civitello, D. J., Venesky, M. D., McMahon, A., & Rohr, J. R. (2019). An interaction between climate change and infectious disease drove widespread amphibian declines. *Global Change Biology*, *25*(3), 927–937. https://doi.org/10.1111/gcb.14489

Combelas, N., Holmblat, B., Joffret, M. L., Colbère-Garapin, F., & Delpeyroux, F. (2011). Recombination between poliovirus and coxsackie A viruses of species C: A model of viral genetic plasticity and emergence. *Viruses*, *3*(8), 1460–1484. https://doi.org/10.3390/v3081460

Crispell, J., Balaz, D., & Gordon, S. V. (2019). Homoplasyfinder: A simple tool to identify homoplasies on a phylogeny. *Microbial Genomics*, *5*(1). https://doi.org/10.1099/mgen.0.000245

Croizier, G., Croizier, L., Argaud, O., & Poudevigne, D. (1994). Extension of Autographa californica nuclear polyhedrosis virus host range by interspecific replacement of a short DNA sequence in the p143 helicase gene. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(1), 48–52. https://doi.org/10.1073/pnas.91.1.48

Croucher, N. J., Page, A. J., Connor, T. R., Delaney, A. J., Keane, J. A., Bentley, S. D., Parkhill, J., & Harris, S. R. (2015). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research*, *43*(3), e15. https://doi.org/10.1093/nar/gku1196

Cunningham, A. A. A., Langton, T. E. S., Bennett, P. M., Lewin, J. F., Drury, S. E. N., Gough, R. E., & Macgregor, S. K. (1996). Pathological and Microbiological Findings from Incidents of Unusual Mortality of the Common Frog (*Rana temporaria*). *Philisophical Trannsactions of the Royal Society B*, *351*(1347), 1539–1557.

Cunningham, A. A., Daszak, P., & Rodriguez, J. (2003). Pathogen pollution: defining a parasitological threat to biodiversity conservation. *The Journal of Parasitology*, *89*, S78–S83.

Cunningham, A. A., Hyatt, A. D., Russell, P., & Bennett, P. M. (2007). Emerging epidemic diseases of frogs in Britain are dependent on the source of ranavirus agent and the route of exposure. *Epidemiology and Infection*, *135*(7),

1200–1212. https://doi.org/10.1017/S0950268806007679

Daszak, P., Berger, L., Cunningham, A. A., Hyatt, A. D., Green, D. E., & Speare, R. (1999). Emerging Infectious Diseases and Amphibian Population Declines. *Emerging Infectious Diseases, 5*(6), 735–748.

Daszak, P., Cunningham, A. A., & Hyatt, A. D. (2000). Emerging infectious diseases of wildlife - Threats to biodiversity and human health. *Science*, *287*(5452), 443–449. https://doi.org/10.1126/science.287.5452.443

Daszak, P., Cunningham, A. A., & Hyatt, A. D. (2001). Anthropogenic environmental change and the emergence of infectious diseases in wildlife. *Acta Tropica*, *78*(2), 103–116. https://doi.org/10.1016/S0001-706X(00)00179-0

Daszak, P., Cunningham, A. A., & Hyatt, A. D. (2003). Infectious disease and amphibian population declines. *Diversity and Distributions*, *9*(2), 141–150. https://doi.org/10.1046/j.1472-4642.2003.00016.x

de Carvalho, J. A., Hagen, F., Fisher, M. C., de Camargo, Z. P., & Rodrigues, A. M. (2020). Genome-wide mapping using new AFLP markers to explore intraspecific variation among pathogenic *Sporothrix* species. *PLoS Neglected Tropical Diseases*, *14*(7), 1–26. https://doi.org/10.1371/journal.pntd.0008330

de Voe, R., Geissler, K., Elmore, S., Rotstein, D., Lewbart, G., & Guy, J. (2004). Ranavirus-associated morbidity and mortality in a group of captive eastern box turtles (*Terrapene carolina carolina*). *Journal of Zoo and Wildlife Medicine*, *35*(4), 534–543. https://doi.org/10.1638/03-037

Didelot, X., Croucher, N. J., Bentley, S. D., Harris, S. R., & Wilson, D. J. (2018). Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Research*, *46*(22). https://doi.org/10.1093/nar/gky783

Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R., & Rodrigo, A. G. (2003). Measurably evolving populations. *Trends in Ecology and Evolution*, *18*(9), 481–488. https://doi.org/10.1016/S0169-5347(03)00216-7

Drummond, D. A., Silberg, J. J., Meyer, M. M., Wilke, C. O., & Arnold, F. H. (2005). On the conservative nature of intragenic recombination. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(15), 5380–5385. https://doi.org/10.1073/pnas.0500729102

Duffus, A. L. J., Garner, T. W. J., Davis, A. R., Dean, Ashley, W., & Nichols, R. (2017). Phylogentic Analysis of 24 *Ranavirus* Isolates from English Amphibians using 2 Partial Loci. *Journal of Emerging Diseases and Virology*, *3*(2), 1–7. https://doi.org/10.16966/2473-1846.131

Duffus, A. L. J., Waltzek, T. B., Stöhr, A. C., Allender, M. C., Gotesman, M., Whittington, R. J., Hick, P., Hines, M. K., & Marschang, R. E. (2015). Distribution and Host Range of Ranaviruses. In M. J. Gray & V. G. Chinchar (Eds.), *Ranaviruses* (pp. 9–58). Springer International Publishing.

Duffus, A. L., Nichols, R. A., & Garner, T. W. J. (2014). Experimental evidence in support of single host maintenance of a multihost pathogen. *Ecosphere*, *5*(November), 1–11. https://doi.org/10.1890/ES14-00074.1

Eaton, H. E., Metcalf, J., Penny, E., Tcherepanov, V., Upton, C., & Brunetti, C. R. (2007). Comparative genomic analysis of the family *Iridoviridae*: Re-annotating and defining the core set of iridovirus genes. *Virology Journal*, *4*, 1–17. https://doi.org/10.1186/1743-422X-4-11

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Elde, N. C., Child, S. J., Eickbush, M. T., Kitzman, J. O., Rogers, K. S., Shendure, J., Geballe, A. P., & Malik, H. S. (2012). Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses. *Cell*, *150*(4), 831–841. https://doi.org/10.1016/j.cell.2012.05.049

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, *30*(7), 1575–1584. https://doi.org/10.1093/nar/30.7.1575

Epstein, B., & Storfer, A. (2016). Comparative Genomics of an Emerging Amphibian Virus. *G3*, *6*(1), 15–27. https://doi.org/10.1534/g3.115.023762

Evans, E., Klemperer, N., Ghosh, R., & Traktman, P. (1995). The vaccinia virus D5 protein, which is required for DNA replication, is a nucleic acid-independent nucleoside triphosphatase. *Journal of Virology*, *69*(9), 5353–5361. https://doi.org/10.1128/jvi.69.9.5353-5361.1995

Faust, C. L., McCallum, H. I., Bloomfield, L. S. P., Gottdenker, N. L., Gillespie, T. R., Torney, C. J., Dobson, A. P., & Plowright, R. K. (2018). Pathogen spillover during land conversion. *Ecology Letters*, *21*(4), 471–483. https://doi.org/10.1111/ele.12904

Ferreira, C. M., Subramaniam, K., de Sousa, R. L. M., Tavares, L. S., Corrêa, T. C., & Waltzek, T. B. (2021). Genomic sequencing of a frog virus 3 strain from cultured American bullfrogs (*Lithobates catesbeianus*) in Brazil. *Archives of Virology*, *166*(7), 1961–1964. https://doi.org/10.1007/s00705-021-05094-y

Ficetola, G. F., Bonin, A., & Miaud, C. (2008). Population genetics reveals origin and number of founders in a biological invasion. *Molecular Ecology*, *17*(3), 773–782. https://doi.org/10.1111/j.1365-294X.2007.03622.x

Ficetola, G. F., Coïc, C., Detaint, M., Berroneau, M., Lorvelec, O., & Miaud, C. (2007). Pattern of distribution of the American bullfrog *Rana catesbeiana* in Europe. *Biological Invasions*, *9*(7), 767–772. https://doi.org/10.1007/s10530-006-9080-y

Fijan, N., Matasin, Z., Petrinec, Z., Valpotic, I., & Zwillenberg, L. O. (1991). Isolation of an iridovirus-like agent from the green frog (*Rana esculenta* L.). *Vet Arch*, *61*, 151–158.

Filée, J. (2009). Lateral gene transfer, lineage-specific gene expansion and the evolution of Nucleo Cytoplasmic Large DNA viruses. *Journal of Invertebrate Pathology*, *101*(3), 169–171. https://doi.org/10.1016/j.jip.2009.03.010

Filée, J. (2013). Route of NCLDV evolution: The genomic accordion. *Current Opinion in Virology*, *3*(5), 595–599. https://doi.org/10.1016/j.coviro.2013.07.003

Fisher, M. C., & Murray, K. A. (2021). Emerging infections and the integrative environment-health sciences: the road ahead. *Nature Reviews Microbiology*, *19*(3), 133–135. https://doi.org/10.1038/s41579-021-00510-1

Fitch, W. M. (1971). Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Biology*, *20*(4), 406–416. https://doi.org/10.1093/sysbio/20.4.406

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Galli, A., & Bukh, J. (2014). Comparative analysis of the molecular mechanisms of recombination in hepatitis C virus. *Trends in Microbiology*, *22*(6), 354–364. https://doi.org/10.1016/j.tim.2014.02.005

Galli, A., Kearney, M., Nikolaitchik, O. A., Yu, S., Chin, M. P. S., Maldarelli, F., Coffin, J. M., Pathak, V. K., & Hu, W.-S. (2010). Patterns of Human Immunodeficiency Virus Type 1 Recombination Ex Vivo Provide Evidence for Coadaptation of Distant Sites, Resulting in Purifying Selection for Intersubtype Recombinants during Replication. *Journal of Virology*, *84*(15), 7651–7661. https://doi.org/10.1128/jvi.00276-10

Gendrault, J. L., Steffan, A. M., Bingen, A., & Kirn, A. (1981). Penetration and uncoating of frog virus 3 (FV3) in cultured rat Kupffer cells. *Virology*, *112*(2),

375–384. https://doi.org/10.1016/0042-6822(81)90284-1

Gibbs, A. J. (2013). Viral taxonomy needs a spring clean; Its exploration era is over. *Virology Journal*, *10*(254), 1–4. https://doi.org/10.1186/1743-422X-10-254

González-del-Pliego, P., Freckleton, R. P., Edwards, D. P., Koo, M. S., Scheffers, B. R., Pyron, R. A., & Jetz, W. (2019). Phylogenetic and Trait-Based Prediction of Extinction Risk for Data-Deficient Amphibians. *Current Biology*, *29*(9), 1557-1563.e3. https://doi.org/10.1016/j.cub.2019.04.005

Goorha, R., & Murti, G. (1982). The genome of frog virus 3, an animal DNA virus, is circularly permuted and terminally redundant. *Proceedings of the National Academy of Sciences USA*, *79*(January), 248–252.

Granoff, A., Came, P. E., & Breeze, D. C. (1966). Viruses and renal carcinoma of *Rana pipiens*. I. The isolation and properties of virus from normal and tumor tissue. *Virology*, *29*(1), 133–148. https://doi.org/10.1016/0042-6822(66)90203-0

Granoff, A., Came, P. E., & Rafferty, K. A. (1965). The isolation and properties of viruses from *Rana pipiens*: their possible relationship to the renal adenocarcinoma of the leopard frog. *Annals New York Academy of Sciences*, *126*, 237–255.

Gratwicke, B., Evans, M. J., Jenkins, P. T., Kusrini, M. D., Moore, R. D., Sevin, J., & Wildt, D. E. (2010). Is the international frog legs trade a potential vector for deadly amphibian pathogens? *Frontiers in Ecology and the Environment*, *8*(8), 438–442. https://doi.org/10.1890/090111

Gray, M. J., & Chinchar, G. V. (2015). Ranaviruses. In M. J. Gray & G. V Chinchar (Eds.), *Ranaviruses: Lethal Pathogens of Ectothermic Vertebrates*. Springer International Publishing. https://doi.org/10.1007/978-3-319-13755-1

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. A., & Holmes, Edward, C. (2004). Unifying the Epidemiological and Evolutionary Dynamics Patogens. *Science*, *303*(2004), 327–332. https://doi.org/10.1126/science.1090727

Grizzle, J. M., Altinok, I., Fraser, W. A., & Francis-Floyd, R. (2002). First isolation of largemouth bass virus. *Diseases of Aquatic Organisms*, *50*(3), 233–235. https://doi.org/10.3354/dao050233

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*(18), 2847–2849. https://doi.org/10.1093/bioinformatics/btw313

Hanisch, S. L., Riley, S. J., & Nelson, M. P. (2012). Promoting wildlife health or

fighting wildlife disease: Insights from history, philosophy, and science. *Wildlife Society Bulletin*, *36*(3), 477–482. https://doi.org/10.1002/wsb.163

He, C.-Q., Ding, N.-Z., He, M., Li, S.-N., Wang, X.-M., He, H.-B., Liu, X.-F., & Guo, H.-S. (2010). Intragenic Recombination as a Mechanism of Genetic Diversity in Bluetongue Virus. *Journal of Virology*, *84*(21), 11487–11495. https://doi.org/10.1128/jvi.00889-10

He, J. G., Lü, L., Deng, M., He, H. H., Weng, S. P., Wang, X. H., Zhou, S. Y., Long, Q. X., Wang, X. Z., & Chan, S. M. (2002). Sequence analysis of the complete genome of an Iridovirus isolated from the tiger frog. *Virology*, *292*(2), 185–197. https://doi.org/10.1006/viro.2001.1245

Hick, P. M., Subramaniam, K., Thompson, P. M., Waltzek, T. B., Becker, J. A., & Whittington, R. J. (2017). Molecular epidemiology of Epizootic haematopoietic necrosis virus (EHNV). *Virology*, *511*(August), 320–329. https://doi.org/10.1016/j.virol.2017.07.029

Hick, P. M., Subramaniam, K., Thompson, P., Whittington, R. J., & Waltzek, T. B. (2016). Complete genome sequence of a Bohle iridovirus isolate from ornate burrowing frogs (*Limnodynastes ornatus*) in Australia. *Genome Announcements*, *4*(4), 13–14. https://doi.org/10.1128/genomeA.00632-16

Hill, W. G., & Robertson, A. (1968). Linkage Disequilibrium in Finite Populations. *Theoretical and Applied Genetics*, *38*, 226–231. https://doi.org/10.1080/03071848408522227

Hoang, D. T., Vinh, L. S., Flouri, T., Stamatakis, A., Von Haeseler, A., & Minh, B. Q. (2018). MPBoot: Fast phylogenetic maximum parsimony tree inference and bootstrap approximation. *BMC Evolutionary Biology*, *18*(1), 1–11. https://doi.org/10.1186/s12862-018-1131-3

Holdich, D. M., Reynolds, J. D., Souty-Grosset, C., & Sibley, P. J. (2009). A review of the ever increasing threat to European crayfish from non-indigenous crayfish species. *Knowledge and Management of Aquatic Ecosystems*, *2009*(394–395), 394–395. https://doi.org/10.1051/kmae/2009025

Holopainen, R., Ohlemeyer, S., Schütze, H., Bergmann, S. M., & Tapiovaara, H. (2009). Ranavirus phylogeny and differentiation based on major capsid protein, DNA polymerase and neurofilament triplet H1-like protein genes. *Disease of Aquatic Organisms*, *85*, 81–91. https://doi.org/10.3354/dao02074

Holopainen, R., Subramaniam, K., Steckler, N. K., Claytor, S. C., Ariel, E., & Waltzek, T. B. (2016). Genome sequence of a ranavirus isolated from pike-

perch *Sander lucioperca*. *Genome Announcements*, *4*(6), 15–16. https://doi.org/10.1128/genomeA.01295-16

Horton, R., Beaglehole, R., Bonita, R., Raeburn, J., McKee, M., & Wall, S. (2014). From public to planetary health: A manifesto. *The Lancet*, *383*(9920), 847. https://doi.org/10.1016/S0140-6736(14)60409-8

Hoverman, J. T., Gray, M. J., Haislip, N. A., & Miller, D. L. (2011). Phylogeny, Life History, and Ecology Contribute to Differences in Amphibian Susceptibility to Ranaviruses. *EcoHealth*, *8*, 301–319. https://doi.org/10.1007/s10393-011-0717-7

Hoverman, J. T., Gray, M. J., & Miller, D. L. (2010). Anuran susceptibilities to ranaviruses: role of species identity, exposure route, and a novel virus isolate. *Diseases of Aquatic Organisms*, *89*, 97–107. https://doi.org/10.3354/dao02200

Huang, Y., Huang, X., Liu, H., Gong, J., Ouyang, Z., Cui, H., Cao, J., Zhao, Y., Wang, X., Jiang, Y., & Qin, Q. (2009). Complete sequence determination of a novel reptile iridovirus isolated from soft-shelled turtle and evolutionary analysis of *Iridoviridae*. *BMC Genomics*, *10*. https://doi.org/10.1186/1471-2164-10-224

Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, *23*(2), 254–267. https://doi.org/10.1093/molbev/msj030

Hutin, S., Ling, W. L., Round, A., Effantin, G., Reich, S., Iseni, F., Tarbouriech, N., Schoehn, G., & Burmeister, W. P. (2016). Domain Organization of Vaccinia Virus Helicase-Primase D5. *Journal of Virology*, *90*(9), 4604–4613. https://doi.org/10.1128/jvi.00044-16

Hyatt, A. D., Gould, A. R., Zupanovic, Z., Cunningham, A. A., Hengstberger, S., Whittingdon, R. J., Kattenbelt, J., & Coupar, B. E. H. (2000). Comparative studies of piscine and amphibian iridoviruses. *Archives of Virology*, *145*, 301–331.

Inouye, K., Yamano, K., Maeno, Y., Nakajima, K., Matsuoka, M., Wada, Y., & Sorimachi, M. (1992). Iridovirus Infection of Cultured Red Sea Bream, *Pagrus major*. *Fish Pathology*, *27*(1), 19–27. https://doi.org/10.3147/jsfp.27.19

International Committee on Taxonomy of Viruses. (2012). Virus Taxonomy: Classification and Nomenclature of Viruses. In A. M. Q. King, M. J. Adams, E. B. Carstens, & E. J. Lefkowitz (Eds.), *Virus Taxonomy*. Elsevier Inc.

https://doi.org/10.1016/B978-0-12-384684-6.00057-4

Iyer, L. M., Balaji, S., Koonin, E. V., & Aravind, L. (2006). Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Research*, *117*(1), 156–184. https://doi.org/10.1016/j.virusres.2006.01.009

Jancovich, J. K., Davidson, E. W., Morado, J. F., Jacobs, B. L., & Collins, J. P. (1997). Isolation of a lethal virus from the endangered tiger salamander *Ambystoma tigrinum stebbinsi*. *Disease of Aquatic Organisms*, *31*, 161–167.

Jancovich, J. K., Davidson, E. W., Seiler, A., Jacobs, B. L., & Collins, J. P. (2001). Transmission of the *Ambystoma tigrinum* virus to alternative hosts. *Diseases of Aquatic Organisms*, *46*(3), 159–163. https://doi.org/10.3354/dao046159

Jancovich, J. K., Mao, J., Chinchar, V. G., Wyatt, C., Case, S. T., Kumar, S., Valente, G., Subramanian, S., Davidson, E. W., Collins, J. P., & Jacobs, B. L. (2003). Genomic sequence of a ranavirus (family *Iridoviridae*) associated with salamander mortalities in North America. *Virology*, *316*(1), 90–103. https://doi.org/10.1016/j.virol.2003.08.001

Jancovich, J. K., Qin, Q., Zhang, Q. Y., & Chinchar, V. G. (2015b). *Ranavirus* Replication: Molecular, Cellular, and Immunological Events. In M. J. Gray & V. G. Chinchar (Eds.), *Ranaviruses* (pp. 71–104). Springer International Publishing.

Jancovich, J. K., Steckler, N. K., & Waltzek, T. B. (2015a). *Ranavirus* taxonomy and phylogeny. In M. J. Gray & V. G. Chinchar (Eds.), *Ranaviruses* (pp. 59–70). Springer International Publishing.

Jensen, N. J., Bloch, B., & Larsen, J. L. (1979). The ulcus-syndrome in cod (*Gadus morhua*). III. A preliminary virological report. *Nordisk Veterinaermedicin*, *31*(10), 436–442. http://europepmc.org/abstract/MED/392451

Jeudy, S., Rigou, S., Alempic, J. M., Claverie, J. M., Abergel, C., & Legendre, M. (2020). The DNA methylation landscape of giant viruses. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-16414-2

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405. https://doi.org/10.1093/bioinformatics/btn129

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, *11*(94). http://www.biomedcentral.com/1471-2156/11/94

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., & Daszak, P. (2008). Global trends in emerging infectious diseases. *Nature*, *451*, 990–994. https://doi.org/10.1038/nature06536

Kahle, D., & Wickham, H. (2013). ggmap: Spatial Visualization with ggplot2. *The R Journal*, *5*, 515–543. https://doi.org/10.1051/ps/2015001

Kamath, P. L., Foster, J. T., Drees, K. P., Luikart, G., Quance, C., Anderson, N. J., Clarke, P. R., Cole, E. K., Drew, M. L., Edwards, W. H., Rhyan, J. C., Treanor, J. J., Wallen, R. L., White, P. J., Robbe-Austerman, S., & Cross, P. C. (2016). Genomics reveals historic and contemporary transmission dynamics of a bacterial disease among wildlife and livestock. *Nature Communications*, *7*(May). https://doi.org/10.1038/ncomms11448

Kamita, S. G., Maeda, S., & Hammock, B. D. (2003). High-Frequency Homologous Recombination between Baculoviruses Involves DNA Replication. *Journal of Virology*, *77*(24), 13053–13061. https://doi.org/10.1128/JVI.77.24.13053

Kanchanakhan, S. (1998). An ulcerative disease of the cultured tiger frog, *Rana tigrina*. *Thailand: Virological Examination. AAHRI Newsl*, *7*, 1–2.

Kass, E. M., & Jasin, M. (2010). Collaboration and competition between DNA double-strand break repair pathways. *FEBS Letters*, *584*(17), 3703–3708. https://doi.org/10.1016/j.febslet.2010.07.057

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kik, M., Martel, A., Sluijs, A. S. Der, Pasmans, F., Wohlsein, P., Gröne, A., & Rijks, J. M. (2011). Ranavirus-associated mass mortality in wild amphibians, The Netherlands, 2010: A first report. *The Veterinary Journal*, *190*(2), 284–286. https://doi.org/10.1016/j.tvjl.2011.08.031

Kilcher, S., Schmidt, F. I., Schneider, C., Kopf, M., Helenius, A., & Mercer, J. (2014). SiRNA screen of early poxvirus genes identifies the AAA+ ATPase D5 as the virus genome-uncoating factor. *Cell Host and Microbe*, *15*(1), 103–112. https://doi.org/10.1016/j.chom.2013.12.008

Kim, H. S., Woo, S. D., Kim, W. J., Choi, J. Y., & Kang, S. K. (2000). High-level expression of a foreign gene by a recombinant baculovirus with an expanded host range. *Cytotechnology*, *32*(2), 87–92. https://doi.org/10.1023/A:1008166310368

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., & Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, *155*(1), 27. https://doi.org/10.1016/j.cell.2013.09.006

Kock, R. A., Wambua, J. M., Mwanzia, J., Wamwayi, H., Ndungu, E. K., Barrett, T., Kock, N. D., & Rossiter, P. B. (1999). Rinderpest epidemic in wild ruminants in Kenya 1993-97. *Veterinary Record*, *145*(10), 275–283. https://doi.org/10.1136/vr.145.10.275

Kohl, T. A., Diel, R., Harmsen, D., Rothgänger, J., Meywald Walter, K., Merker, M., Weniger, T., & Niemann, S. (2014). Whole-genome-based *Mycobacterium tuberculosis* surveillance: A standardized, portable, and expandable approach. *Journal of Clinical Microbiology*, *52*(7), 2479–2486. https://doi.org/10.1128/JCM.00567-14

Kondo, A., & Maeda, S. (1991). Host range expansion by recombination of the baculoviruses *Bombyx mori* nuclear polyhedrosis virus and *Autographa californica* nuclear polyhedrosis virus. *Journal of Virology*, *65*(7), 3625–3632. https://doi.org/10.1128/jvi.65.7.3625-3632.1991

Koonin, E. V., Zerbini, F. M., Dolja, V. V., Kuhng, J. H., Krupovic, M., Varsani, A., Wolf, Y. I., & Yutin, N. (2020). Global Organization and Proposed Megataxonomy of the Virus World. *Microbiology and Molecular Biology Reviews*, *84*(2), 1–33.

Koonin, E. V, & Yutin, N. (2012). Nucleo-cytoplasmic Large DNA Viruses (NCLDV) of Eukaryotes. *ELS*, 1–18. https://doi.org/10.1002/9780470015902.a0023268

Kuhner, M. K., & Yamato, J. (2015). Practical performance of tree comparison metrics. *Systematic Biology*, *64*(2), 205–214. https://doi.org/10.1093/sysbio/syu085

Kwon, S., Park, J., Choi, W. J., Koo, K. S., Lee, J. G., & Park, D. (2017). First case of ranavirus-associated mass mortality in a natural population of the Huanren frog (*Rana huanrenensis*) tadpoles in South Korea. *Animal Cells and Systems*, *21*(5), 358–364. https://doi.org/10.1080/19768354.2017.1376706

Lachenbruch, P. A., & Goldstein, M. (1979). Discriminant Analysis. *Biometrics*, *35*, 69–85.

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

Langdon, J. S., & Humphrey, J. D. (1987). Epizootic haematopoietic necrosis, a new viral disease in redfin perch, *Perca fluviatilis* L., in Australia. *Journal of Fish Diseases*, *10*, 289–297.

Langdon, J. S., Humphrey, J. D., Williams, L. M., Hyatt, A. D., & Westbury, H. A. (1986). First virus isolation from Australian fish: an iridovirus-like pathogen from redfin perch, *Perca fluviatilis* L. *Journal of Fish Diseases*, *9*, 263–268.

Langton, T. E. S., Atkins, W., & Herbert, C. (2011). On the distribution, ecology and management of non-native reptiles and amphibians in the London Area. Part 1. Distribution and predator/prey impacts. *The London Naturalist*, *90*, 83–156.

Lei, X. Y., Ou, T., Zhu, R. L., & Zhang, Q. Y. (2012). Sequencing and analysis of the complete genome of Rana grylio virus (RGV). *Archives of Virology*, *157*(8), 1559–1564. https://doi.org/10.1007/s00705-012-1316-9

Leung, W. T. M., Thomas-Walters, L., Garner, T. W. J., Balloux, F., Durrant, C., & Price, S. J. (2017). A quantitative-PCR based method to estimate ranavirus viral load following normalisation by reference to an ultraconserved vertebrate target. *Journal of Virological Methods*, *249*, 147–155. https://doi.org/10.1016/j.jviromet.2017.08.016

Löytynoja, A., & Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(30), 10557–10562. https://doi.org/10.1073/pnas.0409137102

Lung, O., Nebroski, M., Buchanan, C., Trapp, M., Sojonky, K., & Joseph, T. (2021). Comparative Genomics Analysis between frog Virus 3-like Ranavirus from the First Canadian Reptile Mortality Event and Similar Viruses from Amphibians. *Research Square Preprints*, 1–17.

Maclaine, A., Wirth, W. T., McKnight, D. T., Burgess, G. W., & Ariel, E. (2020). Ranaviruses in captive and wild Australian lizards. *Facets*, *5*(1), 758–768. https://doi.org/10.1139/FACETS-2020-0011

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A. R. N., Potter, S. C., Finn, R. D., & Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*, *47*(W1), W636–W641. https://doi.org/10.1093/nar/gkz268

Majji, S., Lapatra, S., Long, S. M., Sample, R., Bryan, L., Sinning, A., & Chinchar, V. G. (2006). *Rana catesbeiana* virus Z (RCV-Z): a novel pathogenic ranavirus.

Disease of Aquatic Organisms, *73*, 1–11.

Mandrak, N. E., & Cudmore, B. (2010). The fall of native fishes and the rise of non-native fishes in the great Lakes basin. *Aquatic Ecosystem Health and Management*, *13*(3), 255–268. https://doi.org/10.1080/14634988.2010.507150

Mao, J., Green, D. E., Fellers, G., & Chinchar, V. G. (1999). Molecular characterization of iridoviruses isolated from sympatric amphibians and fish. *Virus Research*, *63*(1–2), 45–52. https://doi.org/10.1016/S0168-1702(99)00057-X

Mao, J., Hedrick, R. P., & Chinchar, V. G. (1997). Molecular Characterization, Sequence Analysis, and Taxonomic Position of Newly Isolated Fish Iridoviruses. *Virology*, *229*(1), 212–220. https://doi.org/10.1006/viro.1996.8435

Marano, N., Arguin, P. M., & Pappaioanou, M. (2007). Impact of globalization and animal trade on infectious disease ecology. *Emerging Infectious Diseases*, *13*(12), 1807–1809. https://doi.org/10.3201/eid1312.071276

Marschang, R. E., Braun, S., & Becher, P. (2005). Isolation of a ranavirus from a gecko (*Uroplatus fimbriatus*). *Journal of Zoo and Wildlife Medicine*, *36*(2), 295–300. https://doi.org/10.1638/04-008.1

Martin, D. P., Murrell, B., Golden, M., Khoosal, A., & Muhire, B. (2015). RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution*, *1*(1), 1–5. https://doi.org/10.1093/ve/vev003

Mavian, C., López-Bueno, A., Balseiro, A., Casais, R., Alcamí, A., & Alejo, A. (2012b). The Genome Sequence of the Emerging Common Midwife Toad Virus Identifies an Evolutionary Intermediate within Ranaviruses. *Journal of Virology*, *86*, 3617–3625. https://doi.org/10.1128/JVI.07108-11

Mavian, C., Lopez-Bueno, A., Fernandez Somalo, M. P., Alcami, A., & Alejo, A. (2012a). Complete Genome Sequence of the European Sheatfish Virus. *Journal of Virology*, *86*(20), 11414–11414. https://doi.org/10.1128/JVI.01968-12

Maynard Smith, J., & Smith, N. H. (1998). Detecting recombination from gene trees. *Molecular Biology and Evolution*, *15*(5), 590–599. https://doi.org/10.1093/oxfordjournals.molbev.a025960

Mazzoni, R., De Mesquita, A. J., Fleury, L. F. F., De Brito, W. M. E. D., Nunes, I. A., Robert, J., Morales, H., Coelho, A. S. G., Barthasson, D. L., Galli, L., & Catroxo, M. H. B. (2009). Mass mortality associated with a frog virus 3-like *Ranavirus* infection in farmed tadpoles *Rana catesbeiana* from Brazil. *Diseases of Aquatic Organisms*, *86*(3), 181–191. https://doi.org/10.3354/dao02096

McDonald, S. M., & Patton, J. T. (2011). Assortment and packaging of the segmented rotavirus genome. *Trends in Microbiology*, *19*(3), 136–144. https://doi.org/10.1016/j.tim.2010.12.002

McKenzie, C. M., Piczak, M. L., Snyman, H. N., Joseph, T., Theijin, C., Chow-Fraser, P., & Jardine, C. M. (2019). First report of ranavirus mortality in a common snapping turtle *Chelydra serpentina*. *Diseases of Aquatic Organisms*, *132*(3), 221–227. https://doi.org/10.3354/dao03324

Meng, Y., Ma, J., Jiang, N., Zeng, L. B., & Xiao, H. B. (2014). Pathological and microbiological findings from mortality of the Chinese giant salamander (*Andrias davidianus*). *Archives of Virology*, *159*(6), 1403–1412. https://doi.org/10.1007/s00705-013-1962-6

Mesnard, J. M., Tham, T. N., Tondre, L., Aubertin, A. M., & Kirn, A. (1988). Organization of RNA transcripts from a 7.8-kb region of the frog virus 3 genome. *Virology*, *165*(1), 122–133. https://doi.org/10.1016/0042-6822(88)90665-4

Miaud, C., Pozet, F., Gaudin, N. C. G., Martel, A., Pasmans, F., & Labrut, S. (2016). *Ranavirus* Causes Mass Die-Offs of Alpine Amphibians in the Southwestern Alps (France). *Journal of Wildlife Diseases*, *52*(2), 1–12. https://doi.org/10.7589/2015-05-113

Miller, D., Gray, M., & Storfer, A. (2011). Ecopathology of Ranaviruses Infecting Amphibians. *Viruses*, *3*, 2351–2373. https://doi.org/10.3390/v3112351

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., & Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research*, *41*(12). https://doi.org/10.1093/nar/gkt263

Moody, N. J. G., & Owens, L. (1994). Experimental demonstration of the pathogenicity of a frog virus, Bohle iridovirus, for a fish species, barramundi *Lates calcarifer*. *Diseases of Aquatic Organisms*, *18*(2), 95–102. https://doi.org/10.3354/dao018095

Morrison, E. A., Garner, S., Echaubard, P., Lesbarrères, D., Kyle, C. J., & Brunetti, C. R. (2014). Complete genome analysis of a frog virus 3 (FV3) isolate and

sequence comparison with isolates of differing levels of virulence. *Virology Journal*, *11*(46), 1–13.

Mu, W. H., Geng, Y., Yu, Z. H., Wang, K. Y., Huang, X. L., Ou, Y. P., Chen, D. F., He, C. L., Zhong, Z. J., Yang, Z. X., & Lai, W. M. (2018). FV3-like ranavirus infection outbreak in black-spotted pond frogs (*Rana nigromaculata*) in China. *Microbial Pathogenesis*, *123*(211), 111–114. https://doi.org/10.1016/j.micpath.2018.06.047

Murray, N. E. (2002). Immigration control of DNA in bacteria: Self versus non-self. *Microbiology*, *148*(1), 3–20. https://doi.org/10.1099/00221287-148-1-3

Murrell, B., Weaver, S., Smith, M. D., Wertheim, J. O., Murrell, S., Aylward, A., Eren, K., Pollner, T., Martin, D. P., Smith, D. M., Scheffler, K., & Kosakovsky Pond, S. L. (2015). Gene-wide identification of episodic selection. *Molecular Biology and Evolution*, *32*(5), 1365–1371. https://doi.org/10.1093/molbev/msv035

Nadin-Davis, S. A., Colville, A., Trewby, H., Biek, R., & Real, L. (2017). Application of high-throughput sequencing to whole rabies viral genome characterisation and its use for phylogenetic re-evaluation of a raccoon strain incursion into the province of Ontario. *Virus Research*, *232*, 123–133. https://doi.org/10.1016/j.virusres.2017.02.007

Near, T. J., Eytan, R. I., Dornburg, A., Kuhn, K. L., Moore, J. A., Davis, M. P., Wainwright, P. C., Friedman, M., & Smith, W. L. (2012). Resolution of ray-finned fish phylogeny and timing of diversification. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(34), 13698–13703. https://doi.org/10.1073/pnas.1206625109

O'Hanlon, S. J., Rieux, A., Farrer, R. A., Rosa, G. M., Waldman, B., Bataille, A., Kosch, T. A., Murray, K. A., Brankovics, B., Fumagalli, M., Martin, M. D., Wales, N., Alvarado-rybak, M., Bates, K. A., Berger, L., Böll, S., Brookes, L., Clare, F., Courtois, E. A., … Bo, C. (2018). Recent Asian origin of chytrid fungi causing global amphibian declines. *Science*, *627*, 621–627. https://doi.org/10.1126/science.aar1965

OIE. (2019). *Aquatic Animal Health Code*. https://www.oie.int/en/what-we-do/standards/codes-and-manuals/aquatic-code-online-access/

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., Minchin, P. R., Hara, R. B. O., Simpson, G. L., Solymos, P., & Henry, M. H. (2020). *vegan: Community Ecology Package. R package version 2.5-7*. https://cran.r-project.org/package=vegan

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 1–14. https://doi.org/10.1186/s13059-016-0997-x

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, *31*(22), 3691–3693. https://doi.org/10.1093/bioinformatics/btv421

Page, A. J., Taylor, B., Delaney, A. J., Soares, J., Seemann, T., Keane, J. A., & Harris, S. R. (2016). SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial Genomics*, *2*(4), e000056. https://doi.org/10.1099/mgen.0.000056

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

Park, J., Grajal-Puche, A., Roh, N. H., Park, I. K., Ra, N. Y., & Park, D. (2021). First detection of ranavirus in a wild population of Dybowski's brown frog (*Rana dybowskii*) in South Korea. *Journal of Ecology and Environment*, *45*(1). https://doi.org/10.1186/s41610-020-00179-2

Parks, D. H., Chuvochina, M., Chaumeil, P. A., Rinke, C., Mussig, A. J., & Hugenholtz, P. (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature Biotechnology*, *38*(9), 1079–1086. https://doi.org/10.1038/s41587-020-0501-8

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P. A., & Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, *36*(10), 996. https://doi.org/10.1038/nbt.4229

Patz, J. A., Daszak, P., Tabor, G. M., Aguirre, A. A., Pearl, M., Epstein, J., Wolfe, N. D., Kilpatrick, A. M., Foufopoulos, J., Molyneux, D., Bradley, D. J., Amerasinghe, F. P., Ashford, R. W., Barthelemy, D., Bos, R., Bradley, D. J., Buck, A., Butler, C., Chivian, E. S., … Zakarov, V. (2004). Unhealthy landscapes: Policy recommendations on land use change and infectious disease emergence. *Environmental Health Perspectives*, *112*(10), 1092–1098. https://doi.org/10.1289/ehp.6877

Peel, A. J., Hartley, M., & Cunningham, A. A. (2012). Qualitative risk analysis of introducing *Batrachochytrium dendrobatidis* to the UK through the

importation of live amphibians. *Diseases of Aquatic Organisms*, *98*(2), 95–112. https://doi.org/10.3354/dao02424

Pérez-Losada, M., Arenas, M., Galán, J. C., Palero, F., & González-Candelas, F. (2015). Recombination in viruses: Mechanisms, methods of study, and evolutionary consequences. *Infection, Genetics and Evolution*, *30*, 296–307. https://doi.org/10.1016/j.meegid.2014.12.022

Peterson, A. T. (2014). Defining viral species: Making taxonomy useful. *Virology Journal*, *11*(1), 1–4. https://doi.org/10.1186/1743-422X-11-131

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E., & Lercher, M. J. (2014). PopGenome: An efficient swiss army knife for population genomic analyses in R. *Molecular Biology and Evolution*, *31*(7), 1929–1936. https://doi.org/10.1093/molbev/msu136

Picco, A. M., & Collins, J. P. (2008). Amphibian commerce as a likely source of pathogen pollution. *Conservation Biology*, *22*(6), 1582–1589. https://doi.org/10.1111/j.1523-1739.2008.01025.x

Pisoni, G., Bertoni, G., Puricelli, M., Maccalli, M., & Moroni, P. (2007). Demonstration of Coinfection with and Recombination by Caprine Arthritis-Encephalitis Virus and Maedi-Visna Virus in Naturally Infected Goats. *Journal of Virology*, *81*(10), 4948–4955. https://doi.org/10.1128/jvi.00126-07

Plumb, J. A., Grizzle, J. M., Young, H. E., Noyes, A. D., & Lamprecht, S. (1996). An iridovirus isolated from wild largemouth bass. *Journal of Aquatic Animal Health*, *8*(4), 265–270. https://doi.org/10.1577/1548-8667(1996)008<0265:AIIFWL>2.3.CO;2

Pond, S. L. K., Frost, S. D. W., & Muse, S. V. (2005). HyPhy: Hypothesis testing using phylogenies. *Bioinformatics*, *21*(5), 676–679. https://doi.org/10.1093/bioinformatics/bti079

Pond, S. L. K., Posada, D., Gravenor, M. B., Woelk, C. H., & Frost, S. D. W. (2006a). Automated phylogenetic detection of recombination using a genetic algorithm. *Molecular Biology and Evolution*, *23*(10), 1891–1901. https://doi.org/10.1093/molbev/msl051

Pond, S. L. K., Posada, D., Gravenor, M. B., Woelk, C. H., & Frost, S. D. W. (2006b). GARD: A genetic algorithm for recombination detection. *Bioinformatics*, *22*(24), 3096–3098. https://doi.org/10.1093/bioinformatics/btl474

Pozet, F., Morand, M., Moussa, A., Torhy, C., & De Kinkelin, P. (1992). Isolation and preliminary characterization of a pathogenic icosahedral deoxyribovirus from the catfish *Ictalurus melas*. *Diseases of Aquatic Organisms*, *14*, 35–42. https://doi.org/10.3354/dao014035

Price, S. J. (2013). *Emergence of a virulent wildlife disease: using spatial epidemiology and phylogenetic methods to reconstruct the spread of amphibian viruses*. Doctoral Thesis, Queen Mary Univeristy of London, UK.

Price, S. J. (2015). Comparative Genomics of Amphibian-like Ranaviruses, Nucleocytoplasmic Large DNA Viruses of Poikilotherms. *Evolutionary Bioinformatics*, *11s2*, 71–82. https://doi.org/10.4137/EBO.S33490

Price, S. J., Ariel, E., Maclaine, A., Rosa, G. M., Gray, M. J., Brunner, J. L., & Garner, T. W. J. (2017a). From fish to frogs and beyond: Impact and host range of emergent ranaviruses. *Virology*, *511*, 272–279. https://doi.org/10.1016/j.virol.2017.08.001

Price, S. J., Garner, T. W. J., Cunningham, A. A., Langton, T. E. S., & Nichols, R. A. (2016). Reconstructing the emergence of a lethal infectious disease of wildlife supports a key role for spread through translocations by humans. *Proceedings of the Royal Society B: Biological Sciences*, *283*. https://doi.org/10.1098/rspb.2016.0952

Price, S. J., Garner, T. W. J., Nichols, R. A., Balloux, F., Ayres, C., Mora-Cabello de Alba, A., & Bosch, J. (2014). Collapse of Amphibian Communities Due to an Introduced *Ranavirus*. *Current Biology*, *24*(21), 2586–2591. https://doi.org/10.1016/j.cub.2014.09.028

Price, S. J., Leung, W. T. M., Owen, C. J., Sergeant, C., Cunningham, A. A., Balloux, F., Puschendorf, R., Garner, T. W. J., & Nichols, R. A. (2019). Effects of historic and projected climate change on the range and impacts of an emerging wildlife disease. *Global Change Biology*. https://doi.org/10.1111/gcb.14651

Price, S. J., Wadia, A., Wright, O. N., Leung, W. T. M., Cunningham, A. A., & Lawson, B. (2017b). Screening of a long-term sample set reveals two *Ranavirus* lineages in British herpetofauna. *PloS One*, *12*(9), e0184768. https://doi.org/10.1371/journal.pone.0184768

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020). Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics*, *70*(1), 1–29. https://doi.org/10.1002/cpbi.102

Pybus, O. G., Rambaut, A., & Harvey, P. H. (2000). An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, *155*(3), 1429–1437. https://doi.org/10.1093/genetics/155.3.1429

Qin, Q. W., Chang, S. F., Ngoh-Lim, G. H., Gibson-Kueh, S., Shi, C., & Lam, T. J. (2003). Characterization of a novel ranavirus isolated from grouper *Epinephelus tauvina*. *Diseases of Aquatic Organisms*, *53*(1), 1–9. https://doi.org/10.3354/dao053001

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology*, *67*(5), 901–904. https://doi.org/10.1093/sysbio/syy032

Rambaut, A., Lam, T. T., Carvalho, L. M., & Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution*, *2*(1), 1–7. https://doi.org/10.1093/ve/vew007

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

Rice, P., Longden, L., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, *16*(6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2

Roach, D. J., Burton, J. N., Lee, C., Stackhouse, B., Butler-Wu, S. M., Cookson, B. T., Shendure, J., & Salipante, S. J. (2015). A Year of Infection in the Intensive Care Unit: Prospective Whole Genome Sequencing of Bacterial Clinical Isolates Reveals Cryptic Transmissions and Novel Microbiota. *PLoS Genetics*, *11*(7), 1–21. https://doi.org/10.1371/journal.pgen.1005413

Robert, J., & Jancovich, J. K. (2016). Recombinant ranaviruses for studying evolution of host–pathogen interactions in ectothermic vertebrates. *Viruses*, *8*(7), 1–13. https://doi.org/10.3390/v8070187

Robinson, C. M., Seto, D., Jones, M. S., Dyer, D. W., & Chodosh, J. (2011). Molecular evolution of human species D adenoviruses. *Infection, Genetics and Evolution*, *11*(6), 1208–1217. https://doi.org/10.1016/j.meegid.2011.04.031

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, *53*(1–2), 131–147. https://doi.org/10.1016/0025-

5564(81)90043-2

Rodrigues, R. AL, de Souza, F. G., de Azevedo, B. L., da Silva, L. C., & Abrahão, J. S. (2021). The morphogenesis of different giant viruses as additional evidence for a common origin of *Nucleocytoviricota*. *Current Opinion in Virology*, *49*, 102–110. https://doi.org/10.1016/j.coviro.2021.05.004

Rosa, G. M., Botto, G. A., Mitra, A., de Almeida, J. S., Hofmann, M., Leung, W. T. M., de Matos, A. P. A., Caeiro, M. F., Froufe, E., Loureiro, A., Price, S. J., Owen, C., Rebelo, R., & Soares, C. (2021). Invasive fish disrupt host-pathogen dynamics leading to amphibian declines. *IN REVIEW*, 1–25.

Rosa, G. M., Sabino-Pinto, J., Laurentino, T. G., Martel, A., Pasmans, F., Rebelo, R., Griffiths, R. A., Stöhr, A. C., Marschang, R. E., Price, S. J., Garner, T. W. J., & Bosch, J. (2017). Impact of asynchronous emergence of two lethal pathogens on amphibian assemblages. *Scientific Reports*, *7*(February), 1–10. https://doi.org/10.1038/srep43260

Rosenberg, R., Johansson, M. A., Powers, A. M., & Miller, B. R. (2013). Search strategy has influenced the discovery rate of human viruses. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(34), 13961–13964. https://doi.org/10.1073/pnas.1307243110

Rubnitz, J., & Subramani, S. (1984). The minimum amount of homology required for homologous recombination in mammalian cells. *Molecular and Cellular Biology*, *4*(11), 2253–2258. https://doi.org/10.1128/mcb.4.11.2253-2258.1984

Russel, P. M., Brewer, B. J., Klaere, S., & Bouckaert, R. R. (2019). Model Selection and Parameter Inference in Phylogenetics Using Nested Sampling. *Systematic Biology*, *68*(2), 219–233. https://doi.org/10.1093/sysbio/syy050

Sanjuán, R., & Domingo-Calap, P. (2016). Mechanisms of viral mutation. *Cellular and Molecular Life Sciences*, *73*(23), 4433–4448. https://doi.org/10.1007/s00018-016-2299-6

Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M., & Belshaw, R. (2010). Viral Mutation Rates. *Journal of Virology*, *84*(19), 9733–9748. https://doi.org/10.1128/JVI.00694-10

Saucedo, B., Hughes, J., Beurden, S. J. Van, Suárez, N. M., Haenen, O. L. M., Voorbergen-laarman, M., Gröne, A., & Kik, J. L. (2017). Complete Genome Sequence of Frog virus 3, Isolated from a Strawberry Poison Frog (*Oophaga pumilio*) Imported from Nicaragua into the Netherlands. *American Society for Microbiology*, *5*(35), 97–98.

Saucedo, B., Hughes, J., Spitzen-Van Der Sluijs, A., Kruithof, N., Schills, M., Rijks, J. M., Jacinto-Maldonado, M., Suarez, N., Haenen, O. L. M., Voorbergen-Laarman, M., Van Den Broek, J., Gilbert, M., Gröne, A., Van Beurden, S. J., & Verheije, M. H. (2018). Ranavirus genotypes in Netherlands and their potential association with virulence in water frogs (*Pelophylax* spp.) article. *Emerging Microbes and Infections*, *7*(1). https://doi.org/10.1038/s41426-018-0058-5

Schloegel, L. M., Daszak, P., Cunningham, A. A., Speare, R., & Hill, B. (2010). Two amphibian diseases, chytridiomycosis and ranaviral disease, are now globally notifiable to the World Organization for Animal Health (OIE): An assessment. *Diseases of Aquatic Organisms*, *92*(2–3), 101–108. https://doi.org/10.3354/dao02140

Schloegel, L. M., Picco, A. M., Kilpatrick, A. M., Davies, A. J., Hyatt, A. D., & Daszak, P. (2009). Magnitude of the US trade in amphibians and presence of *Batrachochytrium dendrobatidis* and ranavirus infection in imported North American bullfrogs (*Rana catesbeiana*). *Biological Conservation*, *142*(7), 1420–1426. https://doi.org/10.1016/j.biocon.2009.02.007

Schorsch, I. G. (1933). *Ranaculture: International Frog Farm, Philadelphia, Pennsylvania*. Buchanan.

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Sepúlveda, V. E., Márquez, R., Turissini, D. A., Goldman, W. E., & Matute, D. R. (2017). Genome sequences reveal cryptic speciation in the human pathogen Histoplasma capsulatum. *MBio*, *8*(6). https://doi.org/10.1128/mBio.01339-17

Sharifian-Fard, M., Pasmans, F., Adriaensen, C., Devisscher, S., Adriaens, T., Louette, G., & Martel, A. (2011). Ranavirosis in Invasive Bullfrogs, Belgium. *Emerging Infectious Diseases*, *17*(12), 2371–2372. https://doi.org/10.3354/dao02032

Shaw, L. P., Wang, A. D., Dylus, D., Meier, M., Pogacnik, G., Dessimoz, C., & Balloux, F. (2020). The phylogenetic range of bacterial and viral pathogens of vertebrates. *Molecular Ecology*, *29*(17), 3361–3379. https://doi.org/10.1111/mec.15463

Sijmons, S., Thys, K., Mbong Ngwese, M., Van Damme, E., Dvorak, J., Van Loock, M., Li, G., Tachezy, R., Busson, L., Aerssens, J., Van Ranst, M., & Maes, P. (2015). High-Throughput Analysis of Human Cytomegalovirus

Genome Diversity Highlights the Widespread Occurrence of Gene-Disrupting Mutations and Pervasive Recombination. *Journal of Virology*, *89*(15), 7673–7695. https://doi.org/10.1128/jvi.00578-15

Simmonds, P., Adams, M. J., Benk, M., Breitbart, M., Brister, J. R., Carstens, E. B., Davison, A. J., Delwart, E., Gorbalenya, A. E., Harrach, B., Hull, R., King, A. M. Q., Koonin, E. V., Krupovic, M., Kuhn, J. H., Lefkowitz, E. J., Nibert, M. L., Orton, R., Roossinck, M. J., … Zerbini, F. M. (2017). Consensus statement: Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology*, *15*(3), 161–168. https://doi.org/10.1038/nrmicro.2016.177

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

Skaliter, R., & Lehman, I. R. (1994). Rolling circle DNA replication in vitro by a complex of herpes simplex virus type 1-encoded enzymes. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(22), 10665–10669. https://doi.org/10.1073/pnas.91.22.10665

Smith, M. D., Wertheim, J. O., Weaver, S., Murrell, B., Scheffler, K., & Kosakovsky Pond, S. L. (2015). Less is more: An adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Molecular Biology and Evolution*, *32*(5), 1342–1353. https://doi.org/10.1093/molbev/msv022

Song, W. J., Qin, Q. W., Qiu, J., Huang, C. H., Wang, F., & Hew, C. L. (2004). Functional Genomics Analysis of Singapore Grouper Iridovirus: Complete Sequence Determination and Proteomic Analysis. *Journal of Virology*, *78*(22), 12576–12590. https://doi.org/10.1128/jvi.78.22.12576-12590.2004

Speare, R., & Smith, J. R. (1992). An iridovirus-like agent isolated from the ornate burrowing frog Limnodynastes ornatus in northern Australia. *Diseases of Aquatic Organisms*, *14*, 51–57. https://doi.org/10.3354/dao014051

Sriwanayos, P., Subramaniam, K., Stilwell, N. K., Imnoi, K., Popov, V. L., Kanchanakhan, S., Polchana, J., & Waltzek, T. B. (2020). Phylogenomic characterization of ranaviruses isolated from cultured fish and amphibians in Thailand. *Facets*, *5*(1), 963–979. https://doi.org/10.1139/FACETS-2020-0043

St-Amour, V., Wong, W. M., Garner, T. W. J., & Lesbarrères, D. (2008). Anthropogenic influence on prevalence of 2 amphibian pathogens. *Emerging Infectious Diseases*, *14*(7), 1175–1176. https://doi.org/10.3201/eid1407.070602

Stagg, H. E. B., Guðmundsdóttir, S., Vendramin, N., Ruane, N. M., Sigurðardóttir, H., Christiansen, D. H., Cuenca, A., Petersen, P. E., Munro, E. S., Popov, V. L., Subramaniam, K., Imnoi, K., Waltzek, T. B., & Olesen, N. J. (2020). Characterization of ranaviruses isolated from lumpfish *Cyclopterus lumpus* L. In the North Atlantic area: Proposal for a new ranavirus species (European North Atlantic Ranavirus). *Journal of General Virology*, *101*(2), 198–207. https://doi.org/10.1099/jgv.0.001377

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stöhr, A. C., López-bueno, A., Blahak, S., Caeiro, M. F., Rosa, G. M., de Matos, A. P. A., Martel, A., Alejo, A., & Marschang, R. E. (2015). Phylogeny and Differentiation of Reptilian and Amphibian Ranaviruses Detected in Europe. *PLoS ONE*, *10*(2), 1–24. https://doi.org/10.1371/journal.pone.0118633

Storfer, A., Alfaro, M. E., Ridenhour, B. J., Jancovich, J. K., Mech, S. G., Parris, M. J., & Collins, J. P. (2007). Phylogenetic concordance analysis shows an emerging pathogen is novel and endemic. *Ecology Letters*, *10*(11), 1075–1083. https://doi.org/10.1111/j.1461-0248.2007.01102.x

Subramaniam, K., Toffan, A., Cappellozza, E., Steckler, N. K., Olesen, N. J., Ariel, E., & Waltzek, T. B. (2016). Genomic sequence of a ranavirus isolated from short-finned eel (*Anguilla australis*). *Genome Announcements*, *4*(4), 9–10. https://doi.org/10.1128/genomeA.00843-16

Tan, W. G. H., Barkman, T. J., Chinchar, V. G., & Essani, K. (2004). Comparative genomic analyses of frog virus 3, type species of the genus *Ranavirus* (family *Iridoviridae*). *Virology*, *323*, 70–84. https://doi.org/10.1016/j.virol.2004.02.019

Tapiovaara, H., Olesen, N. J., Lindén, J., Rimaila-Pärnänen, E., & Von Bonsdorff, C. H. (1998). Isolation of an iridovirus from pike-perch *Stizostedion lucioperca*. *Diseases of Aquatic Organisms*, *32*(3), 185–193. https://doi.org/10.3354/dao032185

Teacher, A. G. F., Cunningham, A. A., & Garner, T. W. J. (2010). Assessing the long-term impact of *Ranavirus* infection in wild common frog populations. *Animal Conservation*, *13*, 514–522. https://doi.org/10.1111/j.1469-1795.2010.00373.x

Todone, F., Weinzierl, R. O. J., Brick, P., & Onesti, S. (2000). Crystal structure of RPB5, a universal eukaryotic RNA polymerase subunit and transcription factor interaction target. *Proceedings of the National Academy of Sciences of the*

United States of America, *97*(12), 6306–6310. https://doi.org/10.1073/pnas.97.12.6306

Trewby, H., Wright, D., Breadon, E. L., Lycett, S. J., Mallon, T. R., McCormick, C., Johnson, P., Orton, R. J., Allen, A. R., Galbraith, J., Herzyk, P., Skuce, R. A., Biek, R., & Kao, R. R. (2016). Use of bacterial whole-genome sequencing to investigate local persistence and spread in bovine tuberculosis. *Epidemics*, *14*, 26–35. https://doi.org/10.1016/j.epidem.2015.08.003

Tsai, C.-T., Ting, J.-W., Wu, M.-H., Wu, M.-F., Guo, I.-C., & Chang, C.-Y. (2005). Complete Genome Sequence of the Grouper Iridovirus and Comparison of Genomic Organization with Those of Other Iridoviruses. *Journal of Virology*, *79*(4), 2010–2023. https://doi.org/10.1128/jvi.79.4.2010-2023.2005

Tweedell, K., & Granoff, A. (1968). Viruses and renal carcinoma of *Rana pipiens*. V. Effect of Frog Virus 3 on Developing Frog Embryos and Larvae. *Journal of the National Cancer Institute*, *40*, 407–410. https://doi.org/10.1016/0042-6822(73)90162-1

Une, Y., Sakuma, A., Matsueda, H., Nakai, K., & Murakami, M. (2009). Ranavirus Outbreak in North American Bullfrogs (*Rana catesbeiana*), Japan, 2008. *Emerging Infectious Diseases*, *15*(7), 1146–1147. https://doi.org/10.3354/dao073001

van Dorp, L., Acman, M., Richard, D., Shaw, L. P., Ford, C. E., Ormond, L., Owen, C. J., Pang, J., Tan, C. C. S., Boshier, F. A. T., Ortiz, A. T., & Balloux, F. (2020). Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infection, Genetics and Evolution*, *83*(April). https://doi.org/10.1016/j.meegid.2020.104351

van Dorp, L., Wang, Q., Shaw, L. P., Acman, M., Brynildsrud, O. B., Eldholm, V., Wang, R., Gao, H., Yin, Y., Chen, H., Ding, C., Farrer, R. A., Didelot, X., Balloux, F., & Wang, H. (2019). Rapid phenotypic evolution in multidrug-resistant *Klebsiella pneumoniae* hospital outbreak strains. *Microbial Genomics*, *5*(4), 1–11. https://doi.org/10.1099/mgen.0.000263

Vanham, P. (2019). *A brief history of globalization.* World Economic Forum Online, Accessed December 2021. https://www.weforum.org/agenda/2019/01/how-globalization-4-0-fits-into-the-history-of-globalization/

Vilaça, S. T., Bienentreu, J.-F., Brunetti, C. R., Lesbarrères, D., Murray, D. L., & Kyle, C. J. (2019). Frog Virus 3 Genomes Reveal Prevalent Recombination between Ranavirus Lineages and Their Origins in Canada. *Journal of Virology*, *93*(20), 1–14. https://doi.org/10.1128/jvi.00765-19

Voigt, C. A., Martinez, C., Wang, Z. G., Mayo, S. L., & Arnold, F. H. (2002). Protein building blocks preserved by recombination. *Nature Structural Biology*, *9*(7), 553–558. https://doi.org/10.1038/nsb805

von Essen, M., Leung, W. T. M., Bosch, J., Pooley, S., Ayres, C., & Price, S. J. (2020). High pathogen prevalence in an amphibian and reptile assemblage at a site with risk factors for dispersal in Galicia, Spain. *PLoS ONE*, *15*(7 July), 1–11. https://doi.org/10.1371/journal.pone.0236803

Walker, P. J., Siddell, S. G., Lefkowitz, E. J., Mushegian, A. R., Adriaenssens, E. M., Dempsey, D. M., Dutilh, B. E., Harrach, B., Harrison, R. L., Hendrickson, R. C., Junglen, S., Knowles, N. J., Kropinski, A. M., Krupovic, M., Kuhn, J. H., Nibert, M., Orton, R. J., Rubino, L., Sabanadzovic, S., … Davison, A. J. (2020). Changes to virus taxonomy and the Statutes ratified by the International Committee on Taxonomy of Viruses (2020). *Archives of Virology*, *165*(11), 2737–2748. https://doi.org/10.1007/s00705-020-04752-x

Waltzek, T. B., Miller, D. L., Gray, M. J., Drecktrah, B., Briggler, J. T., Macconnell, B., Hudson, C., Hopper, L., Friary, J., Yun, S. C., Malm, K. V., Scott Weber, E., & Hedrick, R. P. (2014). New disease records for hatchery-reared sturgeon. I. Expansion of frog virus 3 host range into *Scaphirhynchus albus*. *Diseases of Aquatic Organisms*, *111*(3), 219–227. https://doi.org/10.3354/dao02761

Weinert, L. A., Welch, J. J., Suchard, M. A., Lemey, P., Rambaut, A., & Fitzgerald, J. R. (2012). Molecular dating of human-to-bovid host jumps by *Staphylococcus aureus* reveals an association with the spread of domestication. *Biology Letters*, *8*(5), 829–832. https://doi.org/10.1098/rsbl.2012.0290

Weissenberg, R. (1965). Fifty Years of Research on the Lymphocysts Virus Disease of Fishes (1914 - 1964). *Annals of the New York Academy of Sciences*, *126*, 362–374. https://doi.org/10.1111/j.1475-4991.1993.tb00472.x

Weller, S. K., & Sawitzke, J. A. (2014). Recombination promoted by DNA viruses: Phage λ to herpes simplex virus. *Annual Review of Microbiology*, *68*, 237–258. https://doi.org/10.1146/annurev-micro-091313-103424

Weng, S. P., He, J. G., Wang, X. H., Lü, L., Deng, M., & Chan, S. M. (2002). Outbreaks of an iridovirus disease in cultured tiger frog, *Rana tigrina rugulosa*, in southern China. *Journal of Fish Diseases*, *25*(7), 423–427. https://doi.org/10.1046/j.1365-2761.2002.00386.x

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag

New York. https://ggplot2.tidyverse.org

Wilkinson, D. E., & Weller, S. K. (2004). Recruitment of Cellular Recombination and Repair Proteins to Sites of Herpes Simplex Virus Type 1 DNA Replication Is Dependent on the Composition of Viral Proteins within Prereplicative Sites and Correlates with the Induction of the DNA Damage Response. *Journal of Virology*, *78*(9), 4783–4796. https://doi.org/10.1128/jvi.78.9.4783-4796.2004

Williams, T., Barbosa-Solomieu, V., & Chinchar, V. G. (2005). A DECADE OF ADVANCES IN IRIDOVIRUS RESEARCH. *Advances in Virus Research*, *65*, 172–248. https://doi.org/10.1016/S0065-3527(05)65006-3

Willis, D. B., Goorha, R., & Chinchar, V. G. (1985). Macromolecular synthesis in cells infected by frog virus 3. *Current Topics in Microbiology and Immunology*, *116*, 77–106. https://doi.org/10.1007/978-3-642-70280-8_5

Willis, D. B., & Granoff, A. (1980). Frog virus 3 DNA is heavily methylated at CpG sequences. *Virology*, *107*(1), 250–257. https://doi.org/10.1016/0042-6822(80)90290-1

Wirth, W., Lesbarrères, D., & Ariel, E. (2021). Ten years of ranavirus research (2010–2019): an analysis of global research trends. *Facets*, *6*(1), 44–57. https://doi.org/10.1139/facets-2020-0030

Wiuf, C., Christensen, T., & Hein, J. (2001). A simulation study of the reliability of recombination detection methods. *Molecular Biology and Evolution*, *18*(10), 1929–1939. https://doi.org/10.1093/oxfordjournals.molbev.a003733

Wolf, K. (1988). *Fish viruses and fish viral diseases*. Cornell University Press.

Wolffe, E. J., Vijaya, S., & Moss, B. (1995). A Myristylated Membrane Protein Encoded by the Vaccinia Virus L1R Open Reading Frame Is the Target of Potent Neutralizing Monoclonal Antibodies. *Virology*, *211*(1), 53–63. https://doi.org/10.1006/viro.1995.1378

Wombwell, E. L. (2014). *Emerging Infectious Disease and the Trade in Amphibians*. Doctoral Thesis, Univeristy of Kent, UK.

Wombwell, E. L., Garner, T. W. J., Cunningham, A. A., Quest, R., Pritchard, S., Rowcliffe, J. M., & Griffiths, R. A. (2016). Detection of *Batrachochytrium dendrobatidis* in Amphibians Imported into the UK for the Pet Trade. *EcoHealth*, *13*(3), 456–466. https://doi.org/10.1007/s10393-016-1138-4

Woolhouse, M. E. J., & Gowtage-Sequeria, S. (2005). Host Range and Emerging and Reemerging Pathogens. *Emerging Infectious Diseases*, *11*(12), 1842–1847.

Wynne, F. J. (2019). *Disease ecology of two emerging amphibian pathogens in Costa Rica*. Doctoral Thesis, Univeristy of Plymouth, UK.

Xeros, N. (1954). A Second Virus Disease of the Leatherjacket, *Tipula paludosa*. *Nature*, *174*, 562–563.

Xu, K., Zhu, D. Z., Wei, Y., Schloegel, L. M., Chen, X. F., & Wang, X. L. (2010). Broad distribution of ranavirus in free-ranging *Rana dybowskii* in Heilongjiang, China. *EcoHealth*, *7*(1), 18–23. https://doi.org/10.1007/s10393-010-0289-y

Yang, Z., & Rannala, B. (2012). Molecular phylogenetics: Principles and practice. *Nature Reviews Genetics*, *13*(5), 303–314. https://doi.org/10.1038/nrg3186

Yao, X.-D., & Evans, D. H. (2001). Effects of DNA Structure and Homology Length on Vaccinia Virus Recombination. *Journal of Virology*, *75*(15), 6923–6932. https://doi.org/10.1128/jvi.75.15.6923-6932.2001

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T. Y. (2017). GGTREE: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. *Methods in Ecology and Evolution*, *8*(1), 28–36. https://doi.org/10.1111/2041-210X.12628

Yu, Z., Zhang, W., Gu, C., Chen, J., Zhao, M., Fu, L., Han, J., He, M., Xiao, Q., Xiao, W., He, L., & Zhang, Z. (2020). Genomic analysis of *Ranavirus* and exploring alternative genes for phylogenetics. *Transboundary and Emerging Diseases*, *April*, 1–10. https://doi.org/10.1111/tbed.13864

Zhang, D., Iyer, L. M., & Aravind, L. (2011). A novel immunity system for bacterial nucleic acid degrading toxins and its recruitment in various eukaryotic and DNA viral systems. *Nucleic Acids Research*, *39*(11), 4532–4552. https://doi.org/10.1093/nar/gkr036

Zhang, Q. Y., Xiao, F., Li, Z. Q., Gui, J. F., Mao, J., & Chinchar, V. G. (2001). Characterization of an iridovirus from the cultured pig frog *Rana grylio* with lethal syndrome. *Diseases of Aquatic Organisms*, *48*(1), 27–36. https://doi.org/10.3354/dao048027

Zhuang, J., Jetzt, A. E., Sun, G., Yu, H., Klarmann, G., Ron, Y., Preston, B. D., & Dougherty, J. P. (2002). Human Immunodeficiency Virus Type 1 Recombination: Rate, Fidelity, and Putative Hot Spots. *Journal of Virology*, *76*(22), 11273–11282. https://doi.org/10.1128/jvi.76.22.11273-11282.2002

Annex

The proceeding pages provide an annex to this thesis, in which I present of a body of work that I significantly contributed to during my PhD, but which was not directly related to the ranavirus system. My goal in doing so is to provide a context to some of the methodologies which I developed for the analysis of ranaviral data, but may have been techniques initially co-developed in following projects. My exact contributions to these, and any analyses that were inspired for use in the main thesis, are briefly outlined as follows:

van Dorp, L., Acman, M., Richard, D., Shaw, L.P., Ford, C.E., Ormond, L., Owen, C.J., Pang, J., Tan, C.C.S., Boshier, F.A.T., Ortiz, A.T. & Balloux, F., 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infection, Genetics and Evolution*, 83. https://doi.org/10.1016/j.meegid.2020.104351

In this project, we described the genomic diversity acquired by SARS-CoV-2 over the first few months of the pandemic. I provided a systematic literature review of phylogenetically inferred substitution rates of Sarbecoviruses, and provided tMRCA dates where such analysis had be conducted. I further made extensive use in my thesis of the homoplasy detection protocol we developed, for the main (but not exclusive) purpose of using it as a correction for recombination in phylogenetic reconstruction.

van Dorp, L., Tan, C.C.S., Lam, S.D., Richard, D., Owen, C., Berchtold, D., Orengo, C. & Balloux, F., 2020. Recurrent mutations in SARS-CoV-2 genomes isolated from mink point to rapid host-adaptation. *bioRxiv*. https://doi.org/10.1101/2020.11.16.384743

In this work, we characterised recurrent mutations (putatively adaptive sites) that arose in SARS-CoV-2 following its host jump into farmed mink. I performed regression-based estimates of to the substitution rates and tMRCA of the mink-isolated genomes, whilst investigating whether the rates had changed following its jump from human to mink hosts. I used the protocol for these analyses as the basis for temporal signal assessments of substitution rates in ranavirus. I also assisted in the metadata curation from GISAID database, and contributed to the graphics used in the figures.

Richard, D., Owen, C.J., van Dorp, L. & Balloux, F., 2020. No detectable signal for ongoing genetic recombination in SARS-CoV-2. *bioRxiv*. https://doi.org/10.1101/2020.12.15.422866

In this short project, Dr. Damien Richard and I co-developed a pipeline for linkage disequilibrium decay by genomic distance analysis to discern whether the was a detectable signal of recombination amongst the sample of SARS-CoV-2 genomes available at the time. I significantly extended this pipeline for use with more diverse datasets, which contributed a main pillar to the detection of intragenic recombination in ranaviruses.

Tan, C.C.S., Owen, C.J., Tham, C.Y.L., Bertoletti, A., van Dorp, L. & Balloux, F., 2021. Pre-existing T cell-mediated cross-reactivity to SARS-CoV-2 cannot solely be explained by prior exposure to endemic human coronaviruses. *Infection, Genetics and Evolution*, 95. https://doi.org/10.1016/j.meegid.2021.105075

Finally, in this project, we investigated whether published SARS-CoV-2 cross-reactive epitopes in pre-pandemic T-cells could be explained by sequence homology to any other coronaviruses. Not all cross-reactive epitopes could be explained by significant homology to endemic human coronaviruses. As such, I contributed by generating a core phylogeny of 2,572 isolates across the *Coronaviridae* family (as well as pairwise whole-genome Mash Distances). We then assessed whether any of the remaining epitopes may be cross reactive based on a degree of homology to other coronaviruses more than would be expected by chance in the context of genetic distance. This was one project where I contributed my expertise gained from my work with ranaviruses, in generating phylogenies of highly genetically diverse viral datasets.

I, Christopher Owen, confirm that my contributions to the collaborative work presented in this annex are as outlined above.

Signed:

Christopher J Owen
London, December 2021

Counter signed:

Prof. François Balloux
Lausanne, December 2021

Infection, Genetics and Evolution 83 (2020) 104351

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Emergence of genomic diversity and recurrent mutations in SARS-CoV-2

Lucy van Dorp[a,*], Mislav Acman[a,1], Damien Richard[b,c,1], Liam P. Shaw[d,1], Charlotte E. Ford[a], Louise Ormond[a], Christopher J. Owen[a], Juanita Pang[a,e], Cedric C.S. Tan[a], Florencia A.T. Boshier[e], Arturo Torres Ortiz[a,f], François Balloux[a,*]

[a] UCL Genetics Institute, University College London, London WC1E 6BT, UK
[b] Cirad, UMR PVBMT, F-97410, St Pierre, Réunion, France
[c] Université de la Réunion, UMR PVBMT, F-97490, St Denis, Réunion, France
[d] Nuffield Department of Medicine, John Radcliffe Hospital, University of Oxford, Oxford OX3 9DU, UK
[e] Division of Infection and Immunity, University College London, London WC1E 6BT, UK
[f] Department of Infectious Disease, Imperial College, London W2 1NY, UK

ARTICLE INFO

Keywords:
Betacoronavirus
Homoplasies
Mutation
Phylogenetics

ABSTRACT

SARS-CoV-2 is a SARS-like coronavirus of likely zoonotic origin first identified in December 2019 in Wuhan, the capital of China's Hubei province. The virus has since spread globally, resulting in the currently ongoing COVID-19 pandemic. The first whole genome sequence was published on January 5 2020, and thousands of genomes have been sequenced since this date. This resource allows unprecedented insights into the past demography of SARS-CoV-2 but also monitoring of how the virus is adapting to its novel human host, providing information to direct drug and vaccine design. We curated a dataset of 7666 public genome assemblies and analysed the emergence of genomic diversity over time. Our results are in line with previous estimates and point to all sequences sharing a common ancestor towards the end of 2019, supporting this as the period when SARS-CoV-2 jumped into its human host. Due to extensive transmission, the genetic diversity of the virus in several countries recapitulates a large fraction of its worldwide genetic diversity. We identify regions of the SARS-CoV-2 genome that have remained largely invariant to date, and others that have already accumulated diversity. By focusing on mutations which have emerged independently multiple times (homoplasies), we identify 198 filtered recurrent mutations in the SARS-CoV-2 genome. Nearly 80% of the recurrent mutations produced non-synonymous changes at the protein level, suggesting possible ongoing adaptation of SARS-CoV-2. Three sites in Orf1ab in the regions encoding Nsp6, Nsp11, Nsp13, and one in the Spike protein are characterised by a particularly large number of recurrent mutations (>15 events) which may signpost convergent evolution and are of particular interest in the context of adaptation of SARS-CoV-2 to the human host. We additionally provide an interactive user-friendly web-application to query the alignment of the 7666 SARS-CoV-2 genomes.

1. Introduction

On December 31 2019, China notified the World Health Organisation (WHO) about a cluster of pneumonia cases of unknown aetiology in Wuhan, the capital of the Hubei Province. The initial evidence was suggestive of the outbreak being associated with a seafood market in Wuhan, which was closed on January 1 2020. The aetiological agent was characterised as a SARS-like betacoronavirus, later named SARS-CoV-2, and the first whole genome sequence (Wuhan-HU-1) was deposited on NCBI Genbank on January 5 2020 (Wu et al., 2020). Human-to-human transmission was confirmed on January 14 2020, by which time SARS-CoV-2 had already spread to many countries throughout the world. Further extensive global transmission led to the WHO declaring COVID-19 as a pandemic on March 11 2020.

Coronaviridae comprise a large number of lineages that are found in a wide range of mammals and birds (Shaw et al., 2020), including the other human zoonotic pathogens SARS-CoV-1 and MERS-COV. The propensity of Betacoronaviridiae to undergo frequent host jumps supports SARS-CoV-2 also being of zoonotic origin. To date, the genetically closest-known lineage is found in horseshoe bats (BatCoV RaTG13) (Zhou et al., 2020). However, this lineage shares 96% identity with SARS-CoV-2, which is not sufficiently high to implicate it as the

* Corresponding authors.
E-mail addresses: lucy.dorp.12@ucl.ac.uk (L. van Dorp), f.balloux@ucl.ac.uk (F. Balloux).
[1] Equal contribution.

https://doi.org/10.1016/j.meegid.2020.104351
Received 24 April 2020; Received in revised form 30 April 2020
Available online 05 May 2020

immediate ancestor of SARS-CoV-2. The zoonotic source of the virus remains unidentified at the date of writing (April 23 2020).

The analysis of genetic sequence data from pathogens is increasingly recognised as an important tool in infectious disease epidemiology (Rambaut et al., 2008; Grenfell et al., 2004). Genetic sequence data sheds light on key epidemiological parameters such as doubling time of an outbreak/epidemic, reconstruction of transmission routes and the identification of possible sources and animal reservoirs. Additionally, whole-genome sequence data can inform drug and vaccine design. Indeed, genomic data can be used to identify pathogen genes interacting with the host and allows characterisation of the more evolutionary constrained regions of a pathogen genome, which should be preferentially targeted to avoid rapid drug and vaccine escape mutants.

There are thousands of global SARS-CoV-2 whole-genome sequences available on the rapid data sharing service hosted by the Global Initiative on Sharing All Influenza Data (GISAID; https://www.epicov.org) (Elbe and Buckland-Merrett, 2017; Shu and McCauley, 2017). The extraordinary availability of genomic data during the COVID-19 pandemic has been made possible thanks to a tremendous effort by hundreds of researchers globally depositing SARS-CoV-2 assemblies (Table S1) and the proliferation of close to real time data visualisation and analysis tools including NextStrain (https://nextstrain.org) and CoV-GLUE (http://cov-glue.cvr.gla.ac.uk).

In this work we use this data to analyse the genomic diversity that has emerged in the global population of SARS-CoV-2 since the beginning of the COVID-19 pandemic, based on a download of 7710 assemblies. We focus in particular on mutations that have emerged independently multiple times (homoplasies) as these are likely candidates for ongoing adaptation of SARS-CoV-2 to its novel human host. After filtering, we characterise homoplasies at 198 sites in the SARS-CoV-2 genome. We identify a strong signal of recurrent mutation at nucleotide position 11,083 (Codon 3606 Orf1a), together with two further sites in Orf1ab encoding the non-structural proteins Nsp11 and Nsp13. These, together with a mutation in the Spike protein (21,575, Codon 5), comprise the strongest putative regions under selection in our dataset.

The current distribution of genomic diversity as well as ongoing allele frequency changes both between isolates and along the SARS-CoV-2 genome are publicly available as an open access and interactive web-resource available here:

https://macman123.shinyapps.io/ugi-scov2-alignment-screen/.

2. Material and methods

2.1. Data acquisition

7710 SARS-CoV-2 assemblies flagged as “complete (>29,000 bp)”, “high coverage only”, “low coverage excl” were downloaded from the GISAID Initiative EpiCoV platform as of April 19 2020 (11:30 GMT). A full acknowledgements table of those labs which generated and uploaded data is provided in Table S1. Filtering was performed on the downloaded assemblies to exclude those deriving from animals (bat, pangolin), those with more than 1% missing sites, and otherwise spurious assemblies as also listed by nCov-GLUE (http://cov-glue.cvr.gla.ac.uk/#/excludedSeqs). This left a final dataset of 7666 assemblies for downstream analysis. Sequence metadata was obtained from the NextStrain Github repository (https://github.com/nextstrain/ncov/tree/master/data). While results presented here predominately focus on an analysis of the available assemblies as of April 19 2020, equivalent analyses were performed daily from March 24 2020. This allowed tracking of the emergence of genomic variants in public sequence data as assemblies were uploaded during the course of the pandemic.

2.2. Multi-sequence alignment and maximum likelihood tree

Assemblies were aligned against the Wuhan-Hu-1 reference genome (NC_045512.2, EPI_ISL_402125) using MAFFT (Katoh and Standley, 2013) implemented via the rapid phylodynamic alignment pipeline provided by Augur (https://github.com/nextstrain/augur). Sites in the first 130 bp and last 50 bp of the alignment were masked, as were positions 18,529, 29,849, 29,851 and 29,853, following the protocol also advocated by NextStrain and to account for the fact many putatively artefactual SNPs are located at the beginning and ends of the alignment. Resulting alignments were manually inspected in UGene (http://ugene.net). Subsequently a maximum likelihood phylogenetic tree was built using the Augur tree implementation selecting RAxML as the tree-building method (Kozlov et al., 2019). The resulting phylogeny was viewed and annotated using ggtree (Yu et al., 2017) (Figs. S1-S2). Throughout, site numbering and genome structure are given using Wuhan-Hu-1 (NC_045512.2) as reference.

2.3. Phylogenetic dating

The maximum likelihood phylogenetic tree was tested for the presence of significant molecular evolution over the sampling period using the roottotip() function provided in BactDating (Didelot et al., 2018). After confirmation of a significant regression following 1000 random permutations of sampling dates (Fig. S3), temporal calibration of the phylogeny was performed using TreeDater (Volz and Frost, 2017), assuming a strict clock model of evolution, as we do not expect a significant difference in rate variation across lineages at these time scales (Fig. S4). To obtain confidence intervals around each temporal point estimate we conducted a parametric bootstrapping analysis with 50 replicates on the unmasked alignment, keeping the tree topology constant while generating new branch length estimates using a Poisson distribution and running the same model in TreeDater (Volz and Frost, 2017). We also evaluated all currently available estimates for tip-calibration estimates of the tMRCA of SARS-CoV-2 together with rate estimates for other closely related betacoronaviruses (Table 1, Table S2).

2.4. Maximum parsimony tree and homoplasy screen

In parallel a Maximum Parsimony tree was built using the fast tree inference and bootstrap approximation offered by MPBoot (Hoang et al., 2018). MPBoot was run on the alignment to reconstruct the Maximum Parsimony tree and to assess branch support following 1000 replicates (*−bb 1000*). The resulting Maximum Parsimony treefile was used, together with the input alignment, to rapidly identify recurrent mutations (homoplasies) using HomoplasyFinder (Crispell et al., 2019).

HomoplasyFinder provides, for each site, the minimum number of state changes required on the tree to explain the observed character states at the tips, as described by Fitch (Fitch, 1971), and measured via the site specific consistency index. For this analysis all ambiguous sites in the alignment were set to ‘N’. To assess whether any particular Open Reading Frame (ORF) showed evidence of more homoplasies than expected given the length of the ORF, an empirical distribution was obtained by sampling, with replacement, equivalent length windows and recording the number of homoplasies detected (Table S3).

HomoplasyFinder identified 1132 homoplasies (1042 excluding masked sites), which were distributed over the SARS-CoV-2 genome (Fig. S5, Table S4). Of these, 40 sites have a derived allele at >1% of the total isolates. However, homoplasies can arise due to convergent evolution (putatively adaptive), recombination, or via errors during the processing of sequence data. The latter is particularly problematic here due to the mix of technologies and methods employed by different contributing research groups. We therefore filtered identified homoplasies using a set of thresholds attempting to circumvent this problem (filtering scripts and figures are available at https://github.com/liampshaw/CoV-homoplasy-filtering).

In summary, for each homoplasy we computed the proportion of isolates with the homoplasy p_{nn} where the nearest neighbouring isolate in the phylogeny also carried the homoplasy (excluding identical sequences). This metric ranges between $p_{nn} = 0$ (all isolates with the

Table 1
Estimates of SARS-CoV-2 time to most recent common ancestor (tMRCA). BCI: Bayesian Credible Interval; HPD: Highest Posterior Density; CI: Confidence Interval. Asterix * denotes non-peer reviewed estimate of tMRCA. 'N.' denotes the number of whole genomes analysed.

Reference	N.	Substitution Rate (per site per year)	Estimated tMRCA	Method
Li et al. 2020 (Li et al., 2020)	32	1.0 × 10–3 (95% BCI 1.854 × 10–4, 4.0 × 10–3)	October 15, 2019 (95% BCI May 2, 2019; January 17, 2020)	Rate-informed strict clock model (BEAST v1.8.4)
Li et al. 2020 (Li et al., 2020)	32	1.8266 × 10–3 (95% BCI 7.5813 × 10–4, 3.0883 × 10–3)	December 6, 2019 (95% BCI November 16, 2019; December 21, 2019)	Rate-estimated relaxed clock model (BEAST v1.8.4)
Giovanetti et al. 2020 (Giovanetti et al., 2020)	54	6.58 × 10–3 (95% HPD 5.2 × 10–3, 8.1 × 10–3)	November 25, 2019 (95% CI September 28, 2019; December 21, 2019)	Relaxed clock model (BEAST v1.10.4)
Hill & Rambaut 2020*[1]	75	0.92 × 10–3 (95% HPD 0.33 × 10–3–1.46 × 10–3)	November 29, 2019 (95% CI October 28, 2019; December 20, 2019)	Unreported clock model (BEAST v1.7.0)
Hill & Rambaut 2020*[1]	86	0.80 × 10–3 (95% HPD 0.14 × 10–3, 1.31 × 10–3)	November 17, 2019 (95% CI August 27, 2019; December 19, 2019)	Unreported clock model (BEAST v1.7.0)
Hill & Rambaut 2020*[1]	116	1.04 × 10–3 (95% HPD 0.71 × 10–3, 1.40 × 10–3)	December 3, 2019 (95% CI November 16, 2019; December 17, 2019)	Unreported clock model (BEAST v1.7.0)
Lu et al. 2020* (41)	53	–	November 29, 2019 (95% HPD November 14, 2019; December 13, 2019)	Strict clock model (BEAST v1.10.0)
Duchene et al. 2020*[2]	47	1.23 × 10–4(95% HPD 5.63 × 10–4, 1.98 × 10–3)	November 19, 2019 (HPD October 21, 2019; December 11, 2019)	Strict clock model (BEAST v1.10)
Duchene et al. 2020*[2]	47	1.29 × 10–3 (HPD 5.35 × 10–4, 2.15 × 10–3)	November 12, 2019 (HPD September 26, 2019; December 11, 2019)	Relaxed clock model (BEAST v1.10)
Volz et al. 2020*[3]	53	Model constrained between 7 × 10–4 & 2 × 10–3	December 8, 2019 (95% CI November 21, 2019; December 20, 2019)	Strict clock model (BEAST v2.6.0)
Volz et al. 2020*[3]	53	Model constrained between 5 × 10–4 & 1.25 × 10–3	December 5, 2019(95% CI November 6, 2019; December 13, 2019)	Maximum Likelihood regression (*treedater* R package v0.5.0)

[1] http://virological.org/t/phylodynamic-analysis-of-sars-cov-2-update-2020-03-06/420; [2] http://virological.org/t/temporal-signal-and-the-evolutionary-rate-of-2019-n-cov-using-47-genomes-collected-by-feb-01-2020/379; [3] https://doi.org/10.25561/77169

homoplasy present as singletons) and $p_{nn} = 1$ (no singletons i.e. clustering of isolates with the homoplasy in the phylogeny). We reasoned that artefactual sequencing homoplasies would tend to show up as singletons, so excluded all homoplasies with $p_{nn} < 0.1$ from further analysis.

To obtain a set of high confidence homoplasies, we then used the following criteria: ≥0.1% isolates in the alignment share the homoplasy (equivalent to >8 isolates), $p_{nn} > 0.1$, and derived allele found in strains sequenced from >1 originating lab and > 1 submitting lab. We also required the proportion of isolates where the homoplasic site was in close proximity to an ambiguous base (± 5 bp) to be zero. The application of these various filters reduced the number of homoplasies to 198 (Table S5). We also plotted the distributions of cophenetic distances between isolates carrying each homoplasy compared to the distribution for all isolates (Fig. S6), and inspected the distribution of all identified homoplasies in the phylogenies from our own analyses and on the phylogenetic visualisation platform provided by NextStrain. Finally, we examined whether ambiguous bases were seen more often at homoplasic sites than at random bases (excluding masked sites), which was not the case (Fig. S7).

To further validate the homoplasy detection method applied to the alignment of the 7666 SARS-CoV-2 genome assemblies, we took advantage of the genome sequences for which raw reads were available on the Short Read Archive (SRA). A variant calling pipeline (available at https://github.com/DamienFr/CoV-homoplasy) was used to obtain high-confidence alignments for the 348 (out of 889 as of April 19 2020) SRA genomic datasets both meeting our quality criterions and matching GISAID assemblies. The topology of the Maximum Likelihood phylogeny of these 348 samples was compared to that of the corresponding samples from the GISAID genome assemblies using a Mantel test and the Phytools R package (Revell, 2012) (Figs. S8-S9, see Supplementary text).

As discussed, the GISAID dataset comprises assemblies of variable quality, potentially impairing the detection of genuine homoplasies and/or leading to false positive SNPs due to sequencing error or spurious allele assignment during the production of the *de novo* assembly from raw sequence reads. Therefore, to further assess the detection of homoplasies, we applied HomoplasyFinder to the two datasets comprising the same 348 strains (GISAID and SRA) (Table S6). We detected 19 homoplasies on the dataset originating from the SRA, and 21 on the dataset originating from GISAID assemblies. Of these, 19 were detected in both datasets (Table S7). Using the same filters as for the main dataset (with the exception of the ≥0.1% frequency set to ≥1%), 10 and 11 homoplasies were kept in the SRA dataset and in the GISAID dataset, respectively. Nine sites were detected in both datasets. For sites which failed the filtering thresholds, this was largely due to the low number of studied accessions, which increases the probability of an isolated strain displaying a homoplasy e.g. if $n = 2$ isolates have a homoplasy, by definition they cannot be nearest neighbours, so $p_{nn} = 0$.

2.5. Annotation of variant and homoplasic sites

The alignment was translated to amino acid sequences using SeaView V4 (Gouy et al., 2010). Sites were identified as synonymous or non-synonymous and amino acid changes corresponding to these mutations were retrieved via multiple sequence alignment. We assessed the change in hydrophobicity and charge of amino acid residues arising due to homoplastic non-synonymous mutations using the hydrophobicity scale proposed by Janin (Janin, 1979). The ten most hydrophobic residues on this scale were considered hydrophobic and the rest as hydrophilic. In addition, amino acid residues were either classified as positively charged, negatively charged or neutral at pH 7. The charge of each residue can either increase, decrease or remain the same (neutral mutation) due to mutation (Fig. S10).

2.6. Comparison with SARS-CoV-1 and MERS-CoV

SARS-CoV-1 and MERS-CoV are both zoonotic pathogens related to SARS-CoV-2, which underwent a host jump into the human host previously. We investigated whether the major homoplasies we detect in SARS-CoV-2 affect sites which also underwent recurrent mutations in these related viruses as these adapted to their human host. All Coronaviridae assemblies were downloaded (NCBI TaxID:11118) on April 8 2020 and human associated MERS-CoV and SARS-CoV-1 assemblies extracted. This gave a total of 15 assemblies for SARS-CoV-1 and 255 assemblies for MERS-CoV. Following the same protocol (Augur align) as applied to SARS-CoV-2 assemblies, each species was aligned against the respective RefSeq reference genomes: NC_004718.3 for SARS-CoV-1 and NC_019843.3 for MERS-CoV. This produced alignments of 29,751 bp (187 SNPs) and 30,119 bp (1588 SNPs) respectively.

MPBoot (Hoang et al., 2018) was run on both sets of alignments to reconstruct the maximum parsimony tree and to assess branch support following 1000 replicates (*−bb 1000*). The resulting maximum parsimony treefiles were used, together with the input alignment, to rapidly identify homoplasies using HomoplasyFinder (Crispell et al., 2019). For SARS-CoV-1 we detected six homoplasies and for MERS-CoV we detected 350 homoplasies (pre-filtering) (Fig. S11-S12). The distribution of homoplasies was assessed relative to the Genbank annotation files and in the context of the high confidence homoplasies that we detect in SARS-CoV-2.

3. Results

3.1. Emergence of SARS-CoV-2 genomic diversity over time

The 7666 SARS-CoV-2 genomes offer an excellent geographical and temporal coverage of the COVID-19 pandemic (Fig. 1a-b). The genomic diversity of the 7666 SARS-CoV-2 genomes is represented as Maximum Likelihood phylogenies in a radial (Fig. 1c) and linear layout (Fig. S1-S2). There is a robust temporal signal in the data, captured by a statistically significant correlation between sampling dates and 'root-to-tip' distances for the 7666 SARS-CoV-2 (Fig. S3; $R^2 = 0.20, p < .001$). Such positive association between sampling time and evolution is expected to arise in the presence of measurable evolution over the timeframe over which the genetic data was collected. Specifically, more recently sampled strains have accumulated additional mutations in their genome than older ones since their divergence from the Most Recent Common Ancestor (MRCA, root of the tree).

The origin of the regression between sampling dates and 'root-to-tip' distances (Fig. S3) provides a cursory point estimate for the time to the MRCA (tMRCA) around late 2019. Using TreeDater (Volz and Frost, 2017), we observe an estimated tMRCA, which corresponds to the start of the COVID-19 epidemic, of 6 October 2019–11 December 2019 (95% CIs) (Fig. S4). These dates for the start of the epidemic are in broad agreement with previous estimates performed on smaller subsets of the COVID-19 genomic data using various computational methods (Table 1), though they should still be taken with some caution. Indeed, the sheer size of the dataset precludes the use of some of the more sophisticated inference methods available.

The SARS-CoV-2 global population has accumulated only moderate genetic diversity at this stage of the COVID-19 pandemic with an average pairwise difference of 9.6 SNPs between any two genomes, providing further support for a relatively recent common ancestor. We estimated a mutation rate underlying the global diversity of SARS-Cov-2 of ~6×10^{-4} nucleotides/genome/year (CI: 4×10^{-4}–7×10^{-4}) obtained following time calibration of the maximum likelihood phylogeny. This rate is largely unremarkable for an RNA virus (Domingo-Calap et al., 2018; Holmes et al., 2016), despite Coronaviridae having the unusual capacity amongst viruses of proofreading during nucleotide replication, thanks to the non-structural protein nsp14 exonuclease,

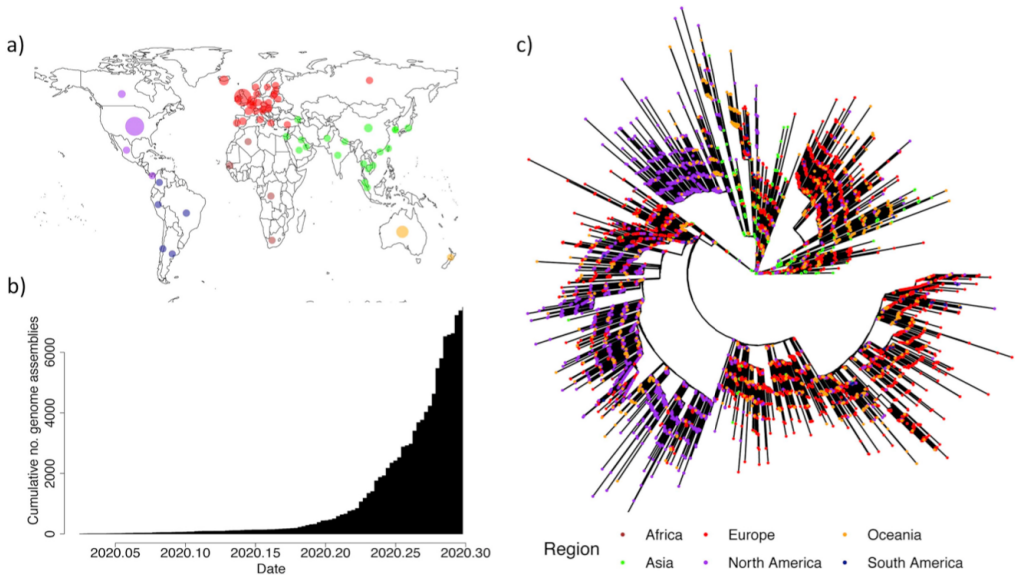


Fig. 1. Global sequencing efforts have contributed hugely to our understanding of the genomic diversity of SARS-CoV-2. a) Viral assemblies available from global regions as of 19/04/2020. b) Cumulative total of viral assemblies uploaded to GISAID included in our analysis. c) Radial Maximum Likelihood phylogeny for 7666 complete SARS-CoV-2 genomes. Colours represent continents where isolates were collected. Green: Asia; Red: Europe; Purple: North America; Orange: Oceania; Dark blue: South America according to metadata annotations available on NextStrain (https://github.com/nextstrain/ncov/tree/master/data). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

which excises erroneous nucleotides inserted by their main RNA polymerase nsp12 (Snijder et al., 2003; Minskaia et al., 2006).

3.2. Everything is everywhere

Some of the major clades in the maximum likelihood phylogeny (Fig. 1c and Fig. S1) are formed predominantly by strains sampled from the same continent. However, this likely represents a temporal rather than a geographic signal. Indeed, the earliest available strains were collected in Asia, where the COVID-19 pandemic started, followed by extensive genome sequencing efforts first in Europe and then in the USA.

The SARS-CoV-2 genomic diversity found in most countries (with sufficient sequences) essentially recapitulates the global diversity of COVID-19 from the 7666-genome dataset. Fig. 2 highlights the proportion of the global genetic diversity found in the UK, the USA, Iceland and China. In the UK, the USA and Iceland, the majority of the global genetic diversity of SARS-CoV-2 is recapitulated, with representatives of all major clades present in each of the countries (Fig. 2A-C). The same is true for other countries such as Australia (Fig. S2a).

This genetic diversity of SARS-CoV-2 populations circulating in different countries points to each of these local epidemics having been seeded by a large number of independent introductions of the virus. The main exception to this pattern is China, the source of the initial outbreak, where only a fraction of the global diversity is present (Fig. 2d). This is also to an extent the case for Italy (Fig. S2b), which was an early focus of the COVID-19 pandemic. However, this global dataset includes only 35 SARS-CoV-2 genomes from Italy, so some of the genetic diversity of SARS-CoV-2 strains in circulation likely remains unsampled. The genomic diversity of the global SARS-CoV-2 population being recapitulated in multiple countries points to extensive worldwide transmission of COVID-19, likely from extremely early on in the pandemic.

3.3. Genetic diversity along the genome alignment and recurrent mutations (homoplasies)

The SARS-CoV-2 alignment can be considered as broken into a large two-part Open Reading Frame (ORF) encoding non-structural proteins, four structure proteins: spike (S), envelope (E), membrane (M) and nucleocapsid (N), and a set of small accessory factors (Fig. 3a). There is variation in genetic diversity across the alignment, with polymorphisms often found in neighbouring clusters (Fig. S5). A simple permutation resampling approach suggests that both Orf3a and N exhibit SNPs which fall in the 95th percentile of the empirical distribution (Table S3). However, not all of these sites can be confirmed as true variant positions, due to the lack of accompanying sequence read data. However, we closely inspected those sites that appear to have arisen multiple times following a maximum parsimony tree building step. We identified a large number of putative homoplasies ($n = 1042$ excluding masked regions), which were filtered to a high confidence cohort of 198 positions (see Methods).

These 198 positions in the SARS-CoV-2 genome alignment (0.67% of all sites) were associated with 290 amino acid changes across all 7666 genomes. Of these amino acid changes, 232 comprised non-synonymous and 58 comprised synonymous mutations. Two non-synonymous mutations involved the introduction or removal of stop codons were found (*13402Y, *26152G). 53 of the remaining 101 non-synonymous mutations involved neutral hydrophobicity changes (Fig. S10a). In addition, 79 of the remaining 101 non-synonymous mutations involved neutral changes (Fig. S10b). Both Orf1ab and N had a four-fold higher frequency of hydrophilic → hydrophobic mutations than

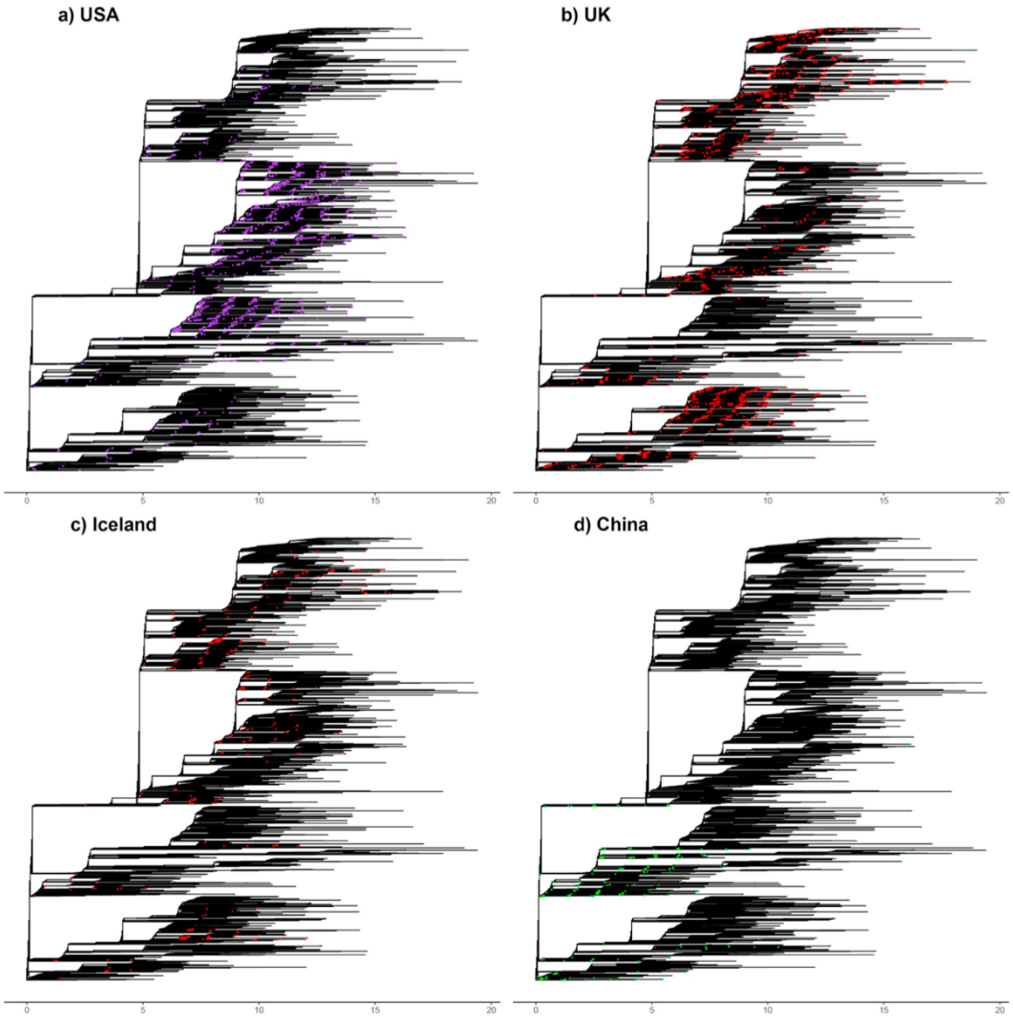


Fig. 2. Genomic diversity of SARSCoV-2 in the USA, UK, Iceland and China. Strains collected from all four countries are highlighted on the global phylogenetic tree. a) Strains collected in the USA shown in purple. b) Strains from the UK shown in red. c) Strains collected in Iceland shown in red. d) Strains collected in China shown in green. Regional colours match to the global phylogeny shown in Fig. 1c. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

hydrophobic → hydrophilic mutations (Fig. S10). In addition, neutral hydrophobic changes were clearly favoured in the S protein. Lastly, 87 of the remaining 110 non-synonymous mutations involved neutral charge changes.

Amongst the strongest filtered homoplasic sites (> 15 change points on the tree), three are found within Orf1ab (nucleotide positions 11,083, 13,402, 16,887) and S (21575). We exemplify the strongest signal and our approach using position 11,083 in Fig. 3 and provide a full list of homoplasic sites, both filtered and unfiltered, in Tables S4–5. The strongest hit in terms of the inferred minimum number of changes required (Fig. 3b-c) at Orf1ab (11,083, Codon 3606) falls over a region encoding the non-structural protein, Nsp6, and is also observed in our analyses of the SRA dataset (Table S7).

We note that some of the hits also overlap with positions identified as putatively under selection using other approaches (http://virological.org/t/selection-analysis-of-gisaid-sars-cov-2-data/448/3, accessed April 23 2020), with Orf1ab consistently identified as a region comprising several candidates for non-neutral evolution. Orf1ab is an orthologous gene with other human-associated betacoronaviruses, in particular SARS-CoV-1 and MERS-CoV which both underwent host jumps into humans from likely bat reservoirs (Lau et al., 2005; Memish et al., 2013). We performed an equivalent analysis on human-associated

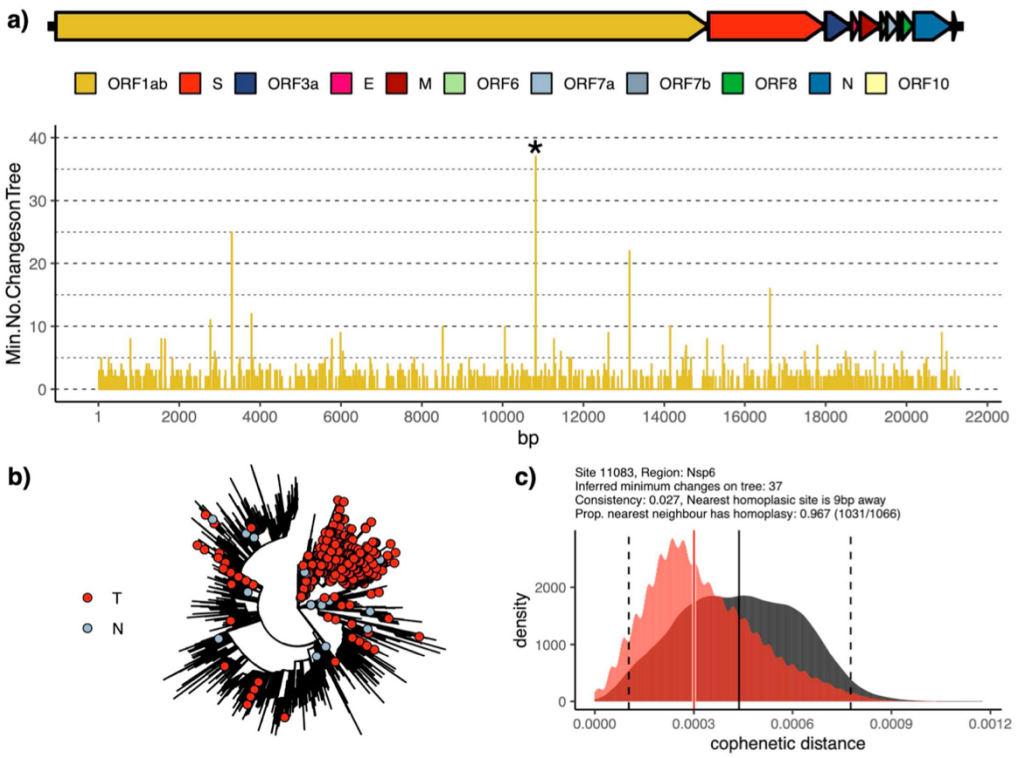


Fig. 3. Inspection of a major homoplastic site in Orf1ab of SARS-CoV-2 genome (position 11,083). Panel A shows a colour-coded schematic of the SARS-CoV-2 genome annotated as per NC_045512.2 and a plot of all potential homoplastic sites in Orf1ab measured as minimal number of character-state changes on a Maximum Parsimony tree (see Methods). Exemplar homoplasy (denoted with *) has been shown on the radial ML phylogenetic tree in panel B. Panel C shows the distribution of cophenetic distances between isolates carrying the identified homoplasy (red) and the distribution for all isolates (grey), showing that isolates with the homoplasy tend to cluster in the phylogeny. Equivalent figures for other filtered homoplasies are generated as part of the filtering method (see Methods). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

virus assemblies available on the NCBI Virus platform. We identified six putative homoplasic sites within SARS-CoV-1, two occurring within the 3c-like proteinase just upstream of Nsp6 (10,384, 10,793) and a further two homoplasies within Orf1ab at Nsp9 and Nsp13 (Fig. S11). In addition, one homoplasy was identified in the spike protein and one in the membrane protein ORFs.

For MERS-CoV, multiple unfiltered homoplasies were detected, consistent with previous observations of high recombination in this species (Dudas and Rambaut, 2016), though only one invoked more than a minimum number of 10 changes on the maximum parsimony tree (Fig. S12). This corresponded to a a further homoplasy identified in Orf1ab Nsp6 (position 11,631). It is of note that this genomic region coincides with the strongest homoplasy in SARS-CoV-2 which also occurs in the Nsp6 encoding region of Orf1ab. Codon 3606 of Orf1ab shares a leucine residue in MERS-CoV and SARS-CoV-2, though a valine in SARS-CoV. The exact role of these and other homoplasic mutations in human associated betacoronaviruses represents an important area of future work, although it appears that the Orf1ab region may exhibit multiple putatively adapted variants across human betacoronavirus lineages.

The genome alignment of the 7666 SARS-CoV-2 genomes can be queried through an open access, interactive web-application (https://macman123.shinyapps.io/ugi-scov2-alignment-screen/). It provides users with information on every SNP and homoplasy detected across our global SARS-CoV-2 alignment and allows visual inspection both within the sequence alignment and across the maximum likelihood tree phylogeny. Fig. 3 illustrates some of the functionalities of the web application using position 11083 in the alignment as an example. This particular homoplasy was observed 1126 times across the genomes and requires a minimum of 37 character-site changes to become congruent with the observed SARS-CoV-2 phylogeny (Fig. 3a and b).

4. Discussion

Pandemics have been affecting humanity for millennia (Balloux and van Dorp, 2017). Over the last century alone, several global epidemics have claimed millions of lives, including the 1957/58 influenza A (H2N2) pandemic, the sixth (1899–1923) and seventh 'El Tor' cholera pandemic (1961–1975), as well as the HIV/AIDS pandemic (1981-today). COVID-19 acts as an unwelcome reminder of the major threat that infectious diseases represent in terms of deaths and disruption.

One positive aspect of the current situation, relative to previous

pandemics, is the unprecedented availability of scientific and technological means to face COVID-19. In particular, the rapid development of drugs and vaccines has already begun. Modern drug and vaccine development are largely based on genetic engineering and an understanding of host-pathogen interactions at a molecular level. The mobilisation to address the COVID-19 pandemic by scientists worldwide has been remarkable. This includes the feat of the global scientific community who has already produced and publicly shared well over 11,000 complete SARS-CoV-2 genome sequences at the time of writing (April 23 2020), which we have used here with gratitude. Further initatives in the United Kingdom (https://www.cogconsortium.uk/data/) have already to date produced over 10,000 genomes, some of which overlap with those already available on GISAID.

To put these numbers of SARS-CoV-2 genomes in context, it is interesting to consider parallels with the 2009 H1N1pdm influenza pandemic, the first epidemic for which genetic sequence data was generated in near-real time (Fraser et al., 2009; Smith et al., 2009). The genetic data available at the time looks staggeringly small in comparison to the amount that has already been generated for SARS-CoV-2 during the early stages of the COVID-19 pandemic. For example, Fraser et al. considered 11 partial hemagglutinin gene sequences two months after the WHO had declared 2009 H1N1pdm influenza a pandemic (Fraser et al., 2009).

This unprecedented genomic resource has already provided strong conclusions about the pandemic. For example, analyses by multiple independent groups place the start of the COVID-19 pandemic towards the end of 2019 (Table 1). This rules out any scenario that assumes SARS-CoV-2 may have been in circulation long before it was identified, and hence have already infected large proportions of the population.

Extensive genomic resources for SARS-CoV-1 should in principle also be key to informing on optimal drug and vaccine design, particularly when coupled with knowledge of human proteome and immune interactions (Gordon et al., 2020). Ideally, drugs and vaccines should target relatively invariant, strongly constrained regions of the SARS-CoV-2 genome, to avoid drug resistance and vaccine evasion. Therefore ongoing monitoring of genomic changes in the virus will be essential to gain a better understanding of fundamental host-pathogen interactions that can inform drug and vaccine design.

As most (but not all) pathogens capable of causing epidemic at a pandemic scale, SARS-CoV-2 is in all likelihood of zoonotic origin. This implies that SARS-CoV-2 may not be fine-tuned to its novel human host. However, it is near-impossible to predict future trajectories for the virulence and transmissibility of horizontally transmitted pathogens (Anderson and May, 1991). It is also possible that the population of SARS-CoV-2 will evolve into different lineages characterised by variable levels of virulence and transmissibility. However, despite existing phylogenetic structure (Rambaut et al., 2020), it is important to stress that there is no evidence for the evolution of distinct phenotypes in SARS-CoV-2 at this stage.

The vast majority of mutations observed so far in SARS-CoV-2 circulating in humans are likely neutral (Cagliani et al., 2020; Dearlove et al., 2020) or even deleterious (Nielsen et al., 2020). Homoplasies, such as those we detect here, can arise by product of neutral evolution or as a result of ongoing selection. Of the 198 homoplasies we detect (after applying stringent filters), some proportion are very likely genuine targets of positive selection which signpost to ongoing adaptation of SARS-CoV-2 to its new human host. Indeed, we do observe an enrichment for non-synonymous changes (80%) in our filtered sites. As such, our provided list (Table S5) contains candidates for mutations which may affect the phenotype of SARS-CoV-2 and virus-host interactions and which require ongoing monitoring. Conversely, the finding that 78% of the homoplasic mutations involve no polarity change could still reflect strong evolutionary constraints at these positions (Hughes, 2007; Yampolsky et al., 2005). The remaining non-neutral changes to amino acid properties at homoplasic sites may be enriched in candidates for functionally relevant adaptation and could warrant further experimental investigation.

One of the strongest homoplasies lies at site 11,083 in the SARS-CoV-2 genome in a region of Orf1a encoding Nsp6. This site passed our stringent filtering cirteria and was also present in our analysis of the SRA dataset (Table S7). Interestingly, this region overlaps a putative immunogenic peptide predicted to result in both CD4+ and CD8+ T-cell reactivity (Grifoni et al., 2020). More minor homoplasies amongst our top candidates, identified within Orf3a (Table S5), also map to a predicted CD4 T cell epitope. While the immune response to SARS-CoV-2 is poorly understood at this point, key roles for CD4 T cells, which activate B cells for antibody production, and cytotoxic CD8 T cells, which kill virus-infected cells, are known to be important in mediating clearance in respiratory viral infections (Kohlmeier and Woodland, 2009). Of note, we also identify a strong recurrent mutation in nucleotide position 21,575, corresponding to the SARS-CoV-2 spike protein (codon 5). While the spike protein is the known mediator of host-cell entry, our detected homoplasy falls outside of the N-terminal and receptor binding domains.

Our analyses presented here provide a snapshot in time of a rapidly changing situation based on available data. Although we have attempted to filter out homoplasies caused by sequencing error with stringent thresholds, and also used available short-read data to validate a subset of homoplasic sites in a smaller dataset, our analysis nevertheless remains reliant on the underlying quality of the publicly available assemblies. As such, it is possible that some results might be artefactual, and further investigation will be warranted as additional raw sequencing data becomes available.

However, given the crucial importance of identifying potential signatures of adaptation in SARS-CoV-2 for guiding ongoing development of vaccines and treatments, we have suggested what we believe to be a plausible approach and initial list in order to facilitate future work and interpretation of the observed patterns. More data continues to be made available, which will allow ongoing investigation by ourselves and others. We believe it is important to continue to monitor SARS-CoV-2 evolution in this way and to make the results available to the scientific community. In this context, we hope that the interactive web-application we provide will help identify key recurrent mutations in SARS-CoV-2 as they emerge and spread.

Author contributions

L.v.D., and F.B. conceived and designed the study; L.v.D., M.A, D.R L.P.S., C.E.F., L.O., C.J.O., J.P., C.C.S.T., F.A.T.B., and A.T.O analysed data and performed computational analyses; L.v.D., and F.B. wrote the paper with inputs from all co-authors.

Acknowledgments and funding

L.v.D and F.B. acknowledge financial support from the Newton Fund UK-China NSFC initiative (grant MR/P007597/1) and the BBSRC (equipment grant BB/R01356X/1). Computational analyses were performed on UCL Computer Science cluster and the South Green bioinformatics platform hosted on the CIRAD HPC cluster. We thank Jaspal Puri for insights and assistance on the development of the alignment visualisation tool and Nicholas McGranahan and Rachel Rosenthal for their comments on the manuscript. We additionally wish to acknowledge the very large number of scientists in originating and submitting labs who have readily made available SARS-CoV-2 assemblies to the research community.

Declaration of Competing Interest

The authors have no competing interests to declare.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2020.104351.

References

Anderson, R.M., May, R.M., 1991. Infectious Diseases of Humans. Dynamics and Control. Oxford University Press, Oxford.

Balloux, F., van Dorp, L., 2017. Q&a: what are pathogens, and what have they done to and for us? BMC Biol. 15, 6.

Cagliani, R., Forni, D., Clerici, M., Sironi, M., 2020. Computational inference of selection underlying the evolution of the novel coronavirus, SARS-CoV-2. J. Virol.

Crispell, J., Balaz, D., Gordon, S.V., 2019. HomoplasyFinder: a simple tool to identify homoplasies on a phylogeny. Microbial Genom. 5 (1), 10.

Dearlove, B.L., Lewitus, E., Bai, H., Li, Y., Reeves, D.B., Joyce, M.G., et al., 2020. A SARS-CoV-2 vaccine candidate would likely match all currently circulating strains. bioRxiv 2020.04.27.064774.

Didelot, X., Croucher, N.J., Bentley, S.D., Harris, S.R., Wilson, D.J., 2018. Bayesian inference of ancestral dates on bacterial phylogenetic trees. Nucleic Acids Res. 46 (22), 11.

Domingo-Calap, P., Schubert, B., Joly, M., Solis, M., Untrau, M., Carapito, R., et al., 2018. An unusually high substitution rate in transplant-associated BK polyomavirus in vivo is further concentrated in HLA-C-bound viral peptides. PLoS Pathog. 14 (10), 18.

Dudas, G., Rambaut, A., 2016. MERS-CoV recombination: implications about the reservoir and potential for adaptation. Virus Evol. 2 (1), 11.

Elbe, S., Buckland-Merrett, G., 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Chall. 1 (1), 33–46.

Fitch, W.M., 1971. Toward defining course of evolution - minimum change for a specific tree topology. Syst. Zool. 20 (4), 406–416.

Fraser, C., Donnelly, C.A., Cauchemez, S., Hanage, W.P., Van Kerkhove, M.D., Hollingsworth, T.D., et al., 2009. Pandemic potential of a strain of influenza A (H1N1): early findings. Science 1557–1561.

Giovanetti, M., Benvenuto, D., Angeletti, S., Ciccozzi, M., 2020. The first two cases of 2019-nCoV in Italy: where they come from? J. Med. Virol. 92 (5), 518–521.

Gordon, D.E., Jang, G.M., Bouhaddou, M., Xu, J., Obernier, K., O'Meara, M.J., et al., 2020. A SARS-CoV-2-human protein-protein interaction map reveals drug targets and potential drug-repurposing. Nature 2020.03.22.002386.

Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView version 4: a multiplatform graphical user Interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 27 (2), 221–224.

Grenfell, B.T., Pybus, O.G., Gog, J.R., Wood, J.L.N., Daly, J.M., Mumford, J.A., et al., 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303 (5656), 327–332.

Grifoni, A., Sidney, J., Zhang, Y., Scheuermann, R.H., Peters, B., Sette, A., 2020. A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. Cell Host Microbe 27 (4), 671–680 e2.

Hoang, D.T., Vinh, L.S., Flouri, T., Stamatakis, A., von Haeseler, A., Minh, B.Q., 2018. MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. BMC Evol. Biol. 18, 11.

Holmes, E.C., Dudas, G., Rambaut, A., Andersen, K.G., 2016. The evolution of Ebola virus: insights from the 2013-2016 epidemic. Nature 538 (7624), 193–200.

Hughes, A.L., 2007. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. Heredity. 99 (4), 364–373.

Janin, J., 1979. Surface and inside volumes in globular proteins. Nature 277 (5696), 491–492.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30 (4), 772–780.

Kohlmeier, J.E., Woodland, D.L., 2009. Immunity to respiratory viruses. Annu. Rev. Immunol. 27, 61–82.

Kozlov, A.M., Darriba, D., Flouri, T., Morel, B., Stamatakis, A., 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 35 (21), 4453–4455.

Lau, S.K.P., Woo, P.C.Y., Li, K.S.M., Huang, Y., Tsoi, H.W., Wong, B.H.L., et al., 2005. Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. Proc. Natl. Acad. Sci. U. S. A. 102 (39), 14040–14045.

Li, X.G., Wang, W., Zhao, X.F., Zai, J.J., Zhao, Q., Li, Y., et al., 2020. Transmission dynamics and evolutionary history of 2019-nCoV. J. Med. Virol. 92 (5), 501–511.

Memish, Z.A., Mishra, N., Olival, K.J., Fagbo, S.F., Kapoor, V., Epstein, J.H., et al., 2013. Middle East respiratory syndrome coronavirus in bats, Saudi Arabia. Emerg. Infect. Dis. 19 (11), 1819–1823.

Minskaia, E., Hertzig, T., Gorbalenya, A.E., Campanacci, V., Cambillau, C., Canard, B., et al., 2006. Discovery of an RNA virus 3 '-> 5 ' exoribonuclease that is critically involved in coronavirus RNA synthesis. Proc. Natl. Acad. Sci. U. S. A. 103 (13), 5108–5113.

Nielsen, R., Wang, H., Pipes, L., 2020. Synonymous mutations and the molecular evolution of SARS-Cov-2 origins. bioRxiv 2020.04.20.052019.

Rambaut, A., Pybus, O.G., Nelson, M.I., Viboud, C., Taubenberger, J.K., Holmes, E.C., 2008. The genomic and epidemiological dynamics of human influenza a virus. Nature 453 (7195), 615–U2.

Rambaut, A., Holmes, E.C., Hill, V., O'Toole, Á., McCrone, J., Ruis, C., et al., 2020. A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. bioRxiv 2020.04.17.046086.

Revell, L.J., 2012. Phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. 3 (2), 217–223.

Shaw, L.P., Wang, A.D., Dylus, D., Meier, M., Pogacnik, G., Dessimoz, C., et al., 2020. The phylogenetic range of bacterial and viral pathogens of vertebrates. bioRxiv 670315.

Shu, Y.L., McCauley, J., 2017. GISAID: global initiative on sharing all influenza data - from vision to reality. Eurosurveillance 22 (13), 2–4.

Smith, G.J.D., Vijaykrishna, D., Bahl, J., Lycett, S.J., Worobey, M., Pybus, O.G., et al., 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature. 459 (7250), 1122–U107.

Snijder, E.J., Bredenbeek, P.J., Dobbe, J.C., Thiel, V., Ziebuhr, J., Poon, L.L.M., et al., 2003. Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. J. Mol. Biol. 331 (5), 991–1004.

Volz, E.M., Frost, S.D.W., 2017. Scalable relaxed clock phylogenetic dating. Virus Evol. 3 (2), 9.

Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., et al., 2020. A new coronavirus associated with human respiratory disease in China. Nature 579 (7798) 265− +.

Yampolsky, L.Y., Kondrashov, F.A., Kondrashov, A.S., 2005. Distribution of the strength of selection against amino acid replacements in human proteins. Hum. Mol. Genet. 14 (21), 3191–3201.

Yu, G.C., Smith, D.K., Zhu, H.C., Guan, Y., Lam, T.T.Y., 2017. GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. 8 (1), 28–36.

Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., et al., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579 (7798) 270− +.

Recurrent mutations in SARS-CoV-2 genomes isolated from mink point to rapid host-adaptation

Lucy van Dorp[1], Cedric CS Tan[1], Su Datt Lam[2,3], Damien Richard[1,4], Christopher Owen[1], Dorothea Berchtold[1], Christine Orengo[3], François Balloux[1]

1) UCL Genetics Institute, University College London, United Kingdom.
2) Department of Applied Physics, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Malaysia.
3) Institute of Structural and Molecular Biology, University College London, United Kingdom.
4) UCL Institute of Child Health, University College London, United Kingdom.

Correspondence: lucy.dorp.12@ucl.ac.uk (Lucy van Dorp), f.balloux@ucl.ac.uk (François Balloux)

Keywords:

Zoonotic disease, SARS-CoV-2, COVID-19, mink, host-adaptation, spike protein

Abstract [234 words]

Severe acute respiratory coronavirus 2 (SARS-CoV-2), the agent of the ongoing COVID-19 pandemic, jumped into humans from an unknown animal reservoir in late 2019. In line with other coronaviruses, SARS-CoV-2 has the potential to infect a broad range of hosts. SARS-CoV-2 genomes have now been isolated from cats, dogs, lions, tigers and minks. SARS-CoV-2 seems to transmit particularly well in mink farms with outbreaks reported in Spain, Sweden, the Netherlands, Italy, the USA and Denmark. Genomic data from SARS-CoV-2 isolated from infected minks provides a natural case study of a secondary host jump of the virus, in this case from humans to animals, and occasionally back again. We screened published SARS-CoV-2 genomes isolated from minks for the presence of recurrent mutations common in mink but infrequent in SARS-CoV-2 genomes isolated from human infections. We identify 23 recurrent mutations including three nonsynonymous mutations in the Receptor Binding Domain of the SARS-CoV-2 spike protein that independently emerged at least four times but are only rarely observed in human lineages. The repeat emergence of mutations across phylogenetically distinct lineages of the virus isolated from minks points to ongoing adaptation of SARS-CoV-2 to a new host. The rapid acquisition and spread of SARS-CoV-2 mutations in minks suggests that if a similar phenomenon of host adaptation had occurred upon its jump into humans, those human-specific mutations would likely have reached fixation already before the first SARS-CoV-2 genomes were generated.

Data Summary

All genome assemblies considered in this manuscript are openly available on registration with GISAID (https://www.gisaid.org). Information on the included assemblies, including the accessions used in the global analysis are provided in **Tables S1-S2**.

Introduction

SARS-CoV-2, the agent of the COVID-19 pandemic, is part of the Coronaviridae family, whose members are widespread among wild populations of birds and mammals and tend to have fairly broad host ranges[1]. Seven coronavirus species are currently known to infect humans; SARS-CoV-1, SARS-CoV-2, MERS-CoV, and four endemic 'common cold' coronaviruses (HCoV-HKU1, HCoV-OC413, HCoV-229E, HCoV-NL63), all of which are of zoonotic origin. The original host of SARS-CoV-2 has not been identified to date, but bats are considered to be a plausible source[2,3]. Bats are also believed to represent the original host reservoir for NL63 and 229E[4–6], SARS-CoV-1[7–10] and MERS-CoV[11–14], whereas the zoonotic source for HCoV-HKU1 and HCoV-OC43 may be rodents, or possibly ruminants for the latter[15–17].

Upon an initial successful host jump, the ability of a pathogen to spread within its novel host population depends on its ability to transmit effectively between individuals. The four endemic HCoVs transmit readily in humans and are highly contagious particularly in children[18,19]. SARS-CoV-1 was probably less transmissible than SARS-CoV-2[20], which likely contributed to the 2003 SARS outbreak being comparably easy to control. MERS-CoV transmits poorly between humans and the repeat outbreaks it causes are primarily due to spillovers into humans from a reservoir it established in camelids in the Arabic Peninsula[21,22]. Finally, SARS-CoV-2 is highly transmissible in humans with an unusually high basic reproductive numbers (R_0) for a respiratory virus having been reported in several setting[23].

Transmissibility of a pathogen is expected to vary in different hosts and may often be limited in the early stages after the host jump. Such initial maladaptation may lead to a phase of intense natural selection when a pathogen acquires key mutations that maximize its transmission ability in the novel host. Efforts to identify mutations associated to SARS-CoV-2's transmissibility failed to identify obvious candidates for adaptation to its human host[24,25]. The only possible exception is the spike D614G mutation, whose role as a driver of transmission, or a neutral genetic marker of the otherwise successful SARS-CoV-2 B.1 lineage remains debated[24,26–28]. One plausible reason for the lack of mutations unambiguously associated to transmission in SARS-CoV-2 is that they may have emerged and spread before the virus was identified. Indeed, if by the time the earliest genomes were being sequenced, all key mutations associated to increased transmission in humans had already reached fixation, it would be impossible to identify those solely by analyzing SARS-CoV-2 genomes in circulation in humans.

Beyond humans and non-human primates, carnivores are predicted to be particularly susceptible to SARS-CoV-2 infection[29–34]. Genomic data is currently available for SARS-CoV-2 isolated from dogs, domestic cats, tigers, lions and minks[31,35–43]. Transmissibility may vary between susceptible carnivore hosts. For example domestic cats (*Felis catus*) are both more susceptible to SARS-CoV-2 and more infectious than dogs (*Canis lupus familiaris*)[41]. SARS-CoV-2 has also been shown to transmit well in experimental mustelid systems[35,39,44,45] and, in real-world settings, minks (*Neovison vison* and *Mustela lutreola*) are highly susceptible to SARS-CoV-2 circulating in humans. Major mink farm outbreaks have been reported in multiple countries since April 2020[37,38,43].

The generation and release of genomic data from SARS-CoV-2 outbreaks in farmed European (*Mustela lutreola*) and American minks (*Neovison vison*) offer opportunities to identify mutations emerging and spreading in mink farms following the host jump(s). The conditions in intensive farming settings may provide highly suitable environments for small fitness selective differentials between strains to overcome the effect of genetic drift, allowing for natural selection to further increase transmissibility. Particularly promising candidates for host adaptation are those mutations occurring recurrently in different mink SARS-CoV-2 lineages, and that are uncommon in human viruses. The secondary host jump from humans into minks offers a glimpse into the window of early viral host adaptation of SARS-CoV-2 to a new host that has likely been missed at the start of the COVID-19 pandemic.

We screened publicly available SARS-CoV-2 genomes isolated from minks for the presence of recurrent mutations. We identified seven nonsynonymous mutations independently arising at least three times as plausible candidates for adaptation to transmission in minks. All seven candidate mutations are at low frequency in large repositories of SARS-CoV-2 strains circulating in humans. We note in particular three recurrent nonsynonymous mutations in the receptor-binding domain of the SARS-CoV-2 spike protein (S-protein), the region essential for binding to host Angiotensin-converting enzyme 2 (ACE2) receptors allowing cell entry[46,47]. We computationally predict the role of candidates in this region on modulating the binding stability of the human and mink spike protein ACE2 complex. Our results highlight the rapid and repeat emergence of mutations in the spike protein, across phylogenetically distinct lineages of the virus isolated from minks in Denmark and the Netherlands, and point to rapid adaptation of SARS-CoV-2 to a new host.

Materials and Methods

Alignment of mink SARS-CoV-2

All publicly available genome assemblies of SARS-CoV-2 isolated from minks were downloaded from both GISAID[48,49] (https://www.gisaid.org) and NCBI (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/) as of 14th October 2020, which included published genomic data associated to two studies of SARS-CoV-2 in minks in the Netherlands[37,38]. A full list of the 239 considered accessions is provided in **Table S1** which includes 227 genome assemblies sampled in the Netherlands and 12 genome assemblies sampled in Denmark covering the period of the 24th of April 2020 to 16th of September 2020. Some of the genomes were flagged as 'low quality' (commented in **Table S1**) but were included in the initial analaysis. All genome assemblies were profile aligned to the SARS-CoV-2 reference genome Wuhan-Hu-1 (NC_045512.2) using MAFFT v7.205[50]. SNPs flagged as putative sequencing errors were masked (a full list of 'masked' sites is available at https://github.com/W-L/ProblematicSites_SARS-CoV2/blob/master/problematic_sites_sarsCov2.vcf, accessed 14/10/2020)[51]. Wuhan-Hu-1 was retained to root the tree. A maximum likelihood phylogenetic tree was built across the alignment (**Figure S1**) using RaxML[52] run via the Augur (https://github.com/nextstrain/augur) tree implementation (**Figure 1**). We informally estimated the substitution rate over the mink SARS-CoV-2 alignment by computing the root-to-tip temporal regression implemented in BactDating v1.0.1[53] (**Figure S2**).

Global context of mink SARS-CoV-2 infections

To place SARS-CoV-2 genomes isolated from minks into a global context we downloaded 56,803 high quality assemblies (high coverage, >29,700bp and with a fraction of 'N' nucleotides <5%) from the worldwide diversity of human SARS-CoV-2 available on GISAID[48,49] on 25/08/2020. All animal strains were removed as well as samples flagged by NextStrain as 'exclude' (https://github.com/nextstrain/ncov/blob/master/defaults/exclude.txt as of 25/08/2020). Following alignment and masking as described above, this left 54,793 assemblies for downstream analysis to which we added the mink SARS-CoV-2 alignment. A full metadata table, list of acknowledgements and exclusions is provided in **Table S2**. We constructured a maximum likelihood phylogenetic tree over the 55,030 included genomes using IQ-TREE v2.1.0 Covid release[54] specifying the fast mode. Following construction of the tree, 57 long branch phylogenetic outliers were detected by TreeShrink[55] and were subsequently removed, together with two low-quality mink SARS-CoV-2 assemblies which

were also considered phylogenetic outliers (EPI_ISL_577816 and EPI_ISL_577819). Trees were queried and plotted using the R packages Ape v5.4[56] and ggtree v1.16.6[57] (see **Figure 1**).

Mutation analysis

All variable positions in the coding regions of the genome were identified and annotated for synonymous or nonsynonymous status (**Figure S3-S4**). This was done by retrieving the amino acid changes corresponding to all SNPs at these positions using a custom Biopython (v1.76) script (https://github.com/cednotsed/mink_homoplasies/blob/main/dnds/snp_to_sav_parser.py) with annotations reported in **Table S3**. The Orf coordinates used (including the Orf1ab ribosomal frameshift site) were obtained from the associated metadata for Wuhan-Hu-1 (NC_045512.2). Assemblies were also uploaded to CoVSurver (https://www.gisaid.org/epiflu-applications/covsurver-mutations-app/) to report the prevalence of mutations and indels relative to SARS-CoV-2 assemblies available on the GISAID database[48,49].

The frequencies of each type of mutation in the human and mink SARS-CoV-2 alignments were calculated using the base.freq function from the R package Ape v5.4[56] (**Figure S5**). We tested whether the mutational frequencies in the human and mink genomes differed using a Monte Carlo simulation of the $\chi 2$ statistic with fixed margins (2000 iterations)[58,59]. This was implemented using the chisq.test function in R with the simulate.p.value flag. CpG dinucleotide frequencies for both SARS-CoV-2 alignments of genomes isolated from human and mink were also calculated using a custom R script (https://github.com/cednotsed/mink_homoplasies/blob/main/CpG/plot_CpG.R) (**Figure S6**). A Wilcoxon rank sum test implemented using the wilcox.test function in R was used to test if the distribution of CpG dinucleotide frequencies differed between the two datasets.

Identification of recurrent mutations

We screened for the presence of recurrent mutations in the mink SARS-CoV-2 masked alignment using HomoplasyFinder v0.0.0.9[60], as described in our previous work[61,62]. HomoplasyFinder employs the method first described by Fitch[63], providing, for each site, the site specific consistency index and the minimum number of independent emergences in the phylogenetic tree. All nucleotide positions with a consistency index <0.5 are considered homoplasic (**Figure 2, Figure S7, Table S4**). Sites identified as homoplasic were screened against the global dataset of 54,793 human SARS-CoV-2 (**Figure S8**). In addition amino acid

replacements were identified in human samples in the CoV-GLUE[64] repository (accessed 16th November 2020, last update from GISAID 9th November 2020), which provides frequently updated screens against all assemblies shared to GISAID[48,49] (**Figure 2, Table S4**).

Structural data used for the analysis

The structure of the SARS-CoV-2 spike protein, reference strain, bound to human ACE2 has been solved at 2.45Å resolution[65] (PDB ID 6M0J). We visualised this structure using PyMOL v2.4.1[66]. We used this as the template to model the structures of the American mink (*Neovision vision*) ACE2 bound to the SARS-CoV-2 reference (Wuhan-Hu-1) and the mink ACE2 protein bound to versions of SARS-CoV-2 mutated for three candidate sites in the RBD. We also built a model of the human ACE2 bound to the mutated SARS-CoV-2.

We generated query–template alignments using HH-suite[67] and predicted 3D models using MODELLER v.9.24[68]. We used the 'very_slow' schedule for model refinement to optimise the geometry of the complex and interface. We generated 10 models for each S-protein:ACE2 complex and selected the model with the lowest nDOPE[69] score, which reflects the quality of the model. Positive scores are likely to be poor models, while scores lower than -1 are likely to be native-like. The sequence similarity of the human ACE2 and the mink ACE2 is fairly high (83% amino-acid sequence identity) and all generated models were of high quality (nDOPE < -1).

Measuring changes in the stability of the S-protein:ACE2 complex following mutation

We calculated the change in stability of the S-protein:ACE2 complex using two independent methods. The first, HADDOCK[70], is one of the top-performing protein-protein docking servers in the CAPRI competition[71]. The HADDOCK scoring function uses linear combination of various energies, van der waals, electrostatics and desolvation. We employed HADDOCK (v2.4 web server) to score the complexes (**Figure S9**). We also calculated the change in the stability of the S-protein:ACE2 complex using mCSM-PPI2[72] (**Figure S10**). This assigns a graph-based signature vector to each mutation, which is then used within machine learning models to predict the binding energy. The signature vector is based upon atom-distance patterns in the wild-type protein, pharmacophore information and available experimental information, evolutionary information and energetic terms. We used the mCSM-PPI2 server (http://biosig.unimelb.edu.au/mcsm_ppi2/) for the simulations. These methods were used

because we found in a previous study that they reported stability changes, following mutations in the S-protein:ACE2 complex, that correlated well with the available *in vivo* and *in vitro* experimental data on susceptibility to infection[29].

In particular we performed the following calculations:

1. Using the model of the mink ACE2:SARS-CoV-2 reference complex we mutated the target residue to those RBD candidates identified in mink SARS-CoV-2.
2. Using the model of the mink ACE2:SARS-CoV-2 mink complex we mutated the target residues to the human SARS-CoV-2 reference strain.
3. Using the structure of the human ACE2:SARS-CoV-2 reference we mutated the target residue to those RBD candidates identified in mink SARS-CoV-2.
4. Using the model of the human ACE2:SARS-CoV-2 mink complex we mutated the target residue to the human SARS-CoV-2 reference strain.

For HADDOCK, a value that is more negative than for the reference Wuhan-Hu-1 S-protein: ACE2 complex suggests stabilisation of the complex. Whilst for mCSM-PPI-2 negative ΔΔG values reflect destabilisation of the complex by the mutation and positive values reflect stabilisation of the complex.

We also evaluated structural changes for all combinations of RBD mutations and receptor complexes (**Figures S11-S13**).

Results

Genomic diversity in mink SARS-CoV-2

At the time of writing (6th November 2020) 239 SARS-CoV-2 genome assemblies isolated from minks were publicly available on GISAID[48,49] and NCBI (**Table S1**) spanning a sampling period of the 24th of April 2020 to the 16th of September 2020. An alignment across all mink SARS-CoV-2 comprised 1073 Single Nucleotide Polymorphisms (SNPs) (**Figure S1**) with a mean pairwise distance of 6.5 (range: 4.1 – 11.7) mutations between any two genomes. No genome deviated from the reference genome, Wuhan-Hu-1 (NC_045512.2), by more than 24 mutations and the alignment exhibited a highly significant temporal signal (**Figure S2**). Dynamic assignment of assemblies to lineages[73] (run 6th November 2020) placed mink SARS-CoV-2 into seven different SARS-CoV-2 lineages. All 12 genomes from mink farms in

Denmark fell into a single lineage: B.1.1. Both lineage assignments and phylogenetic placement of mink SARS-CoV-2 in a large global phylogeny of human isolates are consistent with multiple zoonotic jumps of SARS-CoV-2 from human to mink hosts seeding local farm-related outbreaks[37,38] (**Figure 1, Tables S1-S2**).

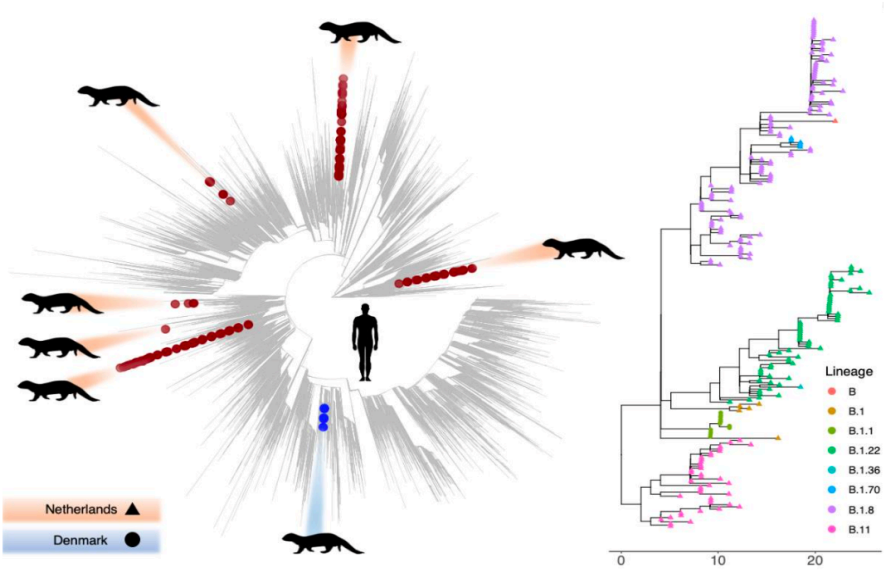


Figure 1: *(left) SARS-CoV-2 radial phylogenetic tree providing the relationships amongst SARS-CoV-2 isolated from humans and minks rooted on Wuhan-Hu-1. Genomes isolated from minks are shown with large tip symbols and coloured blue for those from Denmark and brown for those from the Netherlands. (right) Phylogenetic tree providing the relationships among SARS-CoV-2 genomes from minks with lineage assignments provided by tip colours and location by tip symbols (circle Denmark; triangle Netherlands). Full lineage assignments and a list of included genomes are provided in* ***Table S1*** *and* ***Table S2****.*

Across the mink SARS-CoV-2 alignment we observed a ratio of nonsynonymous to synonymous changes of 1.84, in line with the previous estimate of 1.88 for SARS-COV-2 lineages circulating in humans[24]. A full list of all identified mutations is provided in **Table S3** with nonsynonymous mutations displayed in **Figures S3-S4**. As previously reported for human SARS-CoV-2, we observe marked compositional asymmetries in mink SARS-CoV-2 likely deriving from the mutational action of host RNA-editing mechanisms[24,74,75]. For example, 37% of all nonsynonymous mutations represent C→U changes (**Figure S5**). This results in 87% of

all possible nonsynonymous mutations involving a C being C→U changes, in line with C→U changes also representing by far the most common mutations for SARS-CoV-2 strains circulating in humans[24,74,75]. Though, the overall pattern of nucleotide substitutions is statistically significantly different between strains circulating in humans and minks ($p < 0.001$), with in particular a deficit of G→C and an excess of C→A mutations in minks **(Figure S5)**. SARS-CoV-2 genomes isolated from minks also showed a stronger depletion of CpG sites ($p = 2.2 \times 10^{-16}$) relative to their counterparts circulating in humans **(Figure S6)**.

Recurrent mutations in mink SARS-CoV-2

Across the masked alignment we detect 23 mutations that have appeared independently at least twice in SARS-CoV2 circulating in minks, corresponding to a consistency index (CI) <0.5[60]. Of those 23 mutations, 16 comprise nonsynonymous and seven synonymous changes (**Table S4**). Nonsynonymous mutations which have emerged independently at least three times include one within Orf1ab (12,795 - G4177E), three within Orf3a (25,936 – H182Y; 26,047 – L219V; 26078 – T229I) and three within the spike protein (22,920 – Y453F; 23,018 – F486L; 23,064 – N501T). Of note the latter three recurrent mutations in the spike protein correspond to residue changes in the SARS-CoV-2 spike RBD (**Figure 2, Figure S4, Figure S7**), which may be a particularly important region for adaptation of SARS-CoV-2 to host receptors[47]. In particular, we infer five independent emergences of Y453F across four lineages, five emergences of F486L across three lineages and four emergences of N501T across three lineages.

Mutations specifically increasing transmission in minks can be expected to have emerged only a limited number of times in SARS-CoV-2 strains circulating in humans and to have remained at low frequency in the human SARS-CoV-2 population. Using a phylogenetic dataset of >54,000 human SARS-CoV-2 genomes, we identify some overlap between sites homoplasic in minks and in humans (**Figure S8**). This suggests that mutations having emerged recurrently in viruses circulating in minks may not all reflect host adaptation but could also be the result of mutations induced by vertebrate APOBEC proteins[24,74,75]. This is in line with the large proportion of homoplasies involving C→T changes (12/23) consistent with know patterns of hypermutation in vertebrates[74] (**Figure 2**, **Table S4**). As such, we expect strong candidate mutations for SARS-CoV-2 host adaptation to minks to be at low frequency in SARS-CoV-2 circulating in humans.

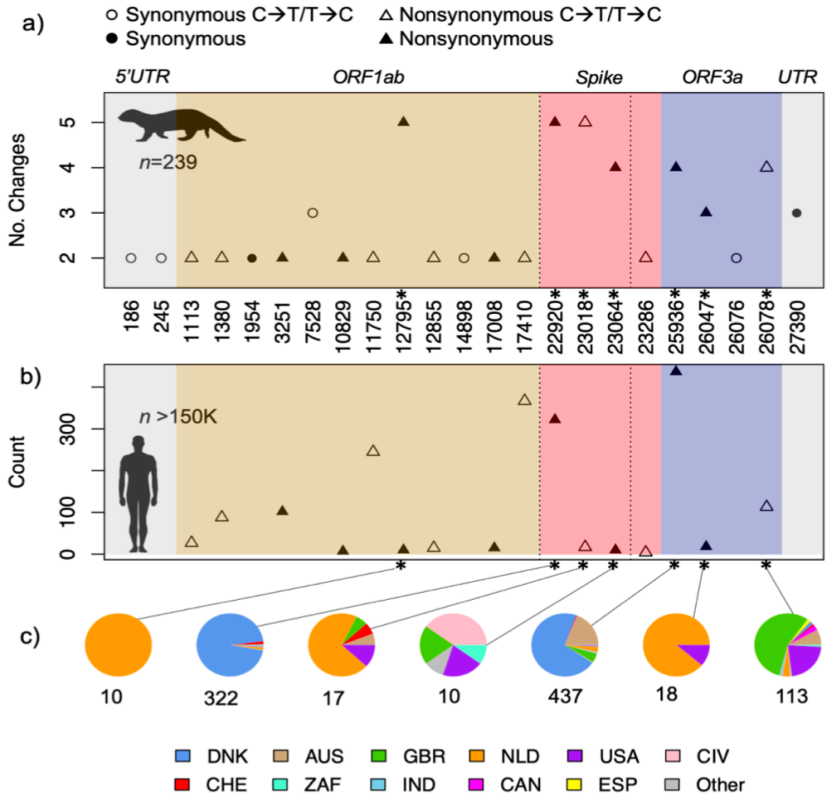


Figure 2: *a) 23 recurrent mutations identified in the mink SARS-CoV-2 alignment. Y-axis provides the minimum number of change points detected by HomoplasyFinder. Colours provide the genome annotations as given at top with different categories of mutation denoted by the symbols. Recurrent mutations in the spike RBD are bounded by vertical dashed lines. b) Number (y-axis) of human associated SARS-CoV-2 genomes annotated as carrying nonsynonymous mutations recurrent in mink associated SARS-CoV-2 as of 9th November 2020. c) Pies providing the geographic distribution of human SARS-CoV-2 assemblies carrying each of the seven candidate changes (corresponding to nonsynonymous mutations with at least three emergences in minks). Legend at bottom provides the country assignments: AUS – Australia, CHE – Switzerland, CAN – Canada, CIV – Ivory Coast, DEN – Denmark, ESP – Spain, GBR – Great Britain, IND – India, NLD – Netherlands, RUS – Russia, USA – United States of America, ZAF – South Africa. A full list is provided in* ***Table S4****.*

Following a screen of amino acid replacements in CoV-GLUE[64] (accessed 16th November 2020) we found that all seven nonsynonymous mutations that have independently emerged at least three times in strains circulating in minks are carried by fewer than ~0.3% of the SARS-CoV-2 genomes isolated from humans available on large publicly available sequencing repositories (**Figure 2, Table S4**). Of the mink homoplasies located in the spike protein, two derive from A→T (nucleotide position 22920 – Y453F) and A→C (nucleotide position 23018 – F486L) changes while one corresponds to a T→C change (nucleotide position 23064 – N501T). The recurrent emergence of these mutations in phylogenetically distant lineages, and their relative scarcity in SARS-CoV-2 samples circulating in humans, further supports these as strong candidates for ongoing host-adaptation of the virus to transmission in minks. Indeed, many of the cases identified in humans seem putatively linked to mink farm outbreaks in the Netherlands and Denmark[37,38,76] (**Figure 2c**).

Predicted impact of recurrent mutations in the spike protein

Having observed the presence of three candidate mutations falling within the SARS-CoV-2 spike RBD we considered their role in receptor binding affinity as putative sites of adaptation to a mink host. The spike RBD (codon positions 319-541[77]) provides a critical region for SARS-CoV-2 to attach to host cells via docking to ACE2 receptors, thereby allowing subsequent SARS-CoV-2 entry into host cells and eventual replication[46,47]. This also makes the RBD the target of neutralising antibodies[78–81], i.e. antibodies preventing cellular infection upon prior exposure to the pathogen or a vaccine. Specific residues within the RBD have been identified as critical for receptor binding[29,82,83], with potential to modulate both infectivity and antigenicity[77]. All three recurrent spike RBD mutations (Y453F, F486L and N501T) suggested by the phylogenetic analyses are in residues directly involved in contacts in the S-protein:ACE2 interface and are therefore relevant to the binding and stability of the complex (**Figure 3**).

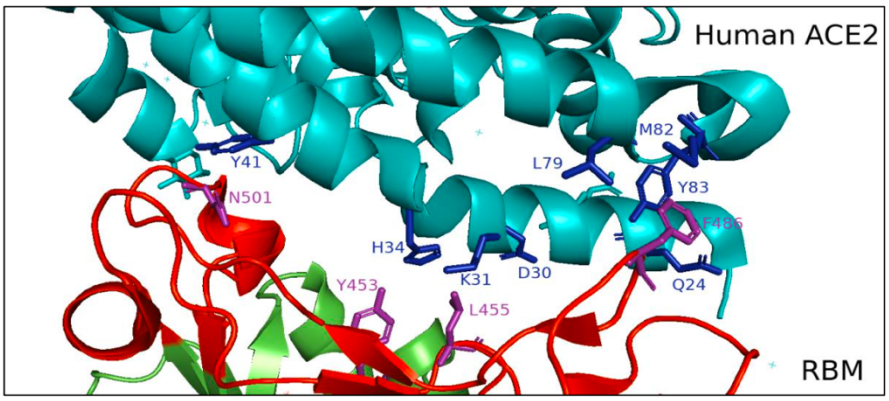


Figure 3: *Protein structure of the receptor binding motif (RBM) of the receptor binding domain (RBD) in complex with human ACE2. Human ACE2, the RBM and the rest of the receptor binding domain are shown in teal, red and green respectively. The wild-type amino acid residues for the recurrent nonsynonymous mutations Y453F, F486L, N501T are shown in purple. The ACE2 residues reported previously to be interacting with the RBM residues L455 (in close proximity to F453), F486 and N501 are shown in blue. The Protein Data Bank (PDB) code for the SARS-CoV-2 RBM–ACE2 complex is 6M0J. This figure was rendered using PyMOL (v2.4.1).*

We used the HADDOCK[70] and mCSM-PPI2[72] protocols to analyse the change in stability of the mink SARS-CoV-2 S-protein:ACE2 complex for the three recurrent spike RBD mutations. To analyse the changes in stability we used 3D-models of the mink ACE2 bound to SARS-CoV-2 mink strain and then mutated the target residue to that found in human-associated reference SARS-CoV-2. We used this approach as previous work showed that it gave results that correlated well with experimental data on susceptibility to infection[29]. However, we got similar results when we performed the calculation by using the complex of the mink ACE2 bound to SARS-CoV-2 Wuhan-Hu-1 and mutated the target SARS-CoV-2 residue to those observed in minks (**Figure S9-S10**). The results indicate marginal changes in the stability of the complex for these mutations and are somewhat conflicting between the methods, but none of the values reported is expected to significantly stabilise or destabilise the complex.

Structural analyses

In light of the small estimated changes in binding energy associated with complex stabilities, additional structural analyses were performed to gain further insights into whether the mutations were likely to affect complex stability. The first of the candidate residues 453

(nucleotide position 22920) lies close to the centre of the S-protein:ACE2 interface. In Wuhan-Hu-1 the tyrosine at this position interacts with H34 in human ACE2 enabling the formation of a polar hydrogen bond which would stabilise the complex. In the American mink ACE2, position 34 has a tyrosine residue which would lead to a clash with the polar tyrosine 453 for Wuhan-Hu-1 SARS-CoV-2. Therefore the emergence of mutation of Y453 to a phenyalanine, observed in 47 mink associated SARS-CoV-2 (**Table S3**), could be favourable as it reduces the clash of polar groups (see **Figure S11**).

The SARS-CoV-2 spike residue 486 (nt position 23018) has been identified as one of the most important locations – 'the receptor binding motif' - for human ACE2 binding to the spike protein[82,83] and lies within a hydrophobic pocket (**Figure S12**). It has been shown that SARS-CoV-2 exploits this pocket better than SARS-CoV-1, as the phenylalanine residue allows for pi stacking of the aromatic rings with Y83 in human ACE2 which enhances affinity[65,82,83]. Our previous analysis[29] showed the Y83 residue to be highly conserved across human and other animals including the American mink. While we note that this residue change arises from a T→C transition (**Figure 2**), this polymorphism has appeared at least five times and was observed in 96 mink associated SARS-CoV-2. Mutation of phenylalanine to leucine replaces the large aromatic ring with an aliphatic leucine which is not involved in any substantive interactions with the Y83 (see **Figure S12**).

Our final candidate, N501T (nt position 23064) is only observed in five mink SARS-CoV-2 but has independently emerged four times (**Table S4**). This residue change has been observed in only ten human SARS-CoV-2 assemblies to date (**Figure 2**). The recurrent mutation leads to the replacement of the asparagine by a threonine residue in the mink associated strain (**Figure S13**). This gives rise to a hydrogen bond to lysine 353 which may contribute to the small increase in stability reported by mCSM-PPI2 (**Figure S10**).

Discussion

The COVID-19 pandemic is understood to have been caused by a unique host jump into humans from a single yet-undescribed zoonotic source in the latter half of 2019 most likely in China[2,3,61]. This created a situation where the entire genetic diversity of the SARS-CoV-2 population was initially negligible despite the virus having rapidly swept across the world. Genetic diversity has been building up since the start of the pandemic through the acquisition of new mutations, but is still limited relative to other RNA viruses with broad distributions. Conversely, the secondary host jump into minks involved multiple independent secondary transmissions into mink farms in different countries. Whilst the data we analysed represents only a subset of infected mink farms from just two countries, the Netherlands and Denmark, we already identify a minimum of seven independent host jumps from humans to minks, involving strains representative of essentially the entire SARS-CoV-2 lineage diversity in circulation (**Figure 1**).

The secondary transmissions of SARS-CoV-2 from humans into minks provides a set of 'natural experiments' to identify mutations involved in the adaptation of the virus to a novel host. By analysing SARS-CoV-2 isolated from minks, we recovered 23 mutations having independently emerged multiple times. By restricting this set to the nonsynonymous mutations that have appeared at least three times in mink, we identify seven variants that are strong candidates for host adaptation to minks (**Figure 2, Table S4**). These seven candidates comprise a recurrent change in nsp9 of Orf1ab, a region involved in mediating viral replication[84], as well as the repeat emergence of three mutations in Orf3a, a protein thought to play an important role in triggering host inflammatory response[85,86].

We particularly focus on the three candidates that have emerged independently in minks at least four times in the RBD of the SARS-CoV-2 spike protein, a region vital for determining host-range[47] (**Figure 2, Figure S4**). We detect at least five independent emergences of Y453F across four phylogenetic lineages, five emergences of F486L across three lineages and four emergences of N501T across three lineages in minks. It is noteworthy that all three mutations are not, or only marginally homoplasic, in humans and all three have also been observed at low frequency in strains circulating in human COVID-19 infections which have been intensively sampled; ~150,000 high quality genome assemblies are available on GISAID at the time of writing (**Table S4, Figure 2**).

Quantifying the consequences of these candidate RBD mutations on the binding affinity of the SARS-CoV-2 spike protein to human and mink ACE2 receptors we predict only subtle changes in S-protein:ACE2 complex stability for Y453F and F486L (**Figure S9-S10**). Consistently, characterisation of mutated variants in the RBD region suggest the region has a high tolerance to mutation in the context of receptor binding[87] and we note that previous studies have shown that animals are susceptible to infection by SARS-CoV2 despite much higher destabilisation of the complex than reported for these mutants[29,30,88]. A published approach based on fluorescent detection of ACE2 binding reports negative binding constants for F486L though suggests a positive effect for Y453F and N501T to human ACE2[87] (https://jbloomlab.github.io/SARS-CoV-2-RBD_DMS/ accessed 12th November 2020).

We also identify structural support for a marginal increase in stability of the mink complex with the N501T mutation relative to wild type (**Figure S10, Figure S13**). In our dataset we detect the presence of this mutation in five mink SARS-CoV-2 though this corresponds to at least four independent emergences, with N501T being exceptionally rare in human infections (**Figure 2**). It may also be relevant that a different residue change in this position (N501Y) has been proposed as a mechanism of host adaption in mice infected with SARS-CoV-2[89] suggesting the broader role of residue changes at 501 as relevant affinity enhancing mutations[87].

In addition to adaptation to a novel host through acquisition of new variants, RNA viruses are also under pressure of their hosts' immune system which comprises antiviral mechanisms such as APOBEC proteins inducing hypermutation at specific sites in their genome[74,75]. While C→U changes, most likely induced by proteins from the APOBEC family, represent the majority of mutations in SARS-CoV-2 circulating both in minks and humans, we also detected subtle shifts in mutational patterns likely due to differences in the innate immune system of humans and minks **(Figure S5-S6)**. The widespread depletion of CpG sites in RNA viruses can also be explained by another defense mechanism in the vertebrate innate immune system, in the form of Zinc-Finger Antiviral Proteins (ZAPs), which target viral RNA for exosome-mediated degradation[90,91]. There have been earlier suggestions that ZAPs from different vertebrate host taxa may exert varying levels of CpG depletion pressure[92,93]. In this context, it is an intriguing observation that SARS-CoV-2 genomes isolated from minks displayed a moderate but statistically highly significant depletion of CpG sites (p = 2.2 x 10^{-16}) relative to their counterparts circulating in humans **(Figure S6)**.

A combination of biological and epidemiological factors make mink farms highly susceptible to SARS-CoV-2 outbreaks with the risk of subsequent transmission back into humans[37,38,94]. As such, mink farms represent reservoirs for the virus which can greatly complicate the containment of COVID-19. At the time of writing a further 214 mink associated human COVID-19 infections were reported in the North Jutland region of Denmark, highlighting the risk of spill-overs from animal reservoirs (announced 5th November 2020, working paper available https://files.ssi.dk/Mink-cluster-5-short-report_AFO2 accessed 12th November 2020)[76]. A phylogenetic grouping of the virus termed 'Cluster 5' circulating in North Jutland has attracted particular attention. This cluster is reported to carry our Y453F candidate mutation, together with three other mutations in the spike protein falling outside the RBD (del69-70, I692V, M1229I), none of which we identify as mink-adapted. The cluster 5 lineage is known to have infected up to 12 people and its spread led to the Danish Government considering closing all mink farms in the country. The putative impact on antigen-escape of the Y453F mutation remains under study, though appears moderate[76].

Our work points to the frequent emergence of adaptive mutations in minks which may be retained during anthroponotic spillovers, at least transiently so. To date, the vast majority of mutations that have been observed in SARS-CoV-2 appear selectively neutral, or even deleterious in humans[24,95] and those which we identify as strong candidates for host-adaptation to minks have remained at low frequency globally in human SARS-CoV-2 to date (**Figure 2**). There is no *a priori* reason to expect that mutations adaptive in minks will lead to any marked change in the dynamic of human COVID-19 infections. Indeed, SARS-CoV-2 lineages carrying candidate mutations adaptive to transmission in minks have been sequenced since the first outbreaks of SARS-CoV-2 on mink farms in April 2020, with the first mink SARS-CoV-2 sample carrying Y453F dating back to the 24th April 2020.

The low prevalence of mink-adapted candidate mutations in strains in circulation in humans to date (November 2020) suggests they are not expected to increase transmission of the virus in humans. However, mutations located within the RBD of the SARS-CoV-2 spike protein do warrant careful monitoring. Genetic variation in the RBD region is common and studies of related Sarbecoviruses have previously identified signals consistent with complex selection histories in this region[2,10,96,97]. The RBD is the most immunodominant region of the SARS-CoV-2 genome and any mutation in this region therefore has potential implications for

antigenic response[78–81]. The spike RBD is also the target of antibody based therapies and vaccines[98–101]. In this context, it may seem concerning that Y453F has been flagged as as a possible vaccine-escape mutation[102]. Though our work suggests this mutation is an adaptive change specifically enhancing transmission in minks. Other RBD variants, already found at higher frequency in human SARS-CoV-2, represent stronger candidates for vaccine-escape mutations[103,104]. Once vaccines will be deployed, careful characterisation and tracking of the frequency of such mutations will be essential to inform if and when vaccines will need to be redesigned. This critical effort will be greatly enabled by open genomic data sharing platforms[48,49,64,105].

References

1. Shaw, L. P. *et al.* The phylogenetic range of bacterial and viral pathogens of vertebrates. *Mol. Ecol.* mec.15463 (2020). doi:10.1111/mec.15463
2. Boni, M. F. *et al.* Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol.* **5**, 1408–1417 (2020).
3. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
4. Corman, V. M. *et al.* Evidence for an Ancestral Association of Human Coronavirus 229E with Bats. *J. Virol.* **89**, 11858–70 (2015).
5. Huynh, J. *et al.* Evidence supporting a zoonotic origin of human coronavirus strain NL63. *J. Virol.* **86**, 12816–25 (2012).
6. Tao, Y. *et al.* Surveillance of Bat Coronaviruses in Kenya Identifies Relatives of Human Coronaviruses NL63 and 229E and Their Recombination History. *J. Virol.* **91**, (2017).
7. Ge, X.-Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).
8. Li, W. *et al.* Bats Are Natural Reservoirs of SARS-Like Coronaviruses. *Science (80-.).* **310**, 676–679 (2005).
9. Lau, S. K. P. *et al.* Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *Proc. Natl. Acad. Sci.* **102**, 14040–14045 (2005).
10. Hu, B. *et al.* Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLOS Pathog.* **13**, e1006698 (2017).
11. Corman, V. M. *et al.* Rooting the phylogenetic tree of middle East respiratory syndrome coronavirus by characterization of a conspecific virus from an African bat. *J. Virol.* **88**, 11297–303 (2014).
12. Anthony, S. J. *et al.* Further Evidence for Bats as the Evolutionary Source of Middle East Respiratory Syndrome Coronavirus. *MBio* **8**, (2017).
13. Lau, S. K. P. *et al.* Receptor Usage of a Novel Bat Lineage C Betacoronavirus Reveals Evolution of Middle East Respiratory Syndrome-Related Coronavirus Spike Proteins for Human Dipeptidyl Peptidase 4 Binding. *J. Infect. Dis.* **218**, 197–207 (2018).
14. Wang, Q. *et al.* Bat Origins of MERS-CoV Supported by Bat Coronavirus HKU4 Usage of Human Receptor CD26. *Cell Host Microbe* **16**, 328–337 (2014).
15. Su, S. *et al.* Epidemiology, Genetic Recombination, and Pathogenesis of

Coronaviruses. *Trends Microbiol.* **24**, 490 (2016).

16. Vijgen, L. *et al.* Complete genomic sequence of human coronavirus OC43: molecular clock analysis suggests a relatively recent zoonotic coronavirus transmission event. *J. Virol.* **79**, 1595–604 (2005).
17. Hogue, B. G., King, B. & Brian, D. A. Antigenic relationships among proteins of bovine coronavirus, human respiratory coronavirus OC43, and mouse hepatitis coronavirus A59. *J. Virol.* **51**, 384–8 (1984).
18. Nickbakhsh, S. *et al.* Epidemiology of Seasonal Coronaviruses: Establishing the Context for the Emergence of Coronavirus Disease 2019. *J. Infect. Dis.* **222**, 17–25 (2020).
19. Dijkman, R. *et al.* Human coronavirus NL63 and 229E seroconversion in children. *J. Clin. Microbiol.* **46**, 2368–73 (2008).
20. Cevik, M., Kuppalli, K., Kindrachuk, J. & Peiris, M. Virology, transmission, and pathogenesis of SARS-CoV-2. *BMJ* **371**, m3862 (2020).
21. Reusken, C. B., Raj, V. S., Koopmans, M. P. & Haagmans, B. L. Cross host transmission in the emergence of MERS coronavirus. *Curr. Opin. Virol.* **16**, 55–62 (2016).
22. Alagaili, A. N. *et al.* Middle East respiratory syndrome coronavirus infection in dromedary camels in Saudi Arabia. *MBio* **5**, e00884-14 (2014).
23. Liu, Y., Gayle, A. A., Wilder-Smith, A. & Rocklöv, J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *J. Travel Med.* **27**, (2020).
24. van Dorp, L. *et al.* No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *bioRxiv* 2020.05.21.108506 (2020). doi:10.1101/2020.05.21.108506
25. MacLean, O. A. *et al.* Natural selection in the evolution of SARS-CoV-2 in bats, not humans, created a highly capable human pathogen. *bioRxiv* 2020.05.28.122366 (2020). doi:10.1101/2020.05.28.122366
26. Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812-827.e19 (2020).
27. Volz, E. M. *et al.* Evaluating the effects of SARS-CoV-2 Spike mutation D614G on transmissibility and pathogenicity. *medRxiv* 2020.07.31.20166082 (2020). doi:10.1101/2020.07.31.20166082
28. Grubaugh, N. D., Hanage, W. P. & Rasmussen, A. L. Making Sense of Mutation: What

D614G Means for the COVID-19 Pandemic Remains Unclear. (2020). doi:10.1016/j.cell.2020.06.040

29. Lam, S. D. D. *et al.* SARS-CoV-2 spike protein predicted to form complexes with host receptor protein orthologues from a broad range of mammals. *Sci. Rep.* **10**, 16471 (2020).
30. Damas, J. *et al.* Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 22311–22322 (2020).
31. Santini, J. M. & Edwards, S. J. L. Host range of SARS-CoV-2 and implications for public health. *The Lancet Microbe* **1**, e141–e142 (2020).
32. Melin, A. D., Janiak, M. C., Marrone, F., Arora, P. S. & Higham, J. P. Comparative ACE2 variation and primate COVID-19 risk. *Commun. Biol.* **3**, 641 (2020).
33. Infection with Novel Coronavirus (SARS-CoV-2) Causes Pneumonia in the Rhesus Macaques. (2020). doi:10.21203/RS.2.25200/V1
34. Rockx, B. *et al.* Comparative pathogenesis of COVID-19, MERS, and SARS in a nonhuman primate model. *Science* **368**, 1012–1015 (2020).
35. Shi, J. *et al.* Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS-coronavirus 2. *Science* **368**, 1016–1020 (2020).
36. McAloose, D. *et al.* From People to Panthera: Natural SARS-CoV-2 Infection in Tigers and Lions at the Bronx Zoo. *MBio* **11**, (2020).
37. Munnink, B. B. O. *et al.* Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science (80-.).* (2020). doi:10.1126/SCIENCE.ABE5901
38. Oreshkova, N. *et al.* SARS-CoV-2 infection in farmed minks, the Netherlands, April and May 2020. *Eurosurveillance* **25**, 2001005 (2020).
39. Kim, Y.-I. *et al.* Infection and Rapid Transmission of SARS-CoV-2 in Ferrets. *Cell Host Microbe* **27**, 704-709.e2 (2020).
40. Sia, S. F. *et al.* Pathogenesis and transmission of SARS-CoV-2 in golden hamsters. *Nature* **583**, 834–838 (2020).
41. Bosco-Lauth, A. M. *et al.* Experimental infection of domestic dogs and cats with SARS-CoV-2: Pathogenesis, transmission, and response to reexposure in cats. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 26382–26388 (2020).
42. Sit, T. H. C. *et al.* Infection of dogs with SARS-CoV-2. *Nature* **586**, 776–778 (2020).

43. Molenaar, R. J. *et al.* Clinical and Pathological Findings in SARS-CoV-2 Disease Outbreaks in Farmed Mink (Neovison vison). *Vet. Pathol.* **57**, 653–657 (2020).
44. Richard, M. *et al.* SARS-CoV-2 is transmitted via contact and via the air between ferrets. *Nat. Commun.* **11**, 3496 (2020).
45. Belser, J. A. *et al.* Ferrets as Models for Influenza Virus Transmission Studies and Pandemic Risk Assessments. *Emerg. Infect. Dis.* **24**, 965–971 (2018).
46. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e8 (2020).
47. Letko, M., Marzi, A. & Munster, V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat. Microbiol.* **5**, 562–569 (2020).
48. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, 30494 (2017).
49. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Challenges* **1**, 33–46 (2017).
50. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
51. De Maio, N. *et al.* Issues with SARS-CoV-2 sequencing data. *Virological* **5**, (2020).
52. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
53. Didelot, X., Croucher, N. J., Bentley, S. D., Harris, S. R. & Wilson, D. J. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res.* **46**, e134–e134 (2018).
54. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
55. Mai, U. & Mirarab, S. TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics* **19**, 272 (2018).
56. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
57. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree : an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).

58. Hope, A. C. A. A Simplified Monte Carlo Significance Test Procedure. *J. R. Stat. Soc. Ser. B* **30**, 582–598 (1968).
59. Patefield, W. M. Algorithm AS 159: An Efficient Method of Generating Random R × C Tables with Given Row and Column Totals. *Appl. Stat.* **30**, 91 (1981).
60. Crispell, J., Balaz, D. & Gordon, S. V. HomoplasyFinder: a simple tool to identify homoplasies on a phylogeny. *Microb. genomics* **5**, (2019).
61. van Dorp, L. *et al.* Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect. Genet. Evol.* **83**, 104351 (2020).
62. Richard, D. *et al.* No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nat. Commun.* doi:10.5281/zenodo.4147272
63. Fitch, W. M. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Syst. Biol.* **20**, 406–416 (1971).
64. Singer, J., Gifford, R., Cotten, M. & Robertson, D. CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation. (2020). doi:10.20944/PREPRINTS202006.0225.V1
65. Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215–220 (2020).
66. Version 2.4.1 Schrödinger, L. The PyMOL Molecular Graphics System.
67. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).
68. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinforma.* **54**, 5.6.1-5.6.37 (2016).
69. Shen, M.-Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15**, 2507–24 (2006).
70. van Zundert, G. C. P. *et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **428**, 720–725 (2016).
71. Koukos, P. I. *et al.* An overview of data-driven HADDOCK strategies in CAPRI rounds 38-45. *Proteins Struct. Funct. Bioinforma.* **88**, 1029–1036 (2020).
72. Rodrigues, C. H. M., Myung, Y., Pires, D. E. V & Ascher, D. B. mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. *Nucleic Acids Res.* **47**, W338–W344 (2019).
73. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to

assist genomic epidemiology. *Nat. Microbiol.* 1–5 (2020). doi:10.1038/s41564-020-0770-5

74. Simmonds, P. Rampant C→U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. *mSphere* **5**, (2020).
75. Giorgio, S. Di, Martignano, F., Torcia, M. G., Mattiuz, G. & Conticello, S. G. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Sci. Adv.* **6**, eabb5813 (2020).
76. Ecdc. *Detection of new SARS-CoV-2 variants related to mink*. (2020).
77. Li, Q. *et al.* The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. (2020). doi:10.1016/j.cell.2020.07.012
78. Jiang, S., Hillyer, C. & Du, L. Neutralizing Antibodies against SARS-CoV-2 and Other Human Coronaviruses. *Trends Immunol.* **41**, 355–359 (2020).
79. Cao, Y. *et al.* Potent Neutralizing Antibodies against SARS-CoV-2 Identified by High-Throughput Single-Cell Sequencing of Convalescent Patients' B Cells. *Cell* **182**, 73-84.e16 (2020).
80. Ju, B. *et al.* Human neutralizing antibodies elicited by SARS-CoV-2 infection. *Nature* **584**, 115–119 (2020).
81. Shi, R. *et al.* A human neutralizing antibody targets the receptor-binding site of SARS-CoV-2. *Nature* **584**, 120–124 (2020).
82. Shang, J. *et al.* Structural basis of receptor recognition by SARS-CoV-2. *Nature* **581**, 221–224 (2020).
83. Brielle, E. S., Schneidman-Duhovny, D. & Linial, M. The SARS-CoV-2 exerts a distinctive strategy for interacting with the ACE2 human receptor. *bioRxiv* 2020.03.10.986398 (2020). doi:10.1101/2020.03.10.986398
84. Littler, D. R., Gully, B. S., Colson, R. N. & Rossjohn, J. Crystal Structure of the SARS-CoV-2 Non-structural Protein 9, Nsp9. *iScience* **23**, 101258 (2020).
85. Siu, K. *et al.* Severe acute respiratory syndrome Coronavirus ORF3a protein activates the NLRP3 inflammasome by promoting TRAF3-dependent ubiquitination of ASC. *FASEB J.* **33**, 8865–8877 (2019).
86. Issa, E., Merhi, G., Panossian, B., Salloum, T. & Tokajian, S. SARS-CoV-2 and ORF3a: Nonsynonymous Mutations, Functional Domains, and Viral Pathogenesis. *mSystems*

5, (2020).

87. Starr, T. N. *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295-1310.e20 (2020).
88. Rodrigues, J. P. *et al.* Insights on cross-species transmission of SARS-CoV-2 from structural modeling. *bioRxiv* 2020.06.05.136861 (2020). doi:10.1101/2020.06.05.136861
89. Gu, H. *et al.* Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. *Science* **369**, 1603–1607 (2020).
90. Zhu, Y. & Gao, G. ZAP-mediated mRNA degradation. *RNA Biol.* **5**, 65–67 (2008).
91. Guo, X., Ma, J., Sun, J. & Gao, G. The zinc-finger antiviral protein recruits the RNA processing exosome to degrade the target mRNA. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 151–6 (2007).
92. Digard, P., Lee, H. M., Sharp, C., Grey, F. & Gaunt, E. Intra-genome variability in the dinucleotide composition of SARS-CoV-2. *Virus Evol.* **6**, (2020).
93. Xia, X. Extreme Genomic CpG Deficiency in SARS-CoV-2 and Evasion of Host Antiviral Defense. *Mol. Biol. Evol.* **37**, 2699–2705 (2020).
94. Edwards, S. J. L. & Santini, J. M. Anthroponotic risk of SARS-CoV-2, precautionary mitigation, and outbreak management. *The Lancet. Microbe* **1**, e187–e188 (2020).
95. Wang, H., Pipes, L. & Nielsen, R. Synonymous mutations and the molecular evolution of SARS-Cov-2 origins. *bioRxiv* 2020.04.20.052019 (2020). doi:10.1101/2020.04.20.052019
96. Frank, H. K., Enard, D. & Boyd, S. D. Exceptional diversity and selection pressure on SARS-CoV and SARS-CoV-2 host receptor in bats compared to other mammals. *bioRxiv* 2020.04.20.051656 (2020). doi:10.1101/2020.04.20.051656
97. MacLean, O. A. *et al.* Natural selection in the evolution of SARS-CoV-2 in bats, not humans, created a highly capable human pathogen. *bioRxiv* 2020.05.28.122366 (2020). doi:10.1101/2020.05.28.122366
98. Chen, W.-H. *et al.* Yeast-Expressed SARS-CoV Recombinant Receptor-Binding Domain (RBD219-N1) Formulated with Alum Induces Protective Immunity and Reduces Immune Enhancement. *bioRxiv Prepr. Serv. Biol.* (2020). doi:10.1101/2020.05.15.098079
99. Quinlan, B. D. *et al.* The SARS-CoV-2 Receptor-Binding Domain Elicits a Potent

Neutralizing Response Without Antibody-Dependent Enhancement. *SSRN Electron. J.* (2020). doi:10.2139/ssrn.3575134

100. Yang, J. *et al.* A vaccine targeting the RBD of the S protein of SARS-CoV-2 induces protective immunity. *Nature* **586**, 572–577 (2020).
101. Krammer, F. SARS-CoV-2 vaccines in development. *Nature* **586**, 516–527 (2020).
102. Baum, A. *et al.* Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. *Science* **369**, 1014–1018 (2020).
103. Thomson, E. C. *et al.* The circulating SARS-CoV-2 spike variant N439K maintains fitness while evading antibody-mediated immunity. *bioRxiv* 2020.11.04.355842 (2020). doi:10.1101/2020.11.04.355842
104. Ortuso, F., Mercatelli, D., Guzzi, P. H. & Giorgi, F. M. Structural Genetics of circulating variants affecting the SARS-CoV-2 Spike / human ACE2 complex. *bioRxiv* 2020.09.09.289074 (2020). doi:10.1101/2020.09.09.289074
105. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).

Data Availability

All analysed data (as of 6th November 2020) is available on registration to GISAID with the IDs provided in **Table S1** (mink associated) and **Table S2** (human associated) together with full acknowledgement of contributing and submitting laboratories. In addition, 12 genomes were included from the NCBI nucleotide archive, as also provided in **Table S1** which were also released to GISAID by the time of writing.

Code Availability

Alignment annotation scripts together with code to obtain genome-wide dinucleotide frequencies are available at GitHub repository https://github.com/cednotsed/mink_homoplasies/.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

L.v.D and F.B designed the study. L.v.D, C.CS.T, D.R, C.Ow performed the genomics analysis. S.D.L and C.Or performed structural analyses. T.B. assisted in literature review. L.v.D and F.B. wrote the manuscript with contributions from all authors.

Acknowledgements and Funding

L.v.D and F.B. acknowledge financial support from the Newton Fund UK-China NSFC initiative (grant MR/P007597/1) and the BBSRC (equipment grant BB/R01356X/1). L.v.D. is supported by a UCL Excellence Fellowship. D.R. is supported by a NIHR Precision AMR award. C.Ow. is funded by a NERC-DTP studentship. SDL is funded by a Fundamental Research Grant Scheme from the Ministry Of Higher Education Malaysia (FRGS/1/2020/STG01/UKM/02/3). We wish to thank Paul Ashford for generously sharing his insight in structural biology and the SARS-CoV-2 Twitter community for alerting us to literature characterising the role of Y453F. Finally, we acknowledge the large number of originating and submitting laboratories openly sharing SARS-CoV-2 genome assemblies with the research community.

No detectable signal for ongoing genetic recombination in SARS-CoV-2

Damien Richard[1,2*], Christopher J. Owen[1], Lucy van Dorp[1], François Balloux[1*]

[1] UCL Genetics Institute, University College London, UK;

[2] Institute of Child Health, University College London, UK;

[*] Correspondence: richarddamienfr@gmail.com (Damien Richard), f.balloux@ucl.ac.uk (François Balloux).

Abstract

The COVID-19 pandemic has led to an unprecedented global sequencing effort of its viral agent SARS-CoV-2. The first whole genome assembly of SARS-CoV-2 was published on January 5 2020. Since then, over 150,000 high-quality SARS-CoV-2 genomes have been made available. This large genomic resource has allowed tracing of the emergence and spread of mutations and phylogenetic reconstruction of SARS-CoV-2 lineages in near real time. Though, whether SARS-CoV-2 undergoes genetic recombination has been largely overlooked to date. Recombination-mediated rearrangement of variants that arose independently can be of major evolutionary importance. Moreover, the absence of recombination is a key assumption behind the application of phylogenetic inference methods. Here, we analyse the extant genomic diversity of SARS-CoV-2 and show that, to date, there is no detectable hallmark of recombination. We assess our detection power using simulations and validate our method on the related MERS-CoV for which we report evidence for widespread genetic recombination.

Introduction

Genetic recombination is widely recognised as an important force in evolution, as it allows for the combination, within a single genome, of variants that arose independently in different genetic backgrounds [1]. Viruses are no exception to this pattern [2]. For example, recombination between its eight genomic segments (reassortment) is the fundamental mechanism behind the emergence of pandemic influenza A strains [3]. Recombination is also key for many viruses to generate new antigenic combinations that allow host immune systems evasion [4]. Moreover, the absence of recombination is also a prerequisite for phylogenetic inference [5]. Indeed, phylogenetic trees are limited to represent a single realisation of the past demography of the samples analysed. In the presence of genetic recombination, different regions of the sequence under scrutiny will support different evolutionary histories for the samples analysed, and hence result in conflicts in the topology of the phylogenetic reconstruction [5]. Not accounting for recombination can also lead to false positive detection of sites undergoing positive selection [6, 7].

Repositories of Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genomes have grown at an unprecedented pace, with over 150,000 high-quality complete genome assemblies currently available on the Global Initiative on Sharing All Influenza Data (GISAID) repository as of 17/11/2020 [8, 9]. This allows for the near-real-time monitoring of the emergence and spread of novel mutations [10, 11] and the description of emerging lineages [12, 13]. Most of these studies rely on phylogenetic reconstructions of genome-wide Single Nucleotide Polymorphisms (SNPs), and implicitly assume the absence of pervasive genetic recombination in SARS-CoV-2. So far there has been limited effort to assess the extent of ongoing recombination in SARS-CoV-2 ([14, 15], https://observablehq.com/@spond/linkage-disequilibirum-in-sars-cov-2) despite its potential relevance to understanding the duration and propensity of co-infections in host.

Conflicts (incongruence) between phylogenies inferred from different genome segments can be indicative of recombination. Some of the numerous methods developed to detect genetic recombination rely on this concept [16-19]. The so-called "compatibility test" checks if all four combinations of alleles of a pair of biallelic sites (00, 01, 10, 11) are present among the sequences. More refined methods relying on this principle have been developed including the pairwise homoplasy index (PHI) [20]. Recombination also has the effect of decorrelating allele frequencies, with this effect increasing with physical distance along the genome. In a population undergoing frequent recombination, this causes linkage disequilibrium of alleles to decay with physical distance on the sequence [4]. The r^2 metric [21] is commonly used to measure linkage disequilibrium [22].

In this work, we aimed to detect signals of genetic recombination within the SARS-CoV-2 global population. We assembled a curated alignment of 6,546 available SARS-CoV-2 genomes enriched for those collected more recently to maximise genetic diversity in the dataset, and hence our ability to detect recombination. We applied two different statistical methods for the detection of genetic recombination and assessed their power using bespoke simulations. We validate our methodology on the related *Betacoronavirus* Middle East respiratory syndrome-related coronavirus (MERS-CoV) [23] responsible for the MERS outbreaks beginning in 2012, for which we find evidence for recombination, consistent with previous reports [24]. Our results do not identify detectable evidence for recombination in the SARS-CoV-2 population as of September 2020.

Results

No signal of recombination in SARS-CoV-2

We compiled an alignment of 6,546 SARS-CoV-2 isolates sampled across six continental regions. In order to maximize our detection power, whilst keeping the dataset computationally manageable, we chose to restrict our analysis to the most recently collected genomes during the month of September 2020. This is expected to maximise the genetic diversity in the dataset, and hence increase power to detect recombination events. We detect over 4,000 polymorphic positions following masking of sites putatively suggested as artefactual ([25] https://github.com/W-L/ProblematicSites_SARS-CoV2/blob/master/problematic_sites_sarsCov2.vcf, accessed 29/10/2020). These include a large fraction of homoplasies (29.6%), thought to largely be induced by host immune system RNA editing [15, 26, 27].

A PHI test applied to the SARS-CoV-2 alignment reported a *p*-value of 0.78 suggesting no signal of recombination. Consistently, LD decay regression coefficients and R squared statistics did not fall outside of the distributions obtained after randomly permuting genome coordinates (Figure 1).

To test our ability to detect low levels of recombination, we simulated alignments with levels of genetic diversity matching that observed in the true SARS-CoV-2 alignment, but using varying recombination rates. We detected recombination in 100% of the datasets simulated with 3e-3 recombination events per genome per viral replication (60% for a rate of 3e-4, Supplementary Table S1). This low detection power is linked to the high homogeneity of the SARS-CoV-2 population, reflected by the mean pairwise distance of 19.4 (95%HPD 3-30) SNPs in the alignment analysed.

In addition, we searched the global SARS-CoV-2 phylogeny for isolates displaying root-to-tip distances in the upper 5% quartile of the distribution. These may offer some of the best candidates for isolates having experienced recombination, which is expected to increase terminal branch length. We detected 24 phylogenetic outliers which grouped into 13 phylogenetic clades. Localisation of their mutations and those of their phylogenetic neighbours in matrices did not support a recombinant origin. Indeed, rather than displaying syntenic groups of private mutations that would suggest a recombination-mediated origin, the 24 outliers mostly displayed a randomly distributed excess of mutations (Supplementary Figures S1-S13). Our results therefore suggest that recombination in SARS-CoV-2 is either absent, or occurring at a rate too low relative to mutation to be detectable under the genetic diversity charactering the SARS-CoV-2 population at this stage.

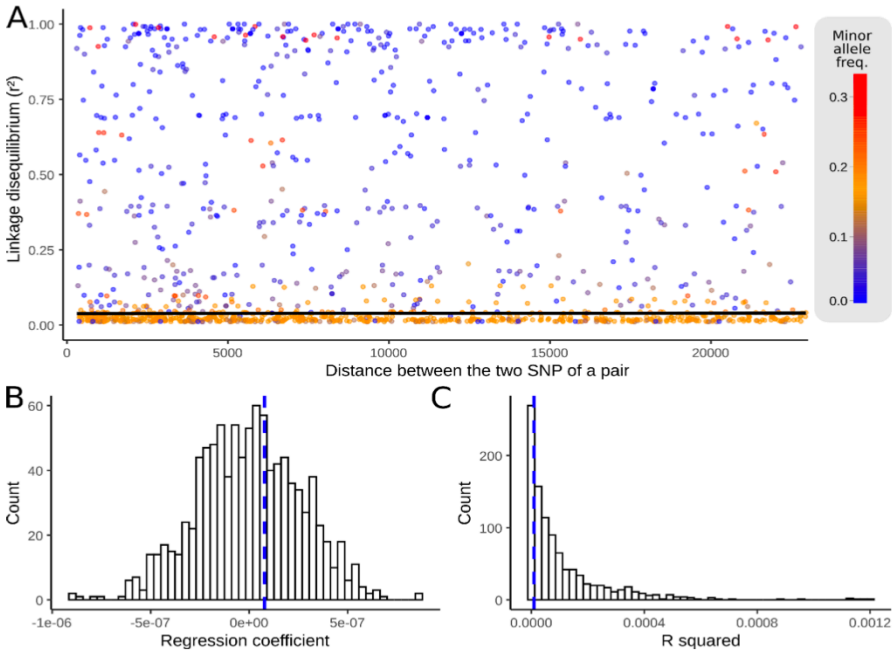


Figure 1: Linkage disequilibrium (r^2) as a function of physical distance on the SARS-CoV-2 genome. (A) Linkage disequilibrium measured by r^2 (y-axis) for all pairs of SNPs represented as a function of the genetic distance separating the SNPs of each pair. Black line: fitted linear model (regression coefficient: 7.84e-8; R-squared: 9.33e-6). (B) Distribution of regression coefficients of the linear models obtained following consideration of 1000 position permuted datasets. Blue dashed line: value obtained for the true SARS-CoV-2 alignment. (C) Distribution of the R squared values of the linear models of 1000 position permuted datasets. Blue dashed line: value of the SARS-CoV-2 true alignment.

Recombination occurs in MERS-CoV

To validate our approach, we applied the same method to an alignment of MERS-CoV, a related *Betacoronavirus* thought to be widely recombining [24, 28]. Counter to observations for SARS-CoV-2, the MERS-CoV dataset yielded detectable evidence of recombination. Beside the decay of the r^2 ~ distance regression slope, values of both R-squared and the regression coefficient largely fall outside of the distributions of the same parameters of the permuted datasets (Figure 2). The PHI test reported a p-value <1e-12. Of note, these tests were repeated after discarding C to T mutations, mostly caused by the host immune RNA editing systems, that might produce an artefactual signal of recombination. Tests on the pruned alignment still provide evidence of significant signals of recombination in MERS-CoV (PHI test p-value <1e-12 and significant decay of r^2, Supplementary Figure S14).

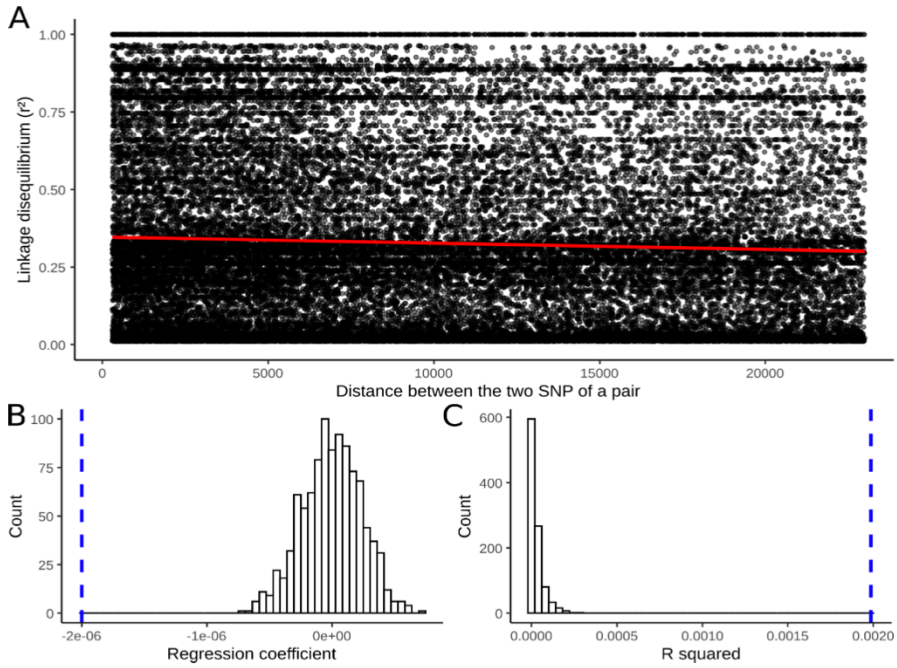


Figure 2: Linkage disequilibrium (r^2) as a function of physical distance on the MERS-CoV genomes. Pairs comprising SNP differing by a frequency >=0.1 have been discarded, lowering the number of pairs from 261,090 to 212,342. (A) Linkage disequilibrium (y-axis) for all pairs of SNPs is represented as a function of the distance separating the SNPs of each pair. Red line: fitted linear model (regression coefficient: -2.00e-6; R-squared: 1.99e-3). (B) Distribution of the regression coefficients of the linear models obtained following consideration of 1000 position permuted datasets. Blue dashed line: value of the MERS-CoV true alignment. (C) Distribution of the R squared values of the linear models of 1000 position permuted datasets. Blue dashed line: value of the MERS-CoV true alignment.

Discussion

In this study, we analyzed a dataset of 6,546 SARS-CoV-2 assemblies sampled during the month of September 2020 across six continental regions. Applying two distinct detection methods, we did not find signal of recombination among the SARS-CoV-2 population tested. Conversely, we detected evidence of recombination in SARS-CoV-2-like simulated recombining datasets as well as in a set of 459 MERS-CoV coronaviruses assemblies, known for being prone to recombination.

A priori it is highly plausible that SARS-CoV-2 has the potential to recombine [28]. Recombination has been suggested to be common in coronaviruses, including both for human [29-33] and animal [34-36] associated lineages, as inferred from genomic approaches [37], observed in cell culture [38, 39] and *in vivo* [40]. It has been claimed that all of the human epidemic coronaviruses: SARS-CoV-1 [41-43], MERS-

CoV [44], and SARS-CoV-2 [45-47] may have evolved through recombination events leading to some genome mosaicism, particular over receptor binding regions.

At this stage we do not detect a genetic hallmark for recombination in SARS-CoV-2. This does not necessarily imply that SARS-CoV-2 lacks the ability to recombine. For genetic recombination to leave a measurable signal in the genetic data, there needs to be sufficient genetic differentiation between the recombining viruses. Given the low intra-host genetic diversity of epidemic viruses such as SARS-CoV-2, this requires mixed infections (i.e. coinfection of the same host by distinct SARS-CoV-2 lineages). While such events are expected to be rare, there have been reports of mixed infections [48]. Though, given the limited genetic diversity of SARS-CoV-2 strains currently in circulation, even mixed infections may often not involve sufficiently differentiated strains to leave a detectable signal following a recombination event. The recent host jump into humans of SARS-CoV-2, most likely through a single transmission to humans from an unknown animal reservoir, created an essentially genetically invariant viral population, with genetic diversity building up through the accumulation of mutations since the begininning of the pandemic. The genomic diversity of SARS-CoV-2 is still far below its mutation-drift equilibrium and remains very low at this stage [10]. As a result, putative recombination events would only be supported by a limited number of SNPs, and would require high detection sensitivity to be identified.

In contrast, we detected recombination in MERS-CoV despite the far smaller sample size of the alignment analysed. Besides the higher genetic diversity of MERS-CoV at this stage, this may also be due to major epidemiological differences between MERS and COVID-19. MERS is mainly a disease of dromedary camels, with spillover events into humans [28]. In camels, the high prevalence of the disease and the mostly mild symptoms it causes is suggested to favour co-infection [24]. On the contrary, the severity of the symptoms in human lowers the probability of co-infection. The camel host, which is known to harbour other coronaviruses, could provide a hub of genetic diversity creation in MERS-CoV through recombination [49]. Human MERS-CoV infection was first documented in 2012, but it is thought the virus had been previously circulating for at least a few years in camels [50]. Our MERS-CoV alignment comprises samples spanning from 2012 to 2019. Mutations accumulating over this time-scale provide more diverse genetic markers that facilitate the detection of putative recombination events.

While we did not detect evidence of genetic recombination in SARS-CoV-2 to date, it remains of importance to repeat such analyses as the genetic diversity of the SARS-CoV-2 population will increase, and to consider its possible impact when conducting phylogenetics studies in the future.

Materials and Methods

SARS-CoV-2 dataset

All 6,546 SARS-CoV-2 high quality genomes (containing less than 5% of "N" and being >29,000 bp long) sampled during the month of September 2020 available on GISAID (as of October 15th 2020) were downloaded and profile aligned to the Wuhan-Hu-1 reference genome (GenBank accession MN908947; GISAID ID EPI_ISL_402125) using MAFFT v7.471 [51]. A full list of acknowledgements together with submitting and originating laboratories is provided in Supplementary Table S2. SNPs flagged as putative sequencing errors were discarded (https://github.com/W-L/ProblematicSites_SARS-CoV2/blob/master/problematic_sites_sarsCov2.vcf, accessed 29/10/2020). The final dataset comprised 4,199 SNPs and a mean pairwise SNP difference of 19.4 (95%HPD 3-30). Following construction of a maximum likelihood tree using IQTree Covid-release [52], 29.6% of SNPs were identified as homoplasic by HomoplasyFinder [53].

MERS dataset

456 high-quality MERS-CoV genomes isolated from both camels and human were downloaded from the NCBI Virus database and profile aligned to the HCoV-EMC/2012 reference genome (GenBank accession NC_019843) using MAFFT v7.471 [51] (Supplementary Table S3). We detected 8,788 SNPs in the alignment, with a mean pairwise SNP count of 123.36 (95%HPD 18-234). Homoplasies were identified as described above and represented 12.4% of the polymorphic positions.

Detection of recombination

Two recombination tests were performed on each dataset. First, a pairwise homoplasy index (PHI) test was used to detect recombination setting the number of permutations to 100 and the window size set to 300 bp with otherwise default parameters [20]. Additionally, we computed the linkage disequilibrium (r^2) for all pairs of bi-allelic SNPs occurring in >=1% of the isolates using tomahawk (https://mklarqvist.github.io/tomahawk/). 90% of the SNP pairs grouped <= 23,000 nucleotides apart. Linkage disequilibrium estimation over distances larger than 23,000 rely on a few SNP pairs only, so we restricted the dataset to those 90% pairs. A linear model was fitted to the distribution of r^2 values as a function of the distance separating the two SNPs in each pair. The regression coefficient of this linear model indicates whether linkage disequilibrium decays with physical distance or not. To formally test for the presence of recombination, we produced 1000 permuted datasets (randomly associating r^2 values with distance values) and fitted a linear model to each one of the permuted datasets. We then assessed whether the real R-squared and regression coefficients values fell either inside or outside of the distributions of the parameters generated by the randomly permuted datasets. A limitation of the use of the r^2 metrics as an estimator of linkage disequilibrium is its dependency on allele frequencies, causing a possible reduction in statistical power [54]. It has therefore been proposed to compute r^2 only for pairs of SNPs that do not differ markedly in frequency in the studied population [55]. Discarding pairs of SNPs is suboptimal in the context of SARS-CoV-2's already restricted genetic diversity. However, we still implemented this approach which led to similar results to tests on non-frequency filtered SNPs (Supplementary Figure S15).

We performed a third test for recombination in SARS-CoV-2 by focusing specifically on isolates that were flagged as phylogenetic outliers in the global phylogeny. Recombinant isolates are expected to

be located at the tip of long terminal branches if there is phylogenetic incongruency between the mutations they carry. We applied TreeShrink to identify the accessions displaying root-to-tip distances in the upper 5% quartile (-q 0.05 parameter) of the root-to-tip distance distribution [56]. The mutations carried by those outliers were visualy compared to that of their neighbours in the phylogeny.

Power of recombination detection

In order to characterise the statistical power of the recombination detection methods employed, we simulated in silico SARS-CoV-2 alignments using MSprime [57]. The simulated mutation rate was set to match that of the real dataset (Supplementary Figure S16). We generated datasets with numbers of recombination events per genome per viral replication of 0, 3e-7, 3e-6, 3e-5, 3e-4, 3e-3 and 3e-2 (ten replicates each). PHI tests and linkage disequilibrium decay tests were performed on those simulated datasets as described previously.

References

1. Muller HJ: **The relation of recombination to mutational advance.** *Mutat Res* 1964, **106:**2-9.
2. Koonin EV, Dolja VV, Krupovic M: **Origins and evolution of viruses of eukaryotes: The ultimate modularity.** *Virology* 2015, **479-480:**2-25.
3. Smith GJD, Bahl J, Vijaykrishna D, Zhang J, Poon LLM, Chen H, Webster RG, Peiris JSM, Guan Y: **Dating the emergence of pandemic influenza viruses.** 2009, **106:**11709-11712.
4. Pérez-Losada M, Arenas M, Galán JC, Palero F, González-Candelas F: **Recombination in viruses: Mechanisms, methods of study, and evolutionary consequences.** *Infection, Genetics and Evolution* 2015, **30:**296-307.
5. Posada D, Crandall KA: **The effect of recombination on the accuracy of phylogeny estimation.** *J Mol Evol* 2002, **54:**396-402.
6. Anisimova M, Nielsen R, Yang Z: **Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites.** *Genetics* 2003, **164:**1229-1236.
7. Shriner D, Nickle DC, Jensen MA, Mullins JI: **Potential impact of recombination on sitewise approaches for detecting positive natural selection.** *Genet Res* 2003, **81:**115-121.
8. Elbe S, Buckland-Merrett G: **Data, disease and diplomacy: GISAID's innovative contribution to global health.** 2017, **1:**33-46.
9. Shu Y, McCauley J: **GISAID: Global initiative on sharing all influenza data – from vision to reality.** 2017, **22:**30494.
10. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ, Pang J, Tan CCS, Boshier FAT, et al: **Emergence of genomic diversity and recurrent mutations in SARS-CoV-2.** *Infection, Genetics and Evolution* 2020, **83:**104351.
11. Thomson EC, Rosen LE, Shepherd JG, Spreafico R, da Silva Filipe A, Wojcechowskyj JA, Davis C, Piccoli L, Pascall DJ, Dillen J, et al: **The circulating SARS-CoV-2 spike variant N439K maintains fitness while evading antibody-mediated immunity.** 2020**:**2020.2011.2004.355842.
12. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, du Plessis L, Pybus OG: **A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology.** *Nature Microbiology* 2020, **5:**1403-1407.
13. Hodcroft EB, Zuber M, Nadeau S, Comas I, González Candelas F, Stadler T, Neher RA: **Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020.** 2020**:**2020.2010.2025.20219063.
14. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, et al: **On the origin and continuing evolution of SARS-CoV-2.** *National Science Review* 2020, **7:**1012-1023.
15. van Dorp L, Richard D, Tan CCS, Shaw LP, Acman M, Balloux F: **No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2.** *Nature Communications* 2020, **11:**5986.
16. Posada DJMb, evolution: **Evaluation of methods for detecting recombination from DNA sequences: empirical data.** 2002, **19:**708-717.
17. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW: **GARD: a genetic algorithm for recombination detection.** *Bioinformatics* 2006, **22:**3096-3098.
18. Boni MF, de Jong MD, van Doorn HR, Holmes EC: **Guidelines for identifying homologous recombination events in influenza A virus.** *PLOS ONE* 2010, **5:**e10434.
19. Lam HM, Ratmann O, Boni MF: **Improved algorithmic complexity for the 3SEQ recombination detection algorithm.** *Molecular Biology and Evolution* 2017, **35:**247-251.
20. Bruen TC, Philippe H, Bryant D: **A simple and robust statistical test for detecting the presence of recombination.** *Genetics* 2006, **172:**2665-2681.
21. Hill WG, Robertson A: **Linkage disequilibrium in finite populations.** *Theor Appl Genet* 1968, **38:**226-231.

22. Haydon DT, Bastos ADS, Awadalla P: **Low linkage disequilibrium indicative of recombination in foot-and-mouth disease virus gene sequence alignments.** 2004, **85:**1095-1100.
23. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA: **Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia.** *N Engl J Med* 2012, **367:**1814-1820.
24. Dudas G, Rambaut A: **MERS-CoV recombination: implications about the reservoir and potential for adaptation.** *Virus evolution* 2016, **2:**vev023-vev023.
25. Turakhia Y, De Maio N, Thornlow B, Gozashti L, Lanfear R, Walker CR, Hinrichs AS, Fernandes JD, Borges R, Slodkowicz G, et al: **Stability of SARS-CoV-2 phylogenies.** *PLOS Genetics* 2020, **16:**e1009175.
26. Simmonds P: **Rampant C→U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses: causes and consequences for their short- and long-term evolutionary trajectories.** 2020, **5:**e00408-00420.
27. Di Giorgio S, Martignano F, Torcia MG, Mattiuz G, Conticello SG: **Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2.** 2020, **6:**eabb5813.
28. Su S, Wong G, Shi W, Liu J, Lai ACK, Zhou J, Liu W, Bi Y, Gao GF: **Epidemiology, genetic recombination, and pathogenesis of coronaviruses.** *Trends in Microbiology* 2016, **24:**490-502.
29. Lau SKP, Lee P, Tsang AKL, Yip CCY, Tse H, Lee RA, So L-Y, Lau Y-L, Chan K-H, Woo PCY, Yuen K-Y: **Molecular epidemiology of human coronavirus OC43 reveals evolution of different genotypes over time and recent emergence of a novel genotype due to natural recombination.** 2011, **85:**11325-11337.
30. Kin N, Miszczak F, Lin W, Gouilh MA, Vabret A, Consortium E: **Genomic analysis of 15 human coronaviruses OC43 (HCoV-OC43s) circulating in France from 2001 to 2013 reveals a high intra-specific diversity with new recombinant genotypes.** *Viruses* 2015, **7:**2358-2377.
31. Pyrc K, Dijkman R, Deng L, Jebbink MF, Ross HA, Berkhout B, van der Hoek L: **Mosaic structure of human coronavirus NL63, one thousand years of evolution.** *Journal of Molecular Biology* 2006, **364:**964-973.
32. Woo PCY, Lau SKP, Yip CCY, Huang Y, Tsoi H-W, Chan K-H, Yuen K-Y: **Comparative analysis of 22 coronavirus HKU1 genomes reveals a novel genotype and evidence of natural recombination in coronavirus HKU1.** 2006, **80:**7136-7145.
33. Graham RL, Baric RS: **Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission.** 2010, **84:**3134-3146.
34. Terada Y, Matsui N, Noguchi K, Kuwata R, Shimoda H, Soma T, Mochizuki M, Maeda K: **Emergence of pathogenic coronaviruses in cats by homologous recombination between feline and canine coronaviruses.** *PLOS ONE* 2014, **9:**e106534.
35. Decaro N, Mari V, Campolo M, Lorusso A, Camero M, Elia G, Martella V, Cordioli P, Enjuanes L, Buonavoglia C: **Recombinant canine coronaviruses related to transmissible gastroenteritis virus of swine are circulating in dogs.** 2009, **83:**1532-1537.
36. Tian P-F, Jin Y-L, Xing G, Qv L-L, Huang Y-W, Zhou J-Y: **Evidence of recombinant strains of porcine epidemic diarrhea virus, United States, 2013.** *Emerging infectious diseases* 2014, **20:**1735-1738.
37. Herrewegh AA, Smeenk I, Horzinek MC, Rottier PJ, de Groot RJ: **Feline coronavirus type II strains 79-1683 and 79-1146 originate from a double recombination between feline coronavirus type I and canine coronavirus.** *J Virol* 1998, **72:**4508-4514.
38. Lai MM, Baric RS, Makino S, Keck JG, Egbert J, Leibowitz JL, Stohlman SA: **Recombination between nonsegmented RNA genomes of murine coronaviruses.** *J Virol* 1985, **56:**449-456.
39. Makino S, Keck JG, Stohlman SA, Lai MM: **High-frequency RNA recombination of murine coronaviruses.** *J Virol* 1986, **57:**729-737.
40. Keck JG, Matsushima GK, Makino S, Fleming JO, Vannier DM, Stohlman SA, Lai MM: **In vivo RNA-RNA recombination of coronavirus in mouse brain.** *J Virol* 1988, **62:**1810-1813.

41. Hon C-C, Lam T-Y, Shi Z-L, Drummond AJ, Yip C-W, Zeng F, Lam P-Y, Leung FC-C: **Evidence of the recombinant origin of a bat Severe Acute Respiratory Syndrome (SARS)-like coronavirus and its implications on the direct ancestor of SARS coronavirus.** 2008, **82:**1819-1826.
42. Hu B, Zeng L-P, Yang X-L, Ge X-Y, Zhang W, Li B, Xie J-Z, Shen X-R, Zhang Y-Z, Wang N, et al: **Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus.** *PLOS Pathogens* 2017, **13:**e1006698.
43. Lau SKP, Li KSM, Huang Y, Shek C-T, Tse H, Wang M, Choi GKY, Xu H, Lam CSF, Guo R, et al: **Ecoepidemiology and complete genome comparison of different strains of severe acute respiratory syndrome-related Rhinolophus bat coronavirus in China reveal bats as a reservoir for acute, self-limiting infection that allows recombination events.** 2010, **84:**2808-2819.
44. Corman VM, Ithete NL, Richards LR, Schoeman MC, Preiser W, Drosten C, Drexler JF: **Rooting the phylogenetic tree of middle East respiratory syndrome coronavirus by characterization of a conspecific virus from an African bat.** *J Virol* 2014, **88:**11297-11303.
45. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF: **The proximal origin of SARS-CoV-2.** *Nature Medicine* 2020, **26:**450-452.
46. Boni MF, Lemey P, Jiang X, Lam TT-Y, Perry BW, Castoe TA, Rambaut A, Robertson DL: **Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic.** *Nature Microbiology* 2020, **5:**1408-1417.
47. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, Hu Y, Tao Z-W, Tian J-H, Pei Y-Y, et al: **A new coronavirus associated with human respiratory disease in China.** *Nature* 2020, **579:**265-269.
48. Lythgoe KA, Hall M, Ferretti L, de Cesare M, MacIntyre-Cockett G, Trebes A, Andersson M, Otecko N, Wise EL, Moore N, et al: **Shared SARS-CoV-2 diversity suggests localised transmission of minority variants.** 2020**:**2020.2005.2028.118992.
49. Sabir JSM, Lam TT-Y, Ahmed MMM, Li L, Shen Y, E. M. Abo-Aba S, Qureshi MI, Abu-Zeid M, Zhang Y, Khiyami MA, et al: **Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia.** 2016, **351:**81-84.
50. Lau SKP, Wong ACP, Lau TCK, Woo PCY: **Molecular evolution of MERS coronavirus: dromedaries as a recent intermediate host or long-time animal reservoir?** *Int J Mol Sci* 2017, **18**.
51. Katoh K, Standley DM: **MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.** *Molecular Biology and Evolution* 2013, **30:**772-780.
52. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R: **IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era.** *Molecular Biology and Evolution* 2020, **37:**1530-1534.
53. Crispell J, Balaz D, Gordon SV: **HomoplasyFinder: a simple tool to identify homoplasies on a phylogeny.** *Microbial genomics* 2019, **5:**e000245.
54. VanLiere JM, Rosenberg NA: **Mathematical properties of the r2 measure of linkage disequilibrium.** *Theoretical population biology* 2008, **74:**130-137.
55. Eberle MA, Rieder MJ, Kruglyak L, Nickerson DA: **Allele Frequency Matching Between SNPs Reveals an Excess of Linkage Disequilibrium in Genic Regions of the Human Genome.** *PLOS Genetics* 2006, **2:**e142.
56. Mai U, Mirarab S: **TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees.** *BMC Genomics* 2018, **19:**272.
57. Kelleher J, Etheridge AM, McVean G: **Efficient coalescent simulation and genealogical analysis for large sample sizes.** *PLoS Comput Biol* 2016, **12:**e1004842.

Data and Code Availability

All analysed SARS-CoV-2 data is available on registration to GISAID with the accession IDS and acknowledgements provided in Table S2. MERS-CoV assemblies are freely available on NCBI with the included accessions provided in Table S3. Scripts used in this study are available at https://github.com/DamienFr/LD_SARS-CoV-2.

Competing Interests

The authors have no competing interests to declare.

Acknowledgements

D.R. is supported by a NIHR Precision AMR award. C.O is funded by a NERC-DTP studentship. L.v.D and F.B. acknowledge financial support from the Newton Fund UK-China NSFC initiative (grant MR/P007597/1) and the BBSRC (equipment grant BB/R01356X/1). L.v.D. is supported by a UCL Excellence Fellowship. We wish to particularly acknowledge all of the large number of contributing and submitting laboratories sharing SARS-CoV-2 assemblies via the GISAID platform, including the UK (COG-UK) consortium (a full list of consortium names and affiliations can be found at https://www.cogconsortium.uk). COG-UK is supported by funding from the Medical Research Council (MRC) part of UK Research & Innovation (UKRI), the National Institute of Health Research (NIHR) and Genome Research Limited, operating as the Wellcome Sanger Institute.

Infection, Genetics and Evolution 95 (2021) 105075

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Pre-existing T cell-mediated cross-reactivity to SARS-CoV-2 cannot solely be explained by prior exposure to endemic human coronaviruses

Cedric C.S. Tan [a,*], Christopher J. Owen [a], Christine Y.L. Tham [b], Antonio Bertoletti [b], Lucy van Dorp [a,1], Francois Balloux [a,1]

[a] UCL Genetics Institute, University College London, Gower Street, London WC1E 6BT, United Kingdom
[b] Emerging Infectious Diseases Program, Duke-NUS Medical School, 8 College Road, Singapore 169857, Singapore

ABSTRACT

T-cell-mediated immunity to SARS-CoV-2-derived peptides in individuals unexposed to SARS-CoV-2 has been previously reported. This pre-existing immunity was suggested to largely derive from prior exposure to 'common cold' endemic human coronaviruses (HCoVs). To test this, we characterised the sequence homology of SARS-CoV-2-derived T-cell epitopes reported in the literature across the full proteome of the *Coronaviridae* family. 54.8% of these epitopes had no homology to any of the HCoVs. Further, the proportion of SARS-CoV-2-derived epitopes with any level of sequence homology to the proteins encoded by any of the coronaviruses tested is well-predicted by their alignment-free phylogenetic distance to SARS-CoV-2 (*Pearson's* $r = -0.958$). No coronavirus in our dataset showed a significant excess of T-cell epitope homology relative to the proportion of expected random matches, given their genetic similarity to SARS-CoV-2. Our findings suggest that prior exposure to human or animal-associated coronaviruses cannot completely explain the T-cell repertoire in unexposed individuals that recognise SARS-CoV-2 cross-reactive epitopes.

1. Introduction

Severe acute respiratory coronavirus 2 (SARS-CoV-2) is a member of a large family of viruses; the *Coronaviridae*, whose members can infect a wide range of mammals and birds (Shaw et al., 2020). Human coronaviruses were first described in the 1960s (Tyrrell and Bynoe, 1965) with SARS-CoV-2 now the seventh coronavirus known to infect humans; joining the epidemic human coronaviruses, SARS-CoV-1 (Ksiazek et al., 2003) and MERS-CoV (Zaki et al., 2012), and the four species of endemic human coronaviruses (HCoVs). Human endemic coronaviruses are associated with mostly mild upper respiratory infections – 'common colds' – and include *Coronaviridae* of the *Alphacoronavirus* genera 229E and NL63 and members of the *Betacoronavirus* genera OC43 and HKU1 (Su et al., 2016) to which MERS-CoV, SARS-CoV-1 and SARS-CoV-2 also belong. Both SARS-CoV-1 and SARS-CoV-2 fall into a subgenus of the *Betacoronavirus* named the *Sarbecovirus* (Boni et al., 2020), with approximately 80% identity at the nucleotide level between SARS-CoV-1 and SARS-CoV-2. All human coronaviruses are thought to be zoonotic in origin, though the exact animal reservoirs remain under debate in some cases (Ye et al., 2020).

SARS-CoV-2 is estimated to have jumped from a currently unknown animal reservoir into the human population towards the end of 2019 (van Dorp et al., 2020) giving rise to the pandemic disease Coronavirus disease 2019 (COVID-19). The symptoms associated with COVID-19 range from fully asymptomatic infections and mild disease through to severe respiratory disease with associated morbidity and mortality. Marked disparities exist in individual risk of severe COVID-19 with gender, ethnicity, metabolic health and age all identified as important determinants (Jordan et al., 2020; Wu et al., 2020; Zhou et al., 2020). Population age structures and heterogeneous burdens in nursing homes only partially explain the variation in infection fatality rates (IFRs) between countries (O'Driscoll et al., 2020). Further important contributors may include climatic variables (e.g. temperature and humidity) and associated seasonal correlates (Walker et al., 2020; Gaunt et al., 2010; Moriyama et al., 2020), the choice of non-pharmaceutical interventions put in place, and more recently vaccination coverage though with a myriad of other possibly unknown contributing factors.

In light of the wide spectrum of symptoms associated to COVID-19, several studies have probed antibody (Lv et al., 2020; Ladner et al., 2021; Ng et al., 2020) or T-cell responses (Mateus et al., 2020; Grifoni et al., 2020a; Weiskopf et al., 2020a; Le Bert et al., 2020; Nelde et al., 2020; Braun et al., 2020; Peng et al., 2020; Schulien et al., 2020; Bacher et al., 2020; Sekine et al., 2020; Steiner et al., 2020; Echeverría et al., 2021; Reynolds et al., 2020; Low et al., 2021) in samples from healthy

* Corresponding author.
E-mail address: cedric.tan.18@ucl.ac.uk (C.C.S. Tan).
[1] Co-last authors.

https://doi.org/10.1016/j.meegid.2021.105075
Received 12 May 2021; Received in revised form 27 August 2021; Accepted 3 September 2021
Available online 10 September 2021

individuals collected prior to the COVID-19 pandemic to test for the presence of pre-existing cross-reactivity to SARS-CoV-2. Collectively, these findings provide evidence for a degree of antibody and T-cell cross-reactivity in unexposed individuals in multiple regions of the world. While the source of this cross-reactivity remains poorly defined, at least some of the cross-reactive T-cell epitopes have been suggested to derive from exposure to the four endemic human coronaviruses (Mateus et al., 2020; Le Bert et al., 2020), which were circulating in most parts of the world prior to the COVID-19 pandemic (Su et al., 2016), typically in seasonal cycles (Neher et al., 2020). Further, SARS-CoV-2 cross-reactive epitopes have been identified in exposed seronegative healthcare workers contributing to abortive infections (Swadling et al., 2021). Such studies have been based, in part, on the degree of homology of detected epitopes to protein sequences found in each of the four HCoVs, though lacked consideration of many other coronaviruses which circulate widely in mammals or the degree of matching expected given the relatedness of these viruses to SARS-CoV-2. As such, the relative contribution of each of the four HCoVs to T-cell cross-reactivity patterns observed in unexposed individuals remains unclear. Notably, Peng et al. (Peng et al., 2020) did not find the presence of cross-reactivity in a cohort of 16 unexposed donors.

To date, it also remains unclear whether detected cross-immunity in unexposed individuals translates into consistently differential COVID-19 pathogenesis. The evidence for a mitigating role of recent HCoV infection on COVID-19 susceptibility and symptom severity upon infection is mixed (Sagar et al., 2020; Gombar et al., 2021). HCoV-reactive T-cells in unexposed individuals have been shown to have only low functional avidity (Bacher et al., 2020), though cohort studies suggest pre-exisiting coronavirus RNA-polymerase-specific T-cells are an important determinant of abortive rather than overt infection (Swadling et al., 2021). As such there has been speculation that cross-immunity with the 'common cold' endemic HCoVs may, in part, explain variation in the COVID-19 case-fatality rate in different parts of the world (Gupta and Misra, 2020; Yaqinuddin, 2020) and that the high incidence of common colds in children and adolescents has contributed to their markedly lower risk of severe disease (Ng et al., 2020). Additionally, the possible unnoticed circulation in the human population of another animal-associated coronavirus, at least in some regions of the world, cannot at this stage be formally ruled out to have contributed to regional heterogeneities in the spread and associated mortality of COVID-19.

In this study, we sought to probe the possible sources of pre-existing T-cell immunity in samples from healthy individuals predating the COVID-19 pandemic. One tractable way to determine the contribution of multiple human or animal-associated coronaviruses to T-cell cross-reactivity is to consider the amino acid sequence homology of experimentally-validated SARS-CoV-2 epitope sequences to proteins encoded by these viruses. The assumption is that viruses that have contributed significantly to cross-reactivity are likely to possess a higher than expected number of protein sequences with reasonable sequence homology to these SARS-CoV-2 epitopes. While we recognise that two epitopes sharing a low sequence homology can be cross-reactive due to structural conservation (Macdonald et al., 2009; Wucherpfennig and Strominger, 1995; Quaratino et al., 1995), the vast majority of cross-reactive epitopes share a high sequence homology (Mateus et al., 2020). That is, epitopes that share a higher sequence homology have a far higher likelihood of being cross-reactive. Therefore, sequence homology offers a good proxy for determining the initial antigen that elicited a T-cell response. We therefore analysed sequence conservation over the SARS-CoV-2 proteome across the *Coronaviridae*, which involved the construction of a core gene family-wide phylogeny. We subsequently assessed the amino acid homology to endemic HCoVs and other members of the *Coronaviridae* of 177 $CD4^+$ and $CD8^+$ epitopes identified in healthy unexposed individuals reported by four independent studies (Mateus et al., 2020; Le Bert et al., 2020; Nelde et al., 2020; Schulien et al., 2020).

We find that more than half of the reported epitopes (54.8%) did not have detectable homology to any of the endemic HCoVs. Further, none of the sequenced members of the *Coronaviridae* could explain a higher proportion of reported epitopes than expected by chance, given the phylogenetic similarity of their entire genome to SARS-CoV-2. Our results suggest that prior exposure to endemic coronaviruses is not the sole explanation of cross-reactivity patterns to SARS-CoV-2 in unexposed individuals. Instead, patterns of pre-existing T-cell cross-reactivity to SARS-CoV-2 seem largely in line with lifelong exposure to a diverse and heterogenous array of primarily microbial antigens. We anticipate that our findings will facilitate further characterisations of the potential sources of pre-existing T-cell immunity.

2. Methods

2.1. Data acquisition

3300 publicly available complete *Coronaviridae* assemblies were downloaded from NCBI Virus using the *taxid*: 1118 together with accompanying metadata on 08/04/2020. We also identified a further set of 41 Sarbecoviruses for inclusion that were released subsequent to January 2021. This dataset includes 12 bat and pangolin Coronavirus sequences from GISAID (Elbe and Buckland-Merrett, 2017) (acknowledged in Table S3). Sequence duplicates were identified and removed from the combined dataset using *seqkit rmdup* (Shen et al., 2016) together with those accessions with >10% of sites set to N. Accessions were later retained in the dataset only for those with a reported host of isolation. This resulted in a final dataset of 2572 assemblies with complete metadata with the latter manually cleaned to ensure consistent reporting of host and viral species.

2.2. Maximum Likelihood phylogeny of Coronaviridae

To reconstruct the core genomic diversity of the entire *Coronaviridae* family, we extracted the shared core genes from the representative genome assemblies across all genera. First, open reading frames (ORFs) were identified using the genome annotation tool *Prokka* v1.14.6 (Seemann, 2014). Next, the *Roary* pipeline v3.11.12 (Page et al., 2015) was used to cluster all *Coronaviridae* ORFs at a minimum amino-acid homology threshold of 30%. Sequences for the four genes ORF1ab, S, M and N were each found to cluster in a minimum of 2572 assemblies, which were then extracted, concatenated and aligned using *MAFFT* v7.453 (Katoh et al., 2002). The resulting alignment was trimmed of gaps found in 20% or more isolates and used to build a Maximum Likelihood phylogeny using *IQTree* v1.6.9 (Nguyen et al., 2015) specifying the *-fast* option. The four core genes in the trimmed concatenated alignment (12,014 bp) corresponds to 43.1% of the average length of all included WGSs (27,867 bp). We provide the curated metadata of the final 2572 viral records used in our analysis in Table S1.

As it was not possible to include an outgroup in the *Coronaviridae* concatenated-core alignment, an alignment-free analysis was used to identify the most basal genus with which to root the family Maximum Likelihood phylogeny. All *RefSeq* genome assemblies belonging to the virus order *Nidovirales* were downloaded, which contained 103 sequences accrsoss the sub-orders *Arnidovirineae*, *Cornidovirineae*, *Mesnidovirineae*, *Nanidovirineae*, *Ronidovirineae* and *Tornidovirineae*. Each assembly contained a ORF1ab CDS annotated ORF, the only gene shared by all members of the *Nidovirales* (Lauber et al., 2013), which were decomposed into 14-mer sequences using *MASH* v2.1.1 (Ondov et al., 2016). Based on pairwise Jaccard Distances of matched 14-mers between all ORF1ab sequences, a Neighbour-Joining tree was constructed to assess the genetic relationship between members of the *Nidovirales*. The genus *Deltacoronavirus* was identified to be the most basal clade of the *Coronaviridae* in the wider context of the taxonomic order and was therefore used to root the family-wide Maximum Likelihood phylogeny.

2.3. Sequence conservation analysis

We decomposed the SARS-CoV-2 proteome (sequences retrieved from *RefSeq*; NC_045512.2) into 9394 15-mer peptides overlapping by 14 amino acids using a custom *R* script. Such a 15-mer sliding window allows for consideration of all possible peptide strings within the SARS-CoV-2 proteome. In addition, we retrieved the sequences of 177 epitopes found to elicit a response in at least one individual unexposed to SARS-CoV-2 from Singapore (Le Bert et al., 2020), the USA (Mateus et al., 2020) and Germany (Nelde et al., 2020; Schulien et al., 2020) from published supplementary tables. The breakdown of the number of epitopes for each T-cell response type is shown in Table S4b. Translated protein sequences of all ORFs from each of the 2572 assemblies were retrieved from *Prokka* (Seemann, 2014) and used to construct a protein BLAST database. Separately, a protein BLAST database was also constructed from the protein annotations associated with the 2572 assemblies, which were downloaded using *NCBI Batch Entrez* (https://www.ncbi.nlm.nih.gov/sites/batchentrez). Subsequently, we used the BLASTP utility from *BLAST*+ v2.11.0 (Camacho et al., 2009) to determine the sequence homology of the 15-mer peptides from the SARS-CoV-2 proteome and the 177 published epitopes using the two databases. Sequence homology (or percentage identity), is defined here as the percentage of amino acid/nucelotide matches between any two sequences. The resultant protein BLAST outputs were merged by retaining only the hit with the maximum percentage identity for each assembly and query combination. To include all tested alignments, we set *-num_alignments* and *-evalue* parameters to 10^9 and 2×10^9, respectively. In addition, to optimise the protein BLAST search for short sequences, *−task* was set to *blastp-short*. Lastly, only alignments involving the full length of the query sequence were considered by setting *-qcov_hsp_perc* as 99. This threshold was employed because the query sequences are short and so sequence identity would only be a meaningful measure of homology in alignments given the whole sequence. Using this BLASTP implementation, we store the sequence homology values when an alignment was produced, and return zero for cases when it was not (referred to as 'no homology').

2.4. Regression analysis

Using the merged output of the protein BLAST search querying the 177 published epitopes, we analysed the proportion of epitopes that had any homology to each virus in our dataset. To do so we additionally calculated the alignment-free genetic distance - 'Mash distance' - of each virus relative to SARS-CoV-2 using *MASH v2.1.1* (Ondov et al., 2016) specifying a *k*-mer size of 14. A least squares regression of the proportion of epitopes with any homology on the natural logarithm of Mash distance was performed using the *lm* function in *R*. This analysis was applied to a representative filtered dataset of all combinations of unique host and virus species requiring a unique Mash distance to SARS-CoV-2 ($n = 365$). Pearson's correlation of the two variables was also calculated using the *cor.test* function in R. The studentised residuals were calculated using the *studres* function as part of the *MASS* package v7.3–53 (Ripley et al., 2013).

2.5. Non-Coronaviridae protein BLAST

To determine if any proteome outside of the *Coronaviridae* had detectable homology to any of the 177 epitopes reported in the literature, we performed a protein BLAST analysis using the online BLASTP *suite* (https://tinyurl.com/y22o4t9z) against the non-redundant protein sequence database (accessed 7/12/2020), while excluding sequences associated with the *Coronaviridae* (taxid: 11118). Protein BLAST searches were conducted in eight batches of 20 and a ninth batch of 17 epitopes with the number of alignments performed set to 1000 per batch. After merging the outputs of the eight batches, we filtered the resultant table to exclude missing organism names, hits with descriptions containing the terms 'synthetic', 'SARS', 'coronavirus', or 'cov', or organism names labelled as 'uncultured bacterium'. Additionally, we excluded hits to the Protein Data Bank accession 6ZGH_A, given it contains a region of the SARS-CoV-2 spike protein sequence.

3. Results

3.1. Conservation analysis across the family-wide phylogeny of Coronaviridae

To reconstruct the shared genomic diversity of the *Coronaviridae* family, we extracted a concatenated alignment of core (shared) genes (ORF1ab, S, M, N) from annotated genome assemblies of 2572 coronaviruses, isolated from human and animal hosts, and constructed a Maximum Likelihood phylogeny (Fig. 1a, Table S1). We then decomposed the SARS-CoV-2 proteome (NC_045512.2) into 15-mer peptide sequences overlapping by 14 amino acids and performed protein *BLAST* searches to determine the homology to protein sequences translated from each of the 2572 coronavirus assemblies isolated from a range of hosts (see Methods). Two sequences are said to have 'no homology' if a protein BLAST alignment of said sequences could not be produced. The proteome-wide homology of 15-mer peptides across the *Coronaviridae* is represented in Fig. 1b. At a 40% amino acid sequence homology cut-off, SARS-CoV-2 peptide sequences were highly conserved across the family at the C-terminal end of ORF1ab. Representations of alternative homology thresholds (66% and 80%) provide qualitatively similar patterns (Fig. S1a and b). This region of homology includes the RNA-dependent RNA polymerase (RdRp) (nsp12) and helicase (nsp13) which are known regions of high conservation across the coronaviruses, with the former frequently used as a taxonomic marker (Latinne et al., 2020).

3.2. Cross-reactivity profiles cannot be completely explained by exposure to endemic HCoVs

We analysed the sequence homology of 177 cross-reactive peptides found to elicit T-cell response in published work on four independent cohorts of healthy unexposed people from Singapore (Le Bert et al., 2020), the USA (Mateus et al., 2020) and Germany (Nelde et al., 2020; Schulien et al., 2020) to endemic HCoV protein sequences (Fig. 2). Without setting any identity threshold to report protein identity, we found that 76.3–83.1% of the SARS-CoV-2 epitopes had no homology to the four endemic HCoV species individually. In addition, 97 of the 177 epitopes (54.8%) had no homology to the proteome of all four endemic HCoVs combined (henceforth 'unexplained' epitopes). To investigate the potential source of 'unexplained' epitopes within the *Coronaviridae* further, we calculated the proportion of the 97 'unexplained' epitopes with any homology to the proteome of each remaining coronavirus in our dataset (excluding SARS-CoV-2) (Fig. S2). The results suggest that a large proportion of 'unexplained' epitopes have homology to at least some of the *Betacoronaviruses* including SARS-CoV-1 and SARS-CoV-1-like coronaviruses within the Sarbecovirus sub-group.

Additionally, given the overrepresentation of some coronavirus species within the dataset, we randomly subset the 2572 viral records to include only representative of each host and viral species that have non-identical Mash distances to the SARS-CoV-2 NCBI reference genome (Wuhan-Hu-1; NC_045512.2). Using the resultant 365 records, we found that the proportion of published epitopes with any homology to coronaviruses is strongly correlated with the natural logarithm of alignment-free Mash distance between the entire genome of each coronavirus relative to SARS-CoV-2 (Pearson's $r = -0.958$, $p < 0.0001$) (Fig. 3a). In fact, none of the 365 viruses in this filtered dataset had studentised residuals exceeding three, indicating that no coronaviruses within the dataset have homology to a significantly higher number of epitopes than expected by chance (Fig. 3b).

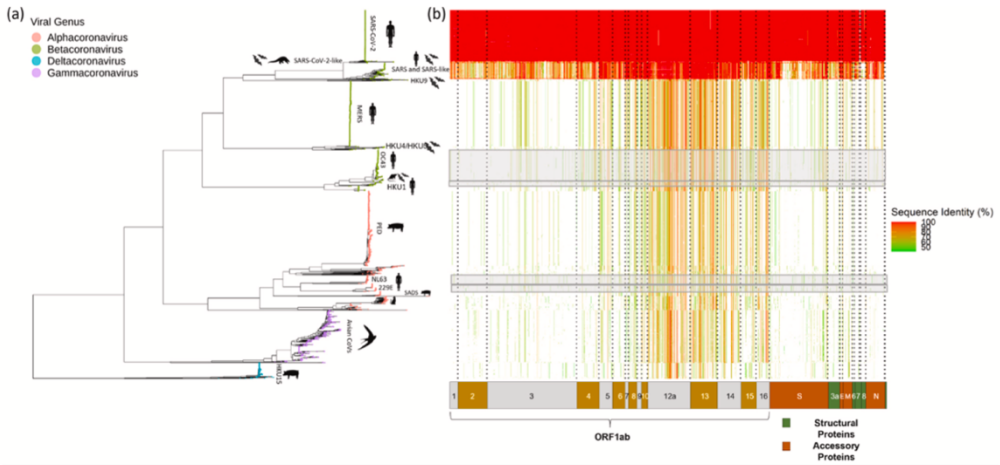


Fig. 1. Conservation analysis of SARS-CoV-2-derived 15-mer peptides across the *Coronaviridae*. (a) Maximum likelihood phylogeny of a concatenated alignment of core genes in the *Coronaviridae* annotated by viral genera (tip colour) and highlighting major hosts (Table S1). (b) Heatmap visualising the homology of SARS-CoV-2-derived 15-mer peptide sequences across the family. Each row and column correspond to a tip on the phylogeny and a single 15-mer peptide, respectively. The fill of each cell provides the level of homology of a particular SARS-CoV-2-derived 15-mer peptide to the proteome of a single genome record as given by the colour scale at right. Grey boxes highlight the rows of the heatmap corresponding to each of the four endemic human coronaviruses. The homology threshold set to report a protein BLAST hit was 40%.

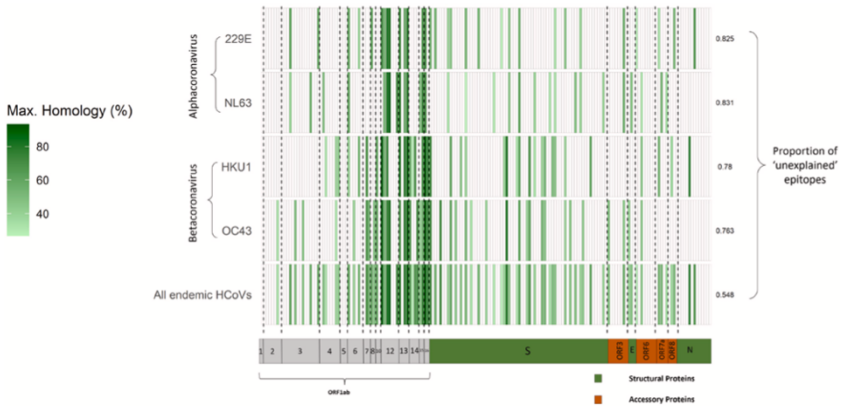


Fig. 2. Sequence homology of deconvoluted peptides from published literature to endemic HCoVs. Heatmap visualising the maximum sequence homology of deconvoluted SARS-CoV-2-derived peptides to the each of the four endemic HCoVs (first four rows) and across all HCoVs combined (last row). The proportion of epitopes that cannot be explained by detectable homology to proteins from each species of HCoV is annotated on the right of the heatmap. Each row and column correspond to a single genome record and a single peptide, respectively. The fill of each cell provides the maximum sequence homology of a particular SARS-CoV-2-derived epitope to the proteome of all genome records for each species. This maximum sequence homology was determined by considering only all viruses isolated from a human host and with species names including the terms '229E', 'NL63', 'HKU1' and 'OC43'.

3.3. Possible sources for T-cell cross-reactivity beyond coronaviruses

To identify possible sources for the T-cell cross-reactivity observed in people unexposed to SARS-CoV-2, we also performed a protein BLAST search for all 177 experimentally validated epitopes against the NCBI non-redundant protein database (excluding the taxon *Coronaviridae*), storing the first 1000 hits in each case. A fraction of the epitopes (10/177) share partial homology with proteins from a very diverse range of taxa, including viruses, bacteria and unicellular eukaryotes (Table S2). However, the lowest Expect (E) value of the protein BLAST hits, which represents the number of similar hits expected by chance given the size of the database used and the length of the query (Tatusova and Madden, 1999), is 7.5. This suggests that all the hits shown in Table S2 could be explained by chance alone. Together with the wide diversity of taxa

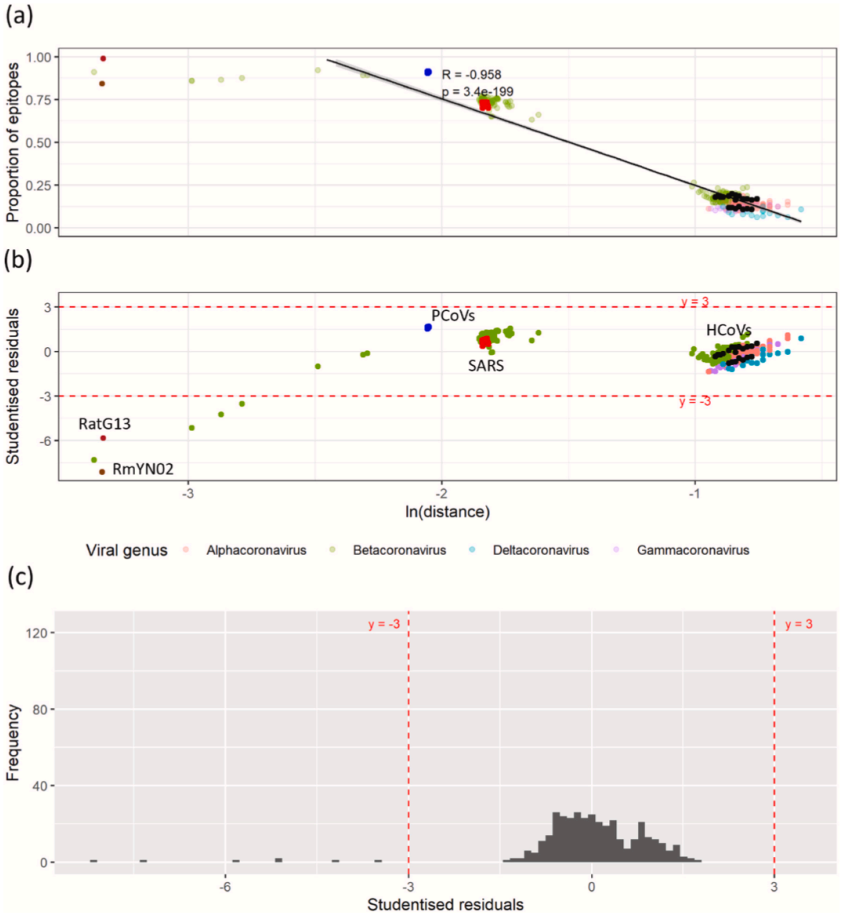


Fig. 3. Relationship between the proportion of unexposed epitopes that have detectable sequence homology and the Mash distance to SARS-CoV-2 in a representative subset of the *Coronaviridae*. (a) Scatter plot and least squares regression line providing the proportion of epitopes with detectable homology to a coronavirus species (y-axis) and the natural logarithm of Mash distance to SARS-CoV-2 (x-axis). The dataset was filtered to only include 365 coronaviruses encompassing all unique host species, viral species and Mash distance combinations. These coronaviruses are coloured by viral genera with key members highlighted.

identified, the results suggest that there is no single candidate for the source(s) of the T-cell cross-reactive repertoire beyond the *Coronaviridae*.

4. Discussion

SARS-CoV-2 cross-reactive T-cells in healthy unexposed individuals have been identified as potentially important contributors to the immunological response to COVID-19. Prior exposure to globally circulating endemic coronaviruses present some of the strongest candidates for eliciting such cross-immunity. Though, the relative contribution of these coronaviruses to the reactive T-cell epitopes identified in multiple cohorts of healthy individuals have been only partially explored. We characterised the amino acid homology of SARS-CoV-2-derived T-cell epitopes reported in COVID-19 unexposed individuals from Singapore (Le Bert et al., 2020), the USA (Mateus et al., 2020) and Germany (Nelde et al., 2020; Schulien et al., 2020) against the entire proteome of the *Coronaviridae* family, including all major mammalian and avian lineages.

Following a comprehensive screen, we found that 54.8% of reported T-cell epitopes had no homology to the four human endemic coronavirus species (HKU1, OC43, NL63 and 229E) (Fig. 2), despite HCoV infections circulating widely in global human populations (Su et al., 2016). We note that the highest conservation to confirmed T-cell epitopes tended to be within members of the *Sarbecovirus* sub-group, which includes SARS-CoV-1, SARS-CoV-2, and a few related species that have been isolated mostly from bats and pangolins but are not known to have been in widespread circulation in humans. However, this homology can be well explained by the core phylogenetic relatedness of these viral species to SARS-CoV-2 (Fig. 3). Furthermore, SARS-CoV-2 infection leads to a

heterogenous pattern of cell-mediated immune responses over the entire SARS-CoV-2 genome, largely falling outside of the spike protein, not enriched in the terminal end of ORF1ab largely conserved among the coronaviruses, and does not consistently lead to cross-reactivity with endemic HCoVs (Ferretti et al., 2020).

Our work adds to a growing suite of evidence that prior HCoV infections are not the only candidates responsible for cross-reactive T-cell epitopes in SARS-CoV-2 unexposed individuals. We argue that previous studies that presented empirical evidence of T-cell cross-reactivity with HCoV-derived peptides did not take into account the genetic relatedness of endemic HCoVs to SARS-CoV-2, placing an over-emphasis on these viruses as the source of pre-existing T-cell immunity. This opens the question as to what other antigens may have primed the intrinsic cross-reactivity identified (Campion et al., 2014) in pre-pandemic samples. A sizeable fraction of cross-reactive T-cell epitopes remains unexplained by prior exposure to any known coronavirus in circulation. It feels fairly implausible that the 'unexplained' cross-reactive epitopes are due to prior exposure to a yet undescribed coronavirus. Indeed, such a hypothetical yet-to-be described coronavirus would have needed to be in circulation globally until very recently and then vanished, which seems highly unlikely. Additionally, since we incorporated the whole known genetic diversity of coronaviruses in our analyses, which has been reasonably well sampled, such an unknown pathogen would likely have to be phylogenetically unrelated to any coronavirus characterised to date. As such, an unknown coronavirus would be an unlikely candidate for a source of this 'unexplained' T-cell cross-reactivity.

Possible alternative agents for the unexplained cross-reactive epitopes may include widespread microbes, or widely administrated vaccines. The tuberculosis bacille Calmette-Guerin (BCG) vaccines have been suggested as candidates providing some cross-immunity against SARS-CoV-2 (Tomita et al., 2020; Escobar et al., 2020). However, our screen of all 177 published T-cell epitopes found no homology to any *Mycobacterium* species (Table S2). As such, the evidence that BCG vaccination is a contributor to the T-cell cross-reactivity observed remains unconvincing. Instead we identify a diverse spread of putative antigens with low detectable homology. The presence of such a broad pre-existing repertoire of $CD4^+$ reactive T-cells in healthy adults has previously been observed in the context of cross-reactivity to HIV and influenza infection, and interpreted as the result of prior exposure to environmental antigens (Su et al., 2013) or proteins in the human microbiome (Campion et al., 2014). It has also been postulated that the cross-reactive profile may take on an increasing role with age and immunological experience (Woodland and Blackman, 2006) which may result in high levels of inter-individual variation based on infection history and HLA type.

Admittedly, sequence homology is an indirect proxy for probing the source of T-cell cross-reactivity. Yin and Mariuzza (Yin and Mariuzza, 2009) reviewed five putative mechanisms of T-cell cross-reactivity, all of which highlight the complex and diverse molecular interactions of peptide, major histocompatibility complex (MHC) and T-cell receptors. In particular, molecular mimicry would suggest that conservation of structure can compensate for lower sequence homology (Macdonald et al., 2009; Wucherpfennig and Strominger, 1995; Quaratino et al., 1995). Deconvolving the relationship between sequence homology and cross-reactivity is evidently non-trivial and remains a limitation of our work. Indeed, we do not rule out the possibility that peptides of lower homology from members of the *Coronaviridae* can result in cross-reactivity. However, it remains evident that a high sequence homology improves the likelihood that structural or chemical characteristics are conserved, with empirical evidence that this is the case. For instance Mateus et al. (Mateus et al., 2020) found that only 1% of SARS-CoV-2:HCoV peptide pairs sharing 33–40% sequence homology were cross-reactive. Meanwhile, 21% of peptide pairs with 47–60% homology and 57% of peptides with >67% homology were cross-reactive. These findings highlight a positive association of sequence homology and the frequency of cross-reactivity, providing strong empirical evidence for our assumption that sequence homology is a good measure for inferring the source of T-cell cross-reactivity. Additionally, Grifoni et al. (Grifoni et al., 2020b) showed that 12 of 17 SARS-CoV-2 peptides with >90% sequence homology to experimentally-validated SARS-CoV epitopes were predicted to elicit a T-cell response. The authors then conclude that these peptides have a high probability of triggering a T-cell response, and could generate responses that are "cross-protective" across *Betacoronaviruses*. This serves as a precedent for using sequence homology to infer T-cell cross-reactivity. Finally, while a sequence homology-based approach may not be able to account for cross-reactivity as a result of structural homology, it offers scalability in that we can screen all known coronaviruses to date, which would not be feasible experimentally.

In conclusion, our results highlight the importance of considering the wider phylogenetic context of circulating antigens contributing to immunological memory to novel pathogens. The widespread and repeated exposure of global human populations to circulating endemic HCoVs is expected to have left an immunological legacy which may modulate COVID-19 pathogenesis. However, our results suggest that the extensive T-cell cross-reactivity previously reported cannot be solely explained by prior exposure to any known coronavirus in global circulation. It is nonetheless clear that the potential cross-reactive repertoire is widespread and present in cohorts of healthy people from multiple countries around the globe (Mateus et al., 2020; Grifoni et al., 2020a; Le Bert et al., 2020; Nelde et al., 2020; Braun et al., 2020; Peng et al., 2020; Schulien et al., 2020; Bacher et al., 2020; Sekine et al., 2020; Weiskopf et al., 2020b), even if perhaps at low avidity (Bacher et al., 2020). It remains to be established to what extent such cross-reactivity translates into immunity to SARS-CoV-2, both in terms of susceptibility to infection and symptom severity upon infection.

Data and code availability

All source code used for the analyses can be found on GitHub (https://github.com/cednotsed/tcell_cross_reactivity_covid.git). Genomic data for the *Coronaviridae* were obtained from publicly available accessions on NCBI Virus. Twelve further bat and pangolin associated coronaviruses were also included downloaded from the GISAID repository, with full acknowledgements provided in Table S3. The list of epitopes used and the frequency table of $CD4^+$ and $CD8^+$ T-cell epitopes stratified by study cohort can be found in Table S4a and b respectively.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2021.105075.

Acknowledgements and funding

L.v.D and F.B. acknowledge financial support from the Newton Fund UK-China NSFC initiative (grant MR/P007597/1) and the BBSRC (equipment grant BB/R01356X/1). L.v.D. is supported by a UCL Excellence Fellowship. C.O. is funded by a NERC-DTP studentship. Finally, we acknowledge the large number of research groups openly sharing SARS-CoV-2 genomic and immunological data with the research community.

Declaration of Competing Interest

A.B. is a cofounder of Lion TCR, a biotechnology company that develops T-cell receptors for the treatment of virus-related diseases and cancers but was not deemed to have any competing interests. The other authors have no competing interests to declare.

References

Allen, M., Poggiali, D., Whitaker, K., Marshall, T.R., Kievit, R.A., 2019. Raincloud plots: a multi-platform tool for robust data visualization. Wellcome Open Res. 4.

Bacher, P., Rosati, E., Esser, D., Martini, G.R., Saggau, C., Schiminsky, E., et al., 2020. Low avidity CD4+ T cell responses to SARS-CoV-2 in unexposed individuals and humans with severe COVID-19. Immunity 50 (6), 1258–1271 [Internet]. Available from: http://www.sciencedirect.com/science/article/pii/S1074761320305033.

Boni, M.F., Lemey, P., Jiang, X., Lam, T.T.-Y., Perry, B., Castoe, T., et al., 2020. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat. Microbiol. 5 (11), 1408–1417.
Braun, J., Loyal, L., Frentsch, M., Wendisch, D., Georg, P., Kurth, F., et al., 2020. SARS-CoV-2-reactive T cells in healthy donors and patients with COVID-19. Nature. 1–5.
Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al., 2009. BLAST+: architecture and applications. BMC Bioinform. 10 (1), 421.
Campion, S.L., Brodie, T.M., Fischer, W., Korber, B.T., Rossetti, A., Goonetilleke, N., et al., 2014 Jun 30. Proteome-wide analysis of HIV-specific naive and memory CD4(+) T cells in unexposed blood donors. J. Exp. Med. 211 (7), 1273–1280 [Internet]. Available from: https://pubmed.ncbi.nlm.nih.gov/24958850.
van Dorp, L., Acman, M., Richard, D., Shaw, L.P., Ford, C.E., Ormond, L., et al., 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect. Genet. Evol. 104351 [Internet]. Available from: http://www.sciencedirect.com/science/article/pii/S1567134820301829.
Echeverría, G., Guevara, Á., Coloma, J., Ruiz, A.M., Vasquez, M.M., Tejera, E., et al., 2021. Pre-existing T-cell immunity to SARS-CoV-2 in unexposed healthy controls in Ecuador, as detected with a COVID-19 Interferon-Gamma Release Assay. Int. J. Infect. Dis. 105, 21–25.
Elbe, S., Buckland-Merrett, G., 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Challenges. 1 (1), 33–46.
Escobar, L.E., Molina-Cruz, A., Barillas-Mury, C., 2020 Jul 28. BCG vaccine protection from severe coronavirus disease 2019 (COVID-19). Proc. Natl. Acad. Sci. 117 (30), 17720 LP–17726 [Internet]. Available from: http://www.pnas.org/content/117/30/17720.abstract.
Ferretti, A.P., Kula, T., Wang, Y., Nguyen, D.M.V., Weinheimer, A., Dunlap, G.S., et al., 2020. Unbiased screens show CD8+ T Cells of COVID-19 patients recognize shared epitopes in SARS-CoV-2 that largely reside outside the spike protein. Immunity 53 (5), 1095–1107.e3 [Internet]. Available from: http://www.sciencedirect.com/science/article/pii/S1074761320304477.
Gaunt, E.R., Hardie, A., ECJ, Claas, Simmonds, P., Templeton, K.E., 2010 Aug. Epidemiology and clinical presentations of the four human Coronaviruses 229E, HKU1, NL63, and OC43 detected over 3 years using a novel multiplex real-time PCR method. J. Clin. Microbiol. 48 (8), 2940 LP–2947 [Internet]. Available from: http://jcm.asm.org/content/48/8/2940.abstract.
Gombar, S., Bergquist, T., Pejaver, V., Hammarlund, N.E., Murugesan, K., Mooney, S., et al., 2021. SARS-CoV-2 infection and COVID-19 severity in individuals with prior seasonal coronavirus infection. Diagn. Microbiol. Infect. Dis. 100 (2), 115338.
Grifoni, A., Weiskopf, D., Ramirez, S.I., Mateus, J., Dan, J.M., Moderbacher, C.R., et al., 2020a. Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. Cell. 187 (7), 1489–1501.
Grifoni, A., Sidney, J., Zhang, Y., Scheuermann, R.H., Peters, B., Sette, A., 2020b. A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. Cell Host Microbe 27 (4), 671–680.e2 [Internet]. Available from: https://www.sciencedirect.com/science/article/pii/S1931312820301669.
Gupta, R., Misra, A., 2020. COVID19 in South Asians/Asian Indians: Heterogeneity of data and implications for pathophysiology and research. Diabetes Res. Clin. Pract. 165, 108267 [Internet]. Available from: http://www.sciencedirect.com/science/article/pii/S0168822720305179.
Jordan, R.E., Adab, P., Cheng, K.K., 2020. Covid-19: Risk Factors for severe Disease and Death. British Medical Journal Publishing Group.
Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30 (14), 3059–3066.
Ksiazek, T.G., Erdman, D., Goldsmith, C.S., Zaki, S.R., Peret, T., Emery, S., et al., 2003. A novel coronavirus associated with severe acute respiratory syndrome. N. Engl. J. Med. 348 (20), 1953–1966.
Ladner, J.T., Henson, S.N., Boyle, A.S., Engelbrektson, A.L., Fink, Z.W., Rahee, F., et al., 2021. Epitope-resolved profiling of the SARS-CoV-2 antibody response identifies cross-reactivity with endemic human coronaviruses. Cell Rep. Med. 2 (1), 100189.
Latinne, A., Hu, B., Olival, K.J., Zhu, G., Zhang, L., Li, H., et al., 2020. Origin and cross-species transmission of bat coronaviruses in China. Nat. Commun. 11 (1), 4235 [Internet]. Available from: https://doi.org/10.1038/s41467-020-17687-3.
Lauber, C., Goeman, J.J., del Carmen, Parquet M., Nga, P.T., Snijder, E.J., Morita, K., et al., 2013. The footprint of genome architecture in the largest genome expansion in RNA viruses. PLoS Pathog. 9 (7), e1003500.
Le Bert, N., Tan, A.T., Kunasegaran, K., Tham, C.Y.L., Hafezi, M., Chia, A., et al., 2020. SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls. Nature. 584 (7821), 457–462.
Low, J.S., Vaqueirinho, D., Mele, F., Foglierini, M., Jerak, J., Perotti, M., et al., 2021. Clonal analysis of immunodominance and cross-reactivity of the CD4 T cell response to SARS-CoV-2. Science 372 (6548), 1336–1341.
Lv, H., Wu, N.C., Tsang, O.T.-Y., Yuan, M., Perera, R.A.P.M., Leung, W.S., et al., 2020. Cross-reactive antibody response between SARS-CoV-2 and SARS-CoV infections. Cell Rep. 31 (9), 107725 [Internet]. Available from: http://www.sciencedirect.com/science/article/pii/S2211124720307026.
Macdonald, W.A., Chen, Z., Gras, S., Archbold, J.K., Tynan, F.E., Clements, C.S., et al., 2009. T cell allorecognition via molecular mimicry. Immunity. 31 (6), 897–908.
Mateus, J., Grifoni, A., Tarke, A., Sidney, J., Ramírez, S.I., Dan, J.M., et al., 2020. Selective and cross-reactive SARS-CoV-2 T cell epitopes in unexposed humans. Science 370 (6512), 89–94.
Moriyama, M., Hugentobler, W.J., Iwasaki, A., 2020 Sep 29. Seasonality of respiratory viral infections. Annu. Rev. Virol. 7 (1), 83–101 [Internet]. Available from. https://doi.org/10.1146/annurev-virology-012420-022445.
Neher, R.A., Dyrdak, R., Druelle, V., Hodcroft, E.B., Albert, J., 2020. Potential impact of seasonal forcing on a SARS-CoV-2 pandemic. Swiss Med. Wkly. 150 (1112).
Nelde, A., Bilich, T., Heitmann, J.S., Maringer, Y., Salih, H.R., Roerden, M., et al., 2020. SARS-CoV-2-derived peptides define heterologous and COVID-19-induced T cell recognition. Nat. Immunol. 1–12.
Ng, K.W., Faulkner, N., Cornish, G.H., Rosa, A., Harvey, R., Hussain, S., et al., 2020 Nov 6. Preexisting and de novo humoral immunity to SARS-CoV-2 in humans. Science 370 (6522), 1339–1343. http://science.sciencemag.org/content/early/2020/11/05/science.abe1107.abstract.
Nguyen, L.-T., Schmidt, H.A., Von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32 (1), 268–274.
O'Driscoll, M., Dos Santos, G.R., Wang, L., Cummings, D.A.T., Azman, A.S., Paireau, J., et al., 2020. Age-specific mortality and immunity patterns of SARS-CoV-2. Nature. 1–9.
Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., et al., 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 17 (1), 132.
Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., et al., 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 31 (22), 3691–3693.
Peng, Y., Mentzer, A.J., Liu, G., Yao, X., Yin, Z., Dong, D., et al., 2020. Broad and strong memory CD4+ and CD8+ T cells induced by SARS-CoV-2 in UK convalescent individuals following COVID-19. Nat. Immunol. 21 (11), 1336–1345 [Internet]. . Available from: https://doi.org/10.1038/s41590-020-0782-6.
Quaratino, S., Thorpe, C.J., Travers, P.J., Londei, M., 1995. Similar antigenic surfaces, rather than sequence homology, dictate T-cell epitope molecular mimicry. Proc. Natl. Acad. Sci. 92 (22), 10398–10402.
Reynolds, C.J., Swadling, L., Gibbons, J.M., Pade, C., Jensen, M.P., Diniz, M.O., et al., 2020. Discordant neutralizing antibody and T cell responses in asymptomatic and mild SARS-CoV-2 infection. Sci. Immunol. 5 (54).
Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A., Firth, D., et al., 2013. Package 'mass'. Cran R. 538.
Sagar, M., Reifler, K., Rossi, M., Miller, N.S., Sinha, P., White, L., et al., 2020 Sep 30. Recent endemic coronavirus infection is associated with less severe COVID-19. J. Clin. Invest. https://doi.org/10.1172/JCI143380 [Internet]. Available from.
Schulien, I., Kemming, J., Oberhardt, V., Wild, K., Seidel, L.M., Killmer, S., et al., 2020. Characterization of pre-existing and induced SARS-CoV-2-specific CD8+ T cells. Nat. Med. 1–8.
Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 30 (14), 2068–2069.
Sekine, T., Perez-Potti, A., Rivera-Ballesteros, O., Strålin, K., Gorin, J.-B., Olsson, A., et al., 2020. Robust T cell immunity in convalescent individuals with asymptomatic or mild COVID-19. Cell 183 (1), 158–168.e14 [Internet]. Available from: http://www.sciencedirect.com/science/article/pii/S0092867420310084.
Shaw, L.P., Wang, A.D., Dylus, D., Meier, M., Pogacnik, G., Dessimoz, C., et al., 2020 May 10. The phylogenetic range of bacterial and viral pathogens of vertebrates. Mol. Ecol. https://doi.org/10.1111/mec.15463 [Internet]. Available from:
Shen, W., Le, S., Li, Y., SeqKit, Hu F., 2016 Oct 5. A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One 11 (10), e0163962 e0163962, Available from. https://doi.org/10.1371/journal.pone.0163962.
Steiner, S., Sotzny, F., Bauer, S., Na, I.-K., Schmueck-Henneresse, M., Corman, V.M., et al., 2020. HCoV-and SARS-CoV-2 cross-reactive T cells in CVID patients. Front. Immunol. 11.
Su, L.F., Kidd, B.A., Han, A., Kotzin, J.J., Davis, M.M., 2013 Feb 21. Virus-specific CD4(+) memory-phenotype T cells are abundant in unexposed adults. Immunity 38 (2), 373–383 [Internet]. Available from: https://pubmed.ncbi.nlm.nih.gov/23395677.
Su, S., Wong, G., Shi, W., Liu, J., Lai, A.C.K., Zhou, J., et al., 2016. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends Microbiol. 24 (6), 490–502.
Swadling, L., Diniz, M.O., Schmidt, N.M., Amin, O.E., Chandran, A., Shaw, E., et al., 2021. Pre-existing polymerase-specific T cells expand in abortive seronegative SARS-CoV-2 infection. medRxiv.
Tatusova, T.A., Madden, T.L., 1999. BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol. Lett. 174 (2), 247–250.
Tomita, Y., Sato, R., Ikeda, T., Sakagami, T., 2020. BCG vaccine may generate cross-reactive T cells against SARS-CoV-2: in silico analyses and a hypothesis. Vaccine 38 (41), 6352–6356 [Internet]. Available from: http://www.sciencedirect.com/science/article/pii/S026441OX20310860.
Tyrrell, D.A.J., Bynoe, M.L., 1965. Cultivation of a novel type of common-cold virus in organ cultures. Br. Med. J. 1 (5448), 1467.
Walker, A.S., Pritchard, E., House T, Robotham, J.V., Birrell, P.J., Bell, I., et al., 2020 Jan 1. Viral load in community SARS-CoV-2 cases varies widely and temporally. medRxiv 12 (10), e64683 [Internet]. Available from: http://medrxiv.org/content/early/2020/10/27/2020.10.25.20219048.abstract.
Weiskopf, D., Schmitz, K.S., Raadsen, M.P., Grifoni, A., Okba, N.M.A., Endeman, H., et al., 2020a. Phenotype and kinetics of SARS-CoV-2-specific T cells in COVID-19 patients with acute respiratory distress syndrome. Sci. Immunol. 5 (48) p.eabd2071.
Weiskopf, D., Schmitz, K.S., Raadsen, M.P., Grifoni, A., Okba, N.M.A., Endeman, H., et al., 2020b. Phenotype of SARS-CoV-2-specific T-cells in COVID-19 patients with acute respiratory distress syndrome. medRxiv.
Woodland, D.L., Blackman, M.A., 2006. Immunity and age: living in the past? Trends Immunol. 27 (7), 303–307.
Wu, C., Chen, X., Cai, Y., Zhou, X., Xu, S., Huang, H., et al., 2020. Risk factors associated with acute respiratory distress syndrome and death in patients with coronavirus disease 2019 pneumonia in Wuhan, China. JAMA Intern. Med. 180 (7), 934–943.

Wucherpfennig, K.W., Strominger, J.L., 1995. Molecular mimicry in T cell-mediated autoimmunity: viral peptides activate human T cell clones specific for myelin basic protein. Cell. 80 (5), 695–705.

Yaqinuddin, A., 2020 Jun 30. Cross-immunity between respiratory coronaviruses may limit COVID-19 fatalities. Med. Hypotheses 144, 110049 [Internet]. Available from: https://pubmed.ncbi.nlm.nih.gov/32758887.

Ye, Z.-W., Yuan, S., Yuen, K.-S., Fung, S.-Y., Chan, C.-P., Jin, D.-Y., 2020. Zoonotic origins of human coronaviruses. Int. J. Biol. Sci. 16 (10), 1686.

Yin, Y., Mariuzza, R.A., 2009. The multiple mechanisms of T cell receptor Cross-reactivity. Immunity 31 (6), 849–851 [Internet]. Available from: http://www.sciencedirect.com/science/article/pii/S1074761309005135.

Zaki, A.M., Van Boheemen, S., Bestebroer, T.M., Osterhaus, A.D.M.E., Fouchier, R.A.M., 2012. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N. Engl. J. Med. 367 (19), 1814–1820.

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., et al., 2020. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet. 395 (10229), 1054–1062.